Character Motion Synthesis: A Survey

Jingtong Xu *

University of Florida, Gainesville, Florida, the United States of America

* Corresponding Author Email: jingtongxu@ufl.edu

Abstract. Character Motion Synthesis is a huge cost when it comes to film, game, and design productions using traditional methods of motion synthesis. However, there are newly designed ways to generate motions, which are efficient and economical compared to traditional methods. This paper presents an overview of different types of novel methods to synthesize motions. This paper will introduce Audio-Driven and Music-Driven Motion Synthesis, Generative Models and Frameworks for Motion Synthesis, Human and Object Interaction. Character and Pet Motion Synthesis, Grasp and Hand Object Interactions, and Motion Retargeting and Editing, assembling and summarizing separate articles.

Keywords: Motion Synthesis, Survey, motion generation.

1. Introduction

Motion Synthesis, a procedure of simulating or creating animations of motions, could be seen in various aspects: films, game, Virtual Reality, cartoon, robotics, etc. Since motion synthesis is so important and versatile, motion synthesis has become a hot topic in the field of computer graphics with an increasing number of researchers focused on improving and developing new methods to better the quality and the efficiency of Motion Synthesis. The topics would cover these contents: 1. Audio-Driven and Music-Driven Motion Synthesis 2. Generative Models and Frameworks for Motion Synthesis 3. Human and Object Interaction 4. Character and Pet Motion Synthesis 5. Grasp and Hand Object Interactions 6. Motion Retargeting and Editing.7 Advanced Motion Synthesis and Interaction

2. Audio-driven and Music Driven Motion synthesis

Simon and etc introduce a new Conformer-based diffusion model for audio-driven probabilistic modeling, improving gesture generation and stylistic control. They propose future research focusing on faster output generation and integrating audio and text inputs [1].

Wenlin Zhuang, and etc presents DabFusion, a model that generates personalized dance videos directly from reference images and music, bypassing the need for keypoint annotations and accommodating various dance styles and rhythms. DabFusion utilizes a novel music encoder and latent optical flow techniques to produce synchronized and diverse dance sequences, demonstrating robust performance on the AIST++ dataset and establishing a new standard for dance video synthesis [2].

Rishabh Dabral and etc make full use of DDPM for advanced 3D human motion synthesis, performing outstandingly in realism and diversity under music or text related signals, which is essential for applications in virtual character animation and crowd simulation [3] (Figure 1).

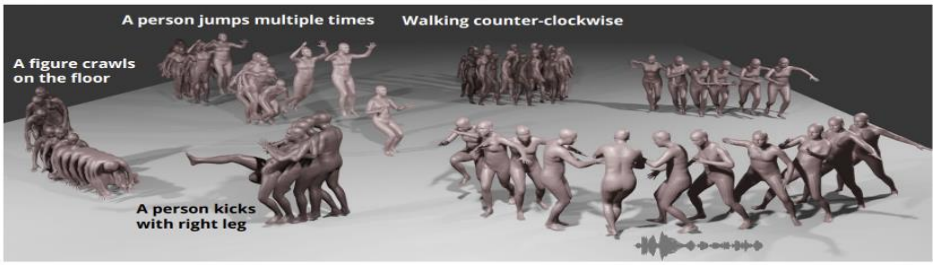


Fig 1. The picture shows the MoFusion approach to synthesizing long sequences of human motions in 3D from textual and audio inputs [3].

3. Generative Models and Frameworks for Motion Synthesis

Peizhuo Li and etc introduce a novel framework that simplifies data collection by generating diverse and realistic motions from a single training sequence, which adapts to different skeletal structures and motion types. In addition, the paper also explores practical applications of this framework, such as synthesizing longer motion sequences with variations and style transfer[4] (Figure 2).

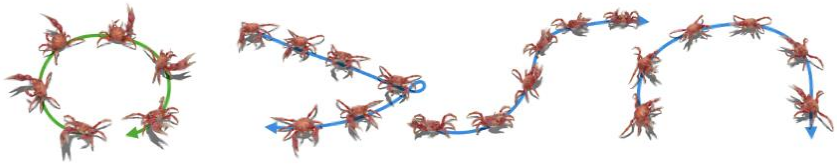


Fig 2. The picture shows a single training motion sequence of skeleton with arbitrary topology, as it is discussed in GANimator: Neural Motion Synthesis from a Single Sequence [4].

