Human Motion Generation From Text

Xinran Li
ShanghaiTech University
lixr@shanghaitech.edu.cn

Yuyao Chen
ShanghaiTech University
chenyy6@shanghaitech.edu.cn

Abstract

Expressive and natural human motion generation is a rewarding area of computer vision, which is because the generation is really a challenging task on account of the complex diversity of human motion, human perception of given text, and the difficulty of accurately describing human motion. However, we have many current methods to generate human motion such as CLIP, diffusion model, they have some limitations on generation based on more specific description, and are maybe not consistent for out-of-domain reference examples. We introduce a new generation model combining CLIP and DDPM training methods to train a more accurate and diverse model from multiple text input. Our goal is to generate a dynamic 3D motion that shows a continuous movement in a short period of time based on the user's text input, and to generate a random diversity of motions based on the same text input, showing randomness and diversity to meet different user needs. We evaluate our model on a dataset of human motion descriptions and compare it with a baseline approach. Our decoder conditional on the motion representation can also produce variants of the motion while preserving its semantics and style, while changing non-essential details that are not present in the motion representation. In addition, CLIP's joint embedding space supports language-guided motion manipulation in a zero-sample fashion. We used a diffusion model for the decoder and experimented with an autoregressive model and a diffusion model for the a priori model, finding that the latter was computationally more efficient and produced higher quality samples. As we demonstrate, our new generation model is a generic approach, enabling different modes of conditioning, and different generation tasks. We show that our model is trained with lightweight resources and yet achieves state-of-the-art results on leading benchmarks for text-to-motion.

1 Introduction

Motion generation from text is a recently emerging field of research. It involves the task of generating a sequence of frames representing a human motion from a text description. This can be used for applications such as virtual reality, animation, and video games. The task is challenging as it requires understanding of natural language, as well as the ability to generate realistic motion.

Recent progress in computer vision has been driven by scaling models on large datasets of captioned images collected from the internet. Within this framework, CLIP[26] has emerged as a successful image representation learner. CLIP embedding has many desirable properties: they are robust to image distribution shifts, have impressive zero-sample capabilities, and are fine-tuned to achieve state-of-the-art results on a variety of visual and linguistic tasks. Meanwhile, diffusion models have emerged as a promising framework for generative modelling, driving the latest developments in image and video generation tasks. To obtain optimal results, diffusion models utilise a bootstrapping technique that improves sample fidelity (for image, photo-level realism) at the expense of sample diversity. In this work, we combine these two approaches to solve the problem of text conditional motion generation. We first train a diffusion decoder to invert the CLIP encoder. Our invertor is non-deterministic and can generate multiple motions corresponding to a given motion embedding.

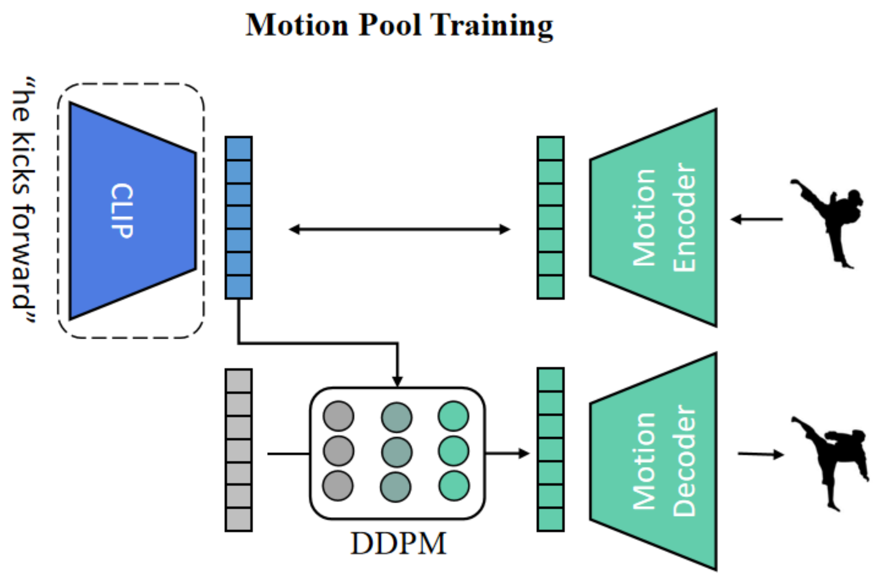

Figure 1: A overview of our human text-to-motion model

The presence of the encoder and its approximate inverse (decoder) allows for the ability to go beyond text-to-motion translation.

