Learning Uncoupled-Modulation CVAE for 3D Action-Conditioned Human Motion Synthesis ——Supplementary Material

Chongyang Zhong[1,2], Lei Hu[1,2], Zihao Zhang[1,2], and Shihong Xia[1,2]

[1] Institute of Computing Technology, Chinese Academy of Sciences
[2] University of Chinese Academy of Sciences
{zhongchongyang, hulei19z, zhangzihao, xsh}@ict.ac.cn

1 Datasets details

HumanAct12 [1] is the in-house dataset adopted from an existing dataset PHSPD [8], consisting of 1,191 motion clips and 90,099 frames in total. All motions are organized into 12 action categories, including drink, throw, warm up, lift dumbbell. Following [1], we use the coarse-grained action annotations. HumanaAct12 has accurate 3D position annotations with 24 joints and stable pose sequences, and has more organized action annotation.

UESTC [3] consists of 25K sequences across 40 action categories and 118 persons collected using Microsoft Kinect v2 sensors. [5] use VIBE on it to obtain SMPL sequences that correspond best to the Kinect skeleton provided, which we use for training and testing. The processed dataset has 10650 sequences for training and 13350 sequences for testing. UESTC has 33 sequences per action on average (10K divided by 8 views, 40 actions).

BABEL [6] leverages the recently introduced AMASS dataset [4] for mocap sequences. BABEL contains action annotations for about 43.5 hours of mocap performed by over 346 subjects from AMASS represented by SMPL-H [7], with 15472 unique language labels. Via a semi-automatic manual categorization, these sequences are organized into 8 broad semantic categories such as throw, martial arts, etc, which are subdivided into 260 action categories such as greet, hop, scratch, dance, play instrument, etc. This dataset has more action types sequences, making it more challenging for action recognition and motion synthesis. We select a subset of BABEL containing 60 categories provided in [6] to train our model, including 19752 clips after our processing. Since BABEL does not provide labels for the testset, we use the extra data they provided as the testset.

2 Implementation details:

We train the model with an input sequence of 64 frames(both our model and other comparison methods). The action-agnostic encoder consists of 6 TCN layers and 1 instance normalization layer. The action-aware encoder consists of 4

GAGCN layers with 4 gating networks and 2 TCN layers. The decoder consists of 4 GRU layers. We use Nvidia 2080ti to train our network for 1500, 5000, and 2000 epochs on UESTC, HumanAct12, and BABEL. We use Adam Optimizer with an initial learning rate of 0.0001. The batch size is 20. The weight of loss terms λ_{pr}, λ_{vr} are 1,1. λ_{kr} is 1e-5 for HumanAct12 and 1e-6 for UESTC and BABEL.

3 Discussions about action-agnostic learning

We have two assumptions on action-conditioned motions: first, a sequence contains both action-aware and action-agnostic information; second, action-aware information is more representative in the spatial branch than the temporal one. Therefore, we adopt three operations in our action-agnostic encoder: 1. Using TCN only and intentionally ignoring spatial features (in contrast, spatial features are emphasized in action-aware encoders); 2. Using only motion sequences as input without action label; 3. According to [2], Instance Normalization(IN) is essentially a kind of style normalization, which can effectively eliminate style-related information in image generation. The label information in motion can also be considered as a kind of style, which is the reason why we add IN to the action-agnostic encoder. Since we input the modulated z' to the decoder during training, directly inputting z will get random label motion in the dataset or some unnatural results that do not match any labels.

4 Testing phase

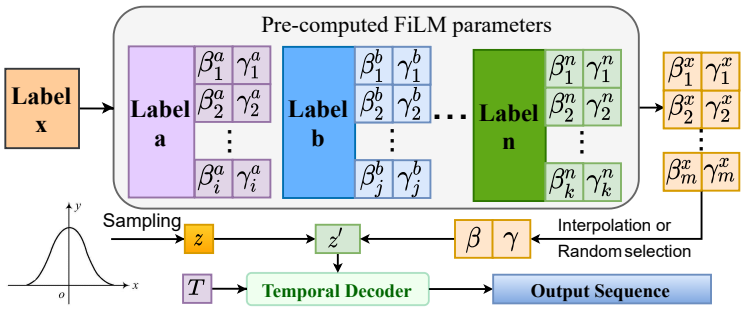

Fig. 1. Testing phase of our method. Please room in for details.

