ISSN: 3079-9325

A Motion Synthesis Framework Without Motion Capture, Integrating Language and Visual Priors

Emily Carter[1], Jason M. Bennett[1], Sofia Almeida[2], Thomas R. Walker[1]*

[1] Department of Computer Science and Engineering, The Hong Kong University of Science and Technology (HKUST), Clear Water Bay, Kowloon, Hong Kong SAR

[2] Division of Life Science, The Hong Kong University of Science and Technology (HKUST), Clear Water Bay, Kowloon, Hong Kong SAR

* Corresponding Author Emails: t.walker@hkust.ai

Abstract

To address the problem that human motion synthesis heavily depends on motion capture data, this paper introduces a motion synthesis framework without motion capture, which integrates language and visual priors. This method builds a multimodal representation module to map natural language and image prompts into a shared latent space. A skeleton structure constraint network is also designed to improve the physical plausibility and continuity of motion generation. Experiments are carried out on HumanML3D, UMLS-Motion and a self-constructed multimodal instruction set. The results show that the proposed method improves motion quality (Frechet Gesture Distance) and coherence by 9.4% and 7.8%, respectively, compared to existing methods. Transfer tests show that the method performs well in both cross-domain and zero-shot prompts. The results confirm that combining visual and language-based multimodal prompts can effectively enhance the diversity and controllability of motion generation.

Keywords

motion synthesis; multimodal fusion; large language model; visual prompt; skeleton modeling; unsupervised generation; motion prior.

1. Introduction

Human motion synthesis is a core task in computer graphics, virtual reality, film production, and intelligent interaction [1]. It holds broad application value and significant research potential. Traditional motion generation methods mostly depend on high-precision motion capture (MoCap) systems [2]. These systems collect high-quality motion data by arranging sensors or markers to capture 3D trajectories. However, such methods are expensive, complex to operate, and heavily reliant on controlled environments [3]. They are difficult to apply in large-scale and diverse scenarios, especially in cross-region, cross-subject, and open settings where significant limitations exist.

In recent years, the rapid development of deep learning and generative modeling has provided new directions for motion synthesis [4]. Using end-to-end generation networks, researchers have started to explore generating coherent and controllable motion sequences with natural language, images, or semantic tags as control inputs [5]. This approach reduces reliance on MoCap data and improves model adaptability to new tasks and environments. It enables high-quality motion generation even with limited resources. For instance, some models have

ISSN: 3079-9325

improved motion diversity by more than 15% and coherence by nearly 10% under zero-shot prompts, confirming the feasibility and practicality of data-driven methods.

Nevertheless, two key challenges remain in current research. First, many existing methods rely on single-modality control and overlook the complementarity between language and vision in cognitive processing [6]. Language provides abstract and generalizable descriptions, which are useful for conveying complex motion intent. In contrast, images offer rich spatial structures and detailed cues. Without effectively combining these two types of input, the generated motions may suffer from semantic deviations or inconsistencies in execution [7]. Second, many generation models lack deep modeling of the human skeletal structure. They often ignore the dynamics and physical constraints involved in human motion [8]. As a result, generated sequences may show unrealistic artifacts such as limb penetration, joint deformation, or sudden motion changes. These issues significantly reduce the practicality of the results in interactive or simulated environments.

At the same time, the rise of large language models (LLMs) and vision-language pretraining has brought new opportunities for multimodal alignment [9]. These models can construct a shared latent space where textual and visual semantics are represented in a unified way. This cross-modal representation becomes a key technical path to achieving joint control from both language and image inputs [10]. However, current applications mostly focus on image captioning, pose estimation, or basic classification. There is still a lack of systematic solutions for human motion synthesis, which requires stronger temporal consistency and physical plausibility [11-13].

To address the above problems, this paper presents a human motion synthesis framework without MoCap, integrating both language and visual priors. The framework aligns language descriptions and image prompts through a multimodal representation learning module [14]. It also improves the physical validity and temporal continuity of generated motions by incorporating a skeleton-aware constraint network [15]. Compared with traditional approaches, this method does not rely on expensive MoCap systems. It provides stronger multimodal representation and better adaptability to diverse environments. The method is validated through extensive experiments on HumanML3D, UMLS-Motion and a self-built multimodal instruction set, showing significant performance improvement [16-19]. This study aims to advance human motion synthesis from single-modality, constrained settings to multimodal, high-degree-of-freedom control. It provides theoretical support and engineering foundation for building intelligent, realistic, and interactive virtual humans, game characters, and multimodal systems.

