Language-guided Human Motion Synthesis with Atomic Actions

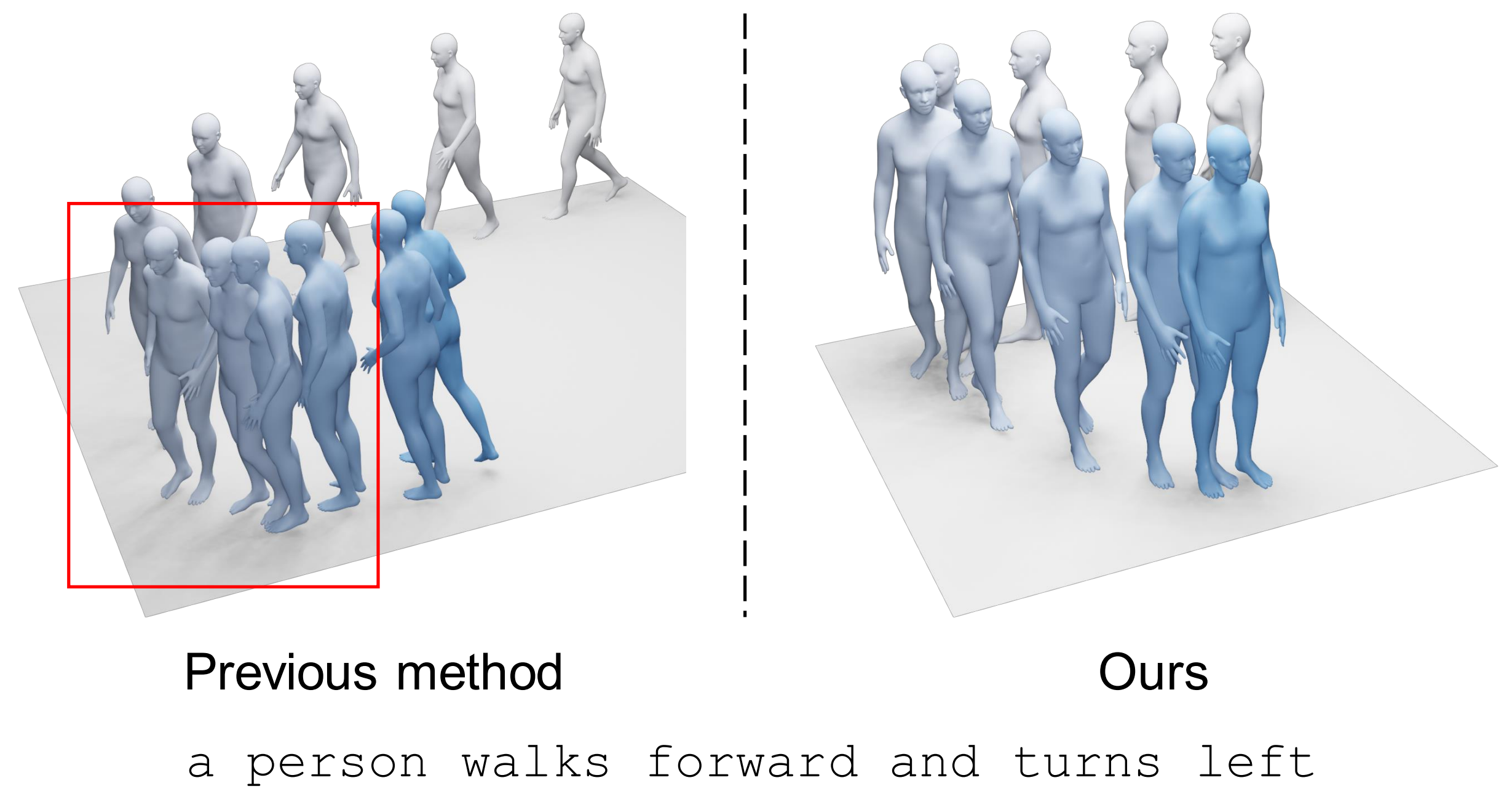
University at Buffalo
The State University of New York

Yuanhao Zhai, Mingzhen Huang, Tianyu Luan, Lu Dong,
Ifeoma Nwogu, Siwei Lyu, David Doermann, Junsong Yuan

