Proceedings of The 6th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/2026.31160

Design of AIGC Content Generation Framework for Gymnastics Teaching Based on Diffusion Model - Creation and Practical Application of Personalized Learning Resources

Jinghui Zhou[1*], Jingxin Zhou[2]

[1]*School of Physical Education, Wuhan University of Science and Technology, Wuhan, China*
[2]*Department of Physical Education, Panlonghu Campus, Hebei Vocational College of Resources and Environment, Shijiazhuang, China*
**Corresponding Author. Email: 2433563462@qq.com*

Abstract. For the strategy of building a sports power, gymnastics teaching is confronted with the core contradiction between the high-precision requirements of movements and the insufficient supply of personalized resources. Generative artificial intelligence (AIGC) provides a technical path for innovation in this field. This study constructs an AIGC content generation framework for gymnastics teaching based on the diffusion model. Experimental verification shows that this framework significantly outperforms the baseline model in terms of generation quality, action coherence and personalized adaptability, enriching the application research of AIGC in the vertical teaching field.

Keywords: Diffusion model Gymnastics teaching AIGC content generation Personalized learning resources

1. Introduction

Against the backdrop of the in-depth advancement of the digitalization strategy in education, the construction of a sports power is making coordinated efforts [1]. The deep integration of new-generation artificial intelligence technology and education and teaching has become the core engine driving the modernization of education [2,3]. At the national level, a series of policy arrangements have been successively introduced, explicitly stating that an innovative education ecosystem should be built under the leadership of AIGC to help achieve "education for everyone", which has pointed out the direction for innovation in the field of physical education [4].

The implementation of the Healthy China strategy has set higher requirements for the high-quality development of physical education. Gymnastics combines the complexity of skills, aesthetic value and educational function [5]. The digital transformation of its teaching is the key to responding to the national strategy. The core pain point in gymnastics teaching is the poor adaptability of the traditional resource model, which is difficult to meet the differentiated learning needs [6].

The breakthrough in AIGC technology provides an innovative path for this field. Diffusion models, with their high-fidelity advantage, have become the preferred choice for generating action-

DOI: 10.54254/2755-2721/2026.31160

related content [7]. Research shows that the efficiency of the time-series aware latent diffusion model is superior to that of ComMDM [8]; The diffusion model-derived frameworks such as SMGDiff were used to achieve controllable generation of sports actions, and their adaptability was verified [9].

The technological evolution of diffusion models provides core support for solving the above problems. The denoising diffusion probability model lays the core mechanism of reverse denoising [10]. The latent diffusion model is compressed to a low-dimensional latent space through a VAE encoder [11], and the conditional diffusion model is injected to achieve generation controllability [12]. The digital deconstruction of gymnastic movements relies on the collaborative technology of skeletal point sequence tensor characteristics, motion trajectory time series modeling and dynamic feature extraction to construct a representation system [13]. Based on this, this study focuses on the core demands of gymnastics teaching, constructs a personalized AIGC content generation framework, fills the research gap in the integration of intelligent technology and personalized gymnastics teaching, and provides technical support for the digital transformation of physical education.

2. Framework design

2.1. Overall framework architecture

The core design concepts include conditional drive, potential diffusion and multimodal decoding. The interaction relationships among each link are shown in Table 1.

Table 1. Interaction relationship of each link

Stage	Input	Core Function	Output
Requirement Input	Teaching Objectives, Movement Types, Learner Profiles	Standardized Validation of Requirements	Structured Requirement Document
Condition Encoding	Structured Requirements, Personalized Parameters	Multimodal Semantic Embedding and Fusion	Conditional Feature Vector
Latent Space Diffusion	Conditional Vector, Initial Noise	Iterative Denoising Generation	Latent Representation of Content
Resource Decoding	Latent Representation	Multimodal Resource Reconstruction	Video/Illustration/Text Resources
Output Application	Multimodal Resources	Personalized Adaptation and Push	Customized Resource Package

2.2. Design of multimodal conditional encoder

The semantic embedding of teaching intention text is based on the CLIP model [14] :

$$v_{text} = CLIP_{text}(t) \in R^d \tag{1}$$

Here, t represents the teaching text, d represents the feature dimension, and the text intent completion mechanism is synchronously integrated.