Prafulla Dhariwal and Alexander Nichol discuss advances in generative models that put an emphasis on GANs' strengths in image production metrics but challenges in diversity and training stability. They focused on improving diffusion model structure and using classifier-guided sampling to boost sample quality and efficiency in order to rival GAN performance in various datasets and missions[5] (Figure 3).



Fig 3. The pictures show the sample from BigGAN-deep with truncation 1.9 versus sample from the diffusion model with guidance mentioned in Diffusion Models Beat GANs on Image Synthesis. In Advances in Neural Information Processing Systems (NeurIPS '21) [5].

Gustav and colleagues introduce MoGlow, which is a novel model by exploiting normalizing flows for motion-data sequences, providing fast sampling, precise probability calculation, and sequential output generation. Experiments prove MoGlow's effectiveness in motion synthesis, providing valuable applications for motion data tasks [6] (Figure 4).



Fig 4. Shows probabilistic motion generation from MoGlow: Probabilistic and Controllable Motion Synthesis Using Normalising Flows [6].

Simon Alexanderson and etc introduce MoGlow, a probabilistic model for speech-driven gesture synthesis, enabling natural variability in gestures without deterministic rules. It enables precise control over gesture styles and extends to generating full-body motions, enhancing realism for interactive applications [7] (Figure 5).

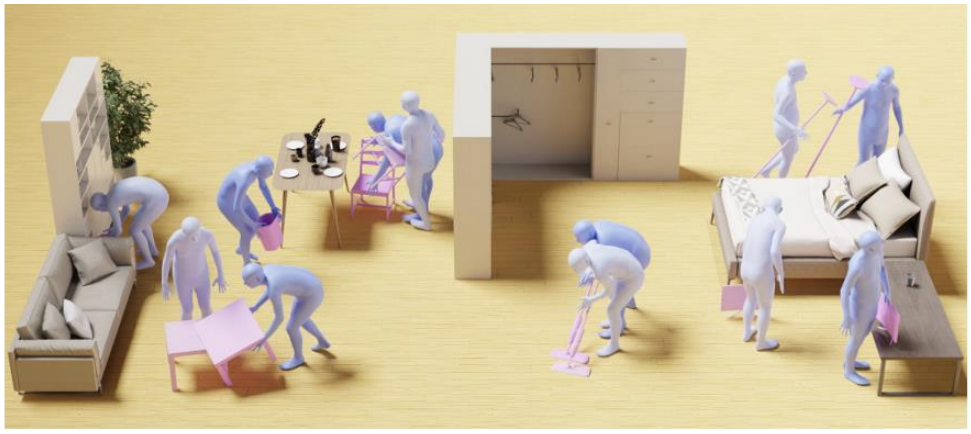


Fig 5. The picture shows the sequence of object states as input and full-body human motion interacting with the given object in OMOMO explored in Object Motion Guided Human Motion Synthesis [7].

Anindita Ghosh and colleagues designed a new framework, IMos, which can highly improve the synthesis of interactions between human characters and objects in virtual 3D environments based on textual instructions. This new framework trumps current methods that automate character motion from inputs like music and speech, resolving complexities in human-object interactions across computer graphics, robotics, movie script visualization, and also game design [8].

Guo, C. and etc presents a new framework for generating realistic 3D human motions using a conditional temporal VAE with Lie algebra representation, which overcomes limitations of existing 2D and joint-coordinate methods by disentangling anatomy and motion dynamics. This approach produces diverse and natural motion sequences and is validated with extensive experiments and a curated dataset [9].

Petrovich, M.; Black, M. J.; and Varol, G present ACTOR, a Transformer-based conditional VAE that generates diverse and realistic 3D human motion sequences from semantic action labels and durations, using a sequence-level latent vector to overcome limitations of autoregressive methods. It demonstrates superior performance in motion synthesis and applies monocular motion estimation to enhance scalability, with promising results in action recognition and MoCap denoising [10] (Figure 6).

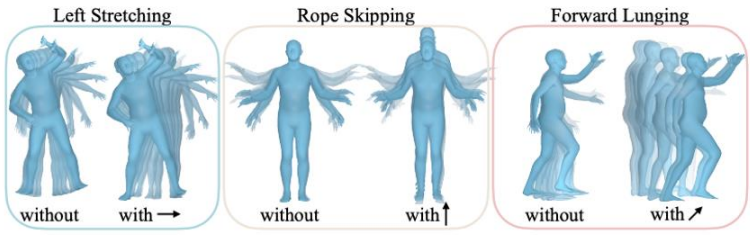


Fig 6. The picture shows the generation of the 3D root translation [10].