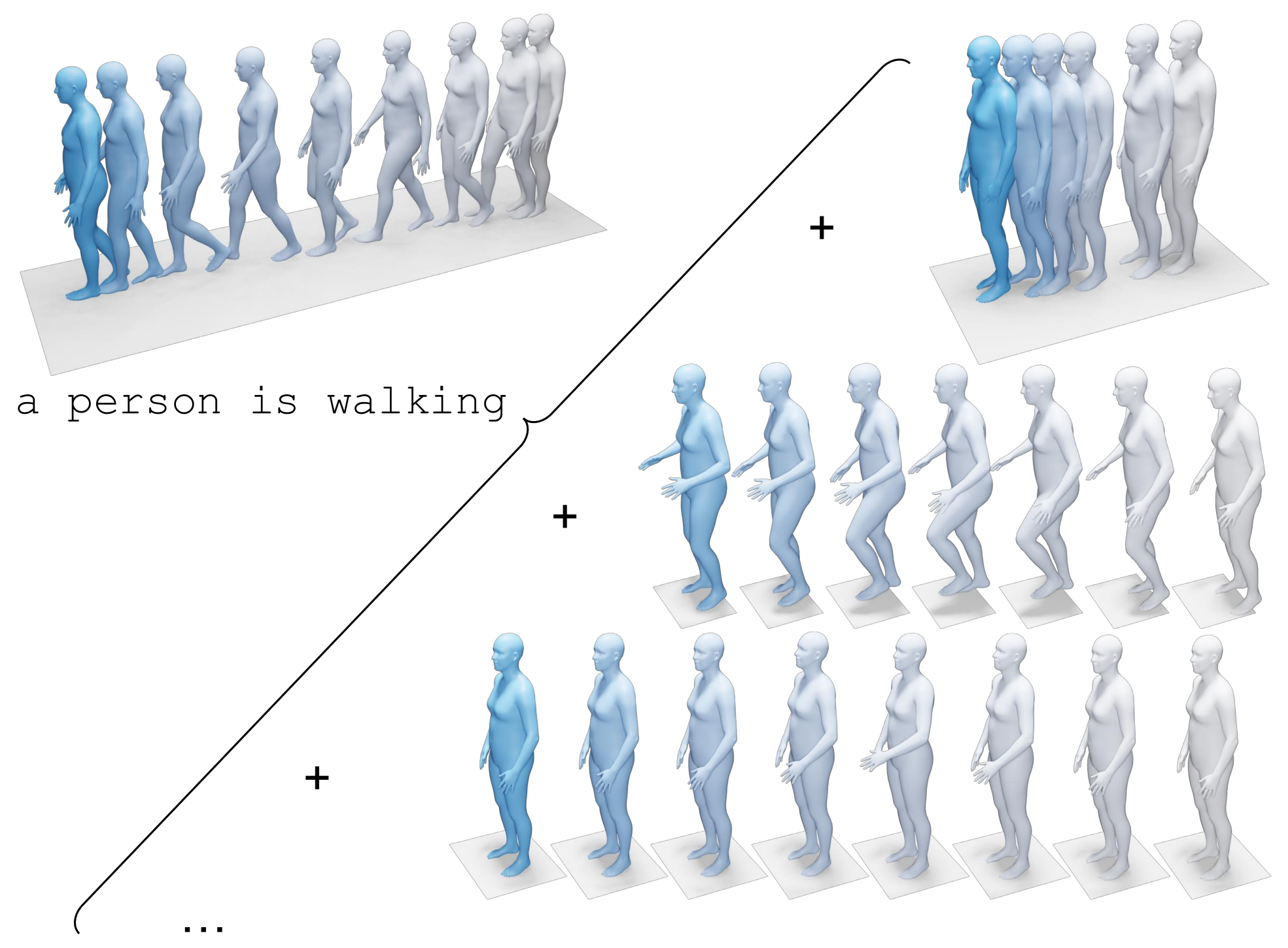
Project page

Motivation

- **Problem**: discontinuities and unrealistic motion transitions from existing methods
  - For rare or unseen actions, this problem leads to abrupt transitions and incoherent movement patterns