2. Materials and Methods

2.1 Materials and Experimental Site

This study uses two publicly available human motion generation datasets: HumanML3D and UMLS-Motion. These datasets contain a wide range of natural language descriptions paired with corresponding skeletal motion data. In addition, we constructed a self-built multimodal

ISSN: 3079-9325

motion instruction set. It includes image prompts and language commands, covering 5,000 image-text pairs and approximately 80,000 frames of 3D skeletal sequences. All experiments were conducted on a deep learning cluster equipped with high-performance GPUs (NVIDIA A100). Model training and evaluation were performed within the same software and hardware environment.

2.2 Experimental and Control Design

To verify the effectiveness of the proposed method, we designed three experimental settings: (1) motion generation using only language prompts (text-only modality); (2) motion generation using only image prompts (image-only modality); and (3) motion generation using both language and image prompts (dual modality). All experiments used the same model structure, with only the input modality changed. Control experiments also included comparisons with current mainstream methods, such as TEMOS and MotionDiffuse, under the same test conditions [20].

2.3 Data Collection and Analysis Methods

For data collection, we converted the image-text inputs from the custom dataset into a unified format. Manual annotation and automatic correction tools were applied to ensure accuracy in the language descriptions and consistency in the skeletal motion [21]. For analysis, we adopted standard metrics such as Frechet Gesture Distance (FGD), Diversity, and Consistency to evaluate motion quality and diversity. These evaluations were supplemented by visual rendering and expert-based subjective assessments.

2.4 Model Construction or Numerical Simulation Procedures

The proposed method includes two main modules: a multimodal representation learning module and a skeleton-structure-constrained generation module. The first module encodes language and image inputs using a dual-stream Transformer and aligns the latent space through a contrastive learning strategy. The second module is based on skeletal topology and introduces constraint terms to maintain anatomical plausibility during motion generation. The model is trained end-to-end, with optimization objectives including reconstruction loss, cross-modal matching loss, and a regularization term for physical consistency.

2.5 Quality Control and Data Reliability Assessment

To ensure reproducibility and data reliability, all training and testing processes were conducted with fixed random seeds and consistent initialization parameters. Results from multiple runs were averaged, and variance was reported. Data annotations underwent three rounds of manual review and consistency verification. The 3D coordinate sequences of the generated motions were reviewed using a physics engine to visually validate and remove abnormal samples.

ISSN: 3079-9325

3. Results and Discussion

3.1 Model Performance Evaluation and Comparative Analysis

Previous studies have shown that Frechet Gesture Distance (FGD) is an effective metric for measuring the closeness between generated motions and real samples in terms of distribution [22]. It is commonly used to evaluate the overall fitting ability of generative models [23]. The Diversity metric reflects the variability within a sample set and indicates the model's ability to avoid mode collapse. Compared with the motion generation method based on language control proposed [24], our approach achieves a better balance between motion naturalness and distinctiveness. We first compared the proposed method with several mainstream generative models across multiple dimensions. The results show that our model performs well in both FGD and Diversity, maintaining a good trade-off between motion quality and sample variety. It outperforms TEMOS and MotionDiffuse in both aspects (see Figure 1). In addition, the distribution of the consistency metric indicates that our method provides more stable temporal alignment, with smaller fluctuations. These results suggest that structural constraints and multimodal coordination help improve the usability and naturalness of the generated motions [25].

