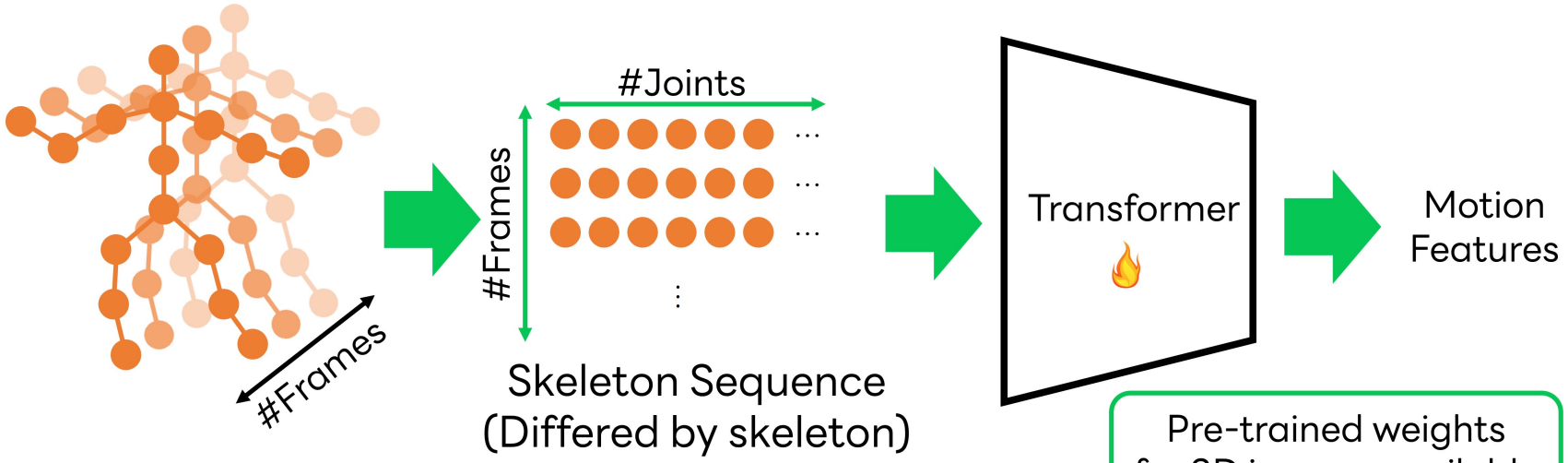
LINEヤフー

Exploring Vision Transformers for 3D Human Motion-Language Models with Motion Patches

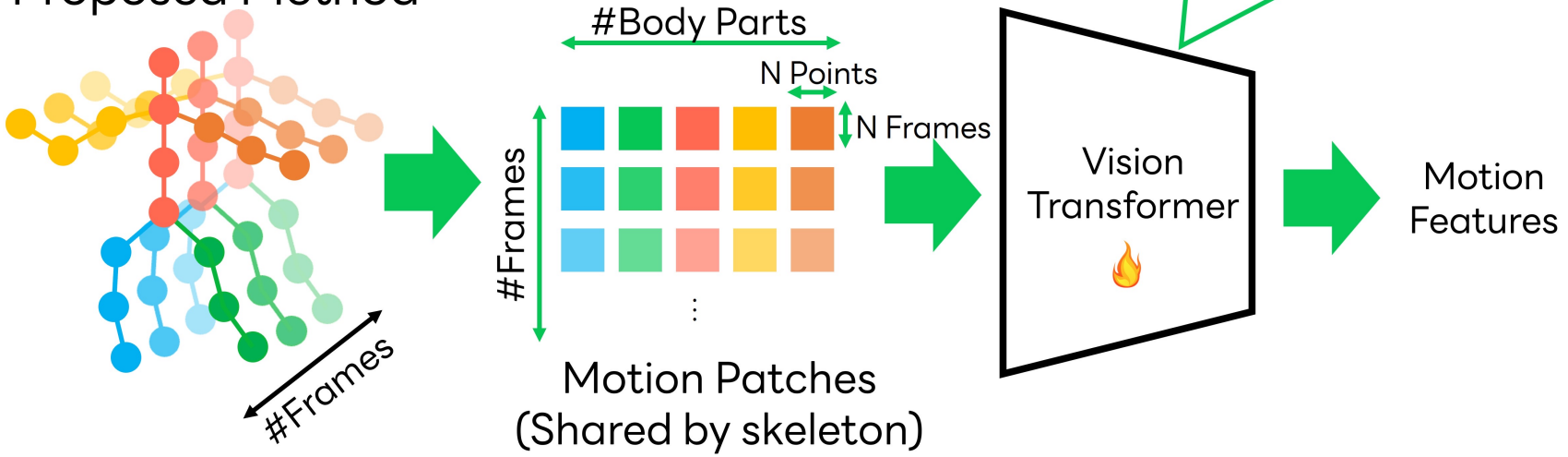
Qing Yu, Mikihiro Tanaka and Kent Fujiwara
LY Corporation