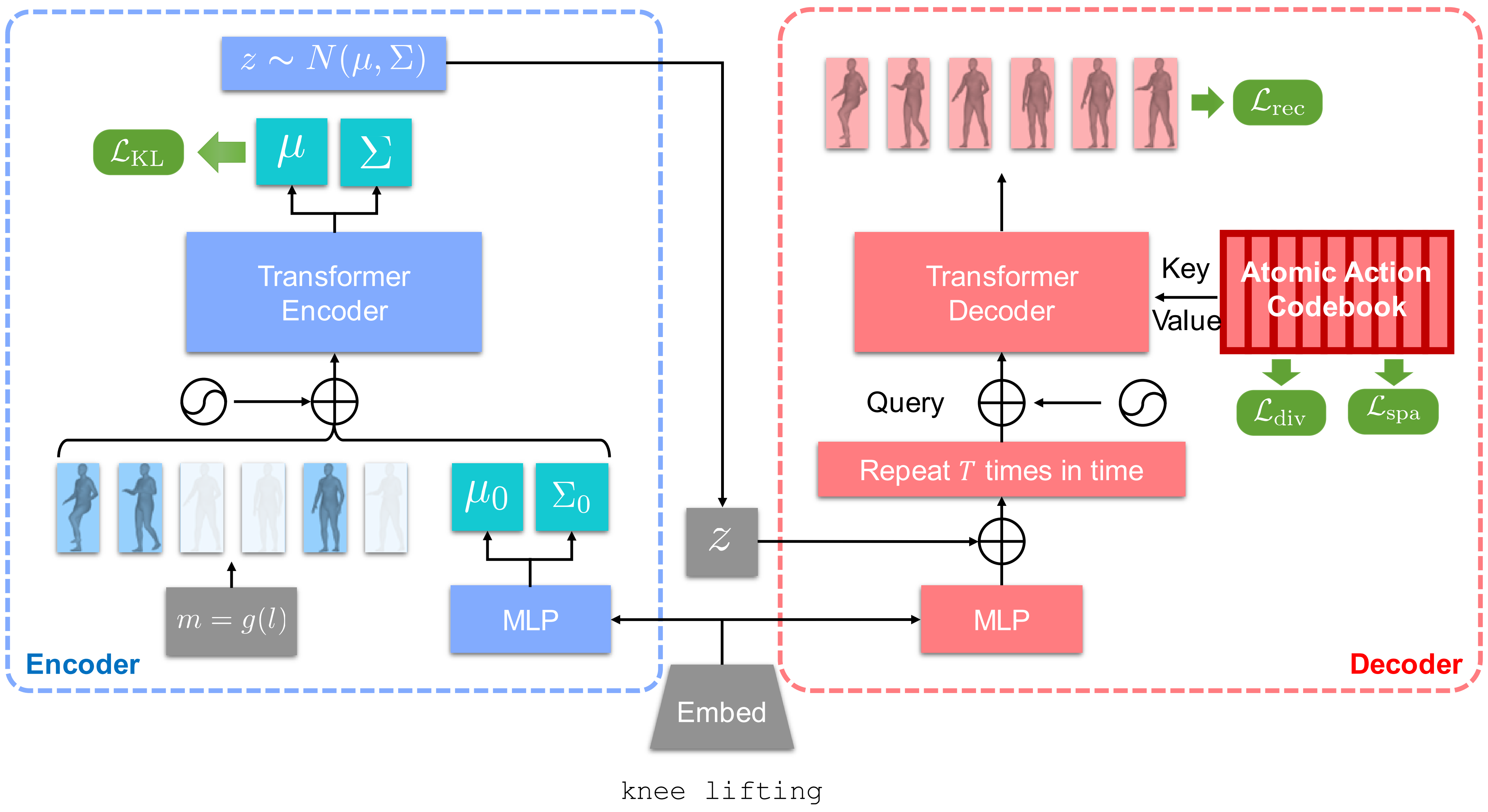


- **Solution**: decompose actions into atomic components, enabling the generation of diverse and coherent motion by assembling the learned atomic actions



Method: ATOM (ATomic mOtion Modeling)

- **Conditional Transformer VAE**
  - Align the motion representation with the conditional language input
- **Atomic action codebook**
  - Decompose complex action into a set of specific, repetitive, and atomic elements
  - A **diversity constraint** ensures that the learned atomic actions are diverse and unique
  - A **sparsity constraint** promotes the use of a sparse set of atomic actions to represent complex motions, enhancing the atomicity and robustness
- **Masked Motion Modeling Curriculum Learning**
  - Temporally mask a random portion of the input motion sequence to learn robust, context-aware motion representations
  - Progressively increasing the mask ratio to enables a more effective and stable learning