Personalized parameter embedding quantifies the learner profile as $M_{user} \in R^{k \times d}$ through the Analytic Hierarchy process, and after L2 normalization, it is weighted and fused with v_{text} :

DOI: 10.54254/2755-2721/2026.31160

$$v_{init} = \alpha \cdot v_{text} + (1-\alpha) \cdot \text{avg}\left(M_{user}\right) \tag{2}$$

Among them, $\alpha \in (0,1)$ are the weight coefficients.

After CLIP encoding, the temporal attention layer is embedded to enhance the temporal relationship weights related to the action sequence.

2.3. Construction of potential diffusion models with temporal perception

The latent space construction is based on VAE. The encoder E maps the high-dimensional action sequence $x \in R^{T\times C\times H\times W}$ to $z = E(x) \in R^{T\times c\times h\times w}$. Decoder D reconstructs $\widehat{x} = D(z)$; Training objective:

$$\mathscr{L}_{\mathscr{V}\mathscr{A}\mathscr{E}} = \mathscr{L}_{e} + \beta \cdot \mathscr{L}_{\mathscr{K}\mathscr{L}} \tag{3}$$

Here, $\mathscr{L}_{e}$ represents the reconstruction loss, $\mathscr{L}_{\mathscr{K}\mathscr{L}}$ is the KL divergence, and β is the equilibrium coefficient.

Optimize the embedding of the temporal attention module in the upper encoding/decoding layer to calculate the inter-frame feature correlation weight:

$$A_{t,i} = \text{softmax}\left(\frac{z_i^T z_i}{\sqrt{d_z}}\right) \tag{4}$$

It is used to enhance inter-frame dependencies. Conditional injection adopts the cross-attention mechanism [15]:

$$Attn(q,k,v) = \text{softmax}\left(\frac{qk^T}{\sqrt{d_k}}\right)v \tag{5}$$

Among them, $q = v_{init}$, k,v are the latent features extracted by U-Net.

3. Experimental design

3.1. Experimental dataset

The public data is sourced from the Human3.6M and AIST++ datasets, and a total of 500 gymnastic movement sequences were screened out.

The self-built dataset was obtained through motion capture of professional gymnasts. Ten athletes of different skill levels were recruited, and 15 types of core gymnastic movements were collected. Each type of movement was repeatedly captured 5 times to generate 375 movement sequences. The individual characteristics of the athletes were simultaneously recorded as personalized labels.

Finally, an experimental dataset containing 875 action sequences was constructed and divided into a training set and a test set in a 7:3 ratio.

3.2. Design of the evaluation index system

The specific indicators are shown in Table 2.

DOI: 10.54254/2755-2721/2026.31160

Table 2. Evaluation indicators

Indicator Type	Indicator Name (Abbreviation)	Core Function
Generation Quality	Fréchet Inception Distance (FID)	Evaluate the distribution similarity between generated and real videos
	Inception Score (IS)	Evaluate the diversity of generated content
Motion Coherence	Percentage of Correct Keypoints (PCK)	Evaluate keypoint localization accuracy
	Joint Angle Error (JAE)	Measure the smoothness of motion transition between adjacent frames
Personalized Adaptation	Cosine Similarity (CS)	Evaluate the adaptation degree between generated resources and learner profiles
Subjective Rating	Likert 5-point Scale Rating	Calculate the average score based on motion standardization, explanation clarity, personalized adaptation degree and teaching practicality

3.3. Experimental scheme design

The experiment consists of 3 groups of comparative experiments and 1 group of ablation experiments. The core protocol is shown in Table 3.

