Generating Diverse and Natural 3D Human Motions from Text

Chuan Guo Shihao Zou Xinxin Zuo Sen Wang Wei Ji Xingyu Li Li Cheng
University of Alberta

CVPR JUNE 19-24 2022 NEW ORLEANS LOUISIANA

Code & Model
https://ericguo5513.github.io/text-to-motion/

Problem Definition and Challenges

Goal: Stochastically generating natural 3D human motions from a given text description.

Motivations & Challenges:

- Previous works typically model this task as a deterministic one-to-one mapping problem.
- The lengths of motion for different or even the same text description may vary as well.
- Representing motions in form of individual poses can be redundant, which also adds on the burdens for generating long sequences.
- Existing human motion-language dataset[1] is limited in both quantity and diversity.

Overview

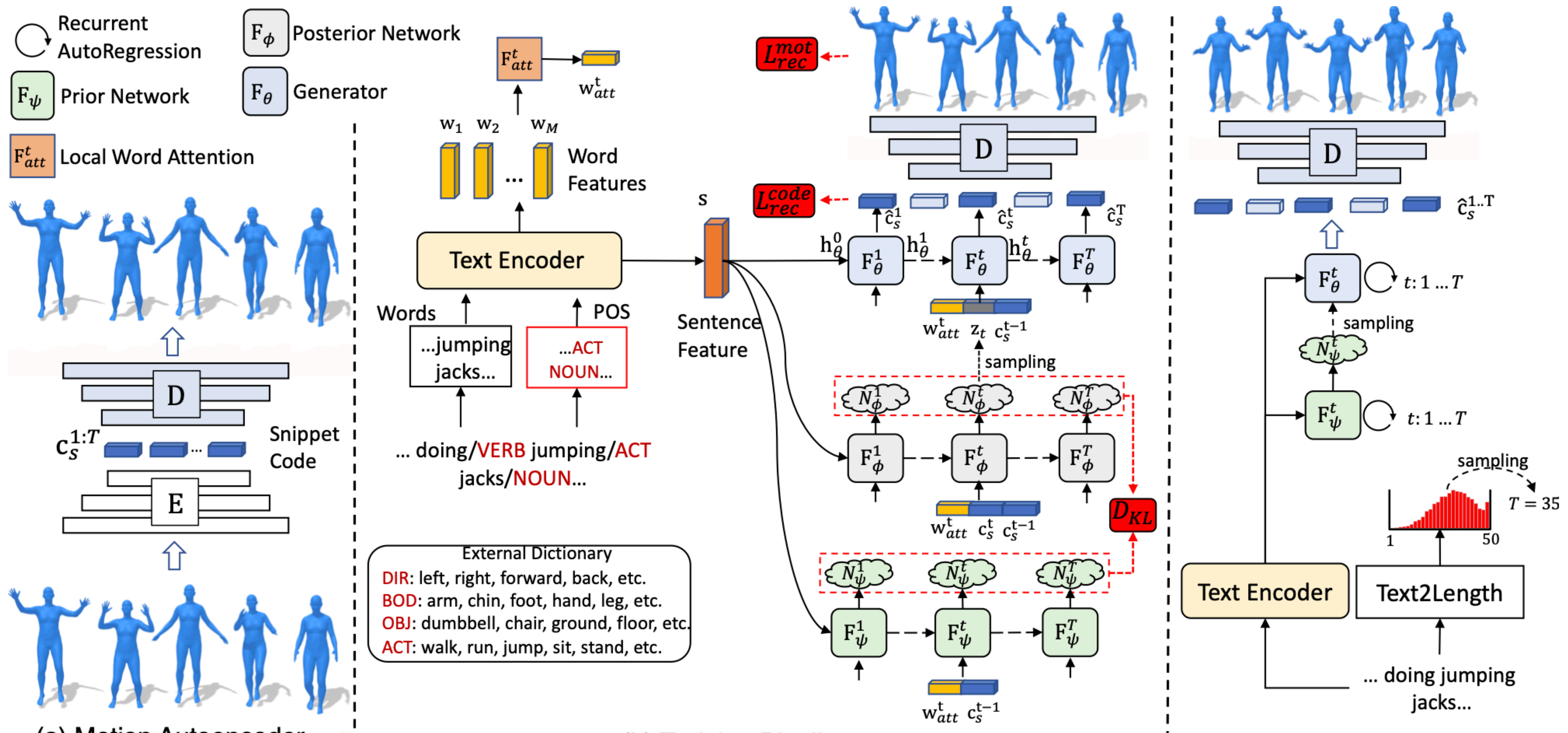
Motion Snippet Code is firstly obtained as the latent sequence from a pre-trained 1-D convolutional motion autoencoder. This would shorten the sequence length by 4 times and leads to a more compact and context-enriched motion representation.

