

Dual Reciprocal Learning of Language-based Human Motion Understanding and Generation

Supplementary Material

In this document, we provide more details of our method and additional experimental results:

- §S1 More Analysis on Evaluation Protocols.
- §S2 More Qualitative Results.
- §S3 Discussion of Linguistic Metric Implementation and Legal/Ethical Considerations.

Dataset	R Precision \uparrow		
	Top 1	Top2	Top3
HumanML3D [1]- <i>Ori.</i> [CVPR22]	0.511	0.703	0.797
Motion-X [2]- <i>Ori.</i> [NeurIPS23]	0.308	0.515	0.629
HumanML3D [1]- <i>Upd.</i> [CVPR22]	0.386	0.568	0.687
Motion-X [2]- <i>Upd.</i> [NeurIPS23]	0.373	0.541	0.676

Table S1. Quantitative analysis of evaluation protocols. *Ori.* represents the original protocol, which is tightly coupled with HumanML3D. *Upd.* represents the updated protocol with improved generalization

S1. Evaluation Protocols

S1.1. Quantitative Analysis on R-Precision

To further illustrate the bias in the *Ori.* evaluation protocol, we use different text/motion encoders to compute R-Precision of text-motion pairs in HumanML3D [1] and Motion-X [2]. Results are shown in Table S1. For pre-provided encoders [1], we obtained 0.629 of RP-TOP3 on Motion-X as results, which are inferior to 0.797 on the HumanML3D test set. This indicates that the text-motion-aligned embedding space learned solely on HumanML3D lacks generalization, making it difficult to accurately assess the text-motion matching in Motion-X. Based on this observation, we updated the evaluation protocol. Specifically, we combined HumanML3D and Motion-X as the training dataset for the encoder and utilized a frozen distilbert-base-uncased model to extract contextual embeddings from text descriptions. We then trained the encoders to align these contextual embeddings with their corresponding motion samples. Subsequently, we recalculated the RP-TOP3 for HumanML3D and Motion-X, obtaining results of 0.687 for HumanML3D and 0.676 for Motion-X, respectively. The R-Precision improvement on Motion-X, along with the closer results between Motion-X and HumanML3D, demonstrates that the updated encoders have better generalization capabilities.

S1.2. Analysis on Scale/Comparability of Metrics

Under the *Upd.* evaluation protocol, we replace the text encoder with a DistilBERT-based network and adjust the batch size and learning rate during contrastive learning. These changes alter the absolute magnitude of the feature vectors, thereby shifting the scale of the learned embedding space. The FID, MMDist, Diversity, and MModality metrics we use are all affected by changes in scale. Accordingly, these metrics are not directly comparable across different evaluation protocols. In contrast, R-Precision measures the relative ranking based on a candidate set, and this internal ranking within the same embedding space is comparable. As seen in Table S1, R-Precision on HumanML3D consistently drops under *Upd.* protocol. This indicates that the *Upd.* protocol challenges text-motion retrieval to generalize beyond the specific domain.

S2. More Qualitative Results

We provide more qualitative comparisons between the baseline and the DRL-bootstrapped model on both motion generation and understanding in Fig. S1 and S2. From visualizations, we infer that the introduction of reciprocity between two tasks and unpaired text and motion data surely enhances model performance. More animation of qualitative results are provided in our project page¹.

S3. Discussion

S3.1. Motion-to-Text Metric Implementation

For Motion-to-Text evaluation, BLEU@1, BLEU@4, ROUGE, and CIDEr metrics can be computed using either the NLG-Eval [3] or NLG-Metricverse [4] toolkit. The results of TM2T and MotionGPT reported in the main paper are calculated using NLG-Eval, consistent with the conference version of the MotionGPT paper. We evaluate our baseline and DRL-enhanced models using NLG-Metricverse, aligning with MotionGPT’s official implementation² as the authors recommended. Other metrics, including R-Precision, MMDist, and BertScore, are implemented in the same way as in TM2T and MotionGPT. We encourage later works to check our codebase and adopt the same settings for fair comparison.

¹<https://github.com/leonnnop/DRL>

²<https://github.com/OpenMotionLab/MotionGPT>

S3.2. Asset License and Consent

In this work, we study human motion understanding and generation with two famous human motion datasets, *i.e.*, HumanML3D [1] and Motion-X [2] that are all publicly and freely available for academic purposes. HumanML3D (<https://github.com/EricGuo5513/HumanML3D>) is released under the [MIT License](#); Motion-X is released at <https://github.com/IDEA-Research/Motion-X>, copyright © 2023 International Digital Economy Academy; NLG-Metricverse (<https://github.com/disi-unibo-nlp/nlg-metricverse>) is released under the [MIT License](#).

S3.3. Broader Impact

This work introduces the Dual Reciprocal Learning framework, utilizing large numbers of unpaired data to optimize both text-to-motion and motion-to-text models. On the positive side, the framework enhances data utilization efficiency in human motion research, improving both the accuracy and diversity of motion generation and understanding. It certainly has broad real-world applications, such as robotics, virtual reality, game development, *etc.* On the negative side, this technique poses a risk of misuse for generating fake videos. For example, a fabricated description could be used to create 3D human motion that never occurred. Though beyond the scope of this paper, we will organize a gated release of our models to make sure that they are not being used beyond academic research purposes.