4. Human Object Interaction

He Zhang, Yuting Ye, Takaaki Shiratori, and Taku Komura explore neural network-based ways to synthesize detailed finger motions for object manipulation tasks, gaining insights from biomechanics and spatial relations between hands and objects. It demonstrates effective generation of various finger gaits of different objects, giving a huge support to applications in gaming, AR, and VR [11].

Daniel Holden and colleagues use deep learning with convolutional autoencoders to synthesize lifelike human motion from large datasets, along with a feedforward neural network for intuitive motion generation. They introduce future improvements like deeper networks and advanced depooling techniques to handle complex body movements in animation more effectively [12] (Figure 7)



Fig 7. The pictures show the top scenes where various 3D objects manipulated at the online demo, as it is described in Manipnet: neural manipulation synthesis with a hand-object spatial representation [12].

Noshaba Cheema and etc explore newly developed technologies in virtual character animation and motion synthesis, putting an emphasis on deep reinforcement learning for generating realistic fatigue animations with not need to capture explicit data. They also discuss a physics-based full-body approach, making it possible to generate efficient real-time interactive animation for virtual characters of varied physiologies [13].

Jiaman Li and etc introduce OMOMO, a two-stage framework for synthesizing human motion aligned by moving objects, refining the realism in human interactions by boosting contact constraints. They also present a new smartphone-based method for capturing human interaction motions, using a large dataset with 3D object geometry and high-quality object motion data [14]. Jihoon Kim, Jiseob Kim, and Sungjoon Choi present Flame model. Flame is a newly developed diffusion-based model for generating and editing motion from free-form text, boosting synthesis quality and flexibility with no need of model fine-tuning. It demonstrates the effectiveness in the field of diverse motion generation and tasks of prediction and in-betweening [15] (Figure 8).

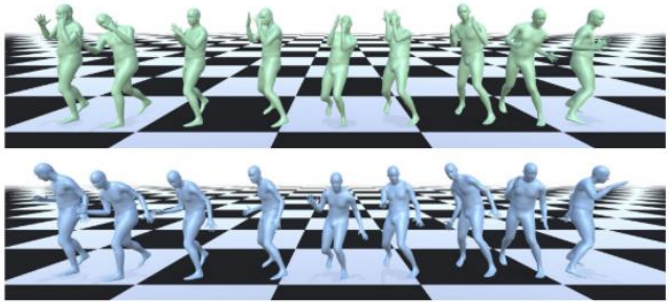


Fig 8. The pictures show the result from FlAME with prompt: "A person dribbles a ball [15].

5. Character and Pet Motion Synthesis

Haimin Luo and colleagues introduce a neural modeling and rendering pipeline that can generate articulated neural pets with realistic appearance and motion synthesis, called ARTEMIS, featuring real-time interactive motion control and high-quality images of animals. Their approach integrates robust real-time animal motion capture and neural character control, improving the experiences of user's interaction with virtual animals [16] (Figure 9).



Fig 9. The picture shows a concept demo of animal and user interaction in a VR environment in Artemis: Articulated Neural Pets with Appearance and Motion Synthesis [16].

Daniel Holden and colleagues use deep learning with convolutional autoencoders to synthesize lifelike human motion from large datasets, along with a feedforward neural network for intuitive motion generation. They introduce future improvements like deeper networks and advanced depooling techniques to handle complex body movements in animation more effectively [17].

Chaitanya Ahuja, Dong Won Lee, Yukiko I. Nakano, and Louis-Philippe Morency introduce Mix-StAGE, a model for generation of co-speech gestures that can change based on multiple speakers' styles using audio inputs. Assessed on the PATs dataset, Mix-StAGE trumps existing methods in gesture quality and style transfer, which is validated through human perception studies [18].

6. Grasp and Hand Object Interactions

LI, YING, and etc presents a shape matching algorithm for creating realistic grasping in virtual environments, exploiting a database of hand poses to make natural interactions with objects smooth and improving grasp quality through task-based metrics[19].

