AttT2M: Text-Driven Human Motion Generation with Multi-Perspective Attention Mechanism–Supplementary Material

Anonymous ICCV submission

Paper ID 6688

1. Details about Body-Part Attention

Firstly, we divide the human body with n joints into five body parts: {Torso, Left Arm, Right Arm, Left Leg, Right Leg}, each containing its own set of joints(seeing Figure. 1). In order to apply the transformer in the spatial dimension, we rearrange the motion representation of T2M [2]. The original representation of each frame is $x_i = \{\dot{r}^a, \dot{r}^x, \dot{r}^z, r^y, j^p, j^v, j^r, c^f\}$, representing root angular velocity along Y-axis, root linear velocities on XZ-plane, root height, local joints positions, velocities, 6D rotations [5], and foot ground contacts, respectively.

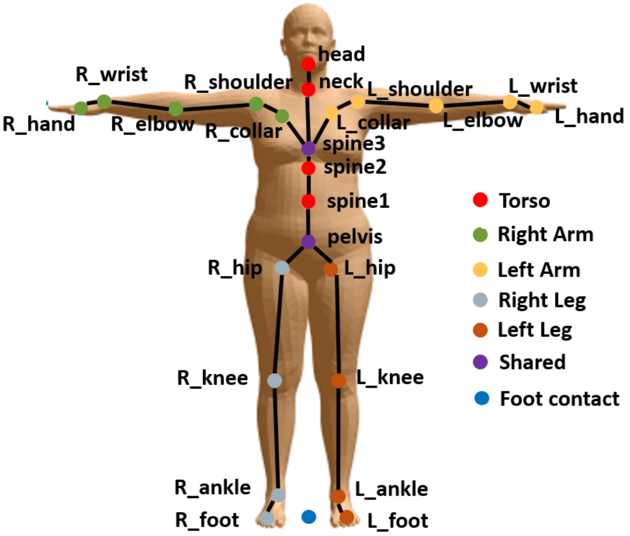

Figure 1. Body-part segmentation on HumanML3D

To gather information from each joint, the new motion representation of the root joint is $j_{root} = \{\dot{r}^a, \dot{r}^x, \dot{r}^z, r^y, j^v_{root}\}$, and the motion representation of other joints is $j_i = \{j_i^p, j_i^v, j_i^r\}$. Specifically, to handle the foot sliding problem, we also treat the foot contact feature as an additional joint of the human body and incorporate it into the attention computation. Therefore, the new motion representation of each frame is $\bar{x}_i = \{j_{root}, j_1, j_2, ..., j_{n-1}, c^f\}$.

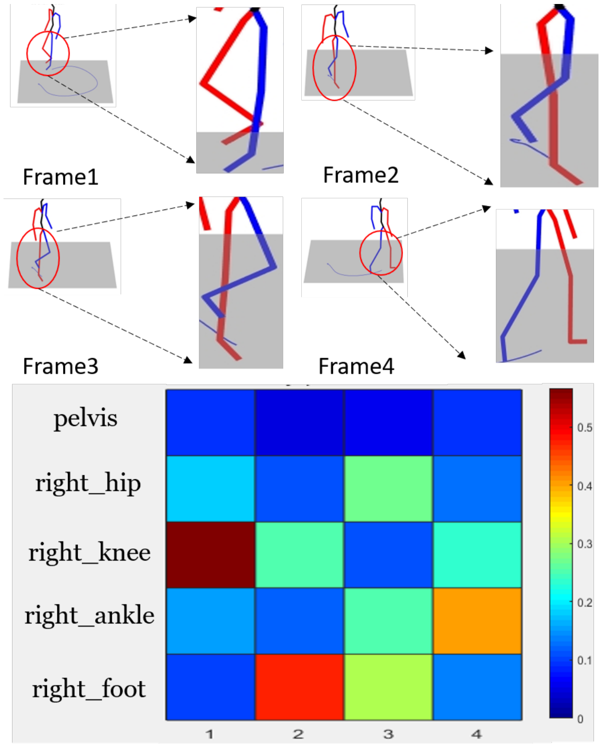

Figure 2. Body-part attention

According to this division, we define a new human adjacency mask $\mathcal{M} = \{m_{i,j}\} \in \mathbb{R}^{(n+1)(n+1)}$. If joint i and j are in the body part, $m_{i,j} = 0$, otherwise $-\infty$. The body-part attention is calculated as described in our paper.

