MVA Computer Vision Project Report
Topic L: Text-Conditioned 3D human motion synthesis

CHAUVIN Paul
paulchauvin97@gmail.com

DELLIAUX Thomas
thomas.delliaux@etu.u-paris.fr

LOISEAU Thibaut
thibaut.loiseau@gmail.com

January 2023

1 Introduction

The objective of this project is to create a generative model that can produce 3D human motion sequences by using a provided textual description. This technology holds potential for various fields such as virtual reality, gaming, and human-robot interaction through speech. Additionally, this technology can also be applied in the film and gaming industries to produce special effects with humans, which are expensive and time consuming. By providing an automatic method of generating new motion data, this technology can save both time and cost. Additionally, the use of text as an input for the model allows for an intuitive and natural way for human-computer interaction.

2 Motivation and definition of the problem

The goal of this project is to investigate the potential of using transformer models to generate 3D human motion sequences based on natural language descriptions. Our research question is: How can we generate 3D human motion sequences from natural language descriptions using transformer models? To answer this question, we will be reproducing the work of previous research studies, such as [1, 2], which have used a Variational Auto-Encoder (VAE) to generate human motion from action labels or textual descriptions. Furthermore, we will also be exploring potential improvements and new directions for this technology, such as changing the data set to retrain the model.

3 Literature review

We present a summary of recent research in the field of human motion synthesis and text-conditioned motion generation, highlighting the use of various methods such as GANs [3, 4], VAEs [5, 6], and normalizing flows [7, 8]. We note that there are two main categories of motion synthesis - unconstrained generation [9, 10] and conditioned synthesis [3, 11, 12, 13] - and that VAEs are seen as more effective and easier to train than GANs. We also discuss text-conditioned motion generation, where methods such as sequence-to-sequence approaches [14, 15, 3]and joint cross-modal embeddings [11, 16] have been used.

We also mention the limitations of many current methods, such as poor body motion representation [17, 18] or unrealistic global trajectory [17, 18]. Additionally, many state-of-the-art methods are deterministic [3, 16]. In this project, we reproduce an already published article that addresses these limitations by using a variational approach with Transformers to generate diverse and realistic motions.

We try to reproduce TEMOS paper [2] and then modifying some of their approaches. TEMOS [2] approach uses a single embedding for the motion sequences and encodes the distribution parameters of the VAE, which is not present in other methods. Furthermore, their approach doesn't require hand crafting the encoding of upper and lower body separately, as is done in other approaches. Our work aims to improve upon the existing methods and contribute to the advancement of the field.

4 Methodology

This work aims to reproduce the paper TEMOS [2] by generating 3D human motion sequences from natural language descriptions using transformer models. We utilize a Variational Auto-Encoder (VAE) and Transformers to learn a joint latent space between motion and text. The model has two encoders for motion and text, the motion encoder takes a sequence of vectors as input and the text encoder takes word embeddings from a pre-trained language model as input. The motion decoder generates 3D human motion sequences non-autoregressively from a single latent vector, obtained from one of the two encoders during training. The model is capable of producing variable durations, adding another source of diversity to the generated motions.

4.1 Data used

In this project, we use the KIT Motion-Language dataset (KIT) [17] and the HumanML3D dataset [19] to train our model.

The KIT Motion-Language dataset which provides raw motion capture data and processed data using the Master Motor Map framework. This dataset consists of 3911 motion sequences with 6353 sequence-level description annotations and an average number of word per sequence equal to 9.5. The project uses the same splits as in Language2Pose [11] by

extracting 1784 training, 566 validation and 587 test motions. On the other hand, the HumanML3D dataset [19] is much bigger with 14616 motion sequences, with 44970 sequence level description annotations and an average number of word per sequence equal to 12.

The project employs the VAE formulation and Transformer models to learn a joint latent space between the two modalities: motion and text. The project will evaluate the results using average positional error (APE) and average variance error (AVE) metrics. It will also generate multiple different motions and evaluate the closest sample to the ground truth. Additionally, the project will rely on additional perceptual studies to assess the quality of the generated motions.

