MoLang: Leveraging General-Purpose Language Models for Human Animation

Emanuel Borges Passinato[1], **Walcy Santos Rezende Rios** [1]
Rafael Teixeira Sousa[2], **Arlindo Rodrigues Galvão Filho**[1]

[1]Universidade Federal de Goiás (UFG)
Goiânia – GO – Brazil

[2]Universidade Federal de Mato Grosso
(UFMT)

{emanuel.passinato,walcy.rios}@discente.ufg.br

rafaelsousa@ufmt.br, arlindogalvao@ufg.br

Abstract. *Applications in VR and interactive media increasingly require methods for generating human motion that is both realistic and controllable. This paper introduces MotionLLM, a work-in-progress framework for text-to-motion synthesis that leverages Large Language Models (LLMs). Our approach first tokenizes continuous 3D motion into a discrete sequence using a Residual Vector Quantized Variational Autoencoder (RQ-VAE), adapting the tokenization strategy from MoMask. We then reframe motion generation as an autoregressive language modeling task, where a pre-trained LLM generates motion tokens conditioned on text. We hypothesize that LLMs are well-suited for producing long, coherent motion sequences, offer a scalable architecture, and enable multilingual, multimodal control.*

Resumo. *Aplicações em realidade virtual (VR) e mídias interativas demandam cada vez mais métodos para gerar movimentos humanos que sejam ao mesmo tempo realistas e controláveis. Este artigo apresenta o MotionLLM, um framework, em desenvolvimento, para síntese de movimento a partir de texto que, aproveita o poder dos Large Language Models (LLMs). Nossa abordagem começa tokenizando movimentos 3D contínuos em uma sequência discreta, utilizando um Residual Vector Quantized Variational Autoencoder (RQ-VAE), adaptando a estratégia de tokenização do MoMask. Em seguida, reformulamos a geração de movimento como uma tarefa de modelagem de linguagem autoregressiva, na qual um LLM pré-treinado gera tokens de movimento condicionados ao texto. Nossa hipótese é que LLMs são especialmente adequados para produzir sequências de movimento longas e coerentes, oferecendo uma arquitetura escalável e possibilitando controle multilíngue e multimodal.*

1. Introduction

The emerging fields of virtual reality (VR) and interactive media demands advanced methods for generating realistic and controllable human motion. High-fidelity digital avatars are central to these experiences, enhancing social presence and non-verbal communication [Chen et al. 2024], with motion realism directly affecting user engagement [Wu et al. 2024, Tseng et al. 2023]. Traditional approaches like manual keyframing and motion capture (MoCap) lack scalability and adaptability [Yun et al. 2023, Zhao et al. 2024], producing static animations unsuited for dynamic contexts. Recent advances in deep learning have enabled automated motion synthesis from diverse control modalities such as text and audio [Gong et al. 2023, Zhang et al. 2022], marking a shift toward interactive, data-driven animation pipelines.

Recent work has pushed this further by reframing motion generation as a language translation task, treating motion as a "foreign language" that can be understood and generated by large-scale models [Tseng et al. 2023]. This approach leverages the vast world knowledge and powerful sequence modeling capabilities of Transformers and LLMs. However, many state-of-the-art models rely on custom Transformer architectures trained from scratch for the motion generation task [Guo et al. 2023].

In this work, we explore a compelling alternative: replacing the custom generative model with a general-purpose, pre-trained Large Language Model. Our project, MotionLLM, investigates this hypothesis. We adopt the powerful motion tokenization scheme from MoMask [Guo et al. 2024], which uses a Residual Vector Quantizer (RVQ), but diverge by feeding these tokens into an LLM for autoregressive generation. We hypothesize that this approach offers several advantages:

1. **Long-Sequence Coherence:** LLMs are inherently designed to handle long-range dependencies, a feature critical for generating natural, extended motion sequences.
2. **Architectural Scalability:** We can readily experiment with a wide range of existing LLM architectures, from smaller models (0.5B parameters) to large ones (70B), without redesigning the core framework.
3. **Multimodal and Multilingual Potential:** Pre-trained LLMs possess rich semantic understanding, which may facilitate zero-shot or few-shot generalization to new languages (e.g., training in English, prompting in Portuguese) and easier integration of other modalities like audio.

