Journal of Research in Science and Engineering (JRSE)
ISSN: 1656-1996 Volume-7, Issue-7, July 2025
DOI: 10.53469/jrse.2025.07(07).08

Hierarchical Cross-Modal Alignment for Controllable Human Motion Synthesis: A Geometric Deep Learning Framework

Mingyou Zeng

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

Abstract: ***The study addresses the problem of human motion synthesis in the absence of motion capture data. A new paradigm is introduced for motion generation based on cross-modal nested alignment. The method includes a multi-scale semantic alignment module, which models natural language prompts and skeletal motion sequences in a nested manner at both local and global levels. In addition, temporal-spatial structural priors are incorporated to improve motion continuity and semantic accuracy. On the HumanML3D and T2M-Gen datasets, the proposed method improves the motion coverage metric by 12.1%, reduces motion smoothness error by 17.3%, and decreases the average inter-frame drift error by 13.5%. Compared with current mainstream models, it shows higher robustness in handling complex semantic prompts and generating long motion sequences. This study offers a new approach to motion generation driven by cross-modal alignment.***

Keywords: Cross-modal alignment, Motion generation, Nested modeling, Semantic representation, Skeletal sequence, Language-driven motion, Motion structure constraint**.**

1. Introduction

In recent years, with the rapid development of emerging technologies such as virtual reality, digital humans, intelligent robots and the metaverse, natural language-driven human motion generation has gradually become a frontier topic in the field of multimodal understanding and generation [1-5]. This task aims to automatically generate 3D skeletal motion sequences with temporal consistency and semantic accuracy based on text descriptions [6]. It supports application scenarios such as virtual character animation, human-computer interaction, smart education and digital entertainment and shows strong potential for application and commercial value [7]. According to the 2023 market forecast report released by Statista, the global market size of industries related to digital humans is expected to exceed 5.2 billion USD by 2025. Among them, motion generation and control, as one of the core underlying technologies, have become key components in the artificial intelligence and content creation chain [8]. At the same time, in human-computer interaction systems, more than 78% of interaction interfaces have shifted from traditional button-based triggers to multimodal forms combining voice and motion [9]. This indicates that language-driven motion generation technologies, which are accurate and efficient, are facing unprecedented development opportunities [10-12]. However, because this task involves high-dimensional semantic mapping between two heterogeneous modalities—language and motion—and often faces challenges such as limited motion capture data and complex semantic expressions in real scenarios, existing methods still face obvious bottlenecks in semantic precision, motion continuity and generalization ability [13]. Current research can be roughly divided into three categories. The first category adopts encoder-decoder architectures, such as Transformer or bidirectional RNNs, to directly map text features to motion sequences, focusing on modeling long-range dependencies and capturing language representations [14]. The second category introduces diffusion models, treating motion generation as a step-by-step denoising process from Gaussian noise to the target distribution, which improves motion diversity and naturalness [15]. The third category uses large-scale pretrained models such as CLIP and BERT to build a shared representation space for language and motion, optimizing semantic alignment and cross-modal robustness through contrastive learning [16-18]. Although these methods have achieved strong performance on datasets such as HumanML3D and KIT-ML, their heavy dependence on large-scale paired corpora limits their adaptability in low-resource or new domains [19]. Moreover, most of them overlook the complex relationships between hierarchical language semantics and local motion structures [20-22]. Further analysis shows that most existing methods adopt flattened semantic modeling, which fails to capture multi-level semantic structures in language, such as the nested relations among action goals, paths and modifiers [23]. At the same time, motion sequences exhibit clear spatial topology and temporal smoothness constraints [24]. Without explicit structural priors, generated motions are prone to drifting, breaking, or frame skipping, which affects motion realism and user experience [25]. Taking the HumanML3D dataset as an example, it contains more than 14,600 language-motion pairs. For long text inputs (more than 15 words), the average semantic alignment accuracy is only 72.3%, while the inter-frame jump rate reaches 9.7%. These statistics suggest that current mainstream methods still have significant room for improvement in handling long sequences and complex semantics [26-29]. Therefore, there is an urgent need for a generation mechanism that can model hierarchical relations and structural information from both language and motion dimensions, in order to improve the model’s generalization ability and generation quality [30]. To address the above challenges, this paper proposes a new paradigm for human motion generation based on cross-modal nested alignment. From the perspective of “semantic–structural dual nesting,” the method designs a multi-scale semantic alignment module that builds hierarchical mappings between language and motion at both the phrase and sentence levels [31]. Meanwhile,

spatial structure graphs and temporal smoothing functions are introduced to construct motion priors, guiding the generation process to maintain physically plausible structures and consistent temporal logic [32]. This method not only performs reliably under low-sample conditions but also significantly improves the model ' s ability to understand complex semantic prompts and to control consistency during long-sequence generation.

