Transformer VAE Based 3D Human Action Generation

Sung Jun Park, Joong Hwan Baek

***Abstract*—**In this paper, we propose a Transformer-based Variational Auto-Encoder (VAE) model for generating 3D human action sequences. The method is designed to address issues of action ambiguity and regression to mean poses commonly found in previous GRU-based models. Our model consists of a real action encoder, a generation encoder, and a decoder, which effectively captures both global and local action dependencies through the self-attention mechanism. The proposed model demonstrates significant improvements in recognition accuracy, diversity, and visual quality metrics, achieving the best performance on the HumanAct12 dataset. This work represents a key advancement in the field of 3D human motion generation and offers potential applications in virtual reality and medical simulations.

***Keywords*—**3D Action Generation, Transformer, Variational Auto-Encoder.

I. Introduction

With the recent emergence of large language model (LLM)-based generative AI services such as ChatGPT [1] and Gemini [2], research into image and video generation [3, 4] has gained considerable attention. However, when generating human actions using these models, distortions in realistic expression and posture often occur. This poses a significant challenge for applications requiring precise human posture representation, such as medical data generation or character control in virtual reality environments. To address these concerns, research on skeleton-based human action generation [5, 6, 7, 8] has been actively pursued.

For instance, Action2Motion [7] introduced a method of generating actions using a Variational Auto-Encoder (VAE), representing stable 3D human joint rotations via Lie algebra, and achieving human-like motion generation. Despite the high performance of the GRU-based VAE model [9, 10], it remains sensitive to sequence length and struggles to effectively process global context information. This often leads to a regression toward mean poses after a certain period, introducing ambiguity in the generated action sequences.

One of the key challenges with GRU-based models is their difficulty in capturing long-range dependencies, especially in human motion sequences that require modeling complex, dynamic changes over time. This limitation becomes more evident in sequences involving varied, multi-step actions or interactions between multiple joints in the skeleton. Additionally, GRU models tend to lose contextual information over extended sequences, making it difficult to maintain realistic, continuous motion. These drawbacks highlight the need for a more robust approach to effectively capture both local and global dependencies in human action generation.

To overcome these limitations, we propose a Transformer-based VAE model. Our Transformer VAE consists of a real action encoder, a generation encoder, and a decoder. The real action encoder is designed to estimate the latent space from real human actions, while the generation encoder estimates the latent space for specific actions starting from initialized poses. The decoder then generates human actions by decoding the latent space along with the features derived from the generation encoder. This approach reduces ambiguous movements and achieves the best performance on the HumanAct12 dataset.

The structure of this paper is as follows. Section 2 explains the proposed model in detail, Section 3 presents the experimental results. Finally, a conclusion is drawn in Section 4.

II. Methods

The overview of the proposed model in this paper is shown in Fig. 1. The structure of the proposed model is divided into three components: the real action encoder, the generation encoder, and the decoder. All linear projections in the diagram are fully connected layers, and the real action encoder is only used during training.

The real action encoder takes 3D action sequences obtained from sensors as input and embeds them into a latent space. First, the action sequence is embedded along the skeleton axis through a linear projection. Positional encoding in the form of sinusoidal functions and sequence axis pooling are then added, after which self-attention is performed through the Transformer encoder. Next, independent fully connected layers that do not share parameters are used to compute the mean (μ_r) and variance (σ_r) of the action. The latent space vector is then calculated using the reparameterization trick [10].

The generation encoder shares the same structure as the real action encoder but is designed with a different number of layers and hidden dimensions. As mentioned earlier, the real action encoder is only used during training, so it is designed with

Manuscript received October 21, 2024. This work was supported by the Gyeonggi-do Regional Research Center (GRRC) Program of Gyeonggi Province, [Development of Intelligent Interactive Media and Space Convergence Application System, under Grant GRRC Aviation 2024-B02].

