



# TEMOS: Generating diverse human motions from textual descriptions

Paper, video,  
PyTorch code,  
pretrained models  
available online

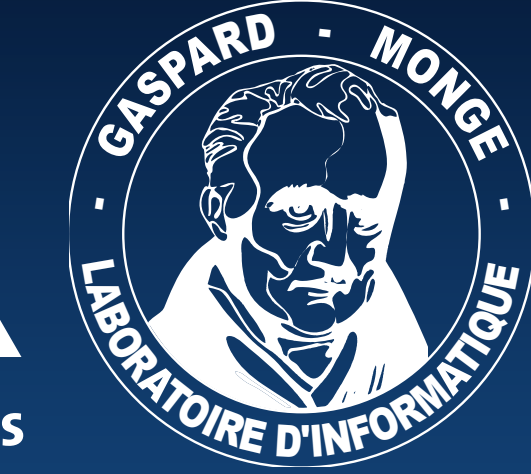


<https://mathis.petrovich.fr/temos>

Mathis Petrovich<sup>1,2</sup> Michael J. Black<sup>2</sup> Gül Varol<sup>1</sup>

<sup>1</sup>LIGM, École des Ponts, Univ Gustave Eiffel, CNRS, France

<sup>2</sup>Max Planck Institute for Intelligent Systems, Tübingen, Germany