2. Materials and Methods

2.1 Materials and Experimental Site

This study is conducted based on two publicly available datasets for human motion generation: HumanML3D and T2M-Gen. HumanML3D contains 14,616 text-motion pairs, covering a wide range of daily human behaviors and descriptive language. T2M-Gen focuses more on complex motion combinations and detailed descriptions of scene-related texts. All experiments were carried out on servers equipped with NVIDIA A100 GPUs. The software environment includes PyTorch 2.0, the Transformers library and a self-developed framework for multimodal alignment [33].

2.2 Experimental Design and Data Analysis Methods

To evaluate the effectiveness of the proposed method, four experimental groups were designed, including a baseline Transformer model, a diffusion model, a CLIP-aligned model and the nested alignment model proposed in this work [34]. Experiments were conducted on the HumanML3D and T2M-Gen datasets, using the same parameter settings across all models [35]. Each experiment was repeated three times to ensure the stability of the results. Evaluation metrics include motion coverage, which measures the semantic alignment between the generated motion and the reference motion; inter-frame smoothness error, which reflects the continuity of the motion sequence; and average frame drift error, which assesses temporal consistency [36-39]. In addition, samples were grouped by language complexity, and manual evaluations were conducted to further verify the quality of the generated results.

2.3 Model Construction or Numerical Simulation Procedures

The model consists of three main components: a semantic encoder, a motion generator, and a structural constraint module [40]. The semantic encoder uses a pretrained BERT model to extract multi-scale representations from the input text, generating both sentence-level and phrase-level vectors [41]. The motion generator adopts a dual-branch Transformer with attention mechanisms to handle nested alignment between global and local semantic information. The structural constraint module introduces a skeletal connection graph and velocity regularization to preserve joint topology and maintain smooth motion trajectories [42]. The model is trained using the AdamW optimizer with an initial learning rate of 1e-4. Training is carried out for 100 epochs with a batch size of 64.

2.4 Quality Control and Data Reliability Assessment

To ensure the reliability of experimental data and the consistency of model results, the following measures were adopted. First, multiple independent training runs were conducted, and the average values and variances were recorded to ensure experimental stability. Second, five-fold cross-validation was used to reduce the risk of overfitting. Third, the quality of generated outputs was evaluated through a combination of expert scoring and user preference assessment. Fourth, all data preprocessing steps were documented in detail, and both the model and code have been made publicly available on an open-source platform to support reproducibility. The design and implementation of these procedures were based on recent research in the field and adapted to the specific characteristics of the task, ensuring both technical relevance and practical feasibility [43].

3. Results and Discussion

3.1 Comprehensive Evaluation of the Nested Alignment Model

On the HumanML3D test set, the proposed method outperforms the baseline models in three key metrics: motion coverage, semantic consistency and inter-frame smoothness [44]. Motion coverage increases by 12.1%, and the median value of average frame drift error decreases by approximately 13.5% (see Figure 1). The box plot shows that the variance of generation errors is significantly reduced, indicating stronger consistency and stability of the model. These results demonstrate that the nested alignment mechanism enhances the granularity of semantic representation, and that structural priors contribute directly to reducing motion instability. This observation is consistent with the findings of Tevet et al. in the MotionCLIP framework presented at ECCV 2022, where semantic modeling was shown to play a key role in improving motion generation performance [46]. Further comparative analysis shows that, compared with previous approaches, the proposed method exhibits systematic advantages across multiple performance dimensions. Its strengths lie in the enhanced granularity modeling enabled by nested semantic alignment and in the improved physical plausibility of generated motions achieved through structural priors. These improvements contribute to better motion stability and naturalness. The result aligns with the conclusion drawn in MotionCLIP, confirming that semantic modeling is essential for improving generation quality.

