Purposer: Putting Human Motion Generation in Context

Nicolas Ugrinovic[1] Thomas Lucas[2] Fabien Baradel[2] Philippe Weinzaepfel[2]
Grégory Rogez[2] Francesc Moreno-Noguer[1]

[1]Institut de Robòtica i Informàtica Industrial, CSIC-UPC, Barcelona, Spain
[2]NAVER LABS Europe

In this supplementary material, we first discuss the attached video which shows samples of motion sequences generated from Purposer (Section 1). We then present additional details about Purposer (Section 2).

1. Supplementary Video

The supplementary video displays generated motions in scenes from HUMANISE [6], PROX [1], and Replica [5]. In most of the generated sequences and unless stated otherwise, our model is conditioned on: action labels, scene, future stream, path and last pose. Please note that we do not use any post-processing optimization step in the results presented in the video. However, if desired such an optimization step could be added to our model's outputs.

2. Model details

In this section we detail hyper-parameters used to train our models and mention the model architecture details. Note that code will be released, thus allowing to access all hyper-parameter details, trained model weights, and evaluation code as well as allowing reproducibility.

Implementation details. We implement our method in PyTorch and use the Adam optimizer for training. The auto-encoder (E and D) is taken from PoseGPT [3]: it was trained on the BABEL [4] dataset using 4 codebooks, of 128 centroids each, with centroids of dimension 256. Our auto-regressive GPT network (G) is trained on the HUMANISE [6] dataset for roughly 800K iterations with a learning rate of 1×10^{-4} and, in the corresponding cases, it is then fine-tuned on the PROX [1] dataset for 8K iterations with a cosine variable learning rate schedule starting from 1×10^{-5} and regularized with elastic decoupled weight decay [2] with weight of 0.01.

The encoder and the decoder are frozen during the training of the generator. We use different data augmentations applied to human motion samples at train time, such as randomly varying the framerate of the sequence by a factor ranging from 0.7 to 1.3, random rotations of the vertical axis of the human motion and the associated scene, and randomly sampling the starting time-steps of the sequence. Empirically, we found these augmentations useful to avoid over-fitting.

When generating long-range motions by concatenating short motion clips, in order to get smoother transitions, we interpolate the SMPL pose parameters on $SO(3)$ between the last pose of the first sequence and the first pose of the second sequence.

Technical details. The GPT is based on a transformer architecture with 8 blocks with 4 attention heads. To enable future conditioning we use an identical additional network architecture for the second branch with the sole difference in the masking of the attention, which in this case is non-causal and thus not designed to preserve causality.

References

[1] Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J. Black. Resolving 3D human pose ambiguities with 3D scene constraints. In *ICCV*, 2019.

[2] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[3] Thomas Lucas, Fabien Baradel, Philippe Weinzaepfel, and Grégory Rogez. PoseGPT: Quantization-based 3D Human Motion Generation and Forecasting. In *ECCV*, 2022.

[4] Abhinanda R. Punnakkal, Arjun Chandrasekaran, Nikos Athanasiou, Alejandra Quiros-Ramirez, and Michael J. Black. BABEL: Bodies, action and behavior with english labels. In *CVPR*, 2021.

[5] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, et al. The replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019.

[6] Zan Wang, Yixin Chen, Tengyu Liu, Yixin Zhu, Wei Liang, and Siyuan Huang. HUMANISE: Language-conditioned human motion generation in 3d scenes. In *NeurIPS*, 2022.