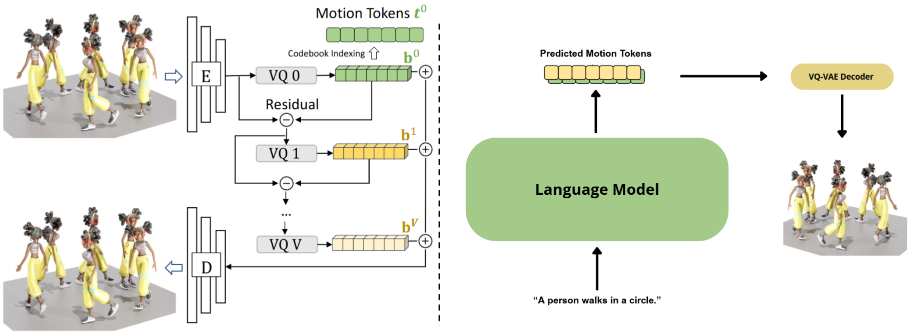
This paper details our proposed methodology, the specifics of our implementation, and our roadmap for evaluation.

2. Related Work

Human motion generation has seen rapid advancements, with comprehensive surveys categorizing tasks, datasets, and model architectures [Zhu et al. 2024, Abootorabi et al. 2025]. Early methods based on RNNs and GANs have largely been replaced by Transformer-based and diffusion-based approaches.

Diffusion Models. Diffusion Probabilistic Models [Ho et al. 2020] have emerged as state-of-the-art in motion generation, producing realistic and diverse results. MotionDiffuse [Zhang et al. 2022] combined diffusion with a Transformer backbone for text-to-motion synthesis. Extensions like FineMoGen [Jiang et al. 2023] and DiffSHEG [Chen et al. 2024] introduced fine-grained editing and multimodal control. However, their iterative inference makes them less suitable for real-time applications.

Motion as Language. Another prominent direction treats motion as a sequence of discrete tokens, typically using a Vector Quantized VAE [Van Den Oord et al. 2017]



(a) Motion Residual VQ-VAE

(b) Motion Inference with LLM

Figura 1. Overview of the MotionLLM pipeline. (a) A continuous 3D motion sequence is converted into discrete motion tokens by the RQ-VAE encoder. (b) Conditioned on a text prompt, an autoregressive LLM, which has been fine-tuned on a vocabulary including these motion tokens, predicts a sequence of new tokens. Finally, the RQ-VAE decoder reconstructs the generated tokens into a 3D animation.

for discretization, followed by an autoregressive model for generation. MotionGPT [Tseng et al. 2023] applied this by fine-tuning an LLM (LLaMA) on multiple motion tasks using prompt-based instructions. MoMask [Guo et al. 2024] introduced a masked modeling approach over motion tokens produced by a Residual VQ-VAE (RQ-VAE), achieving strong performance in generative settings.

Our Approach. MotionLLM builds upon this token-based framework by adopting MoMask's RQ-VAE for motion discretization. Rather than training a Transformer from scratch, we utilize a pre-trained LLM to generate motion tokens conditioned on text. This approach explores the transferability of LLMs to motion synthesis, with potential advantages in scalability, flexibility, and multilingual control.

3. Methodology

Our proposed framework, MotionLLM, consists of two main stages as illustrated in Figure 1: (1) a motion tokenizer based on a Residual Vector Quantized Variational Autoencoder (RQ-VAE) that converts continuous motion data into discrete tokens, and (2) an autoregressive Large Language Model (LLM) that generates these motion tokens conditioned on a text prompt.

3.1. Motion Tokenization with RQ-VAE

To process motion with a language model, we must first convert the continuous, high-dimensional motion sequences into a discrete vocabulary. Following MoMask [**?**], we use an RQ-VAE for this task.

Vector Quantized Variational Autoencoder (VQ-VAE) is an autoencoder that maps continuous inputs to a sequence of discrete latent variables. A standard VQ-VAE consists of an encoder, a discrete codebook (or vocabulary) $E = \{e_k\}_{k=1}^{K}$, and a decoder. The encoder f_{enc} maps an input frame x to a continuous latent vector $z_e(x)$. This vector is then quantized by replacing it with the nearest codebook vector:

$$z_q(x) = e_k, \quad where \quad k = \arg\min_j \|z_e(x) - e_j\|_2$$

The decoder f_{dec} then reconstructs the motion frame from the quantized vector $z_q(x)$.

Residual Vector Quantization (RVQ) enhances this process by using a cascade of quantizers to create a richer, hierarchical representation. Instead of using a single large codebook, RVQ uses multiple smaller codebooks. The first quantizer encodes the input, and subsequent quantizers encode the residual error from the previous stage.

