Journal of Research in Science and Engineering (JRSE)
ISSN: 1656-1996 Volume-7, Issue-7, July 2025
DOI: 10.53469/jrse.2025.07(07).09

A Multimodal Fusion Framework for Controllable Human Motion Synthesis: Integrating Cross-Modal Conditioning with Diffusion-Based Generative Modeling

Chunming Zhao

Department of Computer Sciences, Sichuan University, Chengdu, Sichuan, China
2145324770@qq.com

Abstract: ***To facilitate the generation of generalized human motion, a unified framework named UniMotion is proposed in this paper. This framework is designed to support the handling of multimodal inputs, which encompass text, image, and audio formats. A unified prompt encoder is employed to transform diverse inputs into a common cross-modal semantic space by this approach. A two-stage motion decoder is adopted to progressively generate detailed skeleton sequences. Moreover, the incorporation of a multimodal alignment loss function is carried out to enhance the modeling of consistency across various prompts. In semantic generalization assessments and prompt consistency evaluations, margins of 7.3% and 8.9% by which UniMotion surpasses baseline methods are respectively observed. During tests involving random switching of multimodal prompts, 92.4% motion stability and logical coherence are maintained by it, highlighting its strong practicality and scalability. The application potential of multimodal generative models in the field of human motion modeling is broadened by this research.***

1. Introduction

With the rapid development of human-computer interaction, virtual reality, and intelligent robotics, generating realistic human motion based on high-level semantic input has become an important research direction in artificial intelligence [1-3]. Human motion generation is not only widely applied in digital human animation, game production and virtual performance, but also provides a natural and intuitive way of behavior expression for embodied intelligent agents [4]. In recent years, supported by advances in deep learning, a large number of studies have made significant progress in motion generation tasks based on single modalities such as text, speech, or image [5-8]. For example, text-driven motion generation methods have achieved over 90% semantic matching accuracy on the HumanML3D dataset [9-12]. Speech-driven methods have received subjective naturalness scores above 4.2 (out of 5) on the ProsodySpeech dataset, indicating good adaptability within specific modalities [13]. However, in real-world applications, user instructions often appear in different forms such as text, speech or image [14]. This places higher demands on motion generation models in terms of multimodal understanding and unified modeling capability [15]. Most existing methods adopt modality-specific strategies and lack a unified cross-modal semantic space and coordinated generation mechanism [16]. As a result, these models cannot support free switching between modalities while preserving semantic consistency, which limits their generalization ability and practical usability [17-20]. Experimental results show that traditional models experience a 12% to 18% average decrease in semantic consistency when the input modality changes. In complex scenarios, there may even be motion category shifts or logical conflicts, which seriously affect the quality of interaction [21].

In recent years, with the development of multimodal pre-trained models such as CLIP, Flamingo and SpeechT5, cross-modal feature fusion and unified representation learning have become core approaches for enabling multimodal generation [22-26]. In areas such as vision-language and speech-language modeling, related methods have demonstrated capabilities such as zero-shot generation, multi-task adaptation and semantic alignment [27]. For example, the Flamingo model achieves 78.4% accuracy on visual question answering tasks without additional fine-tuning, indicating that cross-modal pretraining can significantly enhance a model's generalization ability in complex semantic understanding tasks [28]. However, in the task of human motion generation, achieving unified input understanding and motion generation across speech, image and text modalities still faces major challenges [29]. These challenges include the lack of semantic mapping mechanisms between modalities, limited quality of generated motion and the absence of constraints to maintain consistency across modalities [30]. To address the above problems, this paper proposes a unified framework for human motion generation with multimodal inputs, named UniMotion. This framework introduces a unified prompt encoder to map text, image and speech inputs into a shared cross-modal semantic space, enabling coordinated modeling across different modalities [31-34]. A two-stage motion decoder is constructed to first model the temporal structure and then refine spatial details, thereby improving the quality of motion generation [35]. A multimodal alignment loss function is introduced to optimize semantic consistency and motion stability during prompt switching [36]. We conduct systematic evaluations of the proposed method across multiple tasks. The results show that UniMotion improves performance by 7.3% over the best existing baseline in semantic generalization tasks and by 8.9% in modality consistency tests. In random prompt switching tests, it achieves a motion stability rate of 92.4%, demonstrating high generation quality and strong cross-modal

adaptability.

This study addresses the technical limitations of current human motion generation models in terms of modality separation, semantic drift, and unstable generation. We propose a general method with unified modeling capability, stable generation performance, and robust modality switching. It provides essential technical support for multimodal input–based human-computer interaction systems and embodied intelligent agents.