Problem Setting

Existing Methods

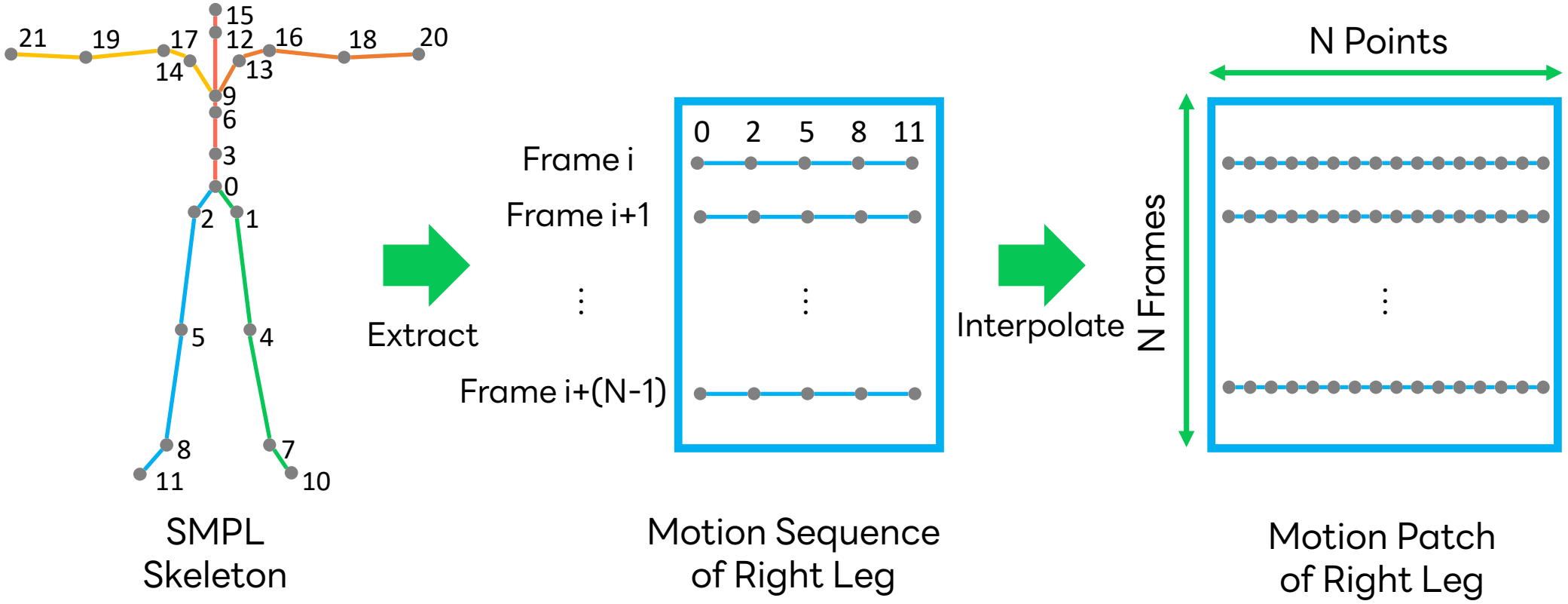


Proposed Method

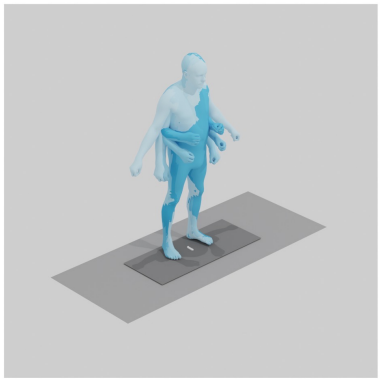


Method

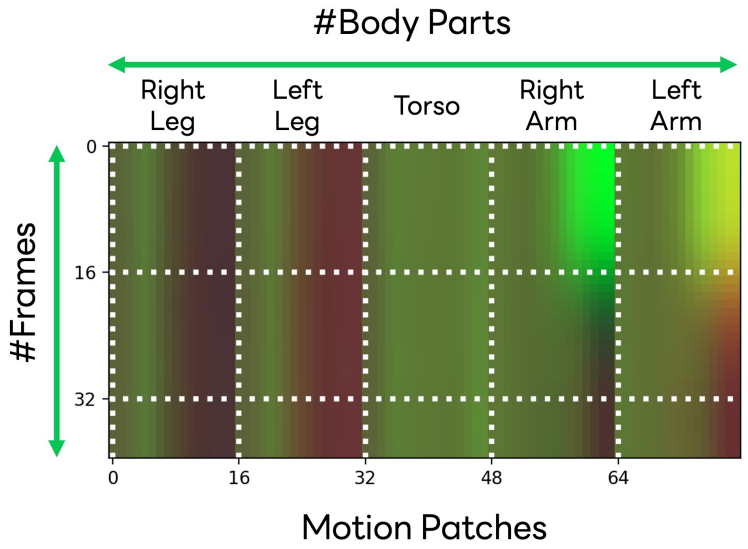
- Motion Patches