In this work, we combine these two approaches to solve the problem of text conditional motion generation. We first train a diffusion decoder to invert the CLIP encoder. Our inverter is non-deterministic and can generate multiple motions corresponding to a given motion embedding. The presence of the encoder and its approximate inverse (decoder) allows for the ability to go beyond text-to-motion translation.

To obtain a complete model for image generation, we combine a CLIP motion embedding decoder with a prior model that generates possible CLIP motion embedding from a given text caption. We also develop methods for training diffusion priors in the latent space and show that they achieve comparable performance to auto regressive priors, while being more computationally efficient.

2 Related Work

2.1 Human Motion Generation

Neural motion generation learned from motion-capture data can be conditioned by any signal describing motion. Any signal that describes motion. So many methods use parts of the movement itself for specific guidance. Some of the related previous work predict human motion from its prefix poses[1, 2, 3, 4]. Others[5, 6, 7, 8] make use of bi-directional GRU[9] and Transformer[10] to solve super-resolution and in-betweening tasks. Other work[11] uses the automatic encoder to learn the latent representation of motion, and then use it to edit and control the motion with spatial constraints, such as root locus and bone length.

Recently, another generative model that has recently attracted a lot of attention is NeRF([33, 34], which has had considerable success in rendering realistic images. An implicit neural representation (INR) is a series of neural networks that optimise their parameters to fit a sample rather than an entire distribution. A major advantage is the technique's ability to generalise extremely well in the spatial or temporal dimension. For example, Cervantes[35] proposed an implicit scheme which models both action categories and timestamps. Similar to the original NeRF, the timestamps are represented by sine values. After supervised training, the proposed method can generate a variable-length motion sequence for each action category.

Apart from input text, human motion can also be controlled with a high-level guidance from natural language[12, 13], action class[14, 15, 16], audio[17]. For instance, recent works[18] generated dance moves conditioned on music and the motion prefix, Edwards[19] generate facial expressions to fit spoken audio sequences. In most cases, the authors recommend a specialized approach to map each regulatory domain to human motion.

2.2 Text to Motion

In recent years, the dominant approach for text-to-motion tasks is to learn a shared latent space of language and motion. A motion-language dataset named KIT[20] offers about 11 hours of motion-capture sequences, each paired with a sentence that clearly describes the action being made. KIT sentences describe movement type, direction, and sometimes speed, but lack details about movement style and do not include abstract descriptions of movement. So Currently most of the researches are based on KIT. Yamada[21] learns both mappings by simultaneously training the text and motion auto-encoders to bind their latent spaces using both text and motion pairs. JL2P[22] learns the KIT motion-language dataset with an auto-encoder, constrained to a one-to-one mapping from text to motion, which has been improved in terms of the subtle concepts of text (i.e. speed, trajectory and type of action). They also learned to joint motion-text latent space and apply training curriculum to ease optimization. Lin[22] has further improved trajectory prediction by adding dedicated layers.

Another dataset BABEL[23] provides per-frame text labels sorted by 260 classes for the larger AMASS dataset [24], including approximately 40 hours of motion capture. Although a clear description of the action is provided, any detail beyond the type of action is usually missing, but this data covers a wider variety of human motions. MotionCLIP[25] extends text-to-motion data limits and enables latent space editing using shared text image latent spaces learned by CLIP[26].