References

- [1] 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 *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 5152–5161, 2022. [S1](#), [S2](#)
- [2] Jing Lin, Ailing Zeng, Shunlin Lu, Yuanhao Cai, Ruimao Zhang, Haoqian Wang, and Lei Zhang. Motion-x: A large-scale 3d expressive whole-body human motion dataset. *Proc. Adv. Neural Inf. Process. Syst.*, 36:25268–25280, 2023. [S1](#), [S2](#)
- [3] Shikhar Sharma, Layla El Asri, Hannes Schulz, and Jeremie Zumer. Relevance of unsupervised metrics in task-oriented dialogue for evaluating natural language generation. *arXiv preprint arXiv:1706.09799*, 2017. [S1](#)
- [4] nlg-metricverse Contributors. nlg-metricverse. <https://github.com/disi-unibo-nlp/nlg-metricverse>, 2023. [S1](#)

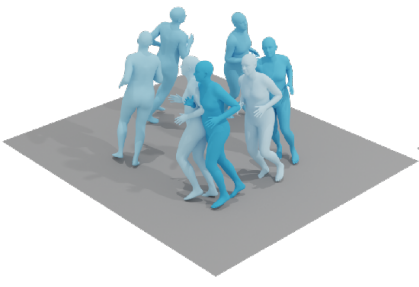

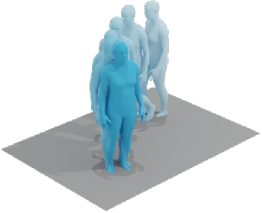
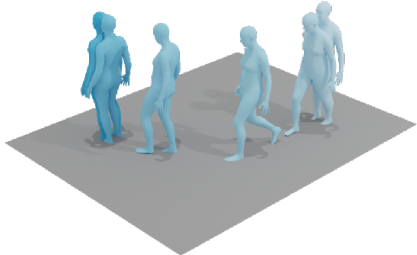
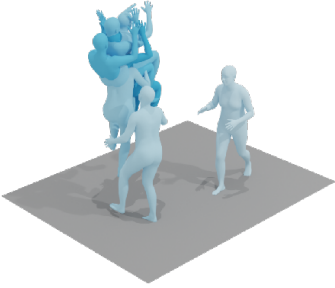
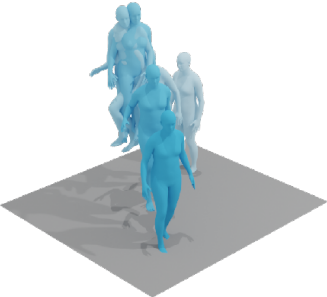
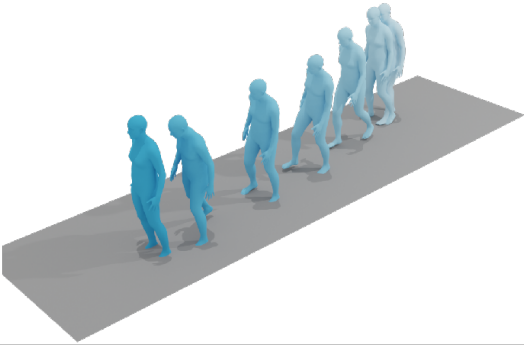

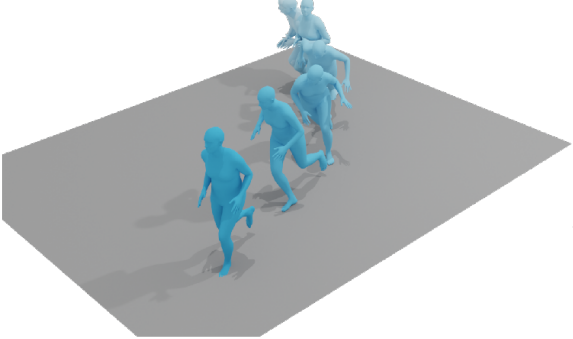

Desc.	TM2T*-T2M	TM2T*-T2M+DRL
A person continuously jogs counter clockwise.		
A person walks in a curve to the right.		
A person walks up stairs, turns left and walks down stairs.		
A person is sneaking around		
A person shoots the ball to the basket.		

Figure S1. More qualitative comparisons on motion generation between TM2T*-T2M and TM2T*-T2M + DRL.

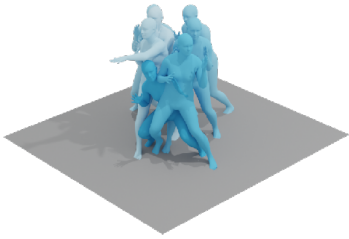



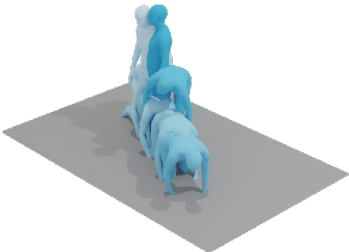
Motion	TM2T*-T2M	TM2T*-T2M+DRL
	A person dances from side to side, swinging arms and crossing feet.	A boxer is dodging his opponent's attacks and launching an attack back.
	A person slowly walked forward and return.	Someone hunches and stalks forward, turns to stalk back, then turns around again .
	A person is raising his right arm.	A sitting person raises his right hand to scratch his head .
	A person ducks from flying objects.	A person squats and ducks to avoid being hit by something with both hands on head .
	A man drops down to the floor and crawls forwards on his hands and knees	A person gets on their hands and knees and crawls forwards, and stands back up .

Figure S2. More qualitative comparisons on motion understanding between TM2T*-M2T and TM2T*-M2T + DRL.