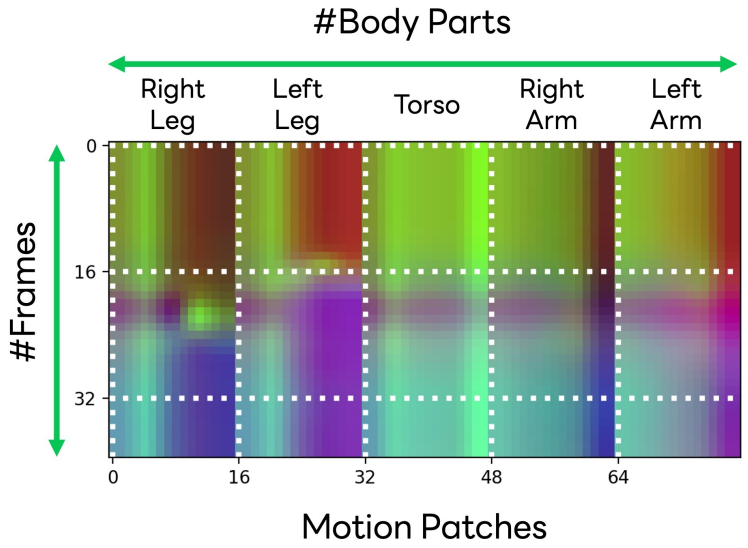
Motion Patches



a person brings their arms down

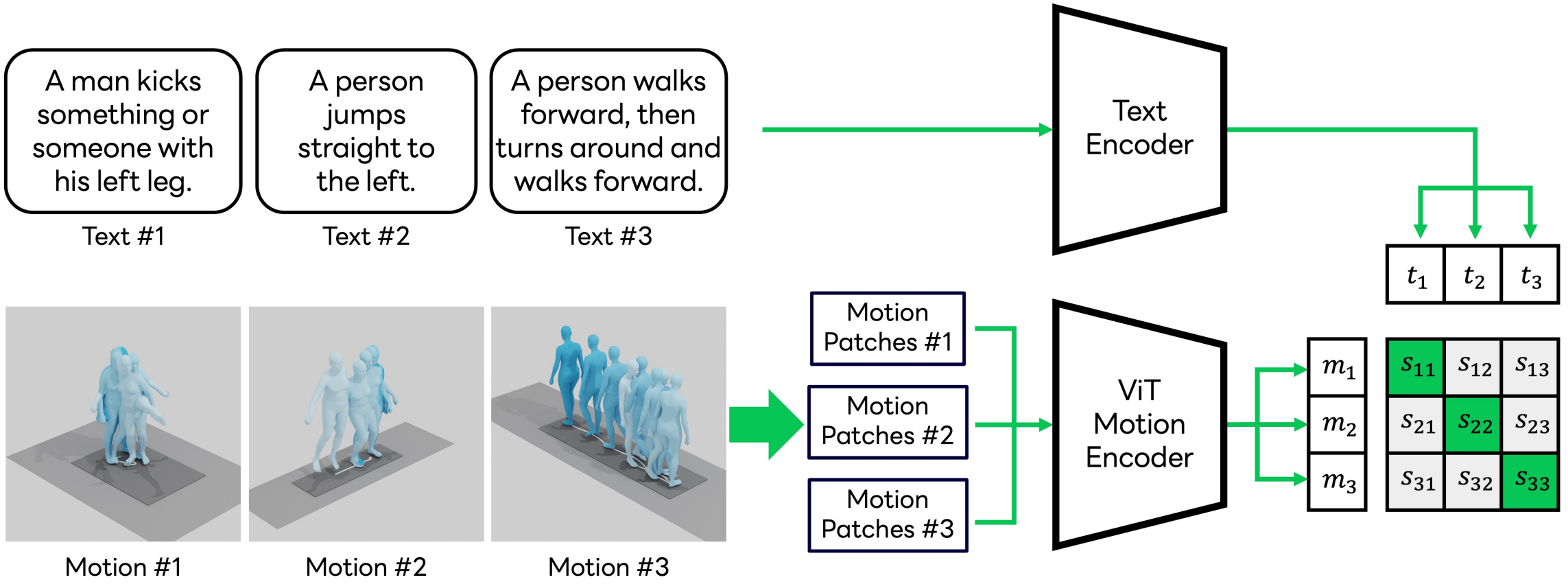


a person takes a small step forward



Method

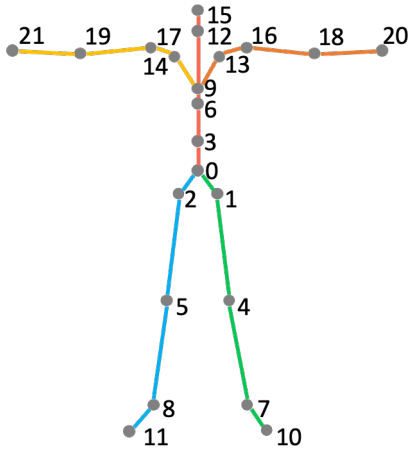
- Training: CLIP



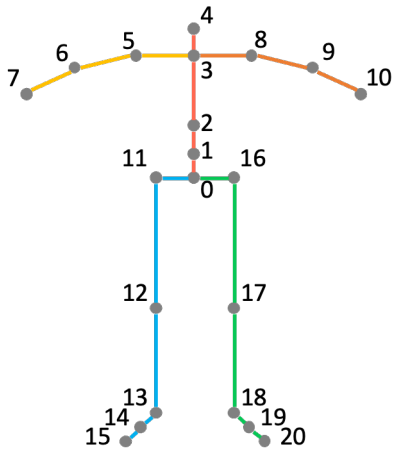
Experiment

- Task: Motion↔Text Retrieval

- Dataset:
  - HumanML3D: #Train 23384, #Test 4380
  - KIT-ML: #Train 4888, #Test 830