2.3 Diffusion Generation Model

Diffusion models[27] are a class of neural generative models based on the stochastic diffusion process modelled in thermodynamics. In this setup, samples from the data distribution are gradually noised by the diffusion process. The neural model then learns the reverse process of gradually denoising the samples. For conditioned generation, Dhariwal & Nichol[29] to enable conditioning on CLIP textual representations. More recently, Zhang[30] and Kim[31] have suggested diffusion models for motion generation. For the MotionDiffuse model proposaled by Zhang et al., unlike autoregressive inference schemes that typically require many long motion sequences for training, MotionDiffuse is able to model correlations between successive movements without introducing additional training costs. MDM[32] is a diffusion-based generative model carefully tuned for the human motion domain. Being diffusion-based, the MDM benefits from the local many-to-many domain representation described above, as evidenced by the quality and diversity of the resulting movements. Furthermore, MDM incorporates insights already established in the field of motion generation, helping it to become lighter and more controllable.

3 Criticism

During recent period of time, the MotionCLIP[25] method generated in 2022, it proposes a motion generation network that makes use of the knowledge encapsulated in CLIP to allow intuitive operations such as text conditional motion generation and editing. But it still have limitations in understanding directions like left, right and counter-clockwise. It also has difficulty in capture some styles(e.g. heavy and proud), and is of course not consistent of some cultural reference examples. For examples, this model fails to produce *Cristiano Ronaldo's* goal celebration and *batman's* signature pose.

MDM[32], a method suitable for a variety of human motion generation tasks. MDM is an atypical classifier-free diffusion model with a transformer encoder backbone and predicts the signal, rather than the noise. A significant limitation of the diffusion approach is the long inference time, with a single result requiring about 1000 forward passes. Since our motion model is small in any case, using a dimension an order of magnitude smaller than the image reduces its inference time from less than a second to about a minute.

Another obvious shortcoming is that although MDM used CLIP in the generation phase, it only masked CLIP randomly for classifier-free learning and did not joint embedding space of CLIP with text, in fact MDM did not link semantics and motion together well, their semantic generation was poor.

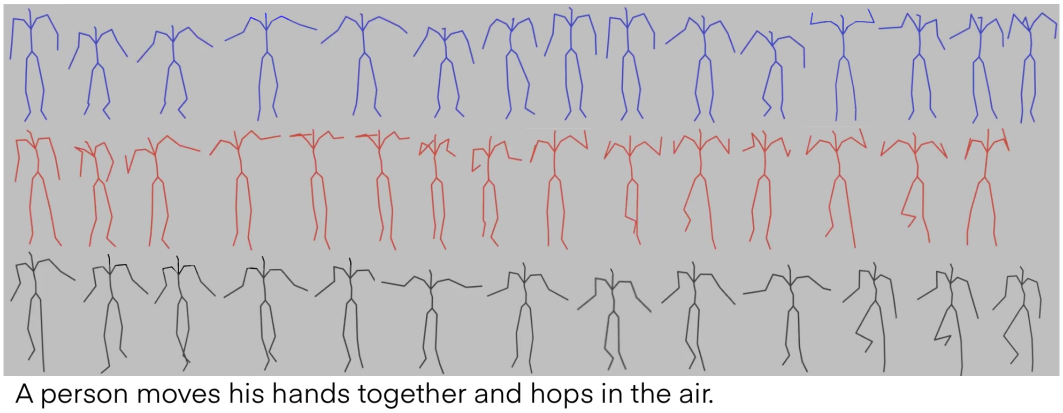

Figure 2: A complex example of motion generation of same input text

MotionDiffuse[30], the first text-driven motion generation method based on a diffusion model. MotionDiffuse demonstrates three main advantages: probabilistic mapping to enhance diversity, realistic synthesis to ensure rationalisation of motion sequences, and multi-level manipulation to allow manipulation of each part and long sequence generation. Although MotionDiffuse pushes the performance boundaries of motion generation tasks forward, a number of issues remain. Firstly, diffusion models require a large number of diffusion steps during inference and generating motion sequences in real time is challenging. Secondly, the current pipeline only accepts a single form of motion representation. A more general pipeline that also adapts to all datasets would be more suitable for a variety of scenarios.

4 Numerical results

Our model successfully generate many diverse and interesting human motions given same input text, I will show some image of this motion in this report(2, 3, 4) and specific demonstration video in our code zip.