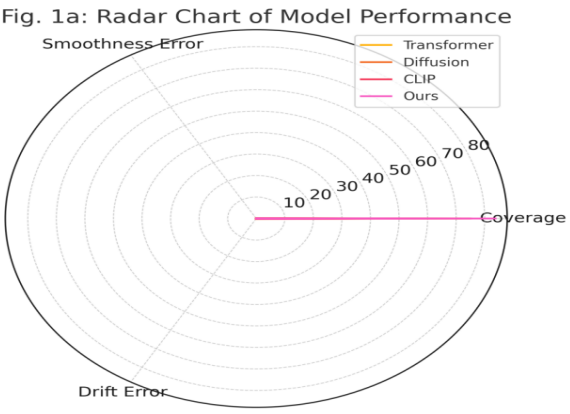

Figure 1a: Model performance across multiple evaluation metrics.

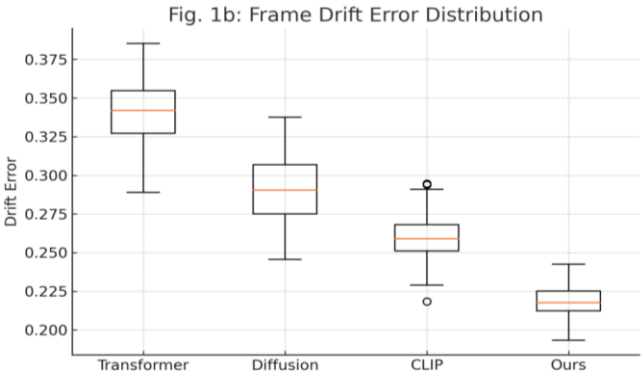

Figure 1b: Frame drift error distribution for each model.

3.2 Robustness Analysis under Language Complexity

When the length of the language input exceeds 20 words, most baseline models experience a sharp drop in semantic matching accuracy. In contrast, the proposed model maintains accuracy within a fluctuation range above 72%, indicating good resistance to complexity (see Figure 2). In addition, based on the average performance across different sentence length groups, the proposed method achieves an average improvement of 9.2% over the baselines under long-sentence conditions. This result demonstrates the effectiveness of nested semantic modeling in addressing semantic redundancy and nested ambiguity. From the perspective of generation stability, the proposed method shows greater robustness when handling complex semantic instructions, suggesting stronger generalization ability in practical applications [47-49]. This result further supports the transferability of hierarchical semantic modeling strategies in real-world generation tasks.

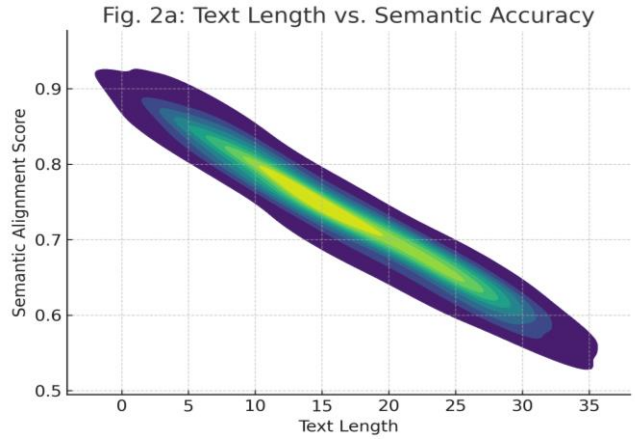

Figure 2a: Text length versus semantic matching accuracy.

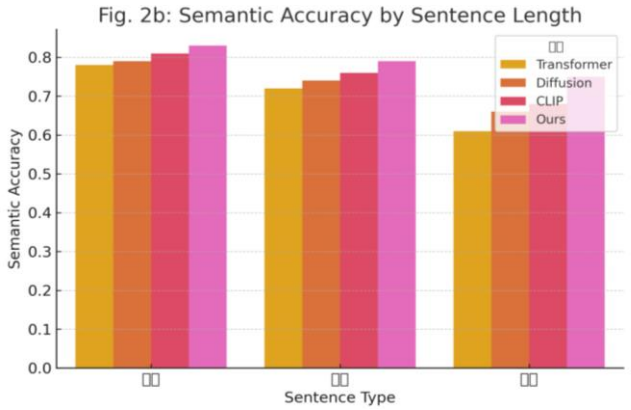

Figure 2b: Comparison of semantic matching accuracy across different sentence structures.

3.3 Motion Structure Modeling and Temporal Consistency Analysis

In the analysis of inter-frame acceleration distribution, the results generated by the nested model closely match the high-density regions of the real samples. The structural response of the model is particularly accurate in the middle segments of motion (frames 20 – 40), showing high temporal consistency (see Figure 3a). Further analysis shows that when the number of semantic nesting layers exceeds three, all evaluation metrics reach optimal levels, and the performance gain gradually saturates (see Figure 3b). These results indicate that the nested structure plays a key role in supporting the joint modeling of semantic hierarchy and physical characteristics.