Result

- **Quantitative comparison**
  - Strong FID, diversity, and multimodality

KIT

Method	FID ↓	Diversity →	MultiModality ↑	R Precision (top3) ↑	MultiModal Dist ↓
Real Motion	$0.031^{\pm.004}$	$11.08^{\pm.097}$	-	$0.779^{\pm.006}$	$2.788^{\pm.012}$
Language2Pose [1]	$6.545^{\pm.072}$	$9.073^{\pm.100}$	-	$0.483^{\pm.005}$	$5.147^{\pm.030}$
Text2Gesture [3]	$12.12^{\pm.183}$	$9.334^{\pm.079}$	-	$0.338^{\pm.005}$	$6.964^{\pm.029}$
Hier [10]	$5.203^{\pm.107}$	$9.563^{\pm.072}$	-	$0.531^{\pm.007}$	$4.986^{\pm.027}$
T2M [13]	$2.770^{\pm.109}$	$10.91^{\pm.119}$	$1.482^{\pm.065}$	$\mathbf{0.693^{\pm.007}}$	$\mathbf{3.401^{\pm.008}}$
MoCoGAN [45]	$82.69^{\pm.242}$	$3.092^{\pm.043}$	$0.250^{\pm.009}$	$0.063^{\pm.003}$	$10.47^{\pm.012}$
Dance2Music [22]	$115.4^{\pm.240}$	$0.241^{\pm.004}$	$0.062^{\pm.002}$	$0.086^{\pm.003}$	$10.40^{\pm.016}$
Ours	$\mathbf{0.472^{\pm.029}}$	$\mathbf{10.957^{\pm.092}}$	$\mathbf{2.049^{\pm.086}}$	$0.390^{\pm.006}$	$9.161^{\pm.027}$