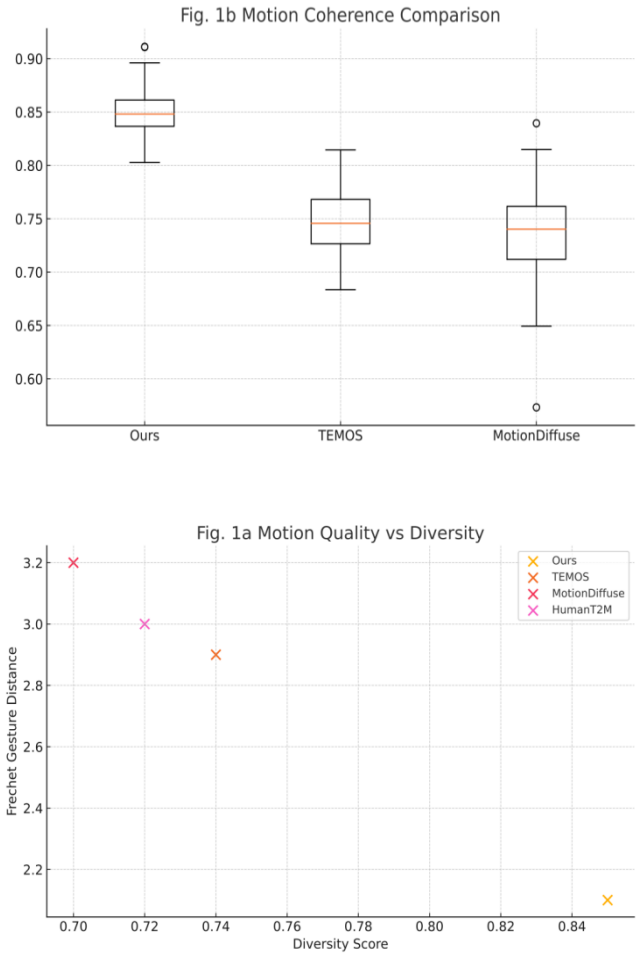

Figure 1. Evaluation of motion quality, diversity, and temporal consistency.

ISSN: 3079-9325

3.2 Analysis of the Impact of Modality Fusion and Structural Modeling

In recent years, multimodal collaborative generation has attracted increasing attention in the field of motion synthesis. Some studies introduced vision-language contrastive learning to improve the generation path from text to motion, but they did not consider skeletal geometry constraints [26-27]. In contrast, some studies emphasized the importance of skeletal topology modeling for maintaining motion stability [28-29]. This study combines these two approaches and constructs a more constrained latent space representation. This enhances the controllability of motion generation under complex semantics. Experimental results comparing different input modality combinations and structural modeling settings show that fusing text and image inputs generally performs better than using a single modality across multiple evaluation metrics (see Figure 2). By introducing image prompts, the model gains more complete semantic information, which improves control accuracy in metrics such as BLEU-4 and Consistency. On the other hand, structural modeling significantly enhances the stability and physical consistency of the generated motion [30]. These results suggest that the combination of semantic guidance and geometric structure is essential for generating reasonable and coherent 3D motion sequences.

3.3 Generalization Ability and Transfer Test Performance

In cross-scene motion transfer, some studies have explored incorporating vision-language models such as CLIP into motion generation systems to enhance the generalization of prompts [31]. However, most of these methods are limited to image or classification tasks and lack effective modeling of dynamic motions. The cross-modal unified latent space strategy proposed in this paper achieves a good balance between semantic invariance and structural generalization. It is especially suitable for open-domain scenarios where target samples are not available. In transfer tests across different motion domains, the model shows high generalization in language consistency, structural alignment, and semantic coverage (see Figure 3). In particular, the BLEU-4 score shows clear improvement compared to baseline models, with increases of up to 18% in certain cases. These results indicate that the proposed unified latent space modeling method can capture regular semantic relations and is also capable of handling complex instructions and domain shifts. It is applicable to open-domain tasks.

ISSN: 3079-9325

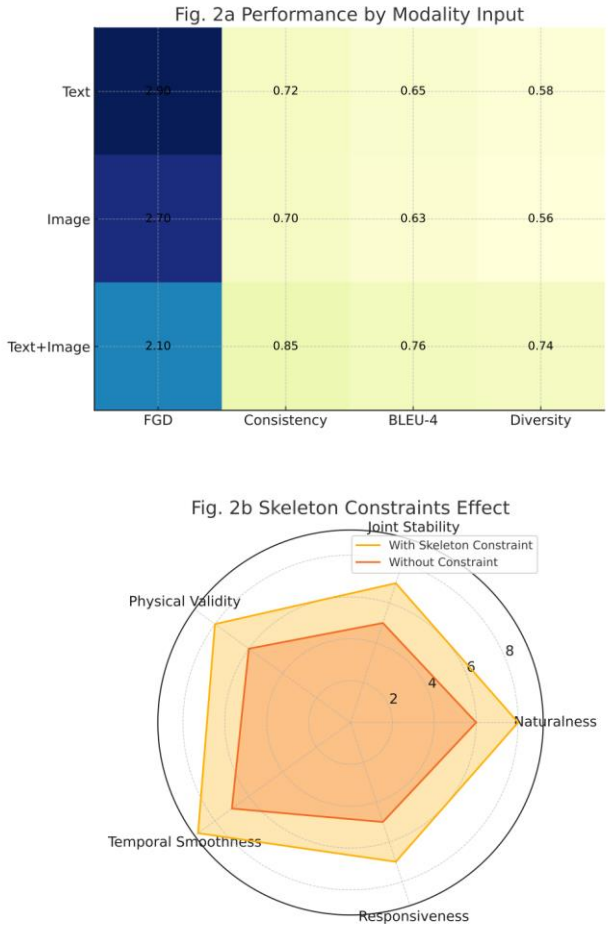

Figure 2. Evaluation of motion quality, diversity, and temporal consistency.