JIANG, HANWEN, and etc introduce a framework for creating vivid human grasps through combining hand grasp synthesis with object contact map estimation. The approach make full advantage of a Conditional Variational AutoEncoder and a ContactNet, showcasing refined grasp realism and ability to adapt across different datasets and novel inputs[20] (Figure 10).

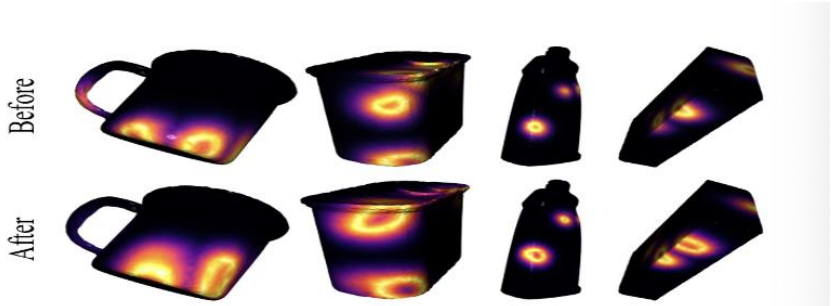


Fig 10. The picture shows the object contact regions where the yellow circles on contact maps reflect the intersection rings between hand-object, as it is discussed in Hand-object contact consistency reasoning for human grasps generation"[20].

7. Advanced Motion Synthesis and Interaction

Weiyu Li and etc explore cutting edge techniques in computer animation, putting an emphasis on data-driven and deep learning approaches like GenMM for high-quality, diverse motion synthesis in complex skeletal structures, combined with future neural generative models [21].

Sebastian Eriksson and etc use SMC methods for real-time human motion synthesis, combining multimodal tracking, trajectory optimization, and MPC for improved realism and efficiency in applications, such as game engines and interactive training[22] (Figure 11).

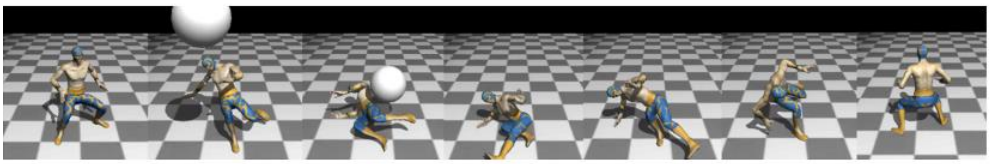


Fig 11. The picture gives an example of the synthesized animation.

Frame 1: balancing in the user-specified ready stance. Frames 2 and 3: the character aniticipates that the ball would hit him and dodges down. Frame 4: anticipation to get leg swing power. Frames 5, 6, and 7: swing the leg around and following with the rest of the body to get back to the ready stance [22].

Anindita Ghosh, and etc introduced a novel method for generating realistic human animations from complex textual descriptions using a hierarchical, two-stream network and contextualized BERT embeddings, significantly improving accuracy and robustness in motion synthesis. It outperforms previous techniques by effectively handling superimposed actions and long-range dependencies, though future work will address generalization and motion quality improvements [23].

Guo, C. and etc present a two-stage pipeline for generating realistic 3D human motions from text, using text2length to determine motion duration and text2motion to create diverse animations with a temporal variational autoencoder. It introduces the HumanML3D dataset, significantly improving on previous methods by handling a broad range of text complexities and motion types, and plans to explore simplification of evaluation metrics and motion captioning[24] (Figure 12).

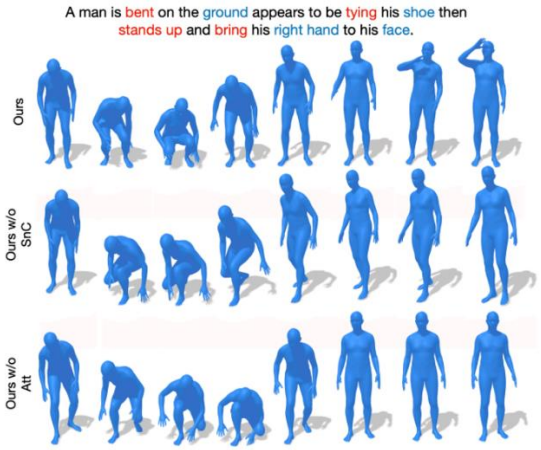


Fig 12. The picture shows visual comparison of motion results generated by their model [24].

Mohamed Hassa and colleagues explore challenges in simulating vivid character interactions with objects in computer graphics and introduce a newly developed framework using Adversarial Motion Priors, which is short as AMP, to improve adaptability in diverse scene contexts [25].