Text2Length Sampling approximates the probability distribution of discrete motion snippet length condition on text. This is learned via cross entropy classification loss. During inference, we would sample a value from the estimated multinomial distribution.

Text2Motion Generation aims to generate 3D human motions from the given text and sampled motion length using temporal VAE as well as dedicated design of local word attention and time-to-arrival signal.

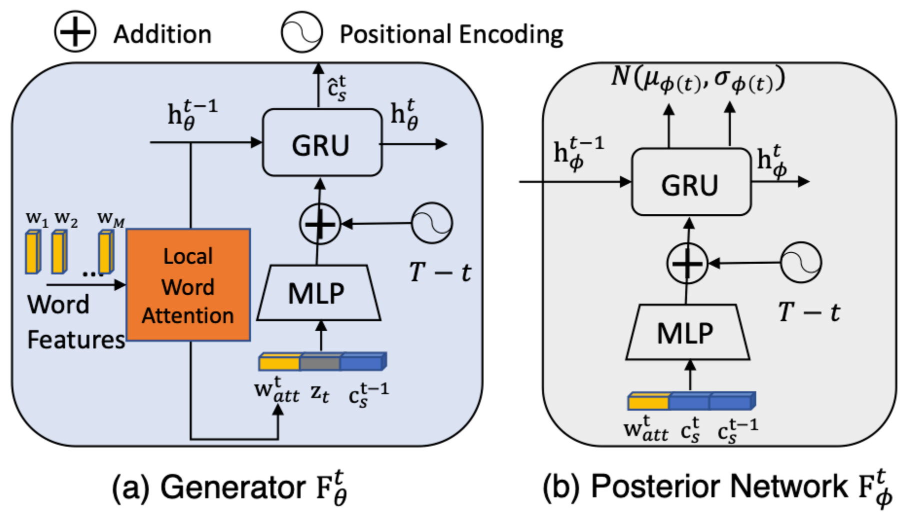
Method

Network Architecture:



(a) Motion Autoencoder (b) Training Pipeline (c) Inference Pipeline

Detailed structure of VAE networks:



(a) Generator F^t_θ (b) Posterior Network F^t_ϕ

Loss Function for Training Pipeline:

$$\mathcal{L} = \mathcal{L}^{code}_{rec} + \lambda_{mot}\mathcal{L}^{mot}_{rec} + \lambda_{KL}\mathcal{L}_{KL}, \quad \text{with}$$

$$\mathcal{L}^{code}_{rec} = \sum_t \|\hat{\mathbf{c}}^t_s - \mathbf{c}^t_s\|_1,$$

$$\mathcal{L}^{mot}_{rec} = \sum_{t'} \|\hat{\mathbf{p}}_{t'} - \mathbf{p}_{t'}\|_1, \qquad (1)$$

$$\mathcal{L}_{KL} = \sum_t \mathrm{KL}(\mathcal{N}(\mu_\phi(t),\sigma_\phi(t))\|\mathcal{N}(\mu_\psi(t),\sigma_\psi(t))).$$

HumanML3D Dataset

A novel large-scale and diverse 3D human motion language dataset:

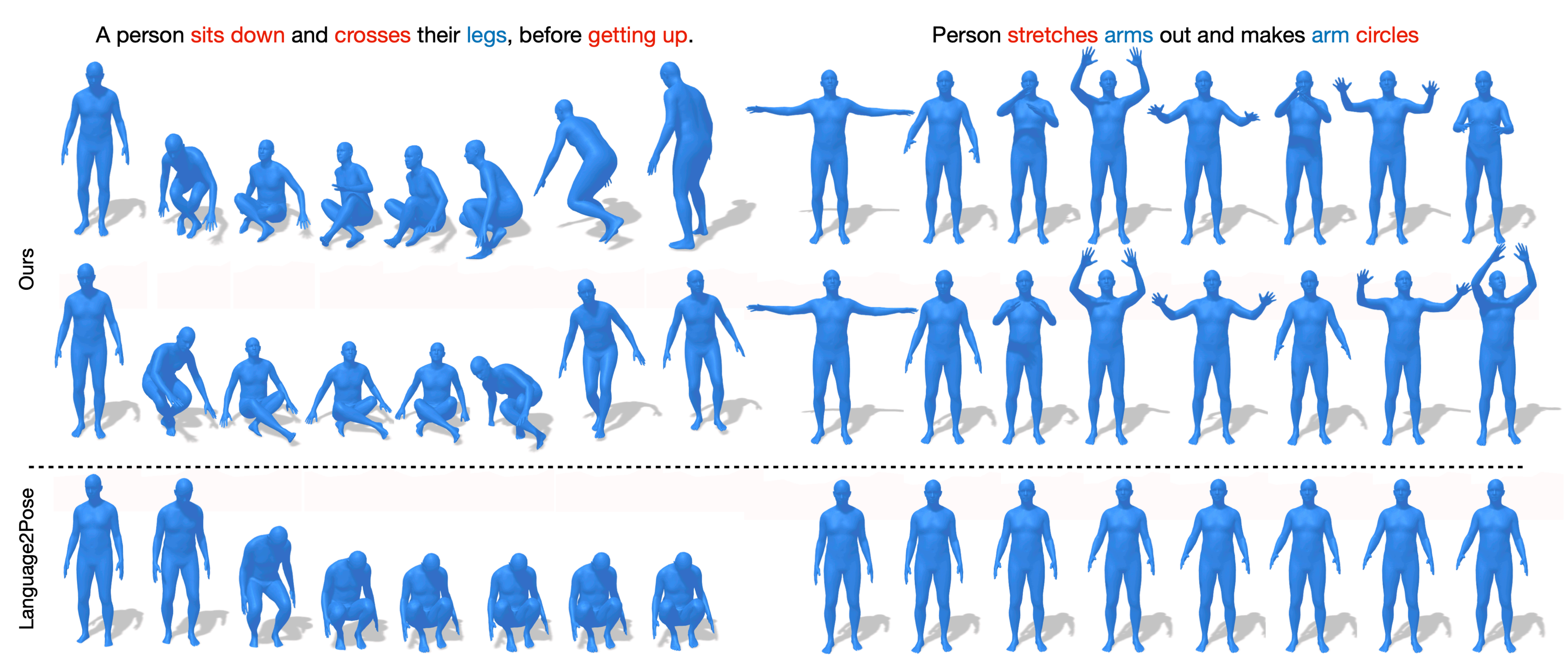
- It consists of 14,616 motions and 44,970 textual descriptions composed by 5,371 distinct words.
- Each motion clip is described by 3 distinct sentences.
- Total and average duration: 28.59 hours and 7.1 seconds respectively.
- Average and median text length: 12 and 10 words respectively.

A Comparison to Existing Dataset:

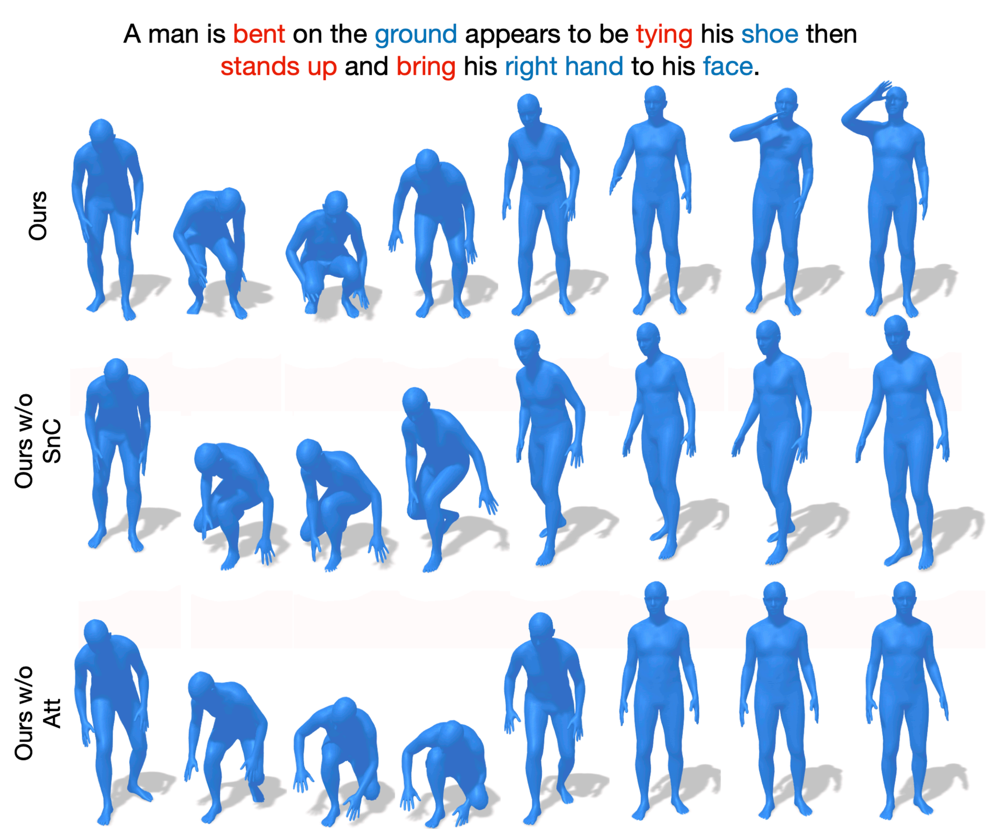
Dataset	#Motions	#texts	Duration	Vocab.
HumanML3D	14,616	44,970	28.59h	5,371
KIT-ML[1]	3,911	6,278	10.33h	1,623

Experiments & Results

Qualitative Results:



Ablation Visualization:



Quantitative Results on HumanML3D Dataset:

Methods	R Precision↑ Top 1	Top 2	Top 3	FID↓	MultiModal Dist↓	Diversity→	MultiModality↑
Real motions	$0.511^{\pm.003}$	$0.703^{\pm.003}$	$0.797^{\pm.002}$	$0.002^{\pm.000}$	$2.974^{\pm.008}$	$9.503^{\pm.065}$	-
Seq2Seq[2]	$0.180^{\pm.002}$	$0.300^{\pm.002}$	$0.396^{\pm.002}$	$11.75^{\pm.035}$	$5.529^{\pm.007}$	$6.223^{\pm.061}$	-
Language2Pose[3]	$0.246^{\pm.002}$	$0.387^{\pm.002}$	$0.486^{\pm.002}$	$11.02^{\pm.046}$	$5.296^{\pm.008}$	$7.676^{\pm.058}$	-
Text2Gesture[4]	$0.165^{\pm.001}$	$0.267^{\pm.002}$	$0.345^{\pm.002}$	$7.664^{\pm.030}$	$6.030^{\pm.008}$	$6.409^{\pm.071}$	-
MoCoGAN[5]	$0.037^{\pm.000}$	$0.072^{\pm.001}$	$0.106^{\pm.001}$	$94.41^{\pm.021}$	$9.643^{\pm.006}$	$0.462^{\pm.008}$	$0.019^{\pm.000}$
Dance2Music[6]	$0.033^{\pm.000}$	$0.065^{\pm.001}$	$0.097^{\pm.001}$	$66.98^{\pm.016}$	$8.116^{\pm.006}$	$0.725^{\pm.011}$	$0.043^{\pm.001}$
Ours w/ real length	$\mathbf{0.457}^{\pm.002}$	$\mathbf{0.639}^{\pm.003}$	$\mathbf{0.740}^{\pm.003}$	$\mathbf{1.067}^{\pm.002}$	$\mathbf{3.340}^{\pm.008}$	$\mathbf{9.188}^{\pm.002}$	$\underline{2.090}^{\pm.083}$
Ours	$\underline{0.455}^{\pm.003}$	$\underline{0.636}^{\pm.003}$	$\underline{0.736}^{\pm.002}$	$\underline{1.087}^{\pm.021}$	$\underline{3.347}^{\pm.008}$	$\underline{9.175}^{\pm.083}$	$\mathbf{2.219}^{\pm.074}$