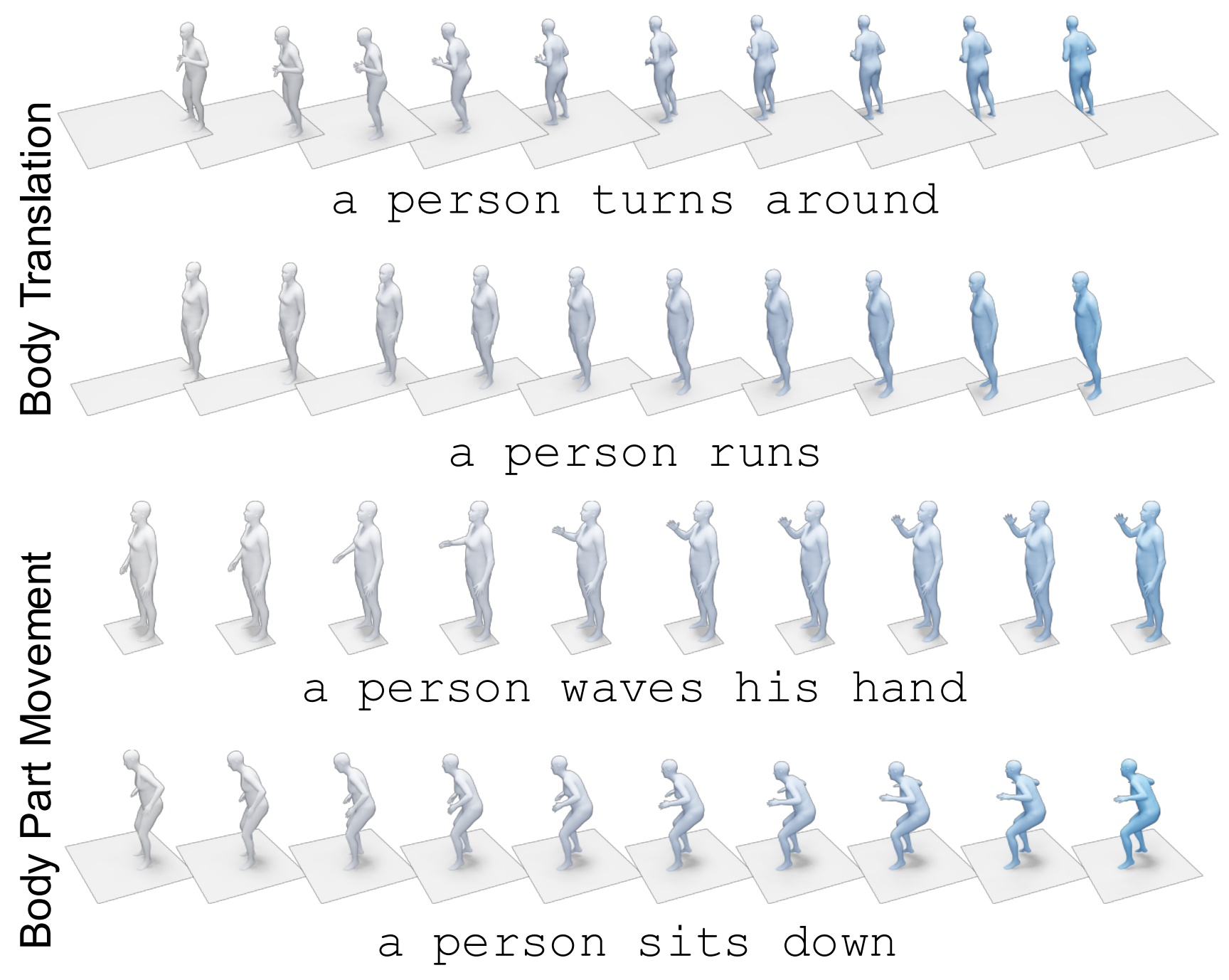
UESTC

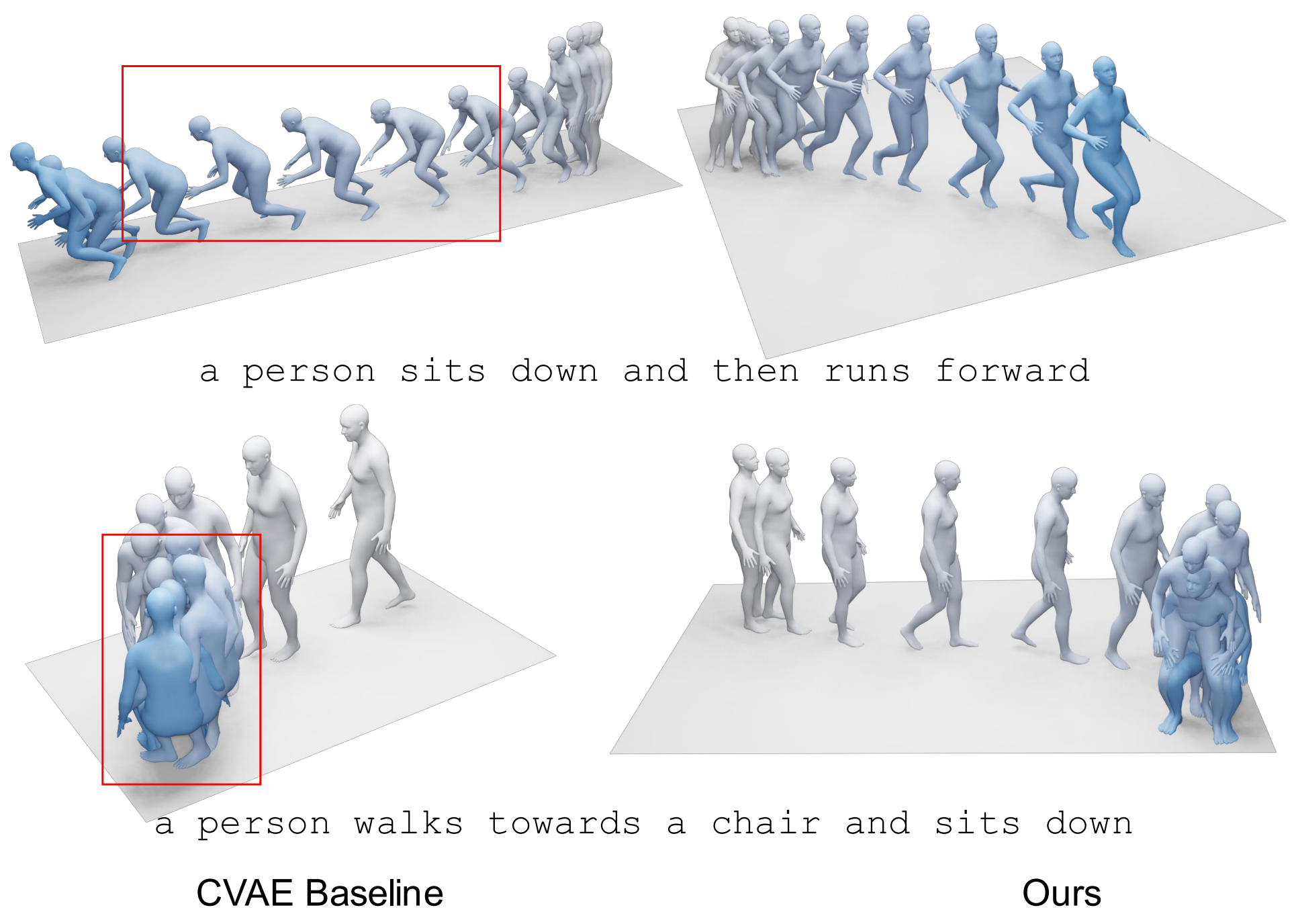
Method	FID (train) ↓	FID (test) ↓	Accuracy ↑	Diversity →	MultiModality →
Real Motion	$2.92^{\pm.26}$	$2.79^{\pm.29}$	$0.988^{\pm.01}$	$33.44^{\pm.320}$	$14.16^{\pm.06}$
Action2Motion [15]	$21.02^{\pm2.51}$	$24.08^{\pm2.17}$	$0.889^{\pm.01}$	$30.47^{\pm.33}$	$13.46^{\pm.03}$
ACTOR [34]	$20.49^{\pm2.31}$	$23.43^{\pm2.20}$	$0.911^{\pm.00}$	$31.96^{\pm.36}$	$\mathbf{14.66^{\pm.03}}$
INR [6]	$9.55^{\pm.06}$	$15.00^{\pm.09}$	$\mathbf{0.941^{\pm.00}}$	$31.59^{\pm.19}$	$14.68^{\pm.07}$
Ours	$\mathbf{6.68^{\pm.04}}$	$\mathbf{9.67^{\pm.17}}$	$0.934^{\pm.01}$	$\mathbf{32.22^{\pm.13}}$	$15.43^{\pm.06}$