We use the HumanML3D[36] dataset for text-tomotion generation, examine MotionCLIP's ability to convert text into animation. Since the latent space of motion is aligned with CLIP, we use CLIP's pre-trained text encoder to process the input text and use MotionCLIP's decoder to convert the resulting latent embedding into motion. We then put the trained latent space into the DDPM model to train the motion decoder, which is better than the original MDM model because we learn the semantics and style of the given text better by using CLIP.

5 Conclusion

In this paper, we proposed a novel approach for human motion generation from text. Our model takes as input a text description of a motion, and outputs a sequence of motion frames that represent the motion described in the text. To obtain a complete motion generation model, we combine the CLIP motion embedding decoder with an a prior model that generates possible CLIP motion embedding from a given text heading, and then this embedding is used to condition a diffusion decoder which produces a final motion. Nevertheless, we see that the same autoencoder with the same data can understand motion manifolds and their semantics significantly better, simply by aligning them with well-behaved, knowledge-rich latent spaces.

References

[1] Alejandro Hernandez, Jurgen Gall, and Francesc Moreno-Noguer. Human motion prediction via spatio-temporal inpainting. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7134–7143, 2019.

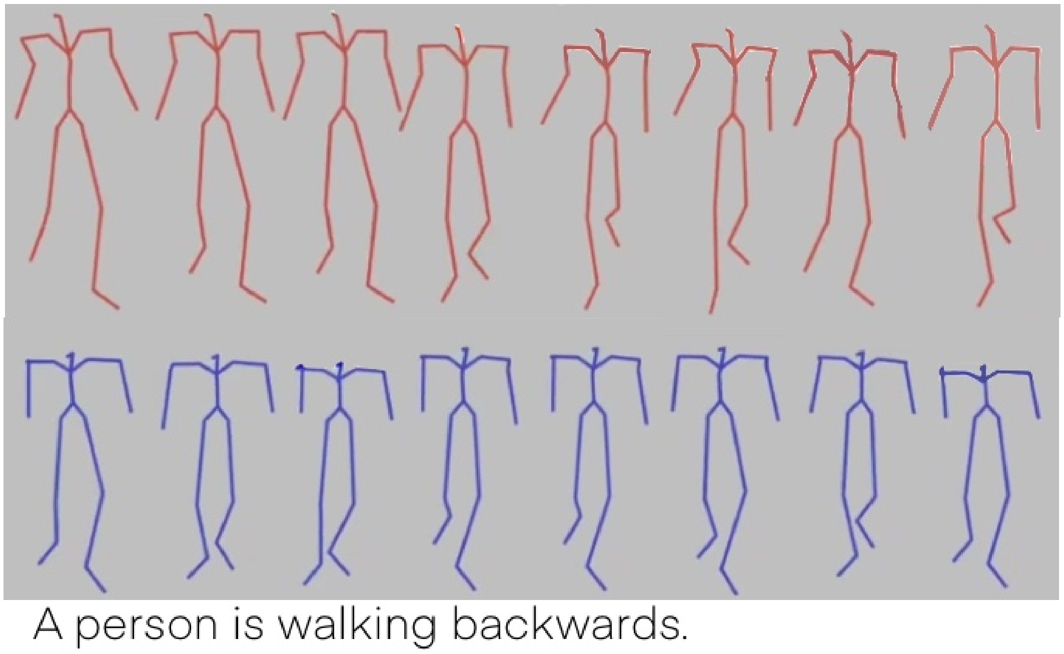

Figure 3: A simple example of motion generation of same input text

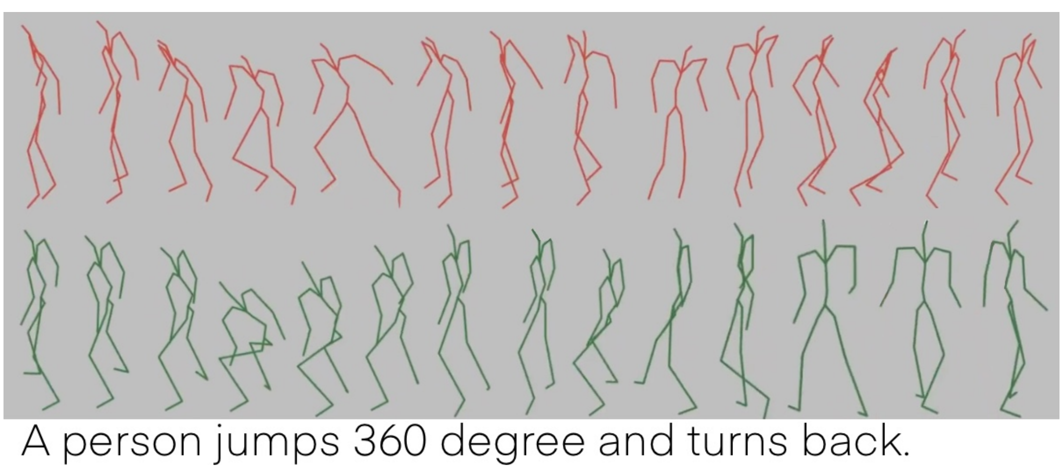

Figure 4: A simple example of motion generation of same input text

[2] Wen Guo, Yuming Du, Xi Shen, Vincent Lepetit, Xavier Alameda-Pineda, and Francesc Moreno Noguer. Back to mlp: A simple baseline for human motion prediction. *arXiv preprint arXiv:2207.01567*, 2022b.

[3] Mathis Petrovich, Michael J. Black and, Gül Varol. Action-Conditioned 3D Human Motion Synthesis with Transformer VAE. *International Conference on Computer Vision (ICCV)*. 10985–10995, 2021.