8. Discussion

Recent advancements in motion synthesis have explored a variety of techniques and applications, ranging from audio-driven motion generation to complex human-object interactions. This field spans multiple focus areas, including the synchronization of motion with audio, the use of generative models for creating new motion sequences, and the development of realistic animations for characters, pets, and grasping tasks. Each of these aspects presents unique challenges and opportunities for innovation, pushing the boundaries of interactive and immersive experiences in gaming, virtual reality, and animation (Table 1).

Table 1. Comparison between different methods of motion synthesis mentioned earlier.

Feature/ Aspect	Audio-Driven & Music-Driven Motion Synthesis	Generative Models & Frameworks for Motion Synthesis	Human & Object Interaction	Character & Pet Motion Synthesis	Grasp & Hand Object Interactions	Motion Retargeting & Editing	Advanced Motion Synthesis & Interaction
Focus	Synchronizing motion with audio input	Generating new motion sequences	Interaction with objects	Realistic motion for characters & pets	Handling and manipulating objects	Adapting motion data between characters	Cutting-edge methods in synthesis & interaction
Techniques	Audio feature extraction, mapping to motion	GANs, VAEs, neural networks	Physics-based simulations	Motion capture, animation frameworks	Kinematics, physics simulations	Algorithms for retargeting, editing tools	Integration of multiple modalities, advanced algorithms
Applications	Music videos, interactive experiences	Game development, virtual reality	Games, simulations	Animation, gaming, virtual assistants	Robotics, virtual environments	Game development, film production	High-fidelity simulations, next-gen interactive experiences
Challenges	Mapping accuracy, synchronization issues	Training data requirements, model complexity	Realism in interactions	Ensuring natural behavior	Realistic hand movements	Maintaining realism across characters	Computational complexity, integration of multiple systems

9. Conclusion

This paper presents a survey of different types of models in the field of character motion synthesis. With the development of technologies and emergence of deep learning and other novel tools of improving images quality and vividness of motions, character motion synthesis has become more robust, efficient, and of more quality and is highly likely to be prevailing in the related fields, such as game design, computational models, robotics, and film making. Future directions of refining motion synthesis could be combining deep learning with physics based controllers to better quality of motions.

References

[1] Simon Alexanderson, Rajmund Nagy, Jonas Beskow, Gustav Eje Henter. Listen, denoise, action! Audio-driven motion synthesis with diffusion models [J]. ACM Trans. Graph., Vol. 42, No. 4

[2] Wenlin Zhuang, Congyi Wang, Jinxiang Chai, Yangang Wang, Ming Shao, and Siyu Xia. Music2dance: Dancenet for music-driven dance generation. ACM Trans. Multimedia Comput. Commun. Appl., 2022. 3, 6

[3] Rishabh Dabral, Muhammad Hamza Mughal, Vladislav Golyanik, and Christian Theobalt. 2023. MoFusion: A Framework for Denoising-Diffusion-based Motion Synthesis. In Computer Vision and Pattern Recognition (CVPR).

[4] PEIZHUO LI, KFIR ABERMAN, ZIHAN ZHANG, RANA HANOCKA, OLGA SORKINE-HORNUNG. GANimator: Neural Motion Synthesis from a Single Sequence [J]. ACM Trans. Graph., Vol. 41, No. 4.

[5] Prafulla Dhariwal and Alexander Nichol. 2021. Diffusion Models Beat GANs on Image Synthesis. In Advances in Neural Information Processing Systems (NeurIPS '21). 8780–8794. https://proceedings. neurips.cc/paper_files/paper/2021/file/49ad23d1ec9fa4bd8d77d02681df5cfa-Paper.pdf

[6] GUSTAV EJE HENTER∗, SIMON ALEXANDERSON∗, JONAS BESKOW. MoGlow: Probabilistic and Controllable Motion Synthesis Using Normalising Flows [J]. ACM Trans. Graph., Vol. 39, No. 4.

[7] Simon Alexanderson, Gustav Eje Henter, Taras Kucherenko, and Jonas Beskow. 2020. Style-controllable speech-driven gesture synthesis using normalising flows. Comput. Graph. Forum 39, 2 (2020), 487–496. https://doi.org/10.1111/cgf.13946

[8] Anindita Ghosh, Rishabh Dabral, Vladislav Golyanik, Christian Theobalt, Philipp Slusallek. IMoS: Intent-Driven Full-Body Motion Synthesis for Human-Object Interactions [J].

