Generative models for human motion

Jose Ribeiro-Gomes[1], Alexandre Bernardino[1], Fernando de la Torre[2]

1) Instituto Superior Técnico, Lisboa
2) Carnegie Mellon University, Pittsburgh

josepgomes@tecnico.ulisboa.pt, alex@isr.tecnico.ulisboa.pt, ftorre@andrew.cmu.edu

CMU PORTUGAL
SUMMIT 2022
NEW FRONTIERS IN TECH

Introduction

This work aims to tackle the problem of generating realistic and diverse human motion sequences from highly semantic textual descriptions. We aim to develop a model that can interpret the prompts being fed and generate believable motions that were not part of the training set. We are using an adapted version of the generative variational autoencoder proposed in [1], combined with a language encoder/model (currently CLIP [2]), and training on the highly annotated motion capture dataset BABEL [3]. This generated motion is then fed through a kinematic model to improve the results and realism.

Contributions

We plan to fully mesh the resulting character, and integrate hand movement in the generated motion, which is often overlooked. Furthermore, we intend to implement physical constraints in our model that will hopefully improve the realism and allow for object and scene interaction.

Methods

- Motion and text are encoded onto latent space
- Both used for training; only text is used for testing
- Generated motion is sampled from latent space
- Generated motion is fine-tuned using physical constraints, adding kinematic models to latent spaces

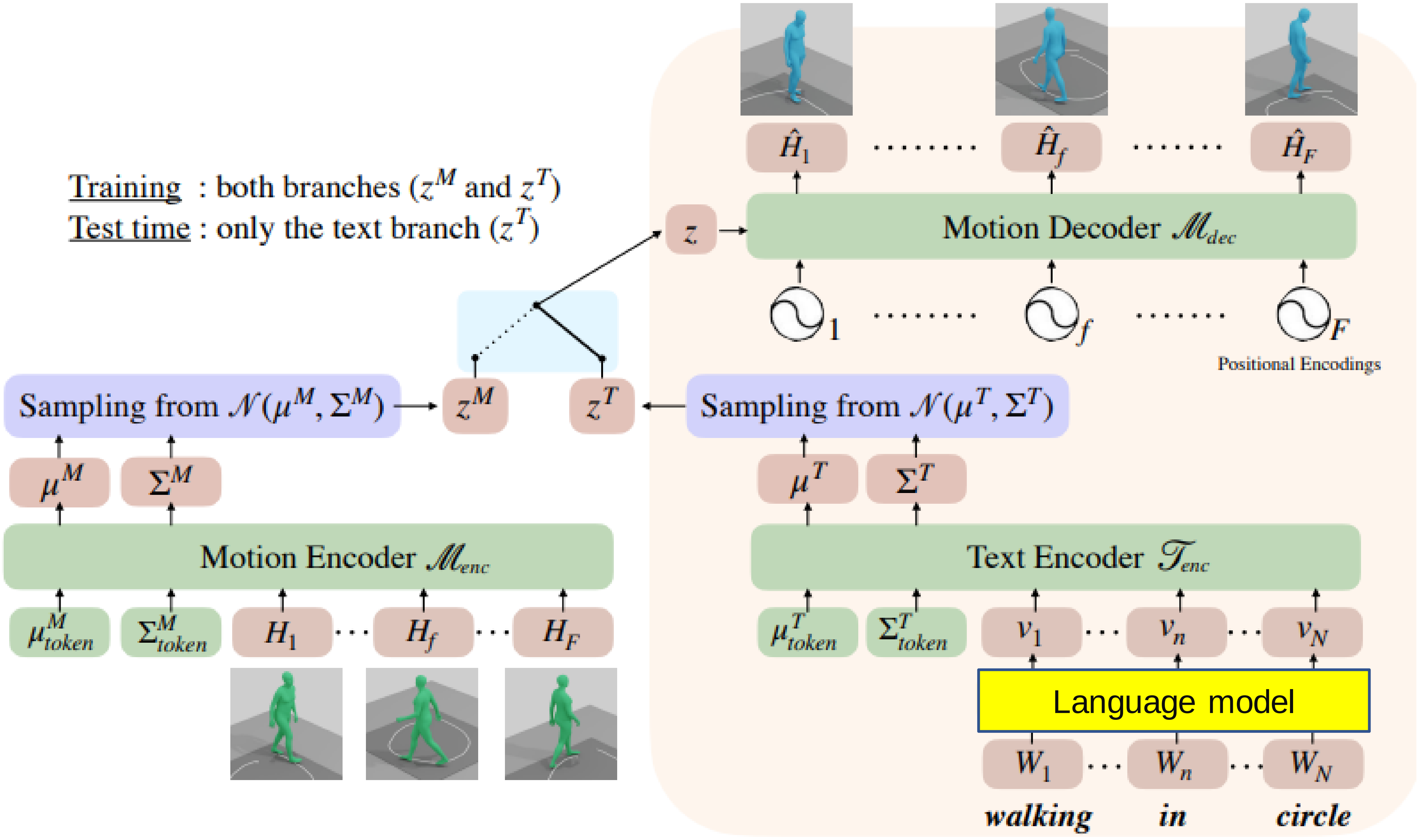
Training : both branches (z^M and z^T)
Test time : only the text branch (z^T)
$\hat{H}_1$
$\hat{H}_f$
$\hat{H}_F$
z
Motion Decoder $\mathcal{M}_{dec}$
1
f
F
Positional Encodings
Sampling from $\mathcal{N}(\mu^M, \Sigma^M)$
z^M
z^T
Sampling from $\mathcal{N}(\mu^T, \Sigma^T)$
μ^M
Σ^M
μ^T
Σ^T
Motion Encoder $\mathcal{M}_{enc}$
Text Encoder $\mathcal{T}_{enc}$
μ^M_{token}
Σ^M_{token}
H_1
H_f
H_F
μ^T_{token}
Σ^T_{token}
v_1
v_n
v_N
Language model
W_1
W_n
W_N
walking
in
circle

Fig. 1: Overview of the current model under analysis, where motion and text are encoded into the latent space at training time. At test time, only the language branch is used to sample from the latent space and generate the motion. Adapted from [4].

References

1) Petrovich, Mathis, Michael J. Black, and Gül Varol. "Action-conditioned 3d human motion synthesis with transformer vae." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021.
2) Radford, Alec, et al. "Learning transferable visual models from natural language supervision." International Conference on Machine Learning. PMLR, 2021.
3) Punnakkal, Abhinanda R., et al. "BABEL: Bodies, action and behavior with english labels." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021.
4) Petrovich, Mathis, Michael J. Black, and Gül Varol. "TEMOS: Generating diverse human motions from textual descriptions." arXiv preprint arXiv:2204.14109 (2022).

Acknowledgements

This work was partially funded by Fundação para a Ciência e a Tecnologia under the scope of the CMU Portugal Visiting Students Program, and the Orient Project ERC grant 693400.

Carnegie Mellon Portugal

fct Fundação para a Ciência e a Tecnologia

TÉCNICO LISBOA

Carnegie Mellon University