[4] Katerina Fragkiadaki, Sergey Levine, Panna Felsen, and Jitendra Malik. Recurrent network models for human dynamics. *Proceedings of the IEEE international conference on computer vision*, pp. 4346–4354, 2015.

[5] Manuel Kaufmann, Emre Aksan, Jie Song, Fabrizio Pece, Remo Ziegler, and Otmar Hilliges. Con volutional autoencoders for human motion infilling. *International Conference on 3D Vision (3DV)*, pp. 918–927. IEEE, 2020.

[6] Felix G Harvey and Christopher Pal. Recurrent transition networks for character locomotion. *SIGGRAPH Asia 2018 Technical Briefs*, pp. 1–4, 2018.

[7] Felix G Harvey, Mike Yurick, Derek Nowrouzezahrai, and Christopher Pal. Robust motion in between-ing.*ACM Transactions on Graphics (TOG)*, 39(4):60–1, 2020.

[8] Yinglin Duan, Tianyang Shi, Zhengxia Zou, Yenan Lin, Zhehui Qian, Bohan Zhang, and Yi Yuan. Single-shot motion completion with transformer. *arXiv preprint arXiv:2103.00776*, 2021.

[9] Kyunghyun Cho, Bart Van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk , and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

[10] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser , and Illia Polosukhin. Attention is all you need.*Advances in neural information processing systems*, 30, 2017.

[11] Daniel Holden, Jun Saito, and Taku Komura. A deep learning framework for character motion synthesis and editing. *ACM Transactions on Graphics (TOG)*, 35(4):1–11, 2016.

[12] Chaitanya Ahuja and Louis-Philippe Morency. Language2pose: Natural language grounded pose forecasting. *International Conference on 3D Vision (3DV)*, pp. 719–728. IEEE, 2019.

[13] Mathis Petrovich, Michael J. Black, and G¨ul Varol. TEMOS: Generating diverse human motions from textual descriptions. *European Conference on Computer Vision (ECCV)*, 2022.

[14] Mathis Petrovich, Michael J. Black, and G¨ul Varol. Action-conditioned 3D human motion synthesis with transformer VAE. *International Conference on Computer Vision (ICCV)*, pp. 10985– 10995, 2021.