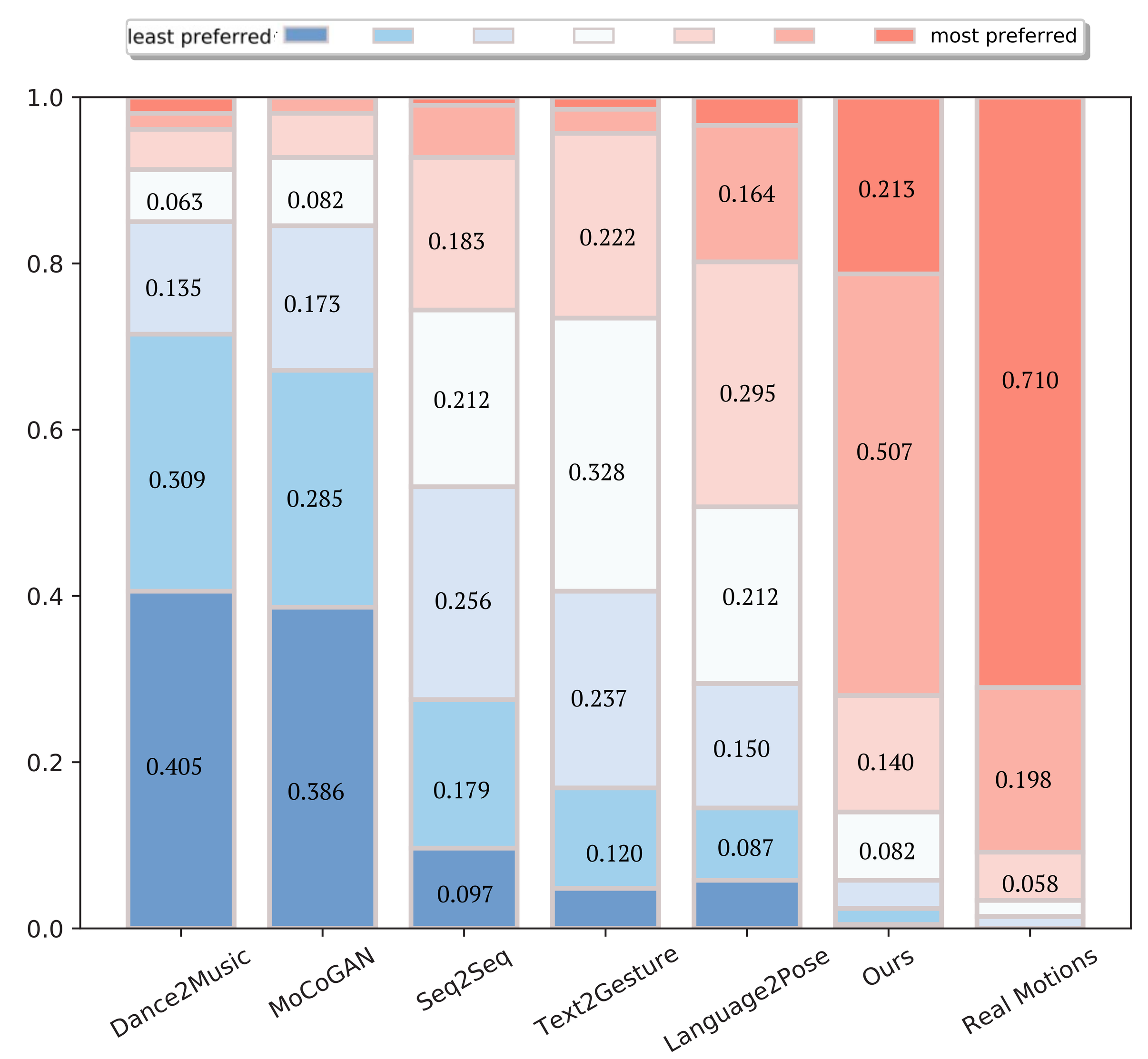
Quantitative Results on KIT-ML Dataset:

Methods	R Precision↑ Top 1	Top 2	Top 3	FID↓	MultiModal Dist↓	Diversity→	MultiModality↑
Real motions	$0.424^{\pm.005}$	$0.649^{\pm.006}$	$0.779^{\pm.006}$	$0.031^{\pm.004}$	$2.788^{\pm.012}$	$11.08^{\pm.097}$	-
Seq2Seq[2]	$0.103^{\pm.003}$	$0.178^{\pm.005}$	$0.241^{\pm.006}$	$24.86^{\pm.348}$	$7.960^{\pm.031}$	$6.744^{\pm.106}$	-
Language2Pose[3]	$0.221^{\pm.005}$	$0.373^{\pm.004}$	$0.483^{\pm.005}$	$6.545^{\pm.072}$	$5.147^{\pm.030}$	$9.073^{\pm.100}$	-
Text2Gesture[4]	$0.156^{\pm.004}$	$0.255^{\pm.004}$	$0.338^{\pm.005}$	$12.12^{\pm.183}$	$6.964^{\pm.029}$	$9.334^{\pm.079}$	-
MoCoGAN[5]	$0.022^{\pm.002}$	$0.042^{\pm.003}$	$0.063^{\pm.003}$	$82.69^{\pm.242}$	$10.47^{\pm.012}$	$3.091^{\pm.043}$	$0.250^{\pm.009}$
Dance2Music[6]	$0.031^{\pm.002}$	$0.058^{\pm.002}$	$0.086^{\pm.003}$	$115.4^{\pm.240}$	$10.40^{\pm.016}$	$0.241^{\pm.004}$	$0.062^{\pm.002}$
Ours w/ real length	$\mathbf{0.370}^{\pm.005}$	$\mathbf{0.569}^{\pm.007}$	$\mathbf{0.693}^{\pm.007}$	$\mathbf{2.770}^{\pm.109}$	$\mathbf{3.401}^{\pm.008}$	$\mathbf{10.91}^{\pm.119}$	$\underline{1.482}^{\pm.065}$
Ours	$\underline{0.361}^{\pm.006}$	$\underline{0.559}^{\pm.007}$	$\underline{0.681}^{\pm.007}$	$\underline{3.022}^{\pm.107}$	$\underline{3.488}^{\pm.028}$	$\underline{10.72}^{\pm.145}$	$\mathbf{2.052}^{\pm.107}$

Ablation Analysis:

Methods	R Precision↑ Top 1	Top 2	Top 3	FID↓
Ours	$\mathbf{0.455}^{\pm.003}$	$\mathbf{0.636}^{\pm.003}$	$\mathbf{0.736}^{\pm.002}$	$\mathbf{1.087}^{\pm.021}$
w/o SnC (Snippet Code)	$0.370^{\pm.002}$	$0.538^{\pm.003}$	$0.642^{\pm.003}$	$1.200^{\pm.027}$
w/o Att (Local Attention)	$0.396^{\pm.002}$	$0.570^{\pm.002}$	$0.674^{\pm.003}$	$1.833^{\pm.032}$
w/o PoS (Part-of-Speech)	$0.443^{\pm.003}$	$0.622^{\pm.003}$	$0.723^{\pm.003}$	$1.157^{\pm.016}$
w/o PoE (Time-to-Arrival)	$0.444^{\pm.005}$	$0.627^{\pm.003}$	$0.729^{\pm.002}$	$1.229^{\pm.020}$

Human Evaluation:



References:

[1] Plappert, Matthias, et al. "The KIT motion-language dataset." Big data 4.4 (2016): 236-252.
[2] Lin, Angela S., et al. "Generating animated videos of human activities from natural language descriptions." Learning 2018 (2018): 1.
[3] Ahuja, Chaitanya, et al. "Language2pose: Natural language grounded pose forecasting." (3DV). IEEE, 2019.
[4] Bhattacharya, Uttaran, et al. "Text2Gestures: A Transformer-Based Network for Generating Emotive Body Gestures for Virtual Agents." (VR), IEEE, 2021.
[5] Tulyakov, Sergey, et al. "Mocogan: Decomposing motion and content for video generation." (CVPR), IEEE, 2018.
[6] Lee, Hsin-Ying, et al. "Dancing to music." Advances in Neural Information Processing Systems 32 (2019).