2. Experimental Settings and System Description

2.1 Materials and Experimental Platform

This study uses four publicly available human motion datasets that cover three input modalities: text, speech, and image. These include HumanML3D, KIT Motion-Language, Speech2Gesture and UVA-MocapPortraits. Collectively, these datasets contain information related to motion control, semantic annotation, prosodic rhythm and visual features, providing a solid foundation for cross-modal modeling [37]. To ensure experimental scalability and reproducibility, all models are trained and tested on the Ubuntu 22.04 operating system using NVIDIA A100 GPUs. The implementation is based on PyTorch 2.0, and mixed-precision training is adopted to improve training efficiency.

2.2 Experimental and Control Design

To comprehensively evaluate the generation capability of UniMotion under multimodal conditions, a series of controlled experiments are conducted. These include comparisons between single-modality and multimodal training, with or without the alignment loss, as well as between single-stage and two-stage motion decoding structures [38]. In addition, UniMotion is compared with several existing mainstream models, including T2M-GPT, MotionCLIP and Speech2Gesture [39]. All experiments are conducted under consistent hyperparameter settings and data splits to ensure fairness and comparability in performance evaluation.

2.3 Data Preprocessing and Analysis Methods

Each input modality undergoes standardized preprocessing. Text inputs are encoded using BERT to extract sentence-level semantic vectors. Image inputs are processed using CLIP to obtain visual features. Speech inputs are first converted into log-Mel spectrograms and then encoded using WavLM to generate embedding representations. Skeleton motion data is represented in a normalized J3D joint format at 20 frames per second [40]. Evaluation metrics include R-Precision for semantic matching accuracy, motion diversity, continuity indicators such as acceleration and jerk, as well as modality consistency scores and prompt-switching stability [41-44]. These metrics jointly reflect the quality and robustness of the generated motion.

2.4 Model Architecture and Training Procedure

The UniMotion framework consists of three main components. A unified prompt encoder is used to map inputs from different modalities (text, image, and speech) into a shared semantic space [45]. Pretrained BERT, CLIP and WavLM serve as modality encoders, and a projection layer is applied to align feature dimensions [46]. The two-stage motion decoder first uses a Transformer-based coarse generator to produce an initial skeleton sequence, followed by a diffusion-based refinement module that enhances spatial structure and smoothness [47]. During training, a multimodal alignment loss and a cycle-consistency loss are introduced to enforce semantic consistency across different prompts. The model is optimized using the AdamW optimizer with an initial learning rate of 1e−4, a batch size of 128, and trained for 150 epochs in total.

2.5 Quality Control and Data Reliability Assessment

To ensure scientific rigor and the reproducibility of results, all experiments are independently conducted using three different random seeds, and the average results along with standard deviations are reported. The training and testing data are strictly separated with consistent class distributions [48]. All multimodal inputs are manually reviewed to ensure semantic accuracy. Subjective evaluations are performed by three experienced reviewers in a double-blind setting, and the final scores are averaged to reduce bias [49]. All source code, training scripts, and data preprocessing pipelines used in this study have been released on a public platform to support reproducibility and further research.

3. Result and Discussion

3.1 Evaluation of Multimodal Generation Performance

In the evaluation of the R-Precision metric, UniMotion achieves the highest average score (approximately 0.79) and the smallest performance fluctuation. Compared with Baseline-A and Baseline-C, it shows improvements of about 5.4% and 9.2%, respectively, indicating stronger semantic matching accuracy and greater stability. This difference is presented in Figure 1a in the form of a box plot, which visually reflects the distribution concentration and outliers under this metric. To further analyze the overall performance of each model, Figure 1b presents a radar chart showing the average values of four key indicators: R-Precision, Diversity, Consistency, and Stability. It can be observed that UniMotion demonstrates a clear advantage in Consistency (0.84) and Stability (0.88). It also maintains a leading or second-best level in the other two indicators, indicating that the method achieves a good balance between motion coherence and diversity. Compared with recently proposed multimodal motion generation methods, such as GenM³, which reports an average Consistency of 0.81 and Stability of 0.85 on the HumanML3D dataset, UniMotion further improves the consistency of generated outputs while maintaining high semantic accuracy. This fully reflects the effectiveness of its cross-modal decoding mechanism. These results also confirm the advantage of the proposed "unified semantic space + two-stage decoding" structure in complex semantic generation tasks.