[15] Chuan Guo, Xinxin Zuo, Sen Wang, Shihao Zou, Qingyao Sun, Annan Deng, Minglun Gong, and Li Cheng. Action2motion: Conditioned generation of 3d human motions. *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 2021–2029, 2020.

[16] Pablo Cervantes, Yusuke Sekikawa, Ikuro Sato, and Koichi Shinoda. Implicit neural representations for variable length human motion generation. *arXiv preprint arXiv:2203.13694*, 2022.

[17] Ruilong Li, Shan Yang, David A. Ross, and Angjoo Kanazawa. Ai choreographer: Music conditioned 3d dance generation with aist++. *The IEEE International Conference on Computer Vision (ICCV)*, 2021.

[18] Andreas Aristidou, Anastasios Yiannakidis, Kfir Aberman, Daniel Cohen-Or, Ariel Shamir, and Yiorgos Chrysanthou. Rhythm is a Dancer: Music-Driven Motion Synthesis with Global Structure. *arXiv preprint arXiv:2111.12159*, 2021.

[19] Pif Edwards, Chris Landreth, Eugene Fiume and Karan Singh. JALI: an animator centric viseme model for expressive lip synchronization. *ACM Transactions on graphics (TOG)* 35, 4 , 1–11, 2016.

[20] Matthias Plappert, Christian Mandery, and Tamim Asfour. The kit motion-language dataset. *Big data*, 4(4):236–252, 2016.

[21] Tatsuro Yamada, Hiroyuki Matsunaga, and Tetsuya Ogata. Paired recurrent autoencoders for bidirectional translation between robot actions and linguistic descriptions. *IEEE Robotics and Automation Letters* 3, 4 (2018), 3441–3448, 2018

[22] Chaitanya Ahuja and Louis-Philippe Morency. Language2pose: Natural language grounded pose forecasting. *2019 International Conference on 3D Vision (3DV)* , pp. 719–728. IEEE, 2019.

[22] Angela S Lin, Lemeng Wu, Rodolfo Corona, Kevin Tai, Qixing Huang, and Raymond J Mooney. Generating animated videos of human activities from natural language descriptions. *Learning2018*, 1, 2018.

[23] Abhinanda R. Punnakkal, Arjun Chandrasekaran, Nikos Athanasiou, Alejandra QuirosRamirez, and Michael J. Black. BABEL: Bodies, Action and Behavior with English Labels. *Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*. 722–731. 2021.

[24] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of Motion Capture as Surface Shapes. *International Conference on Computer Vision*. 5442–5451, 2019.

[25] Guy Tevet, Brian Gordon, Amir Hertz, Amit H Bermano, and Daniel Cohen-Or. Motionclip: Exposing human motion generation to clip space. it arXiv preprint arXiv:2203.08063, 2022.

[26] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. *International Conference on Machine Learning*, pp. 8748–8763. PMLR, 2021.

[27] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. *International Conference on Machine Learning*, pp. 2256–2265. PMLR, 2015.

[28] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021.

[29] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.

[30] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. Motiondiffuse: Text-driven human motion generation with diffusion model. *arXiv preprint arXiv:2208.15001*, 2022.

[31] Jihoon Kim, Jiseob Kim, and Sungjoon Choi. Flame: Free-form language-based motion synthesis & editing. *arXiv preprint arXiv:2209.00349*, 2022.

[32] Tevet G, Raab S, Gordon B, et al. Human motion diffusion model[J]. *arXiv preprint arXiv:2209.14916*, 2022.

[33] Mildenhall B, Srinivasan PP, Tancik M, Barron JT, Ramamoorthi R, Ng R. Nerf: Representing scenes as neural radiance fields for view synthesis. *European conference on computer vision*, Springer, pp 405–421,2020.

[34] Jain A, Tancik M, Abbeel P. Putting nerf on a diet: Semantically consistent few-shot view synthesis. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp 5885–5894, 2021.

[35] Cervantes P, Sekikawa Y, Sato I, Shinoda K. Implicit neural representations for variable length human motion generation. *arXiv preprint arXiv:220313694*, 2022.

[36] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5152–5161, 2022.