Sequence Modeling of Motion-Captured Dance

Emily Napier, Gavia Gray, Sageev Oore
Dalhousie University
Vector Institute
emily.napier@dal.ca, gray@vectorinstitute.ai

Abstract

By treating dance as a long sequence of tokenized human motion data, we build a system that can synthesize novel dance motions. We train a transformer architecture on motion-captured data represented as a sequence of characters. By prompting the model with different sequences or task tokens, we can generate motions conditioned on the movement of a single joint, or the motion of a specific dance move.

1 Introduction

Dance is composed from a vocabulary of movements and poses in sequence. Language models have been demonstrated as an effective method for learning representations of text [1] and other modalities [2–4]. A transformer language model is a natural choice to learn how to compose this form of language.

The practice and understanding of dance can gain from access to a tool that can compose and condition dance motions. In Section 2, we describe a method for the generation of movement conditioned on the movement of subsets of joints, or gestures, which can indicate the trajectory or the target movement quality of a complete movement. We condition these generations on the dancer performing the movement, the genre, and the song it will be paired with.

Treating motion as a generic sequence of tokens is in contrast to existing work, which has mostly focused on treating values in the sequence as continuous [5, 6]. The advantage of a discrete sequence is training becomes identical to training on text with a cross-entropy loss function. We observe that this allows us to avoid problems with "freezing" during generation. Unfortunately, the success of this method is limited by dataset size because it lacks strong priors about motion, such as the laws of physics [7]. The relative size of available motion datasets is compared in Table 1 to other modalities.

2 Methods

Motion data is typically a sequence of poses, each pose is a sequence of joint angles, typically the 24 canonical joints of the SMPL body model [8]. The largest publicly available dataset of human motion is the AMASS [9] dataset.

Following the method described in Janner et al. [4], each dimension of each joint axis-angle vector was binned uniformly. To simplify the task, we only include 13 of the original 24 joints. The resulting integers are matched to arbitrary alphanumeric unicode characters so they can be used in a generic text model as is. Each frame is represented by a "word" with a space placed between frames.

A causal language model with 26 million parameters was pre-trained for 7500 iterations on the AMASS dataset processed with the data splits defined in [12], and the AIST++ [13] dataset. The model was finetuned on the AIST++ dataset with conditioning tokens based on the motion descriptions as illustrated in Figure 1.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

Table 1: Comparison of relative information content of datasets. Size is reported in bits per token for generative models trained on each dataset. The reported bits/frame was trained on all joints rather than the subset used elsewhere in this paper.

Name	Description	Size (bits)
ImageNet	Image Database	179G (3.57 bits/pixel) [10]
The Pile	Text Database	837G (2.45 bits/token) [11]
AMASS	Motion Database	1.6G (186 bits/frame)

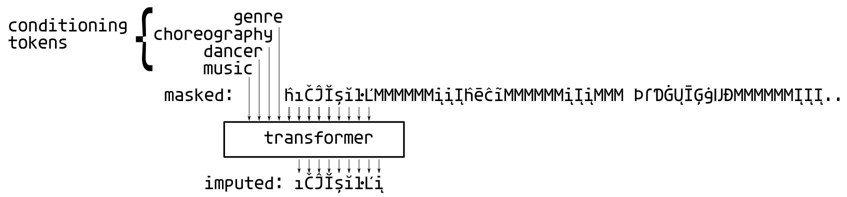

Figure 1: Illustration of how the transformer operates on an example of text with conditioning tokens prepended to the sequence. Masked tokens are denoted with "M", the causal model moves from left to right inferring masked tokens.

The pretrained model was trained to 115 bits/frame, and after finetuning on the AIST++ dataset reached 85 bits/frame, while incurring an absolute quantization error that was not noticeable.

3 Dance Generation

Motions are generated using any number and combination of joint inputs as context for the model, as shown in Figure 2. Generated outputs are conditioned on the tokens identifying the dancer, genre, and target music.

From rendered examples (see this link) we can see that conditioning tokens promote diversity in the generated output, and promote common movements from the target conditions. For example, movements conditioned on "waack" tokens demonstrated more circular arm movements, movements conditioned on "krump" tokens demonstrated more downward movements with the arms and legs, and movements conditioned on "house" tokens demonstrated more bending in the knees, leading to more frequent changes of weight placement.