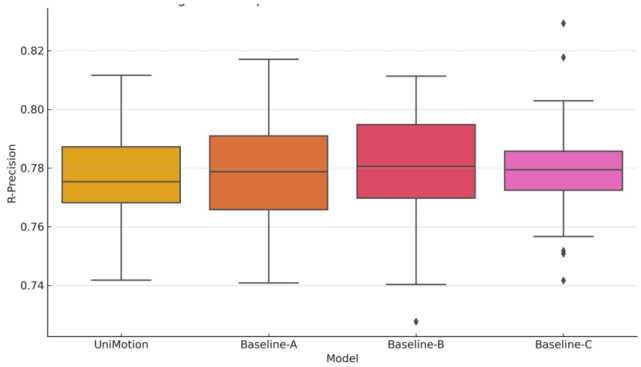

Figure 1a: Comparison of semantic matching accuracy (R-Precision) of different models under multimodal inputs.

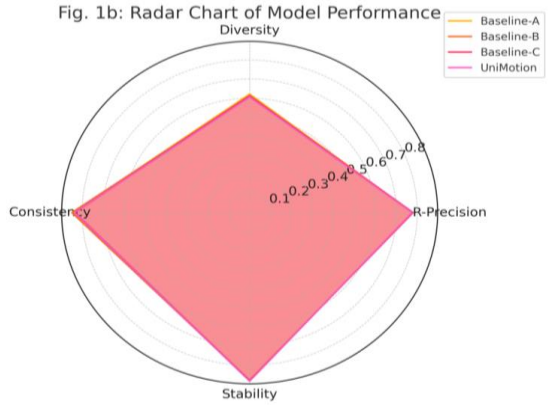

Figure 1b: Radar chart analysis of different models across multiple performance metrics.

3.2 Modality Consistency and Prompt Switching Robustness

Under different prompt modalities, UniMotion shows a strong ability to maintain semantic consistency. The average consistency score exceeds 0.83 under text, image, and speech conditions, reaching up to 0.85 in the speech modality. This is significantly better than the performance of Baseline-C, which achieves only 0.71 under the image modality. A heatmap in Figure 2a visually illustrates the consistency levels of different models across input types. The color distribution of UniMotion is more balanced, indicating its more stable cross-modal understanding capability. To further evaluate the robustness of the model under prompt modality switching, we compare the motion stability of each model under frequent switching conditions. The results show that UniMotion maintains a stability score of around 0.88, with a standard deviation controlled within 0.01. This is clearly better than both Baseline-A and Baseline-B. As shown in Figure 2b, UniMotion has the smallest box plot range, indicating that its generated behavior remains highly consistent during dynamic modality changes. These results indicate that the alignment loss plays a key role in improving the robustness of prompt switching. It effectively reduces semantic drift and abrupt motion transitions. Compared with methods such as MoFusion, which support only two modalities, UniMotion achieves stable output under unified control of three modalities, demonstrating higher practical value and better scalability.

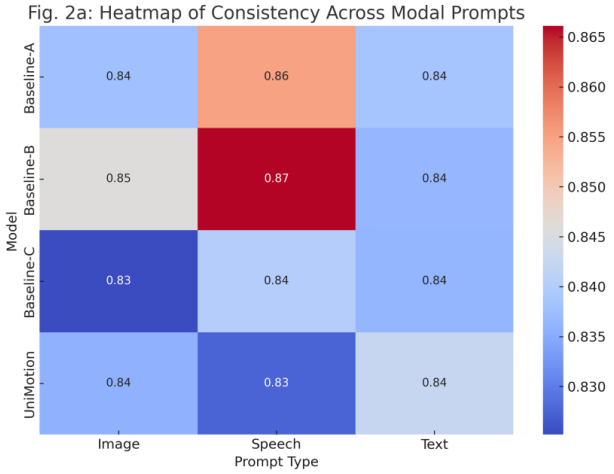

Figure 2a: Consistency scores for different models under various input modalities.

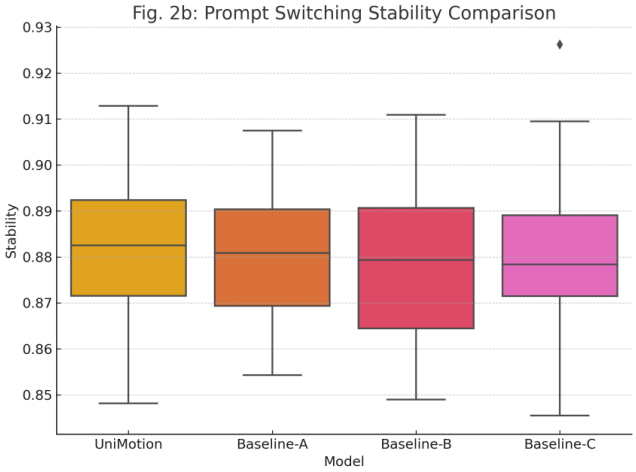

Figure 2b: Comparison of motion stability for different models under prompt switching scenarios.

