


Supplementary Material

ParCo: Part-Coordinating Text-to-Motion Synthesis

Qiran Zou^{*1}, Shangyuan Yuan^{*1}, Shian Du¹, Yu Wang², Chang Liu¹, Yi Xu³, Jie Chen², and Xiangyang Ji¹

¹ Tsinghua University, China

² Peking University Shenzhen Graduate School, China

³ Dalian University of Technology, China

qiranzou@gmail.com, {yuansy21, dsa23}@mails.tsinghua.edu.cn,
2201212856@stu.pku.edu.cn, liuchang2022@tsinghua.edu.cn, yxu@dlut.edu.cn,
chenj@pcl.ac.cn, xyji@tsinghua.edu.cn

This supplementary material provides:

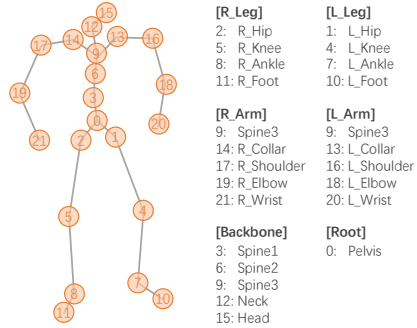
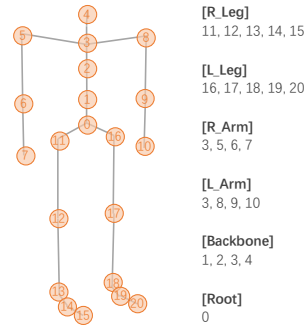
- Sec. A: details of whole-body to part motions discretization.
- Sec. B: details of text-length-based splits for the HumanML3D test set.
- Sec. C: the reconstruction performance of our multiple VQ-VAEs dedicated to different parts.
- Sec. D: additional training details.
- Sec. E: additional qualitative results.

A Whole-body to Part Motions Discretization

The HumanML3D [2] and KIT-ML [5] datasets utilize the SMPL [4] and MMM [6] Human Models, respectively. These datasets include joints related to whole-body motion, excluding hand joints, as depicted in Fig. 1 and Fig. 2. In addition, HumanML3D utilizes 22 joints from the SMPL human model, while the widely-used preprocessed KIT-ML benchmark, provided by [2], comprises 21 joints.

ParCo’s 6-Part Division Our ParCo divides the whole body into six parts: R.Leg, L.Leg, R.Arm, L.Arm, Backbone, and Root. Specific partitioning details for HumanML3D and KIT-ML are illustrated in Fig. 1 and Fig. 2. Both R.Arm and L.Arm include the 9-th joint for HumanML3D (3-th joint for KIT-ML). The inclusion of the joint in both arms is due to its role as a key point connecting the arms and the backbone, providing positional information for the arms relative to this connection point. When reconstructing the whole-body motion from part motions, we obtain three predictions of this joint from R.Arm, L.Arm, and Backbone. We use the average of these three values as the final prediction.

* Equal contribution.

**Fig. 1:** ParCo’s 6-Part Division for SMPL Human Model.**Fig. 2:** ParCo’s 6-Part Division for MMM Human Model.

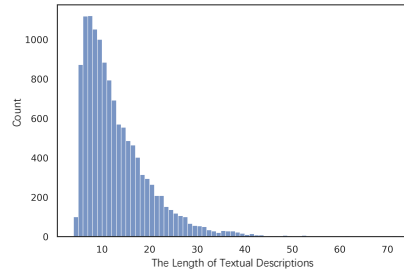
Upper and Lower Body Division The upper-and-lower-body division is proposed by SCA [1], which divides the human body into upper and lower halves, both containing the backbone joints. In our ablation experiments, we perform the upper-and-lower-body division on the HumanML3D as,

- Upper: 9, 14, 17, 19, 21, 13, 16, 18, 20, 0, 3, 6, 12, 15
- Lower: 0, 2, 5, 8, 11, 1, 4, 7, 10, 3, 6, 9, 12, 15

where the numbers denote the joint number. It is noteworthy that, despite SCA dividing whole-body into upper and lower body for motion generation, its generations of upper body motion and lower body motion are entirely independent and lack coordination compared to our method.

Table 1: Statistics of Text-Length-Based Splits.

Statistics	0 – 25%	25 – 50%	50 – 75%	75 – 100%
Min length	4	8	11	16
Max length	7	10	15	72
Avg length	6.0	8.9	12.8	22.3
Total Count	3210	2936	3096	3294
Percentage(%)	25.6	23.4	24.7	26.3

**Fig. 3:** Distribution of Counts for Text Length.

B Details of Text-Length-Based Splits

In order to investigate the synthetic performance given textual descriptions of different lengths, we divide the HumanML3D test set into four splits based on the length of textual descriptions. The test set contains a total of 4,384 motions, each motion is described by multiple textual descriptions. Following [8] and

Table 2: VQ-VAE Reconstruction Performance on HumanML3D and KIT-ML test sets.