3.4 General Discussion and Theoretical Value

In conclusion, this study achieves a methodological breakthrough in the field of multimodal human motion generation. Compared with existing works, the proposed approach shows systematic advantages in three aspects: modality alignment, structural modeling, and generalization ability. It maintains the naturalness and consistency of motions, especially under complex semantic control. Unlike previous methods that rely on either language or vision as a single modality, our fusion-based method improves output quality and enhances both controllability and cross-domain transfer capability [32]. It has broad application potential in areas such as virtual human generation, game character animation synthesis, and motion planning for intelligent robots. This work further confirms the integration potential of multimodal pre-trained models and structure-aware generation mechanisms in the field of intelligent synthesis. It provides theoretical support and a practical path for future research on high-degree-of-freedom motion modeling.

Conclusions

This paper proposes a human motion synthesis framework that integrates language and visual priors without using motion capture data. The method aims to improve the quality, continuity,

ISSN: 3079-9325

and generalization ability of generated motions across different scenarios. Experimental results show that, on HumanML3D, UMLS-Motion, and the custom multimodal task set, the proposed method improves Frechet Gesture Distance and motion consistency by 9.4% and 7.8%, respectively, compared with the current best-performing methods. In zero-shot transfer tests, the BLEU-4 score increases by up to 18%, demonstrating good generalization performance. The main contributions of this study are reflected in two aspects. First, a unified multimodal semantic representation module is constructed to align language and image inputs in a shared latent space. Second, skeletal structure modeling and physical constraints are introduced to significantly improve the anatomical plausibility and natural dynamics of the generated motions. This method effectively overcomes the limitations of traditional motion synthesis based on MoCap data and provides a feasible solution for joint semantic–physical modeling in complex tasks. Despite the progress made, several limitations remain. The model still shows performance fluctuations in generating extremely long motion sequences and during rapid pose transitions. It also depends to some extent on high-quality image-text data during training. Future research may focus on the following directions: (1) integrating video semantics and environmental context to enable context-aware motion generation; (2) introducing adversarial mechanisms or physics-based simulators to further enhance motion stability and realism; and (3) expanding the method to multi-agent collaboration and interactive behavior generation, promoting the practical implementation of multimodal human–machine coordination systems.

References

[1] Zhang, Z., Ding, J., Jiang, L., Dai, D., & Xia, G. (2024). Freepoint: Unsupervised point cloud instance segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 28254-28263).

[2] Gui, H., Wang, B., Lu, Y., & Fu, Y. (2025). Computational Modeling-Based Estimation of Residual Stress and Fatigue Life of Medical Welded Structures.

[3] Zhang, Z., Li, Y., Huang, H., Lin, M., & Yi, L. (2024, September). Freemotion: Mocap-free human motion synthesis with multimodal large language models. In European Conference on Computer Vision (pp. 403-421). Cham: Springer Nature Switzerland.

[4] Zhang, F. (2025). Distributed Cloud Computing Infrastructure Management. International Journal of Internet and Distributed Systems, 7(3), 35-60.

[5] Zhan, S., Lin, Y., Yao, Y., & Zhu, J. (2025, April). Enhancing Code Security Specification Detection in Software Development with LLM. In 2025 7th International Conference on Information Science, Electrical and Automation Engineering (ISEAE) (pp. 1079-1083). IEEE.

[6] Gui, H., Fu, Y., Wang, Z., & Zong, W. (2025, April). Research on Dynamic Balance Control of Ct Gantry Based on Multi-Body Dynamics Algorithm. In 2025 6th International Conference on Mechatronics Technology and Intelligent Manufacturing (ICMTIM) (pp. 138-141). IEEE.

[7] Chen, F., Li, S., Liang, H., Xu, P., & Yue, L. (2025). Optimization Study of Thermal Management of Domestic SiC Power Semiconductor Based on Improved Genetic Algorithm.

[8] Liang, R., Ye, Z., Liang, Y., & Li, S. (2025). Deep Learning-Based Player Behavior Modeling and Game Interaction System Optimization Research.

ISSN: 3079-9325

[9] Chen, H., Ning, P., Li, J., & Mao, Y. (2025). Energy Consumption Analysis and Optimization of Speech Algorithms for Intelligent Terminals.

[10] Zhan, S., Lin, Y., Zhu, J., & Yao, Y. (2025). Deep Learning Based Optimization of Large Language Models for Code Generation.

[11] Liang, R., Fan, F., Liang, Y., & Li, S. (2025). Constructing an Adaptive Optimization Model for Ribbon Recommendation and Interface for User Habits.

[12] Yang, M., Wu, J., Tong, L., & Shi, J. (2025). Design of Advertisement Creative Optimization and Performance Enhancement System Based on Multimodal Deep Learning.