Table 3. Experimental scheme

Experiment Type	Specific Settings
Comparative Experiment 1	Experimental Group: Proposed Framework; Baseline Models: DDPM, Basic LDM, ActionGAN
Comparative Experiment 2	Generate resources for test samples with different skill levels/physical parameters
Comparative Experiment 3	Statistic the single resource generation time and GPU memory usage of each model
Ablation Experiment	ABL1: Removing Temporal Attention; ABL2: Removing Multimodal Condition Fusion; ABL3: Basic VAE replacing Optimized VAE; ABL4: Removing Personalized Parameter Embedding

4. Experimental results

4.1. Overall performance results

The proposed framework demonstrates significant advantages in terms of generation quality, action coherence, and personalized adaptability.

4.2. Personalized generation of effect results

As shown in Table 4, the CS values of the proposed framework for learners of different skill levels all exceed 0.9, and the subjective scores all exceed 4.5 points, showing a significant improvement compared to the baseline model.

DOI: 10.54254/2755-2721/2026.31160

Table 4. Personalized generation effect

Model	Beginner Learners (CS/Subjective Rating)	Intermediate Learners (CS/Subjective Rating)	Advanced Learners (CS/Subjective Rating)
Basic LDM	0.71/3.2	0.75/3.5	0.72/3.3
ActionGAN	0.69/3.0	0.72/3.3	0.70/3.1
Proposed Framework	0.93/4.7	0.91/4.6	0.92/4.8

4.3. Results of operational efficiency

The comparison of the operational efficiency indicators of each model is shown in Figure 1. The proposed framework ensures performance while maintaining an efficient generation capacity. This stems from the efficiency advantages of optimizing VAE's efficient compression of high-dimensional action sequences and its potential spatial diffusion mechanism, achieving a coordinated balance between performance and efficiency, and providing feasibility for large-scale application in actual teaching scenarios.

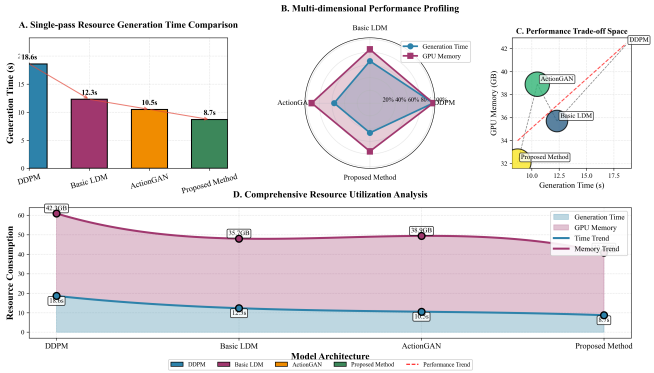

Figure 1. Shows the results of operational efficiency

4.4. Ablation experiment results

As shown in Figure 2, the CS values of ABL2 and ABL4 have dropped sharply by more than 0.2, indicating that these two modules are the core for ensuring personalized adaptation. The decrease in ABL1 and PCK while the increase in FID confirm their crucial role in the generation quality. The performance of ABL3 has significantly declined. The synergy of each core module is the guarantee of the framework's high performance.

DOI: 10.54254/2755-2721/2026.31160

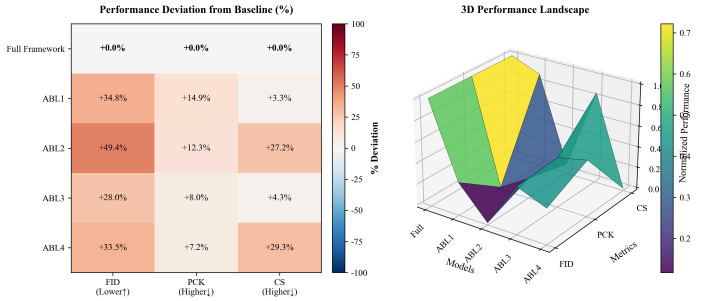

Figure 2. Results of the ablation experiment

5. Conclusion

This study addresses the pain point of insufficient personalized resource supply in gymnastics teaching and designs an AIGC content generation framework for gymnastics teaching based on the diffusion model, achieving efficient generation of personalized teaching resources. The experimental results confirm that this framework significantly outperforms the traditional baseline model in terms of generation quality, action coherence and personalized adaptability. The research not only provides a practical and feasible technical solution for the digital transformation of gymnastics teaching, but also enriches the application paradigms of AIGC in the vertical education field, laying a foundation for subsequent research on the intelligent generation of teaching content for multiple sports.