After model training, we compute a corresponding FiLM parameter for each sequence and save them as pre-computed parameters along with their corresponding labels, as in the top of Fig. 1. When testing, our model no longer requires encoders or any input sequences, but only an action label, desired length T, and a random variable z sampled from a standard Gaussian distribution. As shown in the bottom of Fig. 1, when inputting label x, we select all parameters

corresponding to label x from the pre-computed parameters, and then randomly select a set of parameter from them (or interpolate two sets of parameters) to modulate the sampled variable z to obtain z' as the input to the decoder together with T to generated motion sequences of label x.

5 Motion interpolation

We demonstrate the interpolation results of our method here. We pre-compute FiLM parameters for each action offline and save them with the corresponding labels. Given action label "Throw", we random sampling z in the action-agnostic latent space, then we select two set of FiLM parameters $\{\gamma_a, \beta_a\}$, $\{\gamma_b, \beta_b\}$. Then we compute interpolated parameters $\{\gamma_i, \beta_i\}$ as follows:

$$\gamma_i = \omega_a \gamma_a + \omega_b \gamma_b, \quad \beta_i = \omega_a \beta_a + \omega_b \beta_b \tag{1}$$

where ω_a and ω_b are the interpolation weights. Using the interpolated parameters $\{\gamma_i, \beta_i\}$ to modulate z, we can generate a new motion sequence. The results are shown in Fig. 2. The first row is a two-armed throw, the second row is a one-armed throw(right arm), and the third row is an interpolation of them. We can see that in the interpolated sequence, both hands have a throwing movement, and the right arm has a large amplitude of motion. It is obvious that the interpolation sequence combines the characteristics of the two motion sequences well, and the result is also very natural and smooth.

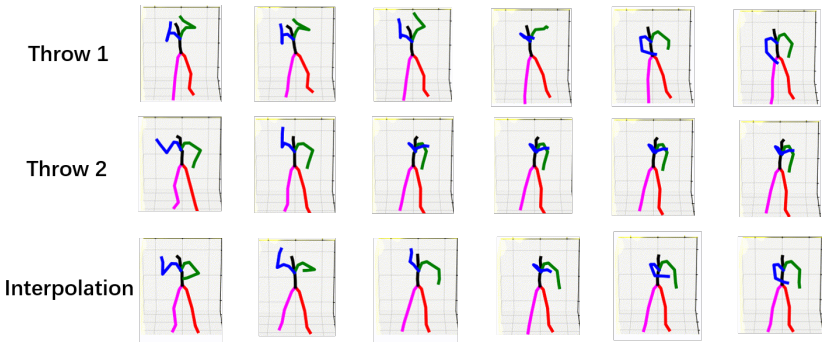

Fig. 2. Motion interpolation of our method. We illustrate our interpolation on "Throw" on HumanAct12. Three motion sequences are generated from the same random sampling latent representation z with different FiLM parameters. The FiLM parameters of the third row are the interpolation of the first and the second row's. In order to make the contrast effect more intuitive, we used the skeleton instead of the shape.

6 More qualitative results

We demonstrate more qualitative results here in Fig. 3, please refer to the supplementary video for more video results.