4.2 Implementation

We first used the code from [2] to train and reproduce the results. We reproduce exactly the paper except for the hardware, where we used T4 instead of V100 for economic reasons.

Then, we went further, by adapting TEMOS paper [2] and using another dataset : HumanML3D. As the dataset is in a different format, we decided to create a script that cleans the data and puts it to the right format. One part of the dataset is from AMASS dataset, which is in smplh format, the other part of the dataset is from HumanAct12. We load the AMASS part with the same tools that in TEMOS, except we used another index to know where to find the right data.

Also we chose to transform the rotations into joints in order to compare our model with the model train on KIT ML with the MMM format. In HumanAct12, there are 22 joints xyz for each skeleton (In KIT there are 21 joints).The extra joint is on the spine, so we removed it before the joints were transformed into features. We also modified the Data Load function. Finally, we also create a configuration file. We trained the model with the split provided on HumanML3D repository [19].

4.3 Experiments

In order to verify our results, we will compare what we have obtained with the baselines from the literature, both for reproducing the results and for the additional dataset. We will also qualitatively assess whether the obtained results are consistent.

5 Results

5.1 Quantitative results

The quantitative results in Table 1 show that the TEMOS original paper [2] had a lower Average Positional Error and Average Variance Error compared to the reproductions with batch size 32 and 24. However, the reproductions of the TEMOS experiment still had relatively low error rates. It can be seen that the global trajectory mean local error and mean global error for the root joint are relatively similar across all experiments, with the HUMAN ML 3D dataset having the lowest error rates.

It is worth noting that the standard deviation of the root joint global trajectory mean local error and mean global error is very low, indicating that the error values are consistent across the experiments.

Overall, these results suggest that TEMOS [2] is a robust model for predicting 3D human motion and that reducing the batch size does not significantly affect the performance of the model. The HUMAN ML 3D dataset also showed to be a good dataset to evaluate the performance of motion prediction models.

5.2 Qualitative results

As per our evaluation on KITSMPL test set, we found that the variant of our model which uses the parametric SMPL representation can generate full body meshes. We provide qualitative examples to illustrate the diversity of our generations for a given text. We observed that the model reproducing TEMOS [2] can generate multiple plausible motions corresponding to the same text, exploring the degrees of freedom remaining from ambiguities in the language description.

We reproduced TEMOS paper [2] and managed to have good qualitative results as we can see on Figure 1, for the input "A person walks in a circle clockwise".

On the other hand, when using HUMAN ML 3D, with text describing a precise action, such as 'A person is walking but paranoid someone is following them' or 'A person is jogging on the spot', the result is not working properly (Figures 2 and 3).

You can click on the following link to see the generated video for a model trained with batchsize 24 : link.

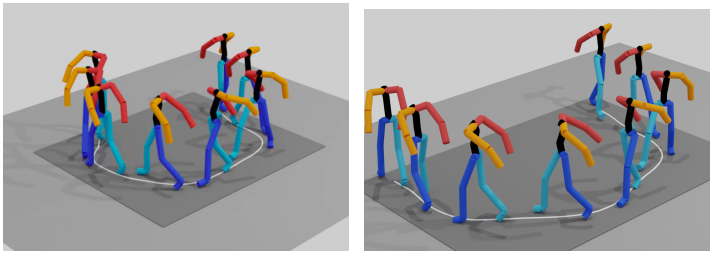

Figure 1: KIT dataset result for input: "A person walks in a circle clockwise". (left) Batchsize is 32 (right) Batchsize is 24

6 Encountered difficulties

During a project, we encountered several difficulties. One of the main difficulties we faced was a complex code architecture in TEMOS paper [2]. It was hard for us to understand, reproduce and change the code in the first place. Another difficulty we encountered was with the use of Google Virtual Machine. It crashed quite often, with hours