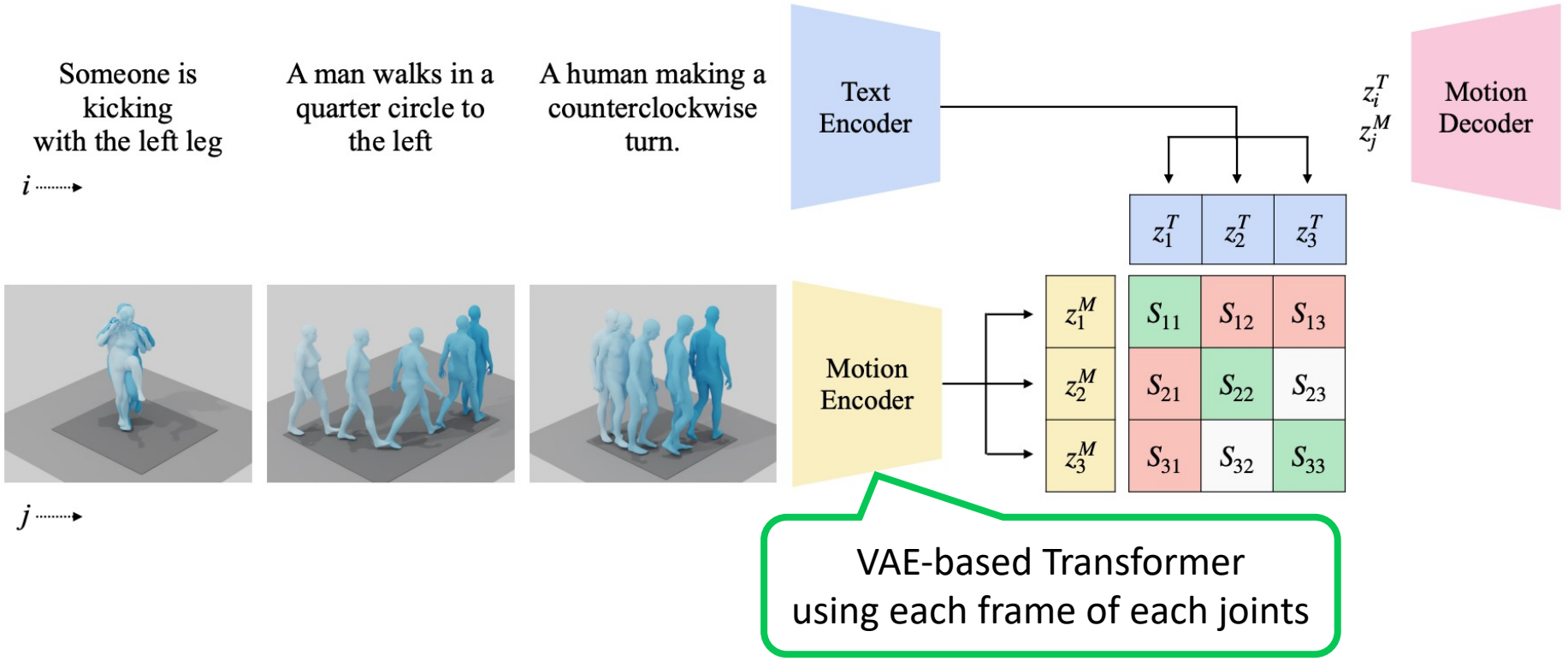
HumanML3D (22 joints)



KIT-ML (21 joints)

Related Works

- TMR: Text-to-Motion Retrieval Using Contrastive 3D Human Motion Synthesis [Petrovich+, ICCV'23]



Result (Text-motion retrieval)

HumanML3D

Method	R@1↑	R@5↑	R@10↑	MedR↓
TEMOS [ECCV'22]	2.12	8.26	13.52	173.0
T2M [CVPR'22]	1.80	7.12	12.47	81.00
TMR [ICCV'23]	8.92	22.06	33.37	25.00
Ours	**10.80**	**26.72**	**38.02**	**19.00**

KIT-ML

Method	R@1↑	R@5↑	R@10↑	MedR↓
TEMOS [ECCV'22]	7.11	24.10	35.66	24.00
T2M [CVPR'22]	3.37	16.87	27.71	28.00
TMR [ICCV'23]	10.05	30.03	44.66	14.00
Ours	**14.02**	**34.10**	**50.00**	**10.50**

Result (Motion-text retrieval)

HumanML3D

Method	R@1↑	R@5↑	R@10↑	MedR↓
TEMOS [ECCV'22]	3.86	9.38	14.00	183.25
T2M [CVPR'22]	2.92	8.36	12.95	81.50
TMR [ICCV'23]	9.44	22.92	32.21	26.00
Ours	**11.25**	**26.86**	**37.40**	**20.50**