The final representation of the input x is the stack of discrete indices $\{k_1, k_2, \ldots, k_{N_q}\}$ from each quantizer. This allows for a fine-grained representation with a compact set of codes. The full reconstruction is the sum of the selected codebook vectors: $\hat{x} = \sum_{i=1}^{N_q} c_i$.

For our implementation, we use an architecture identical to that in MoMask: a temporal encoder with 3 down-sampling layers (for an 8x reduction in sequence length) and an RVQ with $N_q = 6$ quantizers, each containing a codebook of $K = 512$ codes with dimension $D = 512$.

3.2. Autoregressive Generation with an LLM

Once the motion is tokenized into a sequence of discrete codes, we frame the generation task as a language modeling problem. Given a text prompt P, the goal is to generate a sequence of motion token vectors $M = (m_1, m_2, \ldots, m_T)$, where each m_t is a vector of N_q code indices from the RVQ.

We use an autoregressive approach, where an LLM is fine-tuned to predict the next motion token vector m_{t+1} given the text prompt and the preceding motion tokens:

$$p(M|P) = \prod_{t=1}^{T} p(m_t|P, m_{<t})$$

The input to the LLM is a concatenated sequence of the embedded text prompt and the flattened motion token indices. By leveraging a pre-trained LLM, we hypothesize that the model can utilize its inherent understanding of language and sequence patterns to generate semantically relevant and coherent motions. During inference, we provide the model with a text prompt and a start-of-sequence token, and then iteratively sample the next motion tokens until an end-of-sequence token is generated or a maximum length is reached. The resulting sequence of token indices is then passed to the RQ-VAE's decoder to reconstruct the final 3D motion.

4. Experimental Setup and Future Work

This research is currently a work in progress. This section outlines our experimental setup and the future work planned to validate our hypotheses.

4.1. Implementation and Training Details

We are conducting our experiments on the **HumanML3D** dataset [Guo et al. 2022], a standard benchmark for the text-to-motion task. Our implementation and training details for this initial phase are as follows:

- **Motion Data:** Processed at 20 FPS, with 22 joints. Maximum motion length is 55 tokens, following the original paper's setup after down-sampling.

- **Tokenizer:** The RQ-VAE uses $N_q = 6$ residual quantizers, a codebook size of $K = 512$, and a code dimension of $D = 512$. It is trained first on the motion data alone to create the motion vocabulary.
- **LLM Backbone:** We selected a 1.7B parameter model from the Qwen language model family (Qwen-1.7B) as our initial autoregressive generator.
- **Training Procedure:** The LLM is fine-tuned on text-token pairs for 4 epochs using a batch size of 64 and a constant learning rate of 1×10^{-4}. The training was performed on a single NVIDIA RTX 4090 GPU and completed in approximately 8 hours.

At this stage of our work, we have not yet evaluated inference latency, though we recognize its importance for potential real-time applications.

4.2. Planned Evaluation

As we do not have quantitative results yet, we plan a comprehensive evaluation strategy.

- **Quantitative Metrics:** We will use standard objective metrics from the literature, including Frechet Inception Distance (FID) to measure distribution similarity, R-Precision to evaluate text-motion retrieval accuracy, and Diversity to assess the variety of generated motions for the same text prompt.
- **Qualitative Metrics:** We will conduct user studies to evaluate the perceptual quality, realism, and text-motion alignment of the generated animations, comparing our results against baseline models.

4.3. Future Research Directions

Upon establishing a functional baseline, our future work will focus on exploring our core hypotheses:

1. **Scalability Analysis:** We will systematically evaluate the impact of LLM scale on motion quality by fine-tuning models of varying sizes (e.g., 0.5B, 3B, 7B+) and analyzing the trade-offs between performance and computational cost. The computational cost and inference time off LLMs are gonna be compared with the custom transforms architecture, as it could limit the use of this framework in real time applications.
2. **Multilingual Transfer:** A key experiment will be to test the zero-shot multilingual capabilities of the model. We will fine-tune MotionLLM on the English HumanML3D dataset and then evaluate its performance using prompts translated into other languages, such as Portuguese, without any further training.
3. **Multimodal Extension:** We plan to extend the framework to incorporate other conditioning modalities, such as audio. By adding an audio encoder and fine-tuning the LLM to generate motion tokens conditioned on both text and audio, we can explore tasks like co-speech gesture or dance generation.