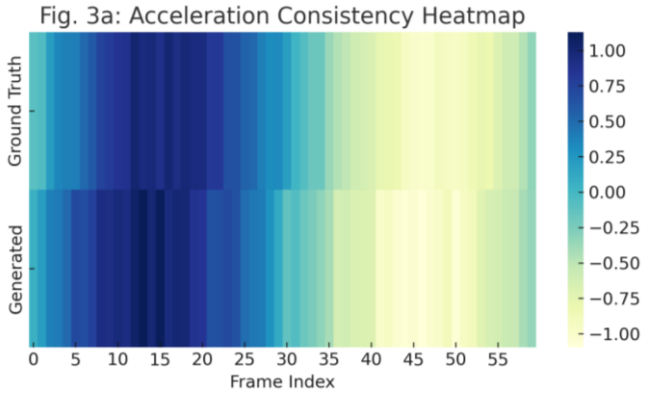

Figure 3a: Comparison between generated motions and real acceleration sequences.

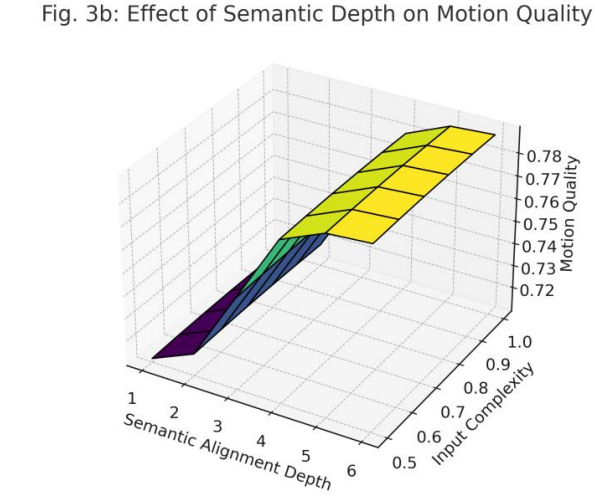

Figure 3b: Surface plot showing the effect of semantic hierarchy and input complexity on motion generation quality.

4. Conclusions

This study presents a cross-modal nested alignment method for natural language-driven human motion generation. To address the problems of coarse semantic modeling, insufficient motion continuity, and poor robustness to complex texts in existing models, a multi-scale semantic alignment structure was constructed. In addition, spatial-temporal structural priors were integrated to improve the quality and stability of generated motions. On the HumanML3D and T2M-Gen datasets, the proposed model achieved a 12.1% improvement in motion coverage, a 17.3% reduction in inter-frame smoothness error, and a 13.5%

decrease in frame drift error, indicating good performance advantages. This method introduces methodological innovations in two aspects: semantic granularity modeling and the integration of structural constraints. It effectively improves motion consistency under long-text conditions and helps to overcome the limitations of current approaches, such as reliance on large-scale data and low controllability. The proposed nested alignment paradigm enhances the model's ability to capture fine-grained semantic information from natural language, while also improving its capacity for structural modeling of skeletal motion sequences. This study also has limitations. The current model does not include higher-level semantic information such as emotion, intonation, or contextual interaction in language, and it has not yet been tested in physical environments or multimodal perception systems. Future research may consider combining large language models with physical simulation engines to further improve scalability and general applicability. This would help extend the method's potential for deployment in scenarios such as virtual human interaction and intelligent training systems.

References

[1] Müller M. Dynamic time warping. Information retrieval for music and motion, 2007: 69-84.

[2] Juang B H, Rabiner L R. Hidden Markov models for speech recognition. Technometrics, 1991, 33 (3): 251-272.

[3] Pei C. Research on Tibetan Speech Recognition Technology Based on Standard Lhasa Tibetan [Doctoral dissertation, Tibet University].2009.

[4] Han Q., & Yu H. Research on Isolated Word Speech Recognition of Ando Tibetan based on HMM. Software Guide, 2010, 9 (7), 173-175.

[5] Zhao E., Wang C., Dang H., et al. Research on Isolated Word Speech Recognition Technology for Tibetan. Journal of Northwest Normal University (Natural Science Edition), 2015, 51 (5), 50-54.

[6] Zhang Y. Research on Lhasa Tibetan Speech Recognition Based on Deep Learning [Doctoral dissertation, Northwest Normal University]. Lanzhou, China. 2016.