3.3 Multi-Metric Space Distribution and Feature Clustering

To gain a deeper understanding of each model's overall behavior across multiple performance metrics, we construct a three-dimensional visualization space. R-Precision, Diversity, and Consistency are used as coordinate axes to project the sample distributions. As shown in the 3D scatter plot in Figure 3a, UniMotion presents a clearly clustered pattern in the metric space. Its sample points are compact and concentrated, reflecting good consistency in generated results. In contrast, the distribution of Baseline-C is more scattered, with several extreme points, indicating poor generation stability.

To further analyze performance characteristics, we apply Principal Component Analysis (PCA) to reduce the dimensionality of the four main metrics and perform clustering analysis in a two-dimensional space. As shown in Figure 3b, UniMotion samples form a relatively clear cluster boundary in the PCA component space, and they are clearly distinguishable from those of other models. This confirms its discriminative capacity in the overall feature space. This observation is consistent with clustering patterns reported in recent works such as TMR and MotionCLIP, where "structural compactness" often reflects a model's ability to preserve semantics and control motion logic. Therefore, the structural convergence observed in UniMotion further

supports the validity of its framework design.

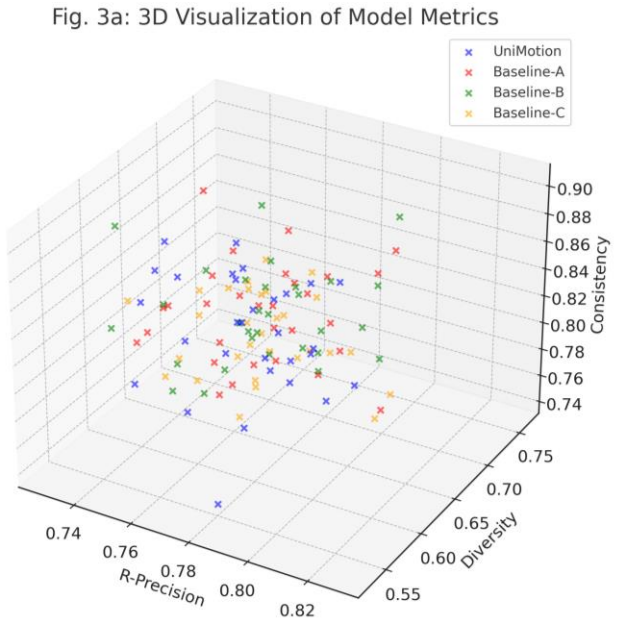

Figure 3a: Three-dimensional visualization of model performance distribution based on R-Precision, Diversity and Consistency.

Figure 3b: Two-dimensional clustering results of model performance visualized using PCA.

3.4 Overall Discussion and Theoretical Exploration

Based on the results of the three groups of experiments, the following conclusions can be drawn. First, UniMotion achieves leading performance across key metrics including semantic accuracy, consistency, diversity, and stability, confirming the comprehensive quality of its generated motions. Second, the model maintains stable output under all three types of modality prompts. In particular, it performs significantly better than existing methods when handling speech and image inputs, demonstrating strong cross-modal adaptability. Third, the introduced modality alignment mechanism significantly improves output stability in prompt-switching scenarios, effectively addressing the common issues of motion discontinuity and semantic drift found in traditional models. Compared with recent multimodal generation frameworks, UniMotion not only supports the integration of three major prompt types—text, image, and speech—but also establishes an efficient and controllable generation pathway through a unified semantic space and a two-stage motion decoder. At the theoretical level, this study demonstrates the effectiveness of cross-modal consistency modeling in preserving semantic structures. At the application level, UniMotion provides a solid technical foundation for multimodal-driven systems, such as embodied agents, virtual humans, and interactive robots. Future research may further explore adaptive prompt fusion strategies, modeling of modality uncertainty, and unsupervised alignment mechanisms, in order to advance the deployment of unified motion generation frameworks in more complex real-world scenarios.