To more intuitively show the effectiveness of this part, we selected four frames from a generated motion generated by the text ”a person walking in a circle counterclockwise”. We visualized the attention of the inner part of the right leg, indicated by the red leg in the skeleton in Figure 2. In the first frame, the person's right leg is bent, preparing to extend forward, and the knee plays an important role, so the weight of the knee is the highest. In the second frame, the right foot is just about to leave the ground, and the weight is highest at the foot position. In the third frame, the entire right leg is

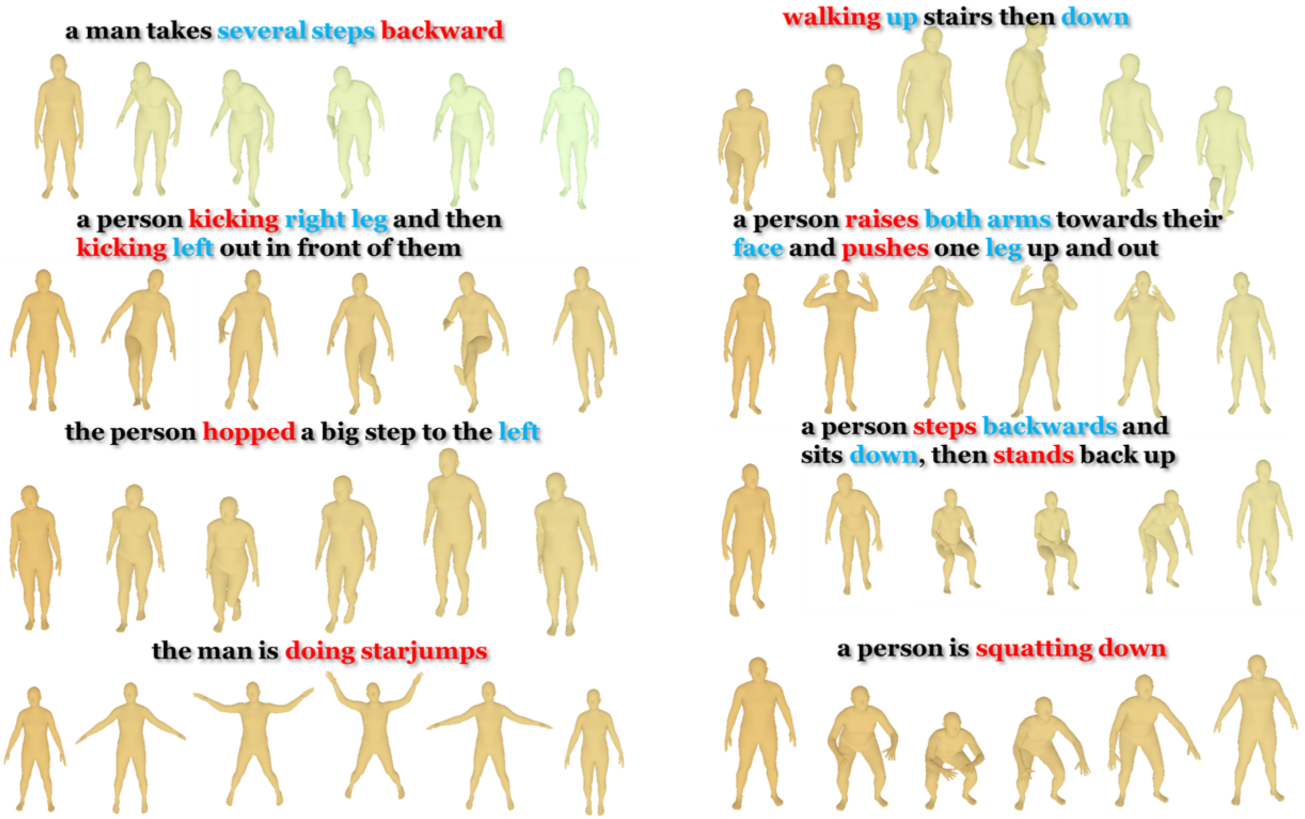
a man takes several steps backward
walking up stairs then down
a person kicking right leg and then kicking left out in front of them
a person raises both arms towards their face and pushes one leg up and out
the person hopped a big step to the left
a person steps backwards and sits down, then stands back up
the man is doing starjumps
a person is squatting down

Figure 3. More visualization results.

in an upright state, and the weight distribution is relatively even. In the fourth frame, the right leg just landed, and the weight is relatively high on both the ankle and the slightly bent knee. From the visualization of the attention mechanism, we can see that our body-part attention well considers the mutual influence of different joints within the body part when extracting spatial features, which can fully adapt to the spatial structure of the human body. This makes our extracted spatial feature expression ability strong and helps us learn a better discrete motion latent space.

2. Details about Implementation

Our code will be released to the public if our submission is accepted.