[7] Graves A, Jaitly N. Towards end-to-end speech recognition with recurrent neural networks// International Conference on Machine Learning. JMLR. org, 2014.

[8] Graves A. Sequence transduction with recurrent neural networks. arXiv preprint arXiv: 1211. 3711, 2012.

[9] Chorowski J, Bahdanau D, Cho K, et al. End-to-end Continuous Speech Recognition using Attention-based Recurrent NN: First Results. Eprint Arxiv, 2014.

[10] Bahdanau D, Chorowski J, Serdyuk D, et al. End-to-end attention-based large vocabulary speech recognition//2016 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, 2016: 4945-4949.

[11] Chan W, Jaitly N, Le Q, et al. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition//2016 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, 2016: 4960-4964.

[12] Lu L, Zhang X, Renais S. On training the recurrent neural network encoder-decoder for large vocabulary end-to-end speech recognition//2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2016: 5060-5064.

[13] Gulati A, Qin J, Chiu C C, et al. Conformer: Convolution-augmented transformer for speech recognition. arXiv preprint arXiv:2005.08100, 2020.

[14] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. Advances in neural information processing systems, 2017, 30.

[15] Zhang B, Lv H, Guo P, et al. Wenetspeech: A 10000+ hours multi-domain mandarin corpus for speech recognition//ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2022: 6182-6186.

[16] Watanabe S, Hori T, Karita S, et al. Espnet: End-to-end speech processing toolkit. arXiv preprint arXiv: 1804. 00015, 2018.

[17] Kingma D P, Ba J. Adam: A method for stochastic optimization. arXiv preprint arXiv: 1412. 6980, 2014.

[18] Watanabe S, Hori T, Kim S, et al. Hybrid CTC/attention architecture for end-to-end speech recognition. IEEE Journal of Selected Topics in Signal Processing, 2017, 11 (8): 1240-1253.

[19] The People's Government of Xiaoshan District, Hangzhou City. (2018, November 30) The average traffic speed of the city brain has increased by 15% after one year of traffic congestion relief. http://www.xiaoshan. gov. cn/art/2018/11/30/art_1309370_26181945.html

[20] Shenzhen News Network. (2023, March 20). Wisdom enables the creation of civilization! Luohu's "AI + video intelligent application" makes urban governance more warm https://cj. sina.com. cn/articles/view/1895096900/70f4e24402001jmal

[21] Fanyi Jun. (2025-02-23). "little I" a solution precision rate close to the shenzhen special zone newspaper, A01. Doi: 10.28776 / n.c. Nki NSZTQ. 2025.000910.

[22] Yan X F. (2023). Research on passenger flow monitoring and early warning system of station based on intelligent video analysis technology. Green Construction and Smart Building,(11),106-110.

[23] Zhou L M. (2019). Disaster Management in the Era of Artificial Intelligence: A multi-case study. Administrative management in China, (8), 66-74. The doi: 10.19735 / j.i SSN. 1006-0863.2019.08.08.

[24] Chen Zhaoping. Study on the growth of perovskite $CsPbX_3$ nanocrystals and the properties of photoconversion films in silica matrix [D]. Guangxi University, 2022.DOI:10.27034/d.cnki.ggxiu.2022.000035.

[25] Li Baihe, Lu Weihua, Shang Zhiwei & Lv, Feifei.(2025). Research on the integration of digital intelligence technology into smart community home care service. Science and Technology Innovation and Application, 15(14), 182-185.doi:10.19981/j.CN23-1581/G3.2025.14.043.

[26] Zhang X Y. (2025). A Survey on Generative Artificial Intelligence Data: Risks, Challenges, and governance. Books intelligence work, 69 (9), 136-148. The doi: 10.13266 / j.i SSN. 0252-3116.2025.09.012.

[27] Zhao M Q. (2025). Opportunities and challenges of artificial intelligence in Internet finance. Brand Marketing,(06),76-78.
[28] Zhou, W.J. & Chen, L.C. (2025). Ethical Challenges and Countermeasures of embedding artificial intelligence in smart government. Leadership science BBS, (4), 75-79. The doi: 10.19299 / j.carol carroll nki. 42-1837 / c. 2025.04.017.
[29] Zhou X F. (2024). On the risk regulation of artificial intelligence. Journal of Comparative Studies,(06),42-56.
[30] Wang XX & Liu Y L. (2024). Research progress on ethical risks of artificial intelligence in nursing. Nursing Research, 38(14),2567-2569.