Method	FID ↓	Diversity →	MultiModality ↑	R Precision (top3) ↑	MultiModal Dist ↓
Real Motion	$0.002^{\pm.000}$	$9.503^{\pm.065}$	-	$0.797^{\pm.002}$	$2.974^{\pm.008}$
Language2Pose [1]	$11.02^{\pm.046}$	$7.676^{\pm.058}$	-	$0.486^{\pm.002}$	$5.296^{\pm.008}$
Text2Gesture [3]	$7.664^{\pm.030}$	$6.409^{\pm.071}$	-	$0.345^{\pm.002}$	$6.030^{\pm.008}$
Hier [10]	$6.532^{\pm.024}$	$8.332^{\pm.042}$	-	$0.552^{\pm.004}$	$5.012^{\pm.018}$
T2M [13]	$\mathbf{0.455^{\pm.003}}$	$9.175^{\pm.002}$	$2.219^{\pm.074}$	$\mathbf{0.736^{\pm.002}}$	$\mathbf{3.347^{\pm.074}}$
MoCoGAN [45]	$94.41^{\pm.021}$	$0.462^{\pm.008}$	$0.019^{\pm.000}$	$0.106^{\pm.001}$	$9.643^{\pm.006}$
Dance2Music [22]	$66.98^{\pm.016}$	$0.725^{\pm.011}$	$0.043^{\pm.001}$	$0.097^{\pm.001}$	$8.116^{\pm.006}$
Ours	$1.691^{\pm.031}$	$\mathbf{9.312^{\pm.011}}$	$\mathbf{2.884^{\pm.130}}$	$0.569^{\pm.004}$	$5.970^{\pm.004}$

HumanML3D

- **Qualitative results**
  - Our ATOM mainly learns two types of atomic actions
    - Whole body translation
    - Body part movement
  - More smooth motion transition




a person sits down and then runs forward
a person walks towards a chair and sits down
CVAE Baseline
Ours