Sung Jun Park is with the Korea Aerospace University, School of Electronics and Information Engineering, Goyang 10540, Korea

Joong Hwan Baek is with Korea Aerospace University, School of Electronics and Information Engineering, Goyang 10540, Korea

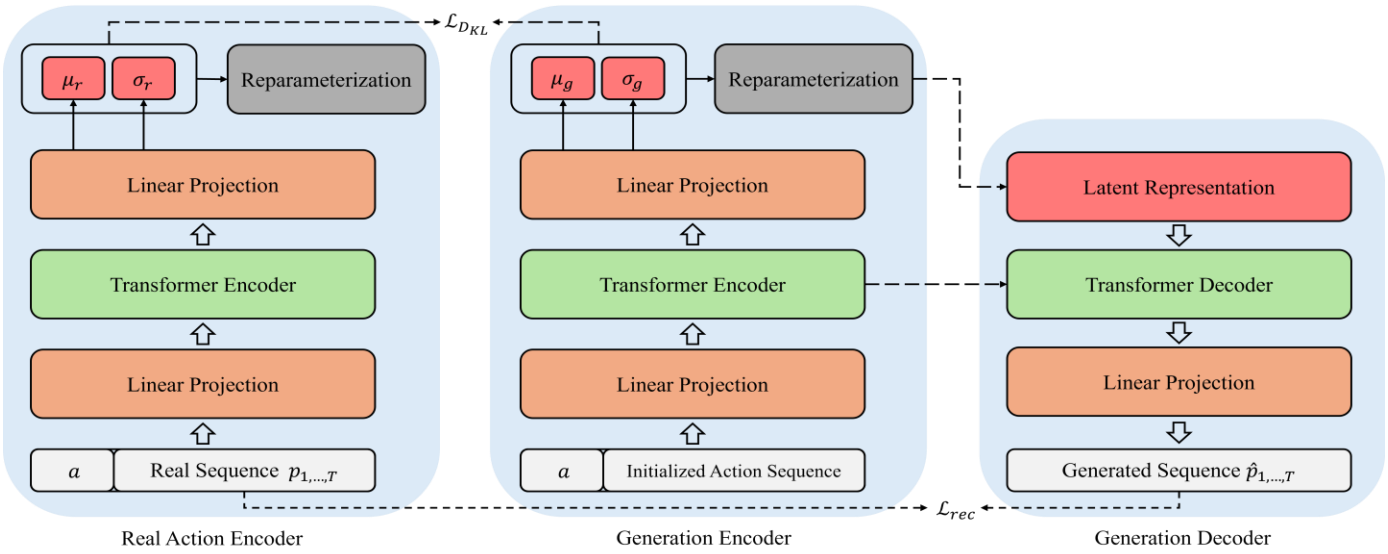

Fig. 1. Proposed model architecture overview.

deeper transformer layers and a hidden dimension set to 2 and 256, respectively, while the generation encoder is set to 1 and 128. The generation encoder takes an initialized action sequence as input and computes the mean and variance of the action label from the initialized action sequence through the same operations as the real action encoder. The generation encoder computes the mean (μ_g) and variance (σ_g) from an initialized action sequence and approximates the values of real actions through KL divergence loss, as described in (1):

$$\mathcal{L}_{D_{KL}} = \{D_{KL}(\mu_r \parallel \mu_g) + D_{KL}(\sigma_r \parallel \sigma_g)\} \quad (1)$$

The choice of Transformer architecture, specifically the self-attention mechanism, is critical in this context. Unlike GRU-based models, which struggle with long-term dependencies, the Transformer's self-attention mechanism allows for efficient capturing of both short-term and long-term dependencies in human motion. This is particularly advantageous for 3D action sequences, where complex dependencies exist not only temporally but also spatially, across different joints. The positional encoding ensures that the sequential nature of the input is preserved, while the self-attention helps capture correlations between distant joints, leading to more natural and fluid motion generation.

In the generation decoder, the inputs to the Transformer decoder TD_G during the training and testing phases differ. During training, as shown in (2) and (3), the latent space from the real sequence and the features output from the generation transformer encoder F_{TE_G} are used as inputs.

$$z_r \sim \mathcal{N}(\mu_r, \sigma_r) \quad (2)$$
$$F_{TD_G} = TD_G(F_{TE_G}, z_r) \quad (3)$$

Like the Transformer encoder, positional embedding is performed before inputting into the Transformer decoder. Action sequences are then generated through a linear projection. During testing, the latent space obtained from the initialized pose is input into the Transformer decoder TD_G, as shown in (4) and (5), to generate an action sequence.

$$z_g \sim \mathcal{N}(\mu_g, \sigma_g) \quad (4)$$
$$F_{TD_G} = TD_G(F_{TE_G}, z_g) \quad (5)$$

The generated action sequence is learned through the reconstruction loss $\mathcal{L}_{rec}$, as shown in (6).

$$\mathcal{L}_{rec} = \frac{1}{T}\sum_{t=1}^{T}(p_t - \hat{p}_t)^2 \quad (6)$$

Here, T represents the sequence length.