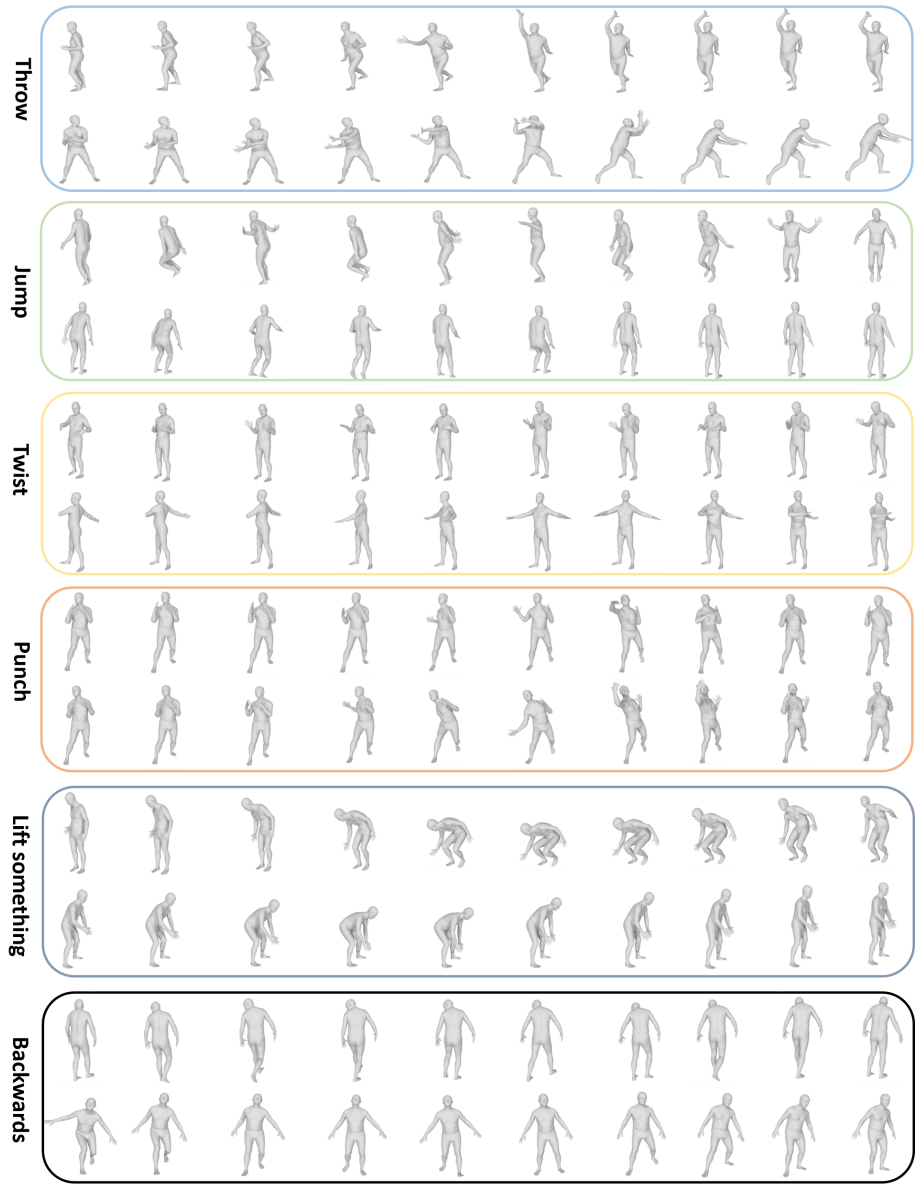

Fig. 3. More qualitative results of our method. We illustrate our generations of "Throw", "Jump", "Twist", "Punch", "Lift something", and "Backwards" actions from BABEL. Each action consists of 2 sequences, please zoom in for details. "Throw" movements include one-handed throwing and two-handed throwing. "Jump" includes jumping in place and spin jumping. And "Twist" includes twisting of the waist and twisting of the wrist. "Punch" includes right hook and uppercuts, and "Lift something" includes one-handed lifting and two-handed lifting. "Backwards" includes moving backwards with/without looking back. These results demonstrate that our method can generate complex, realistic, diverse, and label-compliant motions.

References

1. Guo, C., Zuo, X., Wang, S., Zou, S., Sun, Q., Deng, A., Gong, M., Cheng, L.: Action2motion: Conditioned generation of 3d human motions. In: Proceedings of the 28th ACM International Conference on Multimedia. pp. 2021–2029 (2020)
2. Huang, X., Belongie, S.: Arbitrary style transfer in real-time with adaptive instance normalization. In: Proceedings of the IEEE international conference on computer vision. pp. 1501–1510 (2017)
3. Ji, Y., Xu, F., Yang, Y., Shen, F., Shen, H.T., Zheng, W.S.: A large-scale rgb-d database for arbitrary-view human action recognition. In: Proceedings of the 26th ACM international Conference on Multimedia. pp. 1510–1518 (2018)
4. Mahmood, N., Ghorbani, N., Troje, N.F., Pons-Moll, G., Black, M.J.: Amass: Archive of motion capture as surface shapes. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 5442–5451 (2019)
5. Petrovich, M., Black, M.J., Varol, G.: Action-conditioned 3d human motion synthesis with transformer vae. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10985–10995 (2021)
6. Punnakkal, A.R., Chandrasekaran, A., Athanasiou, N., Quiros-Ramirez, A., Black, M.J.: Babel: Bodies, action and behavior with english labels. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 722–731 (2021)
7. Romero, J., Tzionas, D., Black, M.J.: Embodied hands: Modeling and capturing hands and bodies together. ACM Transactions on Graphics, (Proc. SIGGRAPH Asia) **36**(6) (Nov 2017)
8. Zou, S., Zuo, X., Qian, Y., Wang, S., Xu, C., Gong, M., Cheng, L.: 3d human shape reconstruction from a polarization image. In: European Conference on Computer Vision. pp. 351–368. Springer (2020)