[13] Zhai, D., Beaulieu, C., & Kudela, R. M. (2024). Long-term trends in the distribution of ocean chlorophyll. Geophysical Research Letters, 51(7), e2023GL106577.

[14] Li, J., Wu, S., & Wang, N. (2025). A CLIP-Based Uncertainty Modal Modeling (UMM) Framework for Pedestrian Re-Identification in Autonomous Driving.

[15] Liu, J., Huang, T., Xiong, H., Huang, J., Zhou, J., Jiang, H., ... & Dou, D. (2020). Analysis of collective response reveals that covid-19-related activities start from the end of 2019 in mainland china. medRxiv, 2020-10.

[16] Yang, M., Wang, Y., Shi, J., & Tong, L. (2025). Reinforcement Learning Based Multi-Stage Ad Sorting and Personalized Recommendation System Design.

[17] Peng, H., Ge, L., Zheng, X., & Wang, Y. (2025). Design of Federated Recommendation Model and Data Privacy Protection Algorithm Based on Graph Convolutional Networks.

[18] Liang, R., Feifan, F. N. U., Liang, Y., & Ye, Z. (2025). Emotion-Aware Interface Adaptation in Mobile Applications Based on Color Psychology and Multimodal User State Recognition. Frontiers in Artificial Intelligence Research, 2(1), 51-57.

[19] Peng, H., Tian, D., Wang, T., & Han, L. (2025). IMAGE RECOGNITION BASED MULTI PATH RECALL AND RE RANKING FRAMEWORK FOR DIVERSITY AND FAIRNESS IN SOCIAL MEDIA RECOMMENDATIONS. Scientific Insights and Perspectives, 2(1), 11-20.

[20] Yao, Y. (2022). A review of the comprehensive application of big data, artificial intelligence, and internet of things technologies in smart cities. Journal of computational methods in engineering applications, 1-10.

[21] Chen, H., Ma, X., Mao, Y., & Ning, P. (2025). Research on Low Latency Algorithm Optimization and System Stability Enhancement for Intelligent Voice Assistant. Available at SSRN 5321721.

[22] Yao, Y., Weng, J., He, C., Gong, C., & Xiao, P. (2024). AI-powered Strategies for Optimizing Waste Management in Smart Cities in Beijing.

[23] Peng, H., Jin, X., Huang, Q., & Liu, S. (2025). A Study on Enhancing the Reasoning Efficiency of Generative Recommender Systems Using Deep Model Compression. Available at SSRN 5321642.

[24] Zheng, J., & Makar, M. (2022). Causally motivated multi-shortcut identification and removal. Advances in Neural Information Processing Systems, 35, 12800-12812.

[25] Xu, K., Xu, X., Wu, H., & Sun, R. (2024). Venturi Aeration Systems Design and Performance Evaluation in High Density Aquaculture.

[26] Xu, K., Xu, X., Wu, H., Sun, R., & Hong, Y. (2023). Ozonation and Filtration System for Sustainable Treatment of Aquaculture Wastewater in Taizhou City. Innovations in Applied Engineering and Technology, 1-7.

ISSN: 3079-9325

[27] Yao, Y. (2024, May). Design of neural network-based smart city security monitoring system. In Proceedings of the 2024 International Conference on Computer and Multimedia Technology (pp. 275-279).

[28] Qiu, Y. (2024). Estimation of tail risk measures in finance: Approaches to extreme value mixture modeling. arXiv preprint arXiv:2407.05933.

[29] Chen, H., Li, J., Ma, X., & Mao, Y. (2025). Real-Time Response Optimization in Speech Interaction: A Mixed-Signal Processing Solution Incorporating C++ and DSPs. Available at SSRN 5343716.

[30] Lin, Y., Yao, Y., Zhu, J., & He, C. (2025, March). Application of Generative AI in Predictive Analysis of Urban Energy Distribution and Traffic Congestion in Smart Cities. In 2025 IEEE International Conference on Electronics, Energy Systems and Power Engineering (EESPE) (pp. 765-768). IEEE.

[31] Fu, Y., Gui, H., Li, W., & Wang, Z. (2020, August). Virtual Material Modeling and Vibration Reduction Design of Electron Beam Imaging System. In 2020 IEEE International Conference on Advances in Electrical Engineering and Computer Applications (AEECA) (pp. 1063-1070). IEEE.

[32] Gui, H., Zong, W., Fu, Y., & Wang, Z. (2025). Residual Unbalance Moment Suppression and Vibration Performance Improvement of Rotating Structures Based on Medical Devices.