TM2T: Stochastic and Tokenized Modeling for the Reciprocal Generation of 3D Human Motions and Texts

Chuan Guo Xinxin Zuo Sen Wang Li Cheng
University of Alberta

ECCV TEL AVIV 2022

Code & Model
https://ericguo5513.github.io/TM2T/

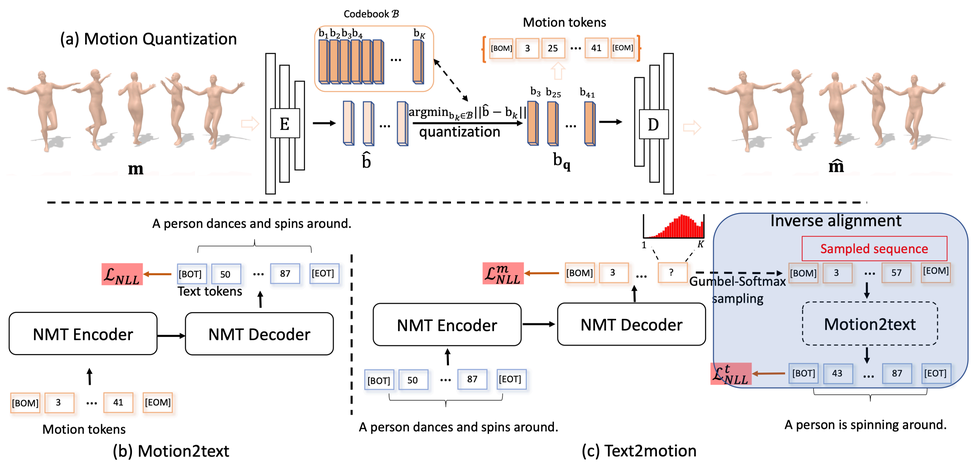
Introduction

Goal: Learning the mutual mappings between texts and motions reciprocally.

Contributions:

- We propose a novel discrete and compact representation of human motion, motion token. This provides one level playing ground when considering both motions and texts.
- Neural machine translator is adopted for stochastic text2motion and motion2text modeling.
- Motion2text is integrated into the training process of text2motion for inverse alignment.
- Extensive experiments and state-of-the-art performance on both two tasks.

Method



Training Quantization: $\mathcal{L}_{vq} = \|\hat{\mathbf{m}} - \mathbf{m}\|_1 + \|\text{sg}[\text{E}(\mathbf{m})] - \mathbf{b_q}\|_2^2 + \beta\|\text{E}(\mathbf{m}) - \text{sg}[\mathbf{b_q}]\|_2^2$

After quantization, we can transform each motion sequence into a discrete token sequence $s \in \{1, ..., |\mathcal{B}|\}^T$, accompanying with its text token sequence $x \in \{1, ..., |\mathcal{V}|\}^N$, where $\mathcal{V}$ is the word vocabulary, and the lengths of motion and text sequence are T and N respectively.

The models of motion2text θ, and text2motion ϕ are optimized as follow:

Training Motion2text: $\mathcal{L}_{m2t} = -\sum_{i=0}^{N-1} \log p_\theta(x_i|x_{<i}, s)$

Training Text2motion: $\mathcal{L}_{t2m} = -\left(\sum_{i=0}^{T-1} \log p_\phi(s_i|s_{<i}, x) + \sum_{i=0}^{N-1} \log p_\theta(x_i|x_{<i}, \hat{s})\right)$

Dataset

HumanML3D[1] is a large 3D human motion dataset that covers a broad range of human actions such as locomotion, sports, and dancing. It consists of **14,616** motions and **44,970** text descriptions. Each motion clip comes with at least 3 descriptions.

KIT-ML[2] contains **3,911** 3D human motion clips and **6,278** text descriptions. For each motion, the corresponding number of text descriptions ranges from one to four.

Both datasets are split into training, testing and validation sets with ratio of 0.8:0.15:0.05.

Experiments & Results

Qualitative Results of Motion2text:

GT: A person perform a golf swing.
RAEs: The person is in a.
Seq2Seq(Att): A man swing his left arm then swing his.
SeqGAN: Person right hand swing.
Ours w/o MT: A person is play a violin.
Ours: A person swing a golf club.