Finally, the total loss function is the sum of the reconstruction loss and the KL divergence loss to imitate the real sequence latent space, as shown in (7).

$$\mathcal{L} = \mathcal{L}_{rec} + \lambda\mathcal{L}_{D_{KL}} \quad (7)$$

where λ is a hyperparameter for tuning the KL divergence loss.

III. Experiments

To validate the performance of the proposed model, we conducted comparative performance evaluation using the HumanAct12 dataset. This dataset, introduced in Action2Motion [7], is derived from PHSPD [12, 13] and consists of 12 action classes and 1,191 motion clips. For optimizing the loss function during training, we used the Adam optimization algorithm with a learning rate set to 0.2×10^{-3}. The parameter λ in the final loss function was set to 0.1×10^{-3}, and the latent space dimension was set to 30.

We adopted four metrics for performance evaluation: recognition accuracy (Acc), Frechet Inception Distance (FID), diversity, and multimodality, following the evaluation method used in [7, 14].

Recognition Accuracy: This metric evaluates the classification accuracy of generated actions, using an RNN action recognition classifier from Action2Motion.

Fig. 2. Qualitative performance evaluation results for the action sequence generated by the proposed model

TABLE I: QUANTITATIVE PERFORMANCE COMPARISON RESULTS WITH EXISTING METHODS

Methods	HumanAct12			
	FID ↓	Acc ↑	Diversity→	Multimodality→
Real motions	$0.092^{\pm.007}$	$0.997^{\pm.001}$	$6.853^{\pm.053}$	$2.449^{\pm.038}$
CondGRU	$40.61^{\pm.144}$	$0.080^{\pm.002}$	$2.381^{\pm.020}$	$2.341^{\pm.0.36}$
Two-stage GAN	$10.48^{\pm.089}$	$0.421^{\pm.006}$	$5.960^{\pm.049}$	$2.805^{\pm.036}$
Act-MoCoGAN	$5.610^{\pm.113}$	$0.793^{\pm.004}$	$6.752^{\pm.071}$	$1.055^{\pm.017}$
Action2Motion	$2.458^{\pm.079}$	$0.923^{\pm.002}$	$7.032^{\pm.038}$	$2.870^{\pm.037}$
Ours	$\mathbf{2.274}^{\pm.083}$	$\mathbf{0.961}^{\pm.006}$	$\mathbf{6.768}^{\pm.034}$	$\mathbf{2.457}^{\pm.024}$

FID (Frechet Inception Distance): FID measures the distance between the distributions of generated and real action sequences, which is a critical indicator of the visual quality of generated sequences.

Diversity: This metric calculates the variance of the generated action sequences across all classes, evaluating the spread of actions within the same class.

Multimodality: This metric assesses the variety of actions generated within the same class when multiple sequences are generated for the same action.

The comparative performance results with existing methods, including CondGRU [7], Two-stage GAN [15], Act-MoCoGAN [16], and Action2Motion [7], are presented in Table I. Our proposed model achieved the best performance across all evaluation metrics. Among the four metrics, FID is crucial for evaluating the overall performance of generative models. The VAE-based method reflects the natural variability of human actions through probabilistic modeling and allows for the generation of diverse sequences through probabilistic sampling. However, GAN-based methods, despite learning to generate human motion sequences via a discriminator, face limitations in modeling variability and diversity, leading to lower performance compared to Action2Motion and our proposed method.

When comparing our model with Action2Motion, we observed superior performance in both FID and recognition accuracy, demonstrating that our model generates more realistic action sequences. The complementary metrics, diversity and multimodality, also achieved values closest to real motion, further validating the effectiveness of our approach.