	Average Positional Error				Average Variance Error			
	root joint	global traj.	mean local	mean global	root joint	global traj.	mean local	mean global
TEMOS original paper with batch size 32 (KIT dataset)	0.963	0.955	0.104	0.976	0.445	0.445	0.005	0.448
TEMOS reproduction with same batch size of 32 (KIT dataset)	1.035	1.025	0.106	1.047	0.474	0.473	0.005	0.477
TEMOS reproduction with batch size of 24 (KIT dataset)	1.050	1.040	0.105	1.064	0.454	0.453	0.005	0.457
HUMAN ML 3D dataset	0.833	0.794	0.087	0.817	0.236	0.236	0.006	0.242

Table 1: Quantitative results

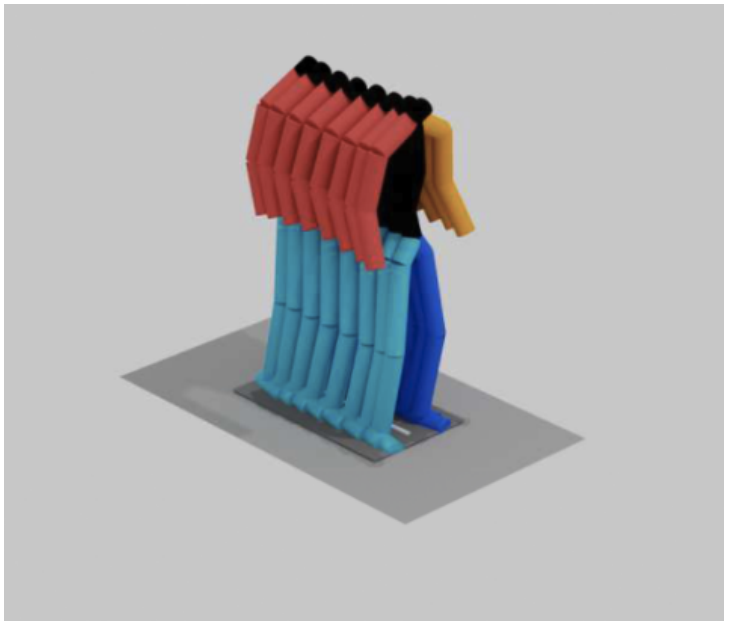

Figure 2: Human ML 3D dataset result for input: "A person is jogging on the spot"

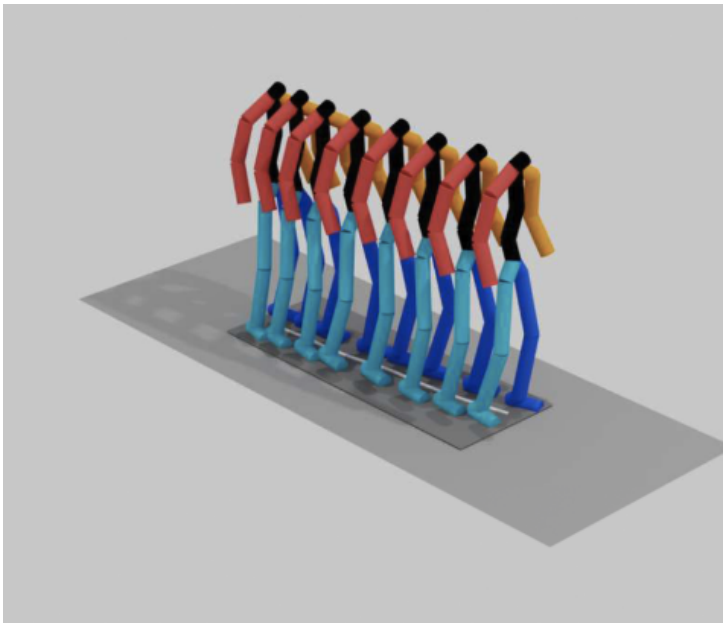

Figure 3: Human ML 3D dataset result for input:"A person is walking but paranoid someone is following them"

of training lost that we had to restart a few times before having successful training. Furthermore, we also struggled with working with difficult and unusual data types, such as MMM and SMPL manipulations. It required specialized knowledge or expertise, and it was challenging for us to implement, test, and debug the code related to such data at first.