5. Conclusion

This paper presented MotionLLM, a work-in-progress framework for text-to-motion generation that combines a Residual Vector Quantized Variational Autoencoder (RQ-VAE) for motion tokenization with a pre-trained Large Language Model (LLM) for autoregressive synthesis. Our approach is motivated by the potential of LLMs to handle long sequences, scale effectively, and facilitate novel applications in multilingual and multimodal control. We have outlined our methodology, which builds upon the tokenization of MoMask while replacing the custom generator with a general-purpose LLM, and detailed our plan for implementation and evaluation on the HumanML3D dataset. We believe this research direction holds significant promise for creating more versatile, controllable, and semantically aware motion generation systems for VR and beyond.

Acknowledgments

This work has been fully/partially funded by the project Research and Development of Algorithms for Construction of Digital Human Technological Components supported by Advanced Knowledge Center in Immersive Technologies (AKCIT), with financial resources from the PPI IoT/Manufatura 4.0 / PPI HardwareBR of the MCTI grant number 057/2023, signed with EMBRAPII.

Referências

Abootorabi, M. M., Ghahroodi, O., Zahraei, P. S., Behzadasl, H., Mirrokni, A., Salimipanah, M., Rasouli, A., Behzadipour, B., Azarnoush, S., Maleki, B., Sadraiye, E., Feriz, K. K., Nahad, M. T., Moghadasi, A., Abianeh, A. E., Nazar, N., Rabiee, H. R., Baghshah, M. S., Ahmadi, M., and Asgari, E. (2025). Generative ai for character animation: A comprehensive survey of techniques, applications, and future directions.

Chen, R., Wang, Z., Jiang, J., Wu, Z., Liu, X., Song, C.-Z., and Liu, F. (2024). Diffsheg: A diffusion-based method for parameterized sign language production. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Gong, Y., Zhao, Z., Zhang, J., Wang, S., Zhu, W., Chen, X., Ma, C., Liu, M., Xu, C., Wen, J., Wu, Y., Chen, C., Yang, J., Jiang, T., Liu, H., Ma, X., and Ci, H. (2023). Text-driven motion generation: Overview, challenges and directions. *arXiv preprint arXiv:2305.09379*.

Guo, C., Mu, Y., Javed, M. G., Wang, S., and Cheng, L. (2024). Momask: Generative masked modeling of 3d human motions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1900–1910.

Guo, C., Wang, X., Zou, S., Zuo, Y., Wang, S., Wu, W., Li, G., and Salsbury, J. K. (2023). Motiongpt: Human motion as a foreign language. *Advances in Neural Information Processing Systems (NeurIPS)*, 36.

Guo, C., Zou, S., Zuo, X., Wang, S., Ji, W., Li, X., and Cheng, L. (2022). Generating diverse and natural 3d human motions from text. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5142–5151.

Ho, J., Jain, A., and Abbeel, P. (2020). Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851.

Jiang, M., Tang, S., Jin, Z., Liu, Z., and Liu, W. (2023). Finemogen: Fine-grained motion generation and editing with spatio-temporal mixture attention. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36.

Tseng, C.-Y., Nakazawa, A., and Harada, T. (2023). Motiongpt: Human motion as a foreign language. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36.

Van Den Oord, A., Vinyals, O., et al. (2017). Neural discrete representation learning. *Advances in neural information processing systems*, 30.

Wu, A., Zhang, K.-Y., Zhang, T.-L., Pan, J.-J., and Zhang, X.-H. (2024). Motioncraft: A unified framework for controllable human motion generation. In *arXiv preprint arXiv:2403.11186*.

Yun, H., Ponton, J. L., Andujar, C., and Pelechano, N. (2023). Animation fidelity in self-avatars: Impact on user performance and sense of agency. In *2023 IEEE Conference Virtual Reality and 3D User Interfaces (VR)*, page 286–296. IEEE.

Zhang, M., Jiang, Z., Liu, S., Zhou, A., Wang, S., and Zhao, Y. (2022). Motiondiffuse: Text-driven human motion generation with diffusion model. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 2516–2525.

Zhao, J., Weng, D., Du, Q., and Tian, Z. (2024). Motion generation review: Exploring deep learning for lifelike animation with manifold.

Zhu, W., Ma, X., Ro, D., Ci, H., Zhang, J., Shi, J., Gao, F., Tian, Q., and Wang, Y. (2024). Human Motion Generation: A Survey . *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 46(04):2430–2449.