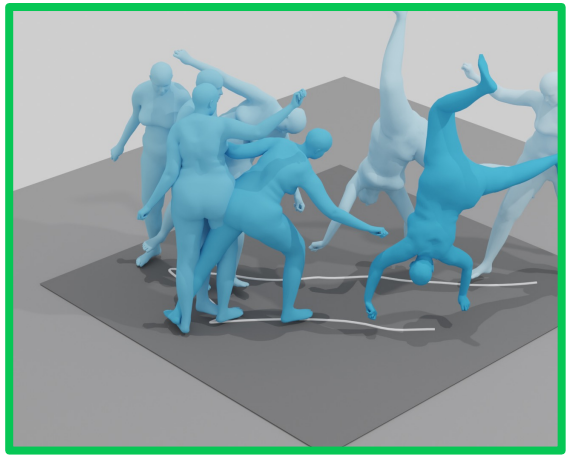
KIT-ML

Method	R@1↑	R@5↑	R@10↑	MedR↓
TEMOS [ECCV'22]	11.69	26.63	36.39	26.50
T2M [CVPR'22]	4.94	16.14	25.30	28.50
TMR [ICCV'23]	11.83	29.39	38.55	16.00
Ours	**13.61**	**33.33**	**44.77**	**13.00**

Qualitative Results

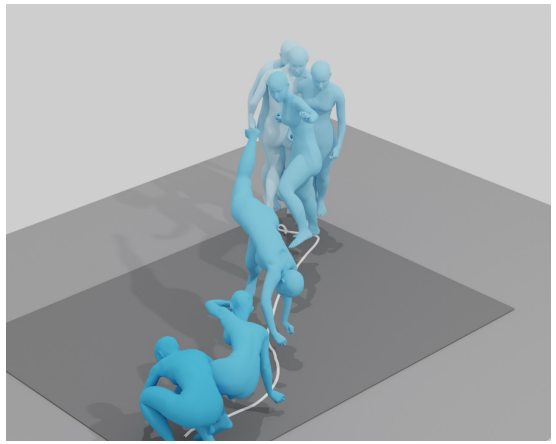
Q The person does 2 cartwheels.

Rank #1



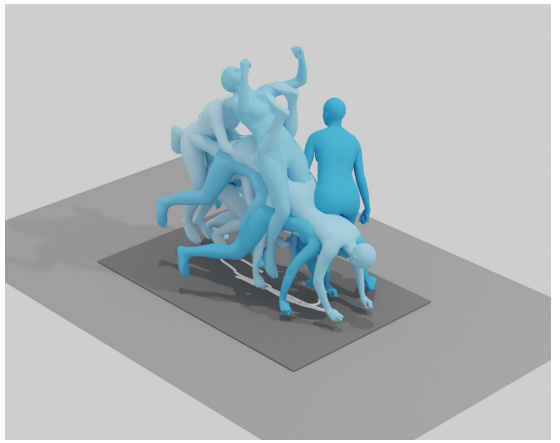
The person does 2 cartwheels.

Rank #2



A person walks forward then turns completely around and does a cartwheel.

Rank #3

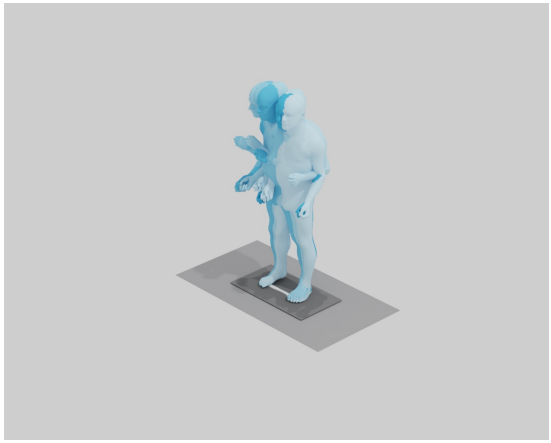


Doing a cartwheel then jumping up and down.

Qualitative Results

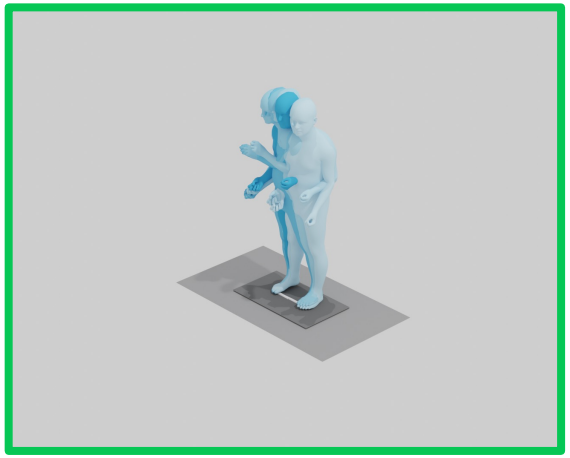
The person pick something up and tilted it onto the right.