7 Conclusion

In conclusion, our evaluation of the variant of our model using the parametric SMPL representation has yielded almost similar results as the reproduction of TEMOS paper [2]. However, we found that the performance of the model decreases when changing the number of epochs and batch size. We were able to obtain some results on the HUMAN ML 3D dataset with good quantitative results. However, the qualitative results were not sufficient. Overall, while the model has shown promise in generating full body meshes, there is still room for improvement in terms of both quantitative and qualitative results. Further research and experimentation is needed to optimize the model's performance.

References

[1] Mathis Petrovich, Michael J Black, and Gül Varol. Action-conditioned 3d human motion synthesis with transformer vae. In *Proceedings of the IEEE/CVF International Conference on*

Computer Vision, pages 10985–10995, 2021.

[2] Mathis Petrovich, Michael J Black, and Gül Varol. Temos: Generating diverse human motions from textual descriptions. *arXiv preprint arXiv:2204.14109*, 2022.

[3] Hyemin Ahn, Timothy Ha, Yunho Choi, Hwiyeon Yoo, and Songhwai Oh. Text2Action: Generative adversarial synthesis from language to action. October 2017.

[4] Xiao Lin and Mohamed R Amer. Human motion modeling using DVGANs. April 2018.

[5] Chuan Guo, Xinxin Zuo, Sen Wang, Shihao Zou, Qingyao Sun, Annan Deng, Minglun Gong, and Li Cheng. Action2Motion: Conditioned generation of 3D human motions. July 2020.

[6] Mathis Petrovich, Michael J Black, and Gül Varol. Action-conditioned 3D human motion synthesis with transformer VAE. April 2021.

[7] Gustav Eje Henter, Simon Alexanderson, and Jonas Beskow. MoGlow. *ACM Trans. Graph.*, 39(6):1–14, December 2020.

[8] Andrei Zanfir, Eduard Gabriel Bazavan, Hongyi Xu, Bill Freeman, Rahul Sukthankar, and Cristian Sminchisescu. Weakly supervised 3D human pose and shape reconstruction with normalizing flows. March 2020.

[9] Yan Zhang, Michael J Black, and Siyu Tang. Perpetual motion: Generating unbounded human motion. July 2020.

[10] Rui Zhao, Hui Su, and Qiang Ji. Bayesian adversarial human motion synthesis. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6224–6233, 2020.

[11] Chaitanya Ahuja and Louis-Philippe Morency. Language2Pose: Natural language grounded pose forecasting. July 2019.

[12] Chuan Guo, Xinxin Zuo, Sen Wang, Shihao Zou, Qingyao Sun, Annan Deng, Minglun Gong, and Li Cheng. Action2Motion: Conditioned generation of 3D human motions. July 2020.

[13] Jiaman Li, Yihang Yin, Hang Chu, Yi Zhou, Tingwu Wang, Sanja Fidler, and Hao Li. Learning to generate diverse dance motions with transformer. August 2020.

[14] Angela S. Lin, Lemeng Wu, and Qixing Huang Raymond J. Mooney Rodolfo Corona, Kevin Tai. Generating animated videos of human activities from natural language descriptions. In *Proceedings of the Visually Grounded Interaction and Language Workshop at NeurIPS 2018*, December 2018.

[15] Matthias Plappert, Christian Mandery, and Tamim Asfour. Learning a bidirectional mapping between human whole-body motion and natural language using deep recurrent neural networks. May 2017.

[16] Anindita Ghosh, Noshaba Cheema, Cennet Oguz, Christian Theobalt, and Philipp Slusallek. Synthesis of compositional animations from textual descriptions. March 2021.

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

[18] 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):3441–3448, October 2018. Publisher Copyright: © 2016 IEEE.

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