References

[1] Shutova, T. N., & Andryushchenko, L. B. (2020). Digitalization of physical education and sports educational process at university. Theory and Practice of Physical Culture, (9), 68-70.
[2] Wu, F., Lu, C., Zhu, M., Chen, H., Zhu, J., Yu, K., ... & Pan, Y. (2020). Towards a new generation of artificial intelligence in China. Nature Machine Intelligence, 2(6), 312-316.
[3] Tang, S. (2025, May). Technology and Education Integration": Reconstruction of the Path of Classroom Teaching Empowered by Intelligent Educational Technology—The Rational Use of Artificial Intelligence Based Technology. In 2025 5th International Conference on Artificial Intelligence and Education (ICAIE) (pp. 845-851). IEEE.
[4] Deng, Z., & Yu, Y. (2025, April). Cultivation Fueled by AI Innovations. In Proceedings of the 2024 3rd International Conference on Educational Science and Social Culture (ESSC 2024) (Vol. 914, p. 345). Springer Nature.
[5] Yang, J. (2025). Curriculum reform in China: physical educators' engagements with, enactments of and reflections on the new Physical Education and Health curriculum.
[6] Vlieghe, J. (2013). Physical education beyond sportification and biopolitics: An untimely defense of Swedish gymnastics. Sport, Education and Society, 18(3), 277-291.
[7] Chen, Z., Sun, W., Tian, Y., Jia, J., Zhang, Z., Jiarui, W., ... & Zhang, W. (2024). Gaia: Rethinking action quality assessment for ai-generated videos. Advances in Neural Information Processing Systems, 37, 40111-40144.
[8] Xing, Z., Feng, Q., Chen, H., Dai, Q., Hu, H., Xu, H., ... & Jiang, Y. G. (2024). A survey on video diffusion models. ACM Computing Surveys, 57(2), 1-42.
[9] Ji, B., Pan, Y., Liu, Z., Tan, S., & Yang, X. (2025). Sport: From zero-shot prompts to real-time motion generation. IEEE Transactions on Visualization and Computer Graphics.
[10] Li, J., Wang, H., Li, Y., & Zhang, H. (2025). A Comprehensive Review of Image Restoration Research Based on Diffusion Models. Mathematics, 13(13), 2079.
[11] Ma, Z., Zhang, Y., Jia, G., Zhao, L., Ma, Y., Ma, M., ... & Zhou, B. (2025). Efficient diffusion models: A comprehensive survey from principles to practices. IEEE Transactions on Pattern Analysis and Machine Intelligence.
[12] Cao, P., Zhou, F., Song, Q., & Yang, L. (2024). Controllable generation with text-to-image diffusion models: A survey. arXiv preprint arXiv: 2403.04279.

DOI: 10.54254/2755-2721/2026.31160

[13] Loi, I., Zacharaki, E. I., & Moustakas, K. (2023). Machine learning approaches for 3d motion synthesis and musculoskeletal dynamics estimation: a survey. IEEE transactions on visualization and computer graphics, 30(8), 5810-5829.
[14] Yang, Q., Ye, M., & Tao, D. (2024, September). Synergy of sight and semantics: visual intention understanding with clip. In European Conference on Computer Vision (pp. 144-160). Cham: Springer Nature Switzerland.
[15] Tian, Q., Liu, X., Zhou, J., Zheng, Y., Wan, J., & Lei, Z. (2025). Cross-Attention with Conditional Matching for Multi-Target Domain Adaptation. IEEE Transactions on Circuits and Systems for Video Technology.