In Fig. 2, we present the qualitative evaluation results of the sequences generated by our model. In the jumping motion (Fig. 2 (a)), the proposed method generates a higher jump compared to Action2Motion and closely mirrors the pre-jump pose of the real motion. In the warm-up motion (Fig. 2 (b)), Action2Motion exhibits a drifting effect toward the mean pose, while our model closely mimics the real motion, with more pronounced arm movements, highlighting its superior fidelity to the real sequence.

IV. CONCLUSION

The Transformer VAE model proposed in this paper for generating 3D action sequences achieved superior performance compared to existing GRU-based VAE and GAN-based models.

https://doi.org/10.17758/EARES19.EAP1224127

By utilizing the Transformer architecture, the model effectively addresses the long-range dependency problem and improves on issues such as action ambiguity and regression to mean poses that often occurred in previous methods. The proposed approach demonstrated the best performance across various metrics, including accuracy and FID, on the HumanAct12 dataset.

In the future, we plan to extend this model to learn user-defined action classes and apply it to programs that control avatars in virtual reality environments through the generated action sequences.

REFERENCES

[1] OpenAI, J. Achiam et al., "Gpt-4 technical report," *arXiv preprint arXiv 2303.08774.*, 2023.

[2] M. Reid et al., "Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context," *arXiv preprint arXiv:2403.05530.*, 2024

[3] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of stylegan," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8110-8119, 2020 https://doi.org/10.1109/CVPR42600.2020.00813.

[4] I. Skorokhodov, W. Menapace, A. Siarohin, and S. Tulyakov, "Hierarchical Patch Diffusion Models for High-Resolution Video Generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7569-7579, 2024. https://doi.org/10.1109/CVPR52733.2024.00723

[5] A. S. Lin, L. Wu, R. Corona, K. Tai, Q. Huang, and R. J. Mooney, "Generating animated videos of human activities from natural language descriptions," in *Proceedings of the Visually Grounded Interaction and Language Workshop at NeurIPS*, 2018.

[6] S. Yan, Z. Li, Y. Xiong, H. Yan, and D. Lin, "Convolutional sequence generation for skeleton based action synthesis," in *International Conference on Computer Vision (ICCV)*, pp. 4393-4401, 2019. https://doi.org/10.1109/ICCV.2019.00449

[7] C. Guo, X. Zuo, S. Wang, S. Zou, Q. Sun, A. Deng, M. Gong, L. Cheng, "Action2motion: Conditioned generation of 3d human motions," in *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 2021-2029, 2020. https://doi.org/10.1145/3394171.3413635

[8] M. Petrovich, M. J. Black, and G. Varol, "Action-conditioned 3d human motion synthesis with transformer vae," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10985-10995, 2021. https://doi.org/10.1109/ICCV48922.2021.01080

[9] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*, 2014.

[10] D. P. Kingma, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[11] A. Vaswani, "Attention is all you need," *Advances in Neural Information Processing Systems*, 2017.

[12] S. Zou, X. Zuo, Y. Qian, S. Wang, C. Xu, M. Gong, and L. Cheng, "3D human shape reconstruction from a polarization image," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. https://doi.org/10.1007/978-3-030-58568-6_21

[13] S. Zou, X. Zuo, Y. Qian, S. Wang, C. Guo, C. Xu, M. Ging, and L. Cheng, "Polarization human shape and pose dataset," *arXiv preprint arXiv:2004.14899*. 2020.

[14] H. Y. Lee, X. Yang, M. Y. Liu, T. C. Wang, Y. D. Lu, M. H. Yang, and J. Kautz, "Dancing to music," *Advances in neural information processing systems*, *32*, 2019.

[15] H. Cai, C. Bai, Y. W. Tai, and C. K. Tang, "Deep video generation, prediction and completion of human action sequences," in *Proceedings of the European conference on computer vision (ECCV)*, pp. 366-382, 2018. https://doi.org/10.1007/978-3-030-01216-8_23

[16] S. Tulyakov, M. Y. Liu, X. Yang, and J. Kautz, "Mocogan: Decomposing motion and content for video generation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1526-1535, 2018. https://doi.org/10.1109/CVPR.2018.00165