Rank #1



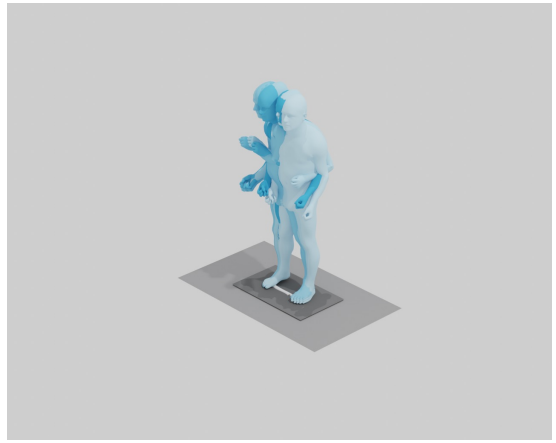
A person picks something up, tilts it, and then places it back.

Rank #2



The person pick something up and tilted it onto the right.

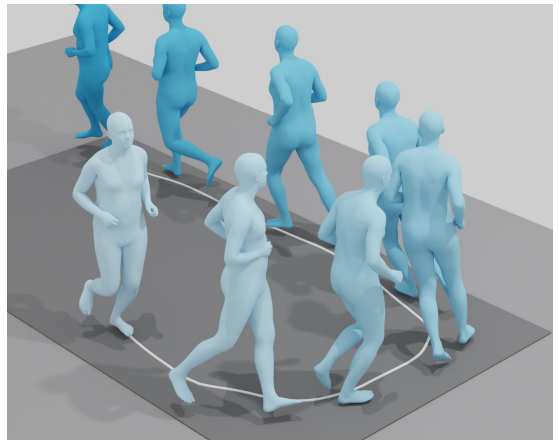
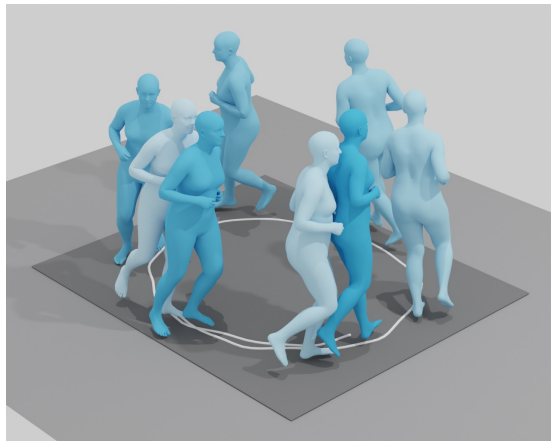
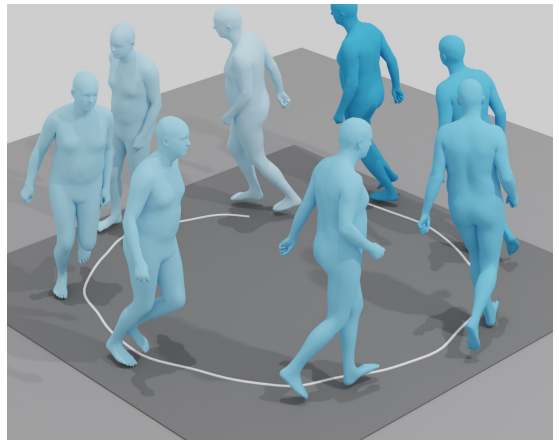
Rank #3



A person picking something up from the left and then placing it right.

Qualitative Results

A person is running in a circle.

Rank #1	Rank #2	Rank #3
		
A person runs moderately in an oval shape.	A person continuously jogs counter clockwise.	A person is jogging around.

More contents

- More Applications:
  - Cross-skeleton Recognition
  - Zero-shot Motion Classification
  - Two-people Interaction Recognition
- More Ablation Studies
- More Visualizations

Paper

Project Page