Datasets	Methods	R-Precision \uparrow			FID \downarrow	MM-Dist \downarrow	Diversity \rightarrow
		Top-1	Top-2	Top-3			
HumanML3D	Real Motion	0.511 \pm .003	0.703 \pm .003	0.797 \pm .002	0.002 \pm .000	2.974 \pm .008	9.503 \pm .065
	T2M-GPT	0.501 \pm .002	0.692 \pm .002	0.785 \pm .002	0.070 \pm .001	3.072 \pm .009	9.593 \pm .079
	Up&Low	0.488 \pm .002	0.683 \pm .002	0.780 \pm .002	0.066 \pm .001	3.100 \pm .007	9.581\pm.062
	ParCo (Ours)	0.503\pm.003	0.693\pm.003	0.790\pm.002	0.021\pm.000	3.019\pm.007	9.411 \pm .086
KIT-ML	Real Motion	0.424 \pm .005	0.649 \pm .006	0.779 \pm .006	0.031 \pm .004	2.788 \pm .012	11.08 \pm .097
	T2M-GPT	0.399 \pm .005	0.614 \pm .005	0.740 \pm .006	0.472 \pm .011	2.986 \pm .027	10.994\pm.120
	ParCo (Ours)	0.407\pm.007	0.629\pm.005	0.760\pm.004	0.311\pm.006	2.892\pm.016	10.987 \pm .081

[2], we set the maximum motion length to 196 and the minimum length to 40, resulting in a total of 12,635 motion-text pairs. The distribution of these pairs, sorted by text length, is shown in Fig. 3. We further divide these pairs into four subsets (0-25%, 25-50%, 50-75%, 75-100%) from short to long. The details are shown in Table. 1, including the shortest/longest text lengths, the average length, the number of pairs, and the percentage.

C VQ-VAE Reconstruction Performance

The reconstruction performance of VQ-VAEs is presented in Table. 2. Specifically, we integrate the reconstructions of part motions into the whole-body motion for evaluation. We conduct the ablation study of reconstruction performance with different partitioning methods on HumanML3D. The results indicate that the performance of our ParCo’s 6 small VQ-VAE for part motion reconstruction surpasses the upper-and-lower-body division, and outperforms the baseline [8] which employs a large-parameter VQ-VAE for whole-body motion.

D Additional Training Details

During training VQ-VAE, we employ the velocity reconstruction auxiliary loss to assist training, following T2M-GPT [8] and SCA [1]. We use the last training checkpoint of VQ-VAE for the subsequent transformer’s training. For the transformer, we select the checkpoint with the lowest FID during training for text-to-motion evaluation. Additionally, we use the decoder of transformer [7] as our text-to-motion generator. The decoder of transformer achieves autoregressive prediction by masking the upper triangle of the self-attention map. To enhance the robustness of synthesis, we utilize the Corrupted Sequence [8] strategy to augment motion sequences. Inspired by MAE [3], we also introduce masked part modeling, a conceptually simple yet effective approach, to enhance part relation learning for coordinated motion generation. Specifically, we randomly replace a portion of body parts at each moment with mask tokens and force the remaining

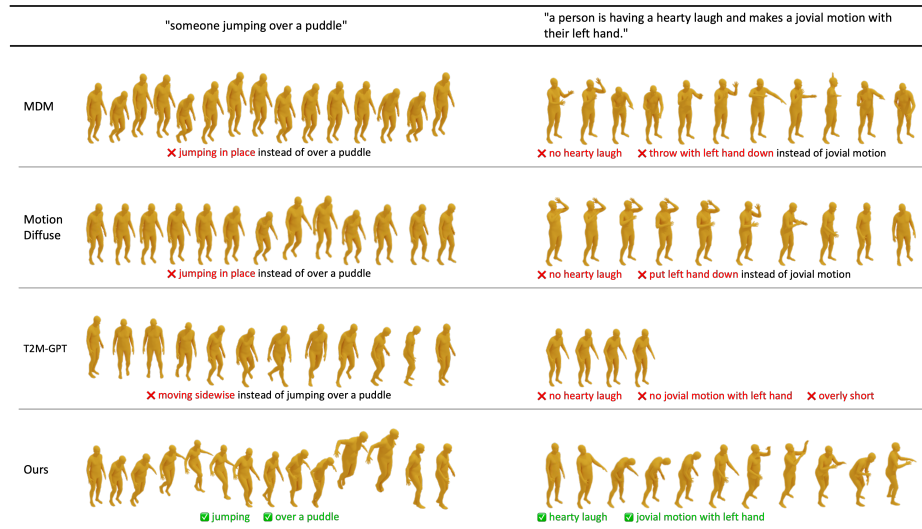


Fig. 4: Additional qualitative comparison with existing methods. Green indicates the motion is consistent with the text description. Red indicates the text description lacks the corresponding motion or got the wrong motion.

parts to predict them. Our ParCo is trained on a single A100 GPU for a total duration of 72.8 hours (20.5 hours for stage 1 and 52.3 hours for stage 2).

E Additional Qualitative Results

Additional qualitative results are presented in Fig. 4. The motions are generated according to text prompts from HumanML3D test set. These results demonstrate that our method can generate realistic and coordinated motions aligned with the text.

References

1. Ghosh, A., Cheema, N., Oguz, C., Theobalt, C., Slusallek, P.: Synthesis of compositional animations from textual descriptions. In: ICCV (2021)
2. Guo, C., Zou, S., Zuo, X., Wang, S., Ji, W., Li, X., Cheng, L.: Generating diverse and natural 3d human motions from text. In: CVPR (2022)
3. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: CVPR (2022)
4. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: Smpl: A skinned multi-person linear model. In: Seminal Graphics Papers: Pushing the Boundaries, Volume 2 (2023)
5. Plappert, M., Mandery, C., Asfour, T.: The kit motion-language dataset. Big data (2016)

6. Terlemez, Ö., Ulbrich, S., Mandery, C., Do, M., Vahrenkamp, N., Asfour, T.: Master motor map (mmm)—framework and toolkit for capturing, representing, and reproducing human motion on humanoid robots. In: 2014 IEEE-RAS International Conference on Humanoid Robots (2014)
7. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. *NeurIPS* (2017)
8. Zhang, J., Zhang, Y., Cun, X., Huang, S., Zhang, Y., Zhao, H., Lu, H., Shen, X.: T2m-gpt: Generating human motion from textual descriptions with discrete representations. *arXiv preprint arXiv:2301.06052* (2023)