[9] Guo, C.; Zuo, X.; Wang, S.; Zou, S.; Sun, Q.; Deng, A.; Gong, M.; and Cheng, L. 2020. Action2motion: Conditioned generation of 3d human motions. In Proceedings of the 28th ACM International Conference on Multimedia, 2021–2029.

[10] Petrovich, M.; Black, M. J.; and Varol, G. 2021. Actionconditioned 3d human motion synthesis with transformer vae. In Proceedings of the IEEE/CVF International Conference on Computer Vision, 10985–10995.

[11] He Zhang, Yuting Ye, Takaaki Shiratori, and Taku Komura. 2021. Manipnet: neural manipulation synthesis with a hand-object spatial representation. ACM Transactions on Graphics (ToG) 40, 4 (2021), 1–14.

[12] Daniel Holden∗, Jun Saito† t, Taku Komura‡, A Deep Learning Framework for Character Motion Synthesis and Editing [J]. ACM Trans. Graph., Vol. 35, No. 4.

[13] Noshaba Cheema, Perttu Hämäläinen, Vladislav Golyanik, Nam Hee Kim, Marc Habermann, Christian Theobalt , Philipp Slusallek. Discovering Fatigued Movements for Virtual Character Animation [J]. SA Conference Papers '23, December 12–15, 2023, Sydney, NSW, Australia.

[14] JIAMAN LI, JIAJUN WU†, C. KAREN LIU†. Object Motion Guided Human Motion Synthesis [J]. ACM Trans. Graph., Vol. 42, No. 6.

[15] Jihoon Kim, Jiseob Kim, and Sungjoon Choi. 2022. FLAME: Free-Form Language Based Motion Synthesis & Editing. arXiv preprint arXiv:2209.00349 (2022). https: //arxiv.org/abs/2209.00349

[16] HAIMIN LUO, TENG XU, YUHENG JIANG, CHENGLIN ZHOU, QIWEI QIU, YINGLIANG ZHANG, WEI YANG, JINGYI YU∗. Artemis: Articulated Neural Pets with Appearance and Motion Synthesis [J]. ACM Trans. Graph., Vol. 41, No. 4.

[17] Daniel Holden∗, Jun Saito† t, Taku Komura‡, A Deep Learning Framework for Character Motion Synthesis and Editing [J]. ACM Trans. Graph., Vol. 35, No. 4.

[18] Chaitanya Ahuja, Dong Won Lee, Yukiko I. Nakano, and Louis-Philippe Morency. 2020b. Style Transfer for Co-Speech Gesture Animation: A Multi-Speaker Conditional Mixture Approach. In Proceedings of the European Conference on Computer Vision (ECCCV '20). 248–265. https://doi.org/10.1007/978-3-030-58523-5_15

[19] LI, YING, FU, JIAXIN L, and POLLARD, NANCY S. "Data-driven grasp synthesis using shape matching and task-based pruning". IEEE Transactions on visualization and computer graphics (2007)

[20] JIANG, HANWEN, LIU, SHAOWEI, WANG, JIASHUN, and WANG, XIAOLONG. "Hand-object contact consistency reasoning for human grasps generation". Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021 2, 3.

[21] WEIYU LI∗†, XUELIN CHEN∗‡, PEIZHUO LI, OLGA SORKINE-HORNUNG, BAOQUAN CHEN. Example-based Motion Synthesis via Generative Motion Matching [J]. ACM Trans. Graph., Vol. 42, No. 4

[22] Perttu H¨am¨al¨ainen1∗, Sebastian Eriksson Esa Tanskanen, Ville Kyrki1, Jaakko Lehtinen. Online Motion Synthesis Using Sequential Monte Carlo [J]. Vol. 33, No. 4.

[23] Anindita Ghosh, Noshaba Cheema, Cennet Oguz, Christian Theobalt, and P. Slusallek. Synthesis of compositional animations from textual descriptions. In ICCV, 2021. 1, 2

[24] 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 Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 5152–5161.

[25] Mohamed Hassan, Yunrong Guo, Tingwu Wang, Michael Black, Sanja Fidler, and Xue Bin Peng. 2023. Synthesizing Physical Character-Scene Interactions. (2023), 1–9.