Conditioning dance generation on a select set of joints allows us to generate long dance sequences that maintain a certain coherence (via the user-controlled joints) despite the limitation of the available model time window. By further conditioning the output on individual dancers, genres, or musical properties, choreographers can explore how specific dancers or dancers with specific dance backgrounds might adapt to their gestures and cues, or generate motion for tasks they are less familiar with.

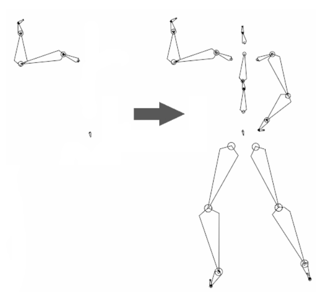

Figure 2: Illustration of the joints used as context for the model, and the generated output. The video of this motion can be found in Demo Video 1.

4 Conclusion

This work provides a foundation for learning human movement from data using the tools of language modeling, providing an interface between these areas of research and enabling exciting new directions in both understanding and composing dance with the help of machine learning.

Acknowledgments and Disclosure of Funding

We would like to thank David Rokeby, Douglas Eacho, Xavier Snelgrove, Michael Thaut, and Tristan Loria. Resources used in preparing this research were provided, in part, by the Province of Ontario, the Government of Canada through CIFAR, and companies sponsoring the Vector Institute. Funding for this research was provided by CIFAR, Research Nova Scotia, and NSERC.

Ethical Implications

There is some small risk this could be used analogous to deepfakes [14] by prompt tuning a conditioning token to a sample from someone's movement. Similarly, the model could be used to generate motion that appears violent or otherwise disturbing.

References

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv*, abs/1810.04805, 2019.

[2] Andrew Jaegle, Felix Gimeno, Andrew Brock, Andrew Zisserman, Oriol Vinyals, and João Carreira. Perceiver: General perception with iterative attention. In *ICML*, 2021.

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2020. URL `https://arxiv.org/abs/2010.11929`.

[4] Michael Janner, Qiyang Li, and Sergey Levine. Offline reinforcement learning as one big sequence modeling problem, 2021. URL `https://arxiv.org/abs/2106.02039`.

[5] Guillermo Valle-Pérez, Gustav Eje Henter, Jonas Beskow, André Holzapfel, Pierre-Yves Oudeyer, and Simon Alexanderson. Transflower: probabilistic autoregressive dance generation with multimodal attention. 2021.

[6] Emre Aksan, Manuel Kaufmann, and Otmar Hilliges. Structured prediction helps 3d human motion modelling. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2019. First two authors contributed equally.

[7] Kevin Xie, Tingwu Wang, Umar Iqbal, Yunrong Guo, Sanja Fidler, and Florian Shkurti. Physics-based human motion estimation and synthesis from videos, 2021. URL `https://arxiv.org/abs/2109.09913`.

[8] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, October 2015.

[9] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of motion capture as surface shapes. In *International Conference on Computer Vision*, pages 5442–5451, October 2019.

[10] Aaron van den Oord, Nal Kalchbrenner, Lasse Espeholt, koray kavukcuoglu, Oriol Vinyals, and Alex Graves. Conditional image generation with pixelcnn decoders. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL `https://proceedings.neurips.cc/paper/2016/file/b1301141feffabac455e1f90a7de2054-Paper.pdf`.

[11] Sid Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, Michael Pieler, USVSN Sai Prashanth, Shivanshu Purohit, Laria Reynolds, Jonathan Tow, Ben Wang, and Samuel Weinbach. GPT-NeoX-20B: An open-source autoregressive language model, 2022. URL `https://arxiv.org/abs/2204.06745`.

[12] Davis Rempe, Tolga Birdal, Aaron Hertzmann, Jimei Yang, Srinath Sridhar, and Leonidas J. Guibas. Humor: 3d human motion model for robust pose estimation. In *International Conference on Computer Vision (ICCV)*, 2021.

[13] Ruilong Li, Shan Yang, David A. Ross, and Angjoo Kanazawa. Learn to dance with AIST++: Music conditioned 3d dance generation, 2021.

[14] Yisroel Mirsky and Wenke Lee. The creation and detection of deepfakes. *ACM Computing Surveys*, 54(1): 1–41, jan 2022. doi: 10.1145/3425780. URL `https://doi.org/10.1145%2F3425780`.