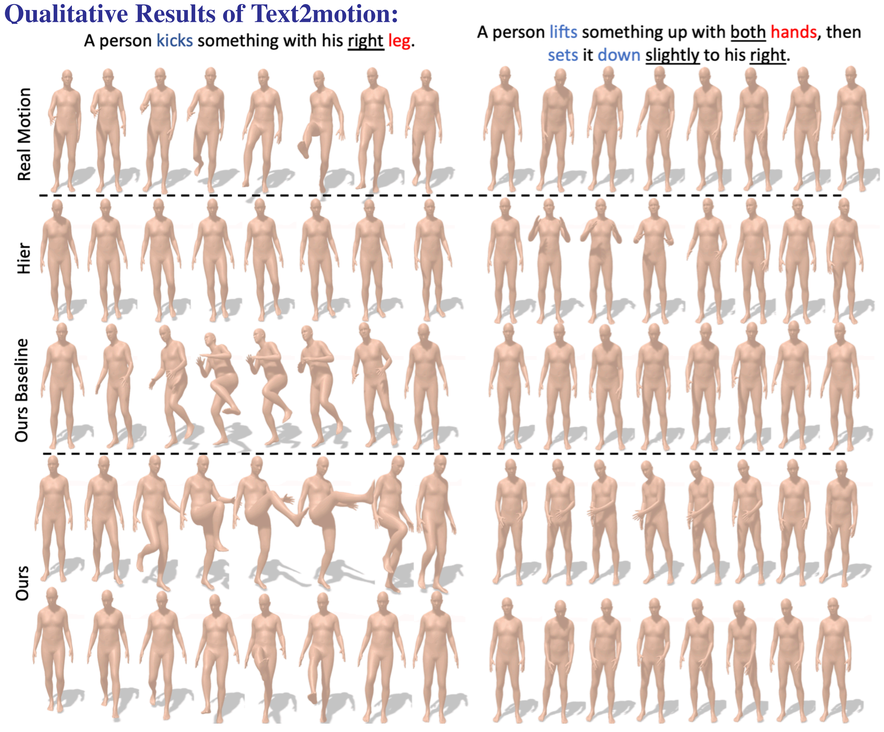
GT: A person is spin arm near chest.
RAEs: The person is in a.
Seq2Seq(Att): The person is stand and move his.
SeqGAN: A person stand with both hand.
Ours w/o MT: A person move his arm in front of him in front of him then move his arm in a circular.
Ours: A person is stand with his arm out in front of him then make a rolling motion with both hand.

GT: A person bend down and touch his toe then reach up and stretch back and forth.
RAEs: The person stand up then.
Seq2Seq(Att): A person bend to the left then.
SeqGAN: A figure stretch bend arm.
Ours w/o MT: A person stretch side to side with his arm above his head.
Ours: A person is stretch his arm over his head and then stretch his body.

Quantitative Results of Motion2text:

Datasets	Methods	R Precision↑ Top 1	Top 2	Top 3	MM Dist↓	Bleu@1↑	Bleu@4↑	Rouge↑	Cider↑	BertScore↑
Human ML3D	**Real Desc**	0.523	0.725	0.828	2.901	-	-	-	-	-
	RAEs[3]	0.100	0.188	0.261	6.337	33.3	10.2	37.5	22.1	10.7
	Seq2Seq(Att)	0.436	0.611	0.706	3.447	51.8	17.9	46.4	58.4	29.1
	SeqGAN[4]	0.332	0.457	0.532	4.895	47.8	13.5	39.2	50.2	23.4
	Ours w/o MT	0.483	0.678	0.783	3.124	59.5	21.2	47.8	68.3	34.9
	Ours	**0.516**	**0.720**	**0.823**	**2.935**	**61.7**	**22.3**	**49.2**	**72.5**	**37.8**
KIT-ML	**Real Desc**	0.399	0.618	0.793	2.772	-	-	-	-	-
	RAEs[3]	0.034	0.063	0.106	9.364	30.6	0.10	25.7	8.00	0.40
	Seq2Seq(Att)	0.293	0.450	0.555	4.455	34.3	9.30	36.3	37.3	5.30
	SeqGAN[4]	0.109	0.345	0.425	6.283	3.12	5.20	32.4	29.5	2.20
	Ours w/o MT	0.284	0.466	0.595	3.979	42.8	14.7	39.9	60.1	18.9
	Ours	**0.359**	**0.561**	**0.668**	**3.298**	**46.7**	**18.4**	**44.2**	**79.5**	**23.0**