The size of the code book used in VQVAE is 512 × 512, with 2 layers of $Trans_{enc}$, a feature dimension of 128, and 4 heads of multi-head attention. The feature motion compression rate l_w is 4, and the parameters α and β in the loss function are set to 0.5 and 1, respectively. During the training of VQVAE, we fix the input length at 64. We use AdamW as the optimizer, with the $[\beta_1, \beta_2]$ settings at $[0.9, 0.99]$ and a batch size of 256. We iterate 300K times on a single GTX 3090Ti graphics card, with a learning rate of 2e-4 for the first 200K iterations, and a learning rate of 1e-5 for the remaining 100K iterations.

During the training of GLAGT, the $Trans_{local}$ has 2 layers, a feature dimension of 1024, and 8 heads of multi-head attention. The $Trans_{global}$ has 9 layers, a feature dimension of 1024, and 16 heads of multi-head attention. The $Trans_{gen}$ has 9 layers, a feature dimension of 1024, and 16 heads of multi-head attention. We fix the input length at 50. We use AdamW as the optimizer, with the $[\beta_1, \beta_2]$ settings at $[0.5, 0.99]$ and a batch size of 128. We iterate HumanML3D 270K times on a single GTX 3090Ti and 290K for KIT-ML, with a learning rate of 1e-4 for the first 150K iterations, and a learning rate of 5e-6 for the rest.

3. More Visualization Results

Here we give more visualization results in Figure. 3 to show the quality of the motion generated by our method. More Video results are in supplemental video.

4. Details about Datasets and Metrics

Here we demonstrate more details about datasets and metrics. Most of these introductions come from T2M [2].

4.1. Datasets

KIT Motion-Language [4] is the first 3D human motion dataset with text labels, consisting of a subset of KIT [3] and CMU [1] datasets. It contains 3911 motion sequences and 6353 sequence-level text annotations, with an average of 9.5 words per annotation. The motions are scaled to 12.5 FPS, and each motion sequence has 1 to 4 text descriptions, with an average length of 8 words. We adopt the same data split as T2M, where 80% is used for training, 5% for validation, and 15% for testing.

HumanML3D [2]: HumanML3D added detailed text labels to the motion in AMASS dataset, creating a motion-language dataset. It contains 14616 motion segments with a total duration of 28.59 hours. The average length of per segment is 7.1 seconds, and the longest and shortest motions are 10 and 2 seconds, respectively. The dataset contains 44970 text descriptions with an average length of 12 words, covering 5317 unique words. The motions are scaled to 20 FPS, and motions longer than 10 seconds are randomly split into 10-second segments. All motion data is aligned to a common default skeleton and the initial orientation is rotated to face the positive Z-axis. The dataset is split into training, validation, and test sets, We follow the split of T2M [2] for our training, validation, and testing on both KIT-ML and HumanML3D.

4.2. Metrics

Frechet Inception Distance (FID): We use the feature extraction module provided by the T2M authors to extract features from the real data in the test set, as well as from the motion generated by our method. We then calculate the FID between these two feature distributions. A smaller FID indicates that the generated motion is closer to real data.

R-precision: For each generated motion, we assign one real text description and 31 randomly selected text descriptions that do not match it. We then calculate the distance between the motion feature and the 32 text features, sort them in ascending order, repeat this process 32 times, and calculate the probabilities of the real text description ranking at the top 1, 2, and 3. A higher probability indicates better retrieval results.

MultiModal distance(MM-D): For each generated motion, we extract the motion feature and the corresponding text feature, and then calculate the Euclidean distance between these two features. A smaller distance indicates a better match between motion and text.

Diversity(Div): We randomly select two subsets of 300 motions from the generated motions. We then extract the motion features for each subset and calculate the mean Euclidean distance between the corresponding motions in the two subsets. This metric measures the diversity of all generated motions, with a larger mean indicating greater differences between different subsets and better diversity.

MultiModality(MM): Given C text descriptions, we generate 20 motion data for each text description and randomly divide them into two subsets of 10. We then calculate the difference in distance between the two subsets in the same way as Diversity and take the average of the results for C text descriptions. This metric measures the diversity of the motions generated for the same text description, with a larger distance indicating better diversity.

References

[1] Carnegie mellon university: Cmu mocap dataset, http://mocap.cs.cmu.edu. 3

[2] 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. 1, 2, 3

[3] Christian Mandery, Ömer Terlemez, Martin Do, Nikolaus Vahrenkamp, and Tamim Asfour. The kit whole-body human motion database. In *International Conference on Advanced Robotics (ICAR)*, pages 329–336, 2015. 3

[4] Matthias Plappert, Christian Mandery, and Tamim Asfour. The kit motion-language dataset. *Big data*, 4(4):236–252, 2016. 3

[5] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5745–5753, 2019. 1