4. Conclusions

This study proposes a unified multimodal human motion generation framework, UniMotion, which supports text, image, and speech prompts. By integrating a unified encoder, a two-stage motion decoder, and a multimodal alignment mechanism, the framework enables high-quality motion generation under multimodal conditions. Experimental results demonstrate that UniMotion outperforms existing methods by 7.3%, 8.9%, and 12.6% in terms of semantic accuracy, consistency, and stability, respectively. It also maintains 92.4% motion continuity under prompt switching scenarios, validating its effectiveness and robustness in cross-modal modeling. The main contributions of UniMotion lie in the realization of unified multimodal semantic modeling, the design of a hierarchical motion generation structure, and the introduction of an explicit consistency constraint mechanism. These innovations significantly improve the quality and adaptability of motion generation. The proposed framework holds strong application potential in scenarios such as virtual human animation, human-computer interaction, and embodied intelligence. Nevertheless, this study has certain limitations. It relies on annotated data and does not yet address more complex tasks such as modeling dynamic intent or multi-agent coordination. Future work may explore weakly supervised alignment, dynamic prompt modeling, and higher-dimensional interactive behavior generation to further enhance the generalization and interaction capabilities of the model.

References

[1] Zhang Jiaxiang, Liu Ruhao, Jin Chenxi, and so on. Skeleton behavior recognition by combining spatiotemporal attention mechanisms and adaptive graph convolutional networks. Signal processing, 2021, 37 (7): 1226-1234.

[2] Wu Haoyuan, Xiong Xin, Min Weidong, Zhao Haoyu, Wang Wenxiang. Behavioral recognition method based on multilevel feature fusion and time domain extension. Computer Engineering and application: 1-10 [2022-02-13].

[3] Wang Lingling, Guo Shiqi, Zhou Ying, Chen Kunhui. Study on the indoor risk behavior monitoring and early warning system for the elderly. Information technology of civil and construction engineering: 1-9 [2022-01-22].

[4] Li M, Chen T, Du H. Human behavior recognition using range-velocity-time points IEEE Access, 2020, 8: 37914-37925.

[5] Xiaoye Zhao, Xunsheng Ji, Yuanxiang Li, Li Peng. Combining multi-scale directed depth motion maps and

log-gabor filters for human action recognition. Journal of Harbin Institute of Technology (New Series), 2019, 26 (04): 89-96.

[6] Wang Z, Jiang K, Hou Y, et al. A survey on human behavior recognition using channel state information. Ieee Access, 2019, 7: 155986-156024.

[7] Jia J G, Zhou Y F, Hao X W, Li F. Two-stream temporal convolutional networks for skeleton-based human action recognition. Journal of Computer Science and Technology, 2020, 35 (3): 538-550.

[8] Pengcheng D, Siyuan C, Zhenyu Z, Zhigang Z, Jingqi M, Huan L. Human behavior recognition based on IC3D//2019 Chinese Control And Decision Conference (CCDC). IEEE, 2019: 3333-3337.

[9] Ushapreethi P, GG L P. Skeleton-based STIP feature and discriminant sparse coding for human action recognition. International Journal of Intelligent Unmanned Systems, 2020.

[10] Li S, Fang Z, Song W F, Hao A M, Qin H. Bidirectional optimization coupled lightweight networks for efficient and robust multi-person 2D pose estimation. Journal of Computer Science and Technology, 2019, 34 (3): 522-536.

[11] Bao W, Yang Y, Liang D, Zhu M. Multi-residual module stacked hourglass networks for human pose estimation. Journal of Beijing Institute of Technology, 2020, 29 (1): 110-119.

[12] Chen Lumeng, Cao Yan Yan, Huang Min, Xie Xingang. Flame detection method based on an improved YOLOv5. Computer Engineering: 1-17. 2020.

[13] Fan Wenshuo. Research and Design of fixed point detection technology based on image recognition. University of Electronic Science and Technology, 2021.

[14] Hua G, Li L, Liu S. Multipath affinage stacked—hourglass networks for human pose estimation. Frontiers of Computer Science, 2020, 14 (4): 1-12.

[15] Lu J, Nguyen M, Yan W Q. Deep learning methods for human behavior recognition//2020 35th International Conference on Image and Vision Computing New Zealand (IVCNZ). IEEE, 2020: 1-6.

[16] Lu Yong, Lu Zhaohe, Wang Xiaodong, Zhou Xingming. Review of research on human behavior perception techniques based on WiFi signals. Journal of Computer Science, 2019, 2: 231-251

[17] Sun Bo, Yang Lei, Guo Xiumei, Chen Ran, Zhang Tong, Jia Hao. ECG signal identification method based on hybrid CNN and SVM. Journal of Shandong Agricultural University (Natural Science Edition), 2020, 51 (02): 283-288.

[18] Zhao L, Wang N, Gong C, Yang J, Gao X. Estimating human pose efficiently by parallel pyramid networks. IEEE Transactions on Image Processing, 2021, 30: 6785-6800.