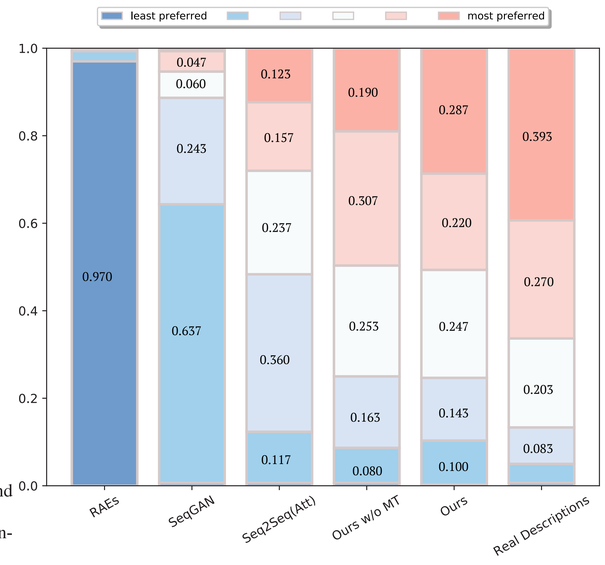
Qualitative Results of Text2motion:



Quantitative Results of Text2motion:

Datasets	Methods	R Precision↑ Top 1	Top 2	Top 3	FID↓	MM Dist↓	Diversity→	MModality↑
Human ML3D	**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[5]	$0.180^{\pm.002}$	$0.300^{\pm.002}$	$0.396^{\pm.002}$	$11.75^{\pm.035}$	$5.529^{\pm.007}$	$6.223^{\pm.061}$	-
	Language2Pose[6]	$0.246^{\pm.002}$	$0.387^{\pm.002}$	$0.486^{\pm.002}$	$11.02^{\pm.046}$	$5.296^{\pm.008}$	$7.676^{\pm.058}$	-
	Text2Gesture[7]	$0.165^{\pm.001}$	$0.267^{\pm.002}$	$0.345^{\pm.002}$	$5.012^{\pm.030}$	$6.030^{\pm.008}$	$6.409^{\pm.071}$	-
	Hier[8]	$0.301^{\pm.002}$	$0.425^{\pm.002}$	$0.552^{\pm.004}$	$6.532^{\pm.024}$	$5.012^{\pm.018}$	$8.332^{\pm.042}$	-
	MoCoGAN[9]	$0.037^{\pm.000}$	$0.072^{\pm.001}$	$0.106^{\pm.001}$	$94.41^{\pm.021}$	$9.643^{\pm.006}$	$0.462^{\pm.008}$	$0.019^{\pm.000}$
	Dance2Music[10]	$0.033^{\pm.000}$	$0.065^{\pm.001}$	$0.097^{\pm.001}$	$66.98^{\pm.016}$	$8.116^{\pm.006}$	$0.725^{\pm.011}$	$0.043^{\pm.001}$
	Ours baseline(T)	$0.351^{\pm.003}$	$0.521^{\pm.003}$	$0.627^{\pm.003}$	$1.669^{\pm.025}$	$4.046^{\pm.018}$	$\mathbf{9.632}^{\pm.072}$	$\mathbf{4.352}^{\pm.149}$
	Ours baseline	$0.351^{\pm.002}$	$0.526^{\pm.002}$	$0.635^{\pm.002}$	$1.739^{\pm.022}$	$3.965^{\pm.010}$	$8.651^{\pm.083}$	$3.139^{\pm.083}$
	Ours	$\mathbf{0.424}^{\pm.003}$	$\mathbf{0.618}^{\pm.003}$	$\mathbf{0.729}^{\pm.002}$	$\mathbf{1.501}^{\pm.017}$	$\mathbf{3.467}^{\pm.011}$	$8.589^{\pm.076}$	$2.424^{\pm.093}$
KIT-ML	**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[5]	$0.103^{\pm.003}$	$0.178^{\pm.005}$	$0.241^{\pm.006}$	$24.86^{\pm.348}$	$7.960^{\pm.031}$	$6.744^{\pm.106}$	-
	Language2Pose[6]	$0.221^{\pm.005}$	$0.373^{\pm.004}$	$0.483^{\pm.005}$	$6.545^{\pm.072}$	$5.147^{\pm.030}$	$9.073^{\pm.100}$	-
	Text2Gesture[7]	$0.156^{\pm.004}$	$0.255^{\pm.004}$	$0.338^{\pm.005}$	$12.12^{\pm.183}$	$6.964^{\pm.029}$	$9.334^{\pm.079}$	-
	Hier[8]	$0.255^{\pm.006}$	$0.432^{\pm.007}$	$0.531^{\pm.007}$	$5.203^{\pm.107}$	$4.986^{\pm.027}$	$9.563^{\pm.072}$	-
	MoCoGAN[9]	$0.022^{\pm.002}$	$0.042^{\pm.003}$	$0.063^{\pm.003}$	$82.69^{\pm.242}$	$10.47^{\pm.012}$	$3.091^{\pm.043}$	$0.250^{\pm.009}$
	Dance2Music[10]	$0.031^{\pm.002}$	$0.058^{\pm.002}$	$0.086^{\pm.003}$	$115.4^{\pm.240}$	$10.40^{\pm.016}$	$0.241^{\pm.004}$	$0.062^{\pm.002}$
	Ours baseline(T)	$0.260^{\pm.005}$	$0.426^{\pm.007}$	$0.538^{\pm.008}$	$4.628^{\pm.126}$	$4.835^{\pm.076}$	$12.16^{\pm.120}$	$4.436^{\pm.106}$
	Ours baseline	$0.251^{\pm.007}$	$0.418^{\pm.008}$	$0.535^{\pm.007}$	$4.814^{\pm.145}$	$4.682^{\pm.048}$	$\mathbf{10.13}^{\pm.117}$	$\mathbf{4.486}^{\pm.117}$
	Ours	$\mathbf{0.280}^{\pm.005}$	$\mathbf{0.463}^{\pm.006}$	$\mathbf{0.587}^{\pm.005}$	$\mathbf{3.599}^{\pm.153}$	$\mathbf{4.591}^{\pm.026}$	$9.473^{\pm.117}$	$3.292^{\pm.081}$

Human Evaluation:



Please visit https://ericguo5513.github.io/TM2T/ for the paper, video, data, code and pre-trained model.

References:

[1] Guo, Chuan, et al. "Generating Diverse and Natural 3D Human Motions From Text." (CVPR). 2022.
[2] Plappert, Matthias, et al. "The KIT motion-language dataset." Big data 4.4 (2016): 236-252.
[3] Yamada, Tatsuro, et al. "Paired recurrent autoencoders for bidirectional translation between robot actions and linguistic descriptions." (RAL), 2018.
[4] Goutsu, Yusuke, et al. "Linguistic Descriptions of Human Motion with Generative Adversarial Seq2Seq Learning."(ICRA). 2021.
[5] Lin, Angela S., et al. "Generating animated videos of human activities from natural language descriptions." Learning 2018 (2018): 1.
[6] Ahuja, Chaitanya, et al. "Language2pose: Natural language grounded pose forecasting." (3DV). IEEE, 2019.
[7] Bhattacharya, Uttaran, et al. "Text2Gestures: A Transformer-Based Network for Generating Emotive Body Gestures for Virtual Agents." (VR), IEEE, 2021.
[8] Ghosh, Anindita, et al. "Synthesis of compositional animations from textual descriptions." (ICCV) 2021.
[9] Tulyakov, Sergey, et al. "Mocogan: Decomposing motion and content for video generation." (CVPR), 2018.
[10] Lee, Hsin-Ying, et al. "Dancing to music." Advances in Neural Information Processing Systems 32 (2019).