Experience Augmentation in Physical Therapy by Simulating Patient-Specific Walking Motions

Md Mustafizur Rahman[1,*], Goshiro Yamamoto[1,2,*], Chang Liu[2], Hiroaki Ueshima[2], Isidro Butaslac[1], Taishi Sawabe[1] and Hirokazu Kato[1]

[1]*Nara Institute of Science and Technology (NAIST), 8916-5 Takayama-cho, Ikoma, Nara 630-0192, JAPAN*
[2]*Kyoto University, 54 Kawahara-cho, Shogoin, Sakyo-ku, Kyoto 606-8507, JAPAN*

Abstract

In physical therapy, understanding and analyzing patient movements, especially impaired gait patterns, is crucial for effective rehabilitation. Traditionally, trainee therapists acquire these skills through hands-on experience with real patients and textbooks. However, these methods are limited by the availability of patients and the variability of impaired motions that therapists can observe. To address these limitations, we propose a novel system that allows therapists to learn from a wide range of impaired gait motions without being restricted by time, place, or patient availability. This system utilizes the **HumanML3D** dataset and a two-step framework combining **text2length** sampling and **text2motion** generation. In the first step, a classification model predicts motion length based on the input textual descriptions. For the second step, we use a temporal variational autoencoder (VAE) for generating varied and consistent 3D motion sequences. A key component of our approach is the utilization of residual vector quantization (RVQ) from the **MoMask** framework, which minimizes errors and enhances the precision of motion generation. Furthermore, a Masked Transformer ensures that the synthesized motion tokens are temporally consistent and contextually accurate. Our system, validated through the **HumanML3D** dataset, provides an immersive and interactive tool for physical therapists, enabling dynamic, patient-specific motion simulations in mixed reality environments. By bridging the gap between conventional methods and MR-assisted training, this approach uses interactive 3D representations to transform how therapists learn. It aims to revolutionize therapeutic training, making rehabilitation strategies more effective and personalized.

Keywords

Physical Therapists, Motion Generation, Therapeutic Training, Rehabilitation, Mixed Reality

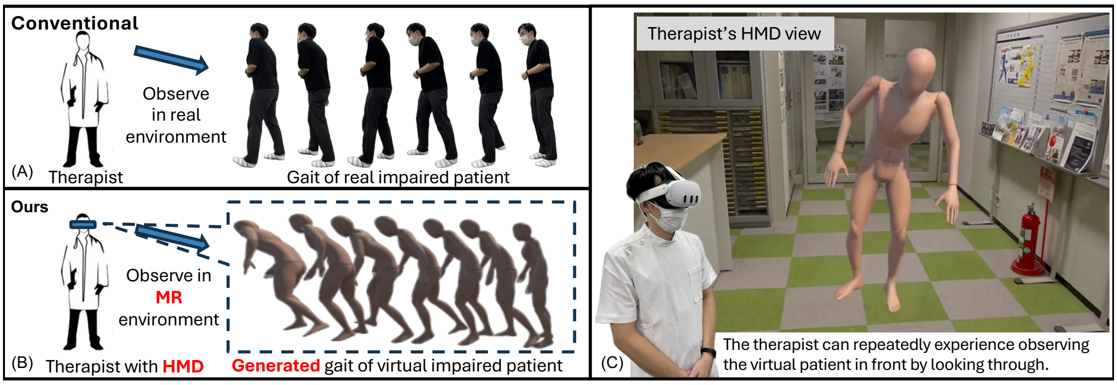

Conventional
Observe
in real
environment
(A)
Therapist
Gait of real impaired patient
Ours
Observe in
MR
environment
(B)
Therapist with HMD
Generated gait of virtual impaired patient
Therapist's HMD view
(C)
The therapist can repeatedly experience observing
the virtual patient in front by looking through.

Figure 1: Comparison of traditional and our mixed-reality-based gait observation. (A) The therapist observes a real patient's gait in a conventional setting. (B, C) A therapist views a virtual patient's gait on a head-mounted display, which enhances learning and diagnosis.

APMAR'24: The 16th Asia-Pacific Workshop on Mixed and Augmented Reality, Nov. 29-30, 2024, Kyoto, Japan

[*] Corresponding author.

rahman.md_mustafizur.rp6@naist.ac.jp (M. M. Rahman); goshiro@kuhp.kyoto-u.ac.jp (G. Yamamoto); liuchang@kuhp.kyoto-u.ac.jp (C. Liu); h_ueshima@kuhp.kyoto-u.ac.jp (H. Ueshima); isidro.b@naist.ac.jp (I. Butaslac); t.sawabe@is.naist.jp (T. Sawabe); kato@is.naist.jp (H. Kato)

0000-0001-7422-1525 (M. M. Rahman); 0000-0002-2014-7195 (G. Yamamoto); 0000-0002-0515-5226 (C. Liu); 0000-0001-5488-185X (H. Ueshima); 0000-0003-1172-7611 (I. Butaslac); 0000-0001-9244-479X (T. Sawabe); 0000-0003-3921-2871 (H. Kato)

1. Introduction

One of the main goals of stroke rehabilitation programs is the recovery of gait, which is often an important goal for patients as well. Post-stroke functional recovery typically involves both natural processes and therapeutic interventions. While the majority of stroke survivors regain the ability to walk, many fail to achieve sufficient endurance, speed, or stability to perform daily activities independently and safely. After a stroke, falls are still a major problem for people who live in the community [1].

In conventional therapy training, trainee therapists rely primarily on textbooks and hands-on experience with real patients to understand impaired gait motion. More recent studies used mixed reality (MR) to develop the therapist's 'clinical eye,' enhancing their assessment through overlaid visualizations of patient data during rehabilitation [2, 3]. However, a limitation of this method is the availability of patients restricts the variety and frequency of learning opportunities, limiting exposure to different types of impairments and hindering skill development. Building upon our prior work on a virtual reality (VR)-based medical training simulator [4], which demonstrated the efficacy of immersive 3D modeling and robotic systems in enhancing medical training and remote surgery, this study extends these principles to augment therapeutic learning with patient-specific motion simulations in MR. In contrast to traditional methodologies, this research investigates how the incorporation of MR-based simulations of impaired gait motion can improve learning outcomes for therapists undergoing training by providing increased exposure to a variety of gait impairments.

As illustrated in Fig. 1, when provided with the input description, "a man walks forward with a noticeable limp due to pain, favoring his right leg as he moves, his steps are uneven, and his body tilts slightly with each step, reflecting discomfort," our system generates multiple unique three-dimensional (3D) impaired human motions that closely correspond to the given textual input. This approach significantly enhances traditional training methods by offering immersive, repeatable learning experiences, leading to improved diagnostic accuracy and therapeutic outcomes. The system aims to faithfully replicate a wide range of realistic 3D human motion dynamics that precisely adhere to the specified directions, actions, timing, speed, and style described in the text.

Applications in robotics, human-machine interface, and virtual content creation, among others, could be greatly impacted by this automated process. Making use of different approaches such as motion capture has its negative aspects which are the high costs and long time taken, therefore the automatic text to motion generation is more feasible and cost effective. Despite this, such a task is quite difficult due to the nature of words and motion data being heterogeneous in many aspects. With this, a number of attempts have been made in recent years, such as the use of an encoder with recurrent neural networks (RNNs)[5], variational autoencoders (VAE)[6], and transformer networks aiming to embed the language and motion in the same space converting them into a unified approach[7, 8]. Although these methods have proved effective with small units of text, the downside is that text of a larger length projecting complex ideas does not produce good sequences of motion. Moreover, while existing diffusion processes have shown effectiveness for image generation and motion generation from text descriptions[9, 10], it remains unclear whether such improvements within one architecture come at a reasonable cost compared to more traditional Vector Quantized Variational Autoencoder (VQ-VAE) based approaches.

In this work, we leverage the MoMask method introduced by Guo et al. [7], which combines hierarchical quantization with generative transformer models to address the limitations of previous techniques. While traditional methods like Residual Vector Quantization (RVQ)[6] attempt to reduce quantization errors by embedding motion tokens multiple times, MoMask offers a more advanced solution. As the first generative masked modeling framework for text-to-motion generation, MoMask features a hierarchical quantization generative model and a dedicated mechanism for precise residual quantization, base token generation, and residual token prediction. Additionally, we integrate the HumanML3D[11] dataset, which contains 14,616 annotated motion clips and 44,970 text descriptions, providing a comprehensive resource for generating and evaluating human motions. To facilitate seamless impaired humanoid motion retargeting, we develop a headless Blender Python API script that enables mapping between different humanoid rigs and allows for local saving of bone mappings. Moreover, we implement a FastAPI backend that allows users to stream data directly from Unity3D and use them for real-time humanoid animation visualization with an HMD, ensuring smooth integration and display.

2. Related Work

Existing work relating to our research mostly fall into domains of (2.1) 3D human motion generation, (2.2) text-motion generation, and (2.3) language models and human motion captioning.

2.1. 3D Human Motion Generation

Significant advancements have been made in 3D human motion generation, utilizing various approaches that leverage action learning, audio, and text inputs. Traditional methods often employ a hidden state vector to

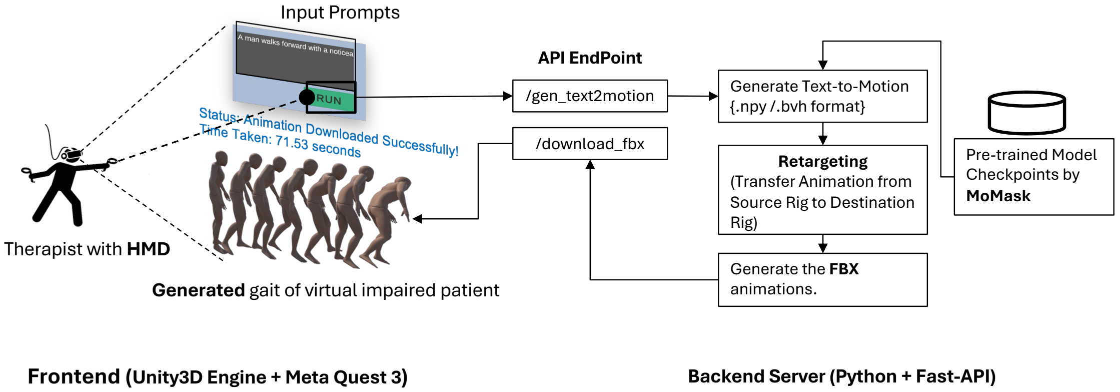

Input Prompts
A man walks forward with a noticea
RUN
Status: Animation Downloaded Successfully!
Time Taken: 71.53 seconds
Therapist with HMD
Generated gait of virtual impaired patient
API EndPoint
/gen_text2motion
/download_fbx
Generate Text-to-Motion {.npy /.bvh format}
Retargeting (Transfer Animation from Source Rig to Destination Rig)
Generate the FBX animations.
Pre-trained Model Checkpoints by MoMask
Frontend (Unity3D Engine + Meta Quest 3)
Backend Server (Python + Fast-API)

Figure 2: System Overview: The system converts therapist's input description into impaired motion using MoMask models in a FastAPI backend. Retargeted animations are converted to FBX and visualized in Unity3D with Meta Quest 3 for real-time analysis and therapeutic decision-making.

generate sequential states. Basic approaches, such as those by Cai et al. [12] and Wang et al. [13], utilized GAN algorithms to extend partial sequences with newly generated states. In contrast, more advanced methods like Yu et al. [14] employed GCNs to capture the spatial and temporal dynamics of human motion. Furthermore, VAE and transformer-based models have been applied to better capture temporal dependencies, as demonstrated by Guo et al. [15, 16] and Petrovich et al. [17]. For audio-driven motion generation, techniques often transform acoustic features into human poses. Studies such as Takeuchi et al. [18] utilized two-way LSTMs to generate gestures from speech, while Shlizerman and Tang et al. [19, 20] investigated song and dance motion generation. Recent models, such as Lee et al. [21], also focused on the stochastic aspects of movement, which introduced uncertainty in dance movements.

2.2. Text-motion Generation

Text-motion generation has become increasingly popular due to the ease of using natural language input. Previous studies [22, 23, 24, 25, 26] used mainly deterministic models, which typically average or blur the motion output. More recent stochastic models, such as those in T2M[27] and TEMOS[28], introduced more realism and variety into motion generation by using VAE structures and transformers to provide the shared transition between speech and motion [29, 30, 31, 32, 33, 34]. Recent innovations, such as autoregressive models [35, 36, 8, 37, 38] have gradually increased the quality of motion synthesis dramatically through denoising or motion suspension. Generative masked modeling inspired by BERT [39] have also been developed for human motion generation, using techniques such as residual quantization [40, 41, 42] to improve motion discretization and reduce quantization errors.

2.3. Language Models and Human Motion Captioning

The translation from natural language to human motion have evolved from mathematical models [43] to advanced neural networks like TM2T [36], which provides two-way visualization between text and movement. Major language models such as BERT [39], T5 [44], and Instruct-GPT [45] have pushed the boundaries of understanding across sectors. In multimodal learning, models like CLIP [46] have linked images with text, inspiring similar advancements in human motion tasks, such as MotionCLIP [47]. Despite this progress, language models still remain underutilized in human motion tasks. Our research seeks to integrate them into motion generation, leveraging pre-trained models to create diverse motions.

Moreover, while existing work predominantly focuses on generating normal human motions, our system specifically targets the generation of impaired motions crucial for physical therapy training and rehabilitation.

3. Method

3.1. System Overview

The proposed system combines text-to-motion generation, motion retargeting, and 3D animation export to create realistic human motion sequences from textual descriptions. As illustrated in Fig. 2, it features a backend powered by a Python server using FastAPI and a frontend in Unity3D, allowing therapists to interact with animations via a Meta Quest 3 headset. The backend processes input prompts with MoMask model checkpoints

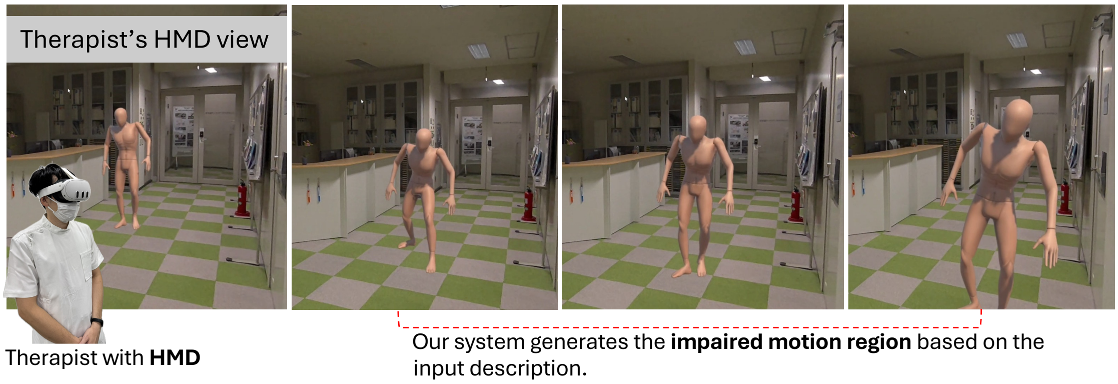

Therapist's HMD view
Therapist with **HMD**
Our system generates the **impaired motion region** based on the input description.

Figure 3: The system generates the impaired motion region based on the therapist's input description, allowing the therapist to interactively observe and analyze the impaired motion for enhanced both diagnostic precision and therapeutic decision making.

to generate patient-specific impaired motions, saved in '.npy' or '.bvh' format.

In general, our system is comprised of the following core components:

Backend: The backend features a FastAPI-based Python server responsible for processing the motion generation pipeline. It utilizes pre-trained models from the MoMask framework, which transform therapist input prompts into impaired motion representations. This includes generating motion sequences that reflect specific conditions or impairments, ensuring realistic and applicable outputs for therapeutic use.

Frontend: The frontend is built on the Unity3D engine, which is employed to visualize the generated animations. It is designed to interface seamlessly with the Meta Quest 3 headset, enabling immersive interaction for therapists. This integration allows users to experience the animations in a three-dimensional space, providing a more intuitive understanding of the impaired motions.

API Endpoints: The system utilizes two key API endpoints to manage communication between the frontend and backend. The first endpoint, "/gen_text2motion", takes the therapist's input prompt and triggers the motion generation process. The backend processes the prompt through the MoMask model, which translates the text description into a motion representation in formats like '.npy' or '.bvh'. Once the motion is generated and retargeted, the second endpoint, "/download_fbx", allows the frontend to retrieve the final FBX animation file. This file is then used to visualize the impaired motions in the MR interface. These API endpoints ensure smooth and efficient interaction between the components, allowing the system to generate and deliver animations in real-time based on simple text input, thereby enhancing the rehabilitation experience for therapists.

3.2. Text-to-Motion Generation

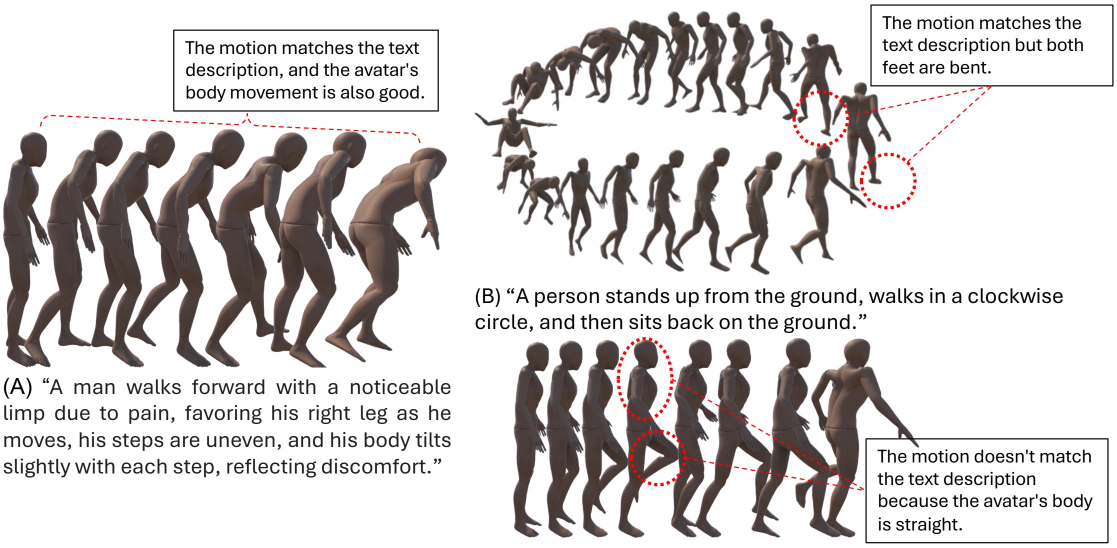
Our system builds upon the state-of-the-art techniques for text-driven motion generation, particularly drawing inspiration from the MoMask framework. The text-to-motion process is detailed below:

Tokenization of Motion Sequences: The textual descriptions are transformed into a sequence of discrete motion tokens using a vector quantization process. This process tokenizes complex human motion into a hierarchical structure of motion segments, each capturing different facets of the described action.

Masked Motion Prediction: A Masked Transformer is employed to predict masked motion tokens conditioned on the input text. During the training phase, the model is trained to fill in randomly masked tokens from incomplete motion sequences. In the inference phase, it generates entire motion sequences by iteratively predicting missing tokens, ensuring global consistency and fidelity to the input description.

Residual Refinement: After the base-layer motion is generated, a Residual Transformer is used to progressively refine the motion by predicting additional motion tokens that capture higher-order details. This step is crucial for enhancing the granularity and subtlety of the generated motion, ensuring fine control over aspects such as posture and movement transitions.

Motion Generation Output: The final output is a continuous 3D human motion sequence generated in '.npy' or '.bvh' formats. These motion sequences represent high-quality, realistic animations that can be further processed or directly visualized.


The motion matches the text description, and the avatar's body movement is also good.
The motion matches the text description but both feet are bent.
The motion doesn't match the text description because the avatar's body is straight.

(A) “A man walks forward with a noticeable limp due to pain, favoring his right leg as he moves, his steps are uneven, and his body tilts slightly with each step, reflecting discomfort.”

(B) “A person stands up from the ground, walks in a clockwise circle, and then sits back on the ground.”

(C) “A man walks forward, they swing their left leg outward, causing their body to lean slightly to the right, after the left foot touches the ground, the right leg smoothly swings forward.”

Figure 4: Impaired motion animations generated from textual descriptions using the HumanML3D dataset. The animation in (A) aligns well with the text description, with the avatar’s movements accurately reflecting the specified actions. In (B), while the motion generally corresponds to the description, there is a minor discrepancy in the animation, as both feet appear bent. In (C), although the described leg movement is captured, the avatar’s posture does not match the expected slight lean, remaining too straight.

3.3. Motion Retargeting and Animation Transfer

After generating the motion sequences, the system applies a motion retargeting process to map the generated motion onto the target 3D model’s skeleton. This is done using ‘keemap.rig.transfer’, a precision retargeting tool within Blender, which ensures accurate bone mapping and preserves key motion attributes such as foot positioning and overall balance. By retargeting the motion, we ensure that the animations are properly transferred to any humanoid rig, allowing seamless integration into various 3D models.

3.4. Export and Visualization in Unity

Once the motion has been retargeted, it is exported in FBX format, which contains both the 3D model and its associated animation. This format is then imported into Unity 3D, where additional adjustments to model positioning and animation playback can be made. A custom Unity frontend was developed to provide an intuitive interface where users, such as physical therapists, can interact with and visualize the generated motion sequences in real-time, enhancing their ability to analyze and evaluate generated impaired motions.

In Fig. 3, the interactive process of generating impaired motion regions based on the therapist’s input description is demonstrated. The sequence of images illustrates a physical therapist utilizing a head-mounted display to observe the virtual patient’s motion in real-time. The leftmost frame captures the therapist’s perspective, showcasing the virtual patient within a simulated environment. The subsequent frames display the progression of the patient’s movement, with impaired regions clearly highlighted to indicate abnormalities in motion. This system enables therapists to interact with and analyze the impaired motion sequences in real-time, providing an immersive, hands-on approach that significantly enhances both diagnostic precision and therapeutic decision making.

4. Pilot Experiment

In the pilot experiment, we assessed the ability of our text-to-motion model to generate impaired gait motions. Fig. 4 highlights several outcomes from the experiment, showcasing both well-matched and mismatched animations.

In well-matched animations Fig. 4(A), the motion aligned well with the text description. The avatar showed a noticeable limp and discomfort, with uneven steps and a tilted posture—accurately portraying the described impairment. In some animations Fig. 4(B), the generated motion mostly matched the description, but the avatar's bent feet introduced a small inconsistency, detracting from realism. In mismatched animations taking Fig. 4(C) as an example, although the left leg's outward movement is captured, the avatar's body remained too straight, failing to show the expected slight lean.

The integration of the HumanML3D dataset, paired with pre-trained model checkpoints from MoMask, greatly improved the quality of the generated motions. These findings highlight the effectiveness of text-to-motion generation techniques. Additionally, the combination of Residual Vector Quantization-VAE (RVQ-VAE) and Transformer models contributed to the model's ability to capture both coarse-grained and fine-grained motion details, further enhancing the fidelity and accuracy of the animations.

5. Future Work

Looking ahead, we plan to broaden our research by developing a custom dataset composed of textual descriptions extracted from the Electronic Health Records (EHRs). This domain-specific dataset will enable the model to generate motions that are more relevant to medical and therapeutic applications. Once this dataset is constructed, we will retrain our model to improve its performance in these specialized contexts.

Currently, we have not tested the effectiveness of the virtual motions in MR environments, which is another key feature of our system. In the future, we also intend to conduct more extensive user studies with a large cohort of therapists to evaluate the long-term impact of using our MR system on their training outcomes. Furthermore, we will incorporate user feedback to refine the user interface and enhance the system's overall functionality.

Finally, we aim to develop additional features within the MR environment, such as the simulation of real-world therapeutic scenarios and the ability to track the therapists' performance over time. These advancements will ensure that our system continues to be a valuable tool for therapist training and ultimately contributes to improved patient care.

6. Conclusion

This study have introduced a system for enhancing physical therapy training through MR simulations of patient-specific walking motions. By utilizing the HumanML3D dataset and advanced techniques like RVQ and Masked Transformers, the system generates realistic impaired gait patterns from textual descriptions. The system is aimed to provide therapists with immersive and repeatable training experiences, leading to improved diagnostic accuracy and therapeutic outcomes.

Future work will focus on developing a dataset from EHR and conducting user studies to assess the system's effectiveness in therapist training. Overall, our system has the potential to significantly improve how therapists learn and analyze gait impairments.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Number JP23K24888.

References

[1] J.-M. Belda-Lois, S. Mena-del Horno, I. Bermejo-Bosch, J. C. Moreno, J. L. Pons, D. Farina, M. Iosa, M. Molinari, F. Tamburella, A. Ramos, et al., Rehabilitation of gait after stroke: a review towards a top-down approach, Journal of neuroengineering and rehabilitation 8 (2011) , pp. 1–20.

[2] M. De Cecco, A. Luchetti, I. Butaslac, F. Pilla, G. M. A. Guandalini, J. Bonavita, M. Mazzucato, K. Hirokazu, Sharing augmented reality between a patient and a clinician for assessment and rehabilitation in daily living activities, Information 14 (2023) , Art. No. 204.

[3] A. Luchetti, I. Butaslac, M. Rosi, D. Fruet, G. Nollo, P. G. Ianes, F. Pilla, B. Gasperini, G. M. Achille Guandalini, J. Bonavita, H. Kato, M. De Cecco, Multidimensional assessment of daily living activities in a shared augmented reality environment, in: 2022 IEEE International Workshop on Metrology for Living Environment, 2022, pp. 60–65.

[4] M. M. Rahman, M. F. Ishmam, M. T. Hossain, M. E. Haque, Virtual reality based medical training simulator and robotic operation system, in: 2022 International Conference on Recent Progresses in Science, Engineering and Technology (ICRPSET), IEEE, 2022, pp. 1–4.

[5] L. R. Medsker, L. Jain, et al., Recurrent neural networks, Design and Applications 5 (2001) , pp. 2.

[6] D. P. Kingma, M. Welling, et al., An introduction to variational autoencoders, Foundations and Trends® in Machine Learning 12 (2019) , pp. 307–392.

[7] C. Guo, Y. Mu, M. G. Javed, S. Wang, L. Cheng, Momask: Generative masked modeling of 3d human motions, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 1900–1910.

[8] B. Jiang, X. Chen, W. Liu, J. Yu, G. Yu, T. Chen, Motiongpt: Human motion as a foreign language,

Advances in Neural Information Processing Systems 36 (2024).

[9] C. Ahuja, L.-P. Morency, Language2pose: Natural language grounded pose forecasting, in: 2019 International Conference on 3D Vision, IEEE, 2019, pp. 719–728.

[10] U. Bhattacharya, N. Rewkowski, A. Banerjee, P. Guhan, A. Bera, D. Manocha, Text2gestures: A transformer-based network for generating emotive body gestures for virtual agents, in: 2021 IEEE virtual reality and 3D user interfaces (VR), IEEE, 2021, pp. 1–10.

[11] C. Guo, S. Zou, X. Zuo, S. Wang, W. Ji, X. Li, L. Cheng, Generating diverse and natural 3d human motions from text, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 5152–5161.

[12] H. Cai, C. Bai, Y.-W. Tai, C.-K. Tang, Deep video generation, prediction and completion of human action sequences, in: Proceedings of the European conference on computer vision (ECCV), 2018, pp. 366–382.

[13] Z. Wang, P. Yu, Y. Zhao, R. Zhang, Y. Zhou, J. Yuan, C. Chen, Learning diverse stochastic human-action generators by learning smooth latent transitions, in: Proceedings of the AAAI conference on artificial intelligence, volume 34, 2020, pp. 12281–12288.

[14] P. Yu, Y. Zhao, C. Li, J. Yuan, C. Chen, Structure-aware human-action generation, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16, Springer, 2020, pp. 18–34.

[15] C. Guo, X. Zuo, S. Wang, X. Liu, S. Zou, M. Gong, L. Cheng, Action2video: Generating videos of human 3d actions, International Journal of Computer Vision 130 (2022) , pp. 285–315.

[16] 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, 2020, pp. 2021–2029.

[17] M. Petrovich, M. J. Black, G. Varol, Action-conditioned 3d human motion synthesis with transformer vae, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 10985–10995.

[18] K. Takeuchi, D. Hasegawa, S. Shirakawa, N. Kaneko, H. Sakuta, K. Sumi, Speech-to-gesture generation: A challenge in deep learning approach with bidirectional lstm, in: Proceedings of the 5th International Conference on Human Agent Interaction, 2017, pp. 365–369.

[19] E. Shlizerman, L. Dery, H. Schoen, I. Kemelmacher-Shlizerman, Audio to body dynamics, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7574–7583.

[20] T. Tang, J. Jia, H. Mao, Dance with melody: An lstm-autoencoder approach to music-oriented dance synthesis, in: Proceedings of the 26th ACM international conference on Multimedia, 2018, pp. 1598–1606.

[21] H.-Y. Lee, X. Yang, M.-Y. Liu, T.-C. Wang, Y.-D. Lu, M.-H. Yang, J. Kautz, Dancing to music, Advances in neural information processing systems 32 (2019).

[22] R. Huang, H. Hu, W. Wu, K. Sawada, M. Zhang, D. Jiang, Dance revolution: Long-term dance generation with music via curriculum learning, arXiv preprint arXiv:2006.06119 (2020).

[23] A. Ghosh, N. Cheema, C. Oguz, C. Theobalt, P. Slusallek, Synthesis of compositional animations from textual descriptions, in: Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 1396–1406.

[24] A. S. Lin, L. Wu, R. Corona, K. Tai, Q. Huang, R. J. Mooney, Generating animated videos of human activities from natural language descriptions, Learning 1 (2018) 1.

[25] M. Plappert, C. Mandery, T. Asfour, Learning a bidirectional mapping between human whole-body motion and natural language using deep recurrent neural networks, Robotics and Autonomous Systems 109 (2018) , pp. 13–26.

[26] C. Ahuja, L.-P. Morency, Language2pose: Natural language grounded pose forecasting, in: 2019 International Conference on 3D Vision, IEEE, 2019, pp. , pp. 719–728.

[27] C. Guo, S. Zou, X. Zuo, S. Wang, W. Ji, X. Li, L. Cheng, Generating diverse and natural 3d human motions from text, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 5152–5161.

[28] M. Petrovich, M. J. Black, G. Varol, Temos: Generating diverse human motions from textual descriptions, in: European Conference on Computer Vision, Springer, 2022, pp. 480–497.

[29] X. Chen, B. Jiang, W. Liu, Z. Huang, B. Fu, T. Chen, G. Yu, Executing your commands via motion diffusion in latent space, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 18000–18010.

[30] J. Kim, J. Kim, S. Choi, Flame: Free-form language-based motion synthesis & editing, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 37, 2023, pp. 8255–8263.

[31] H. Kong, K. Gong, D. Lian, M. B. Mi, X. Wang, Priority-centric human motion generation in discrete latent space, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 14806–14816.

[32] Y. Lou, L. Zhu, Y. Wang, X. Wang, Y. Yang, Di-

versemotion: Towards diverse human motion generation via discrete diffusion, arXiv preprint arXiv:2309.01372 (2023).

[33] J. Tseng, R. Castellon, K. Liu, Edge: Editable dance generation from music, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 448–458.

[34] M. Zhang, Z. Cai, L. Pan, F. Hong, X. Guo, L. Yang, Z. Liu, Motiondiffuse: Text-driven human motion generation with diffusion model, arXiv preprint arXiv:2208.15001 (2022).

[35] K. Gong, D. Lian, H. Chang, C. Guo, Z. Jiang, X. Zuo, M. B. Mi, X. Wang, Tm2d: Bimodality driven 3d dance generation via music-text integration, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 9942–9952.

[36] C. Guo, X. Zuo, S. Wang, L. Cheng, Tm2t: Stochastic and tokenized modeling for the reciprocal generation of 3d human motions and texts, in: European Conference on Computer Vision, Springer, 2022, pp. 580–597.

[37] J. Zhang, Y. Zhang, X. Cun, Y. Zhang, H. Zhao, H. Lu, X. Shen, Y. Shan, Generating human motion from textual descriptions with discrete representations, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2023, pp. 14730–14740.

[38] Z. Zhou, B. Wang, Ude: A unified driving engine for human motion generation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 5632–5641.

[39] J. Devlin, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).

[40] Z. Borsos, R. Marinier, D. Vincent, E. Kharitonov, O. Pietquin, M. Sharifi, D. Roblek, O. Teboul, D. Grangier, M. Tagliasacchi, et al., Audiolm: a language modeling approach to audio generation, IEEE/ACM transactions on audio, speech, and language processing 31 (2023) , pp. 2523–2533.

[41] J. Martinez, H. H. Hoos, J. J. Little, Stacked quantizers for compositional vector compression, arXiv preprint arXiv:1411.2173 (2014).

[42] N. Zeghidour, A. Luebs, A. Omran, J. Skoglund, M. Tagliasacchi, Soundstream: An end-to-end neural audio codec, IEEE/ACM Transactions on Audio, Speech, and Language Processing 30 (2021) , pp. 495–507.

[43] W. Takano, Y. Nakamura, Statistical mutual conversion between whole body motion primitives and linguistic sentences for human motions, The International Journal of Robotics Research 34 (2015) , pp. 1314–1328.

[44] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, Journal of machine learning research 21 (2020) , pp.1–67.

[45] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al., Training language models to follow instructions with human feedback, Advances in neural information processing systems 35 (2022) , pp. 27730–27744.

[46] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: International conference on machine learning, PMLR, 2021, pp. 8748–8763.

[47] G. Tevet, B. Gordon, A. Hertz, A. H. Bermano, D. Cohen-Or, Motionclip: Exposing human motion generation to clip space, in: European Conference on Computer Vision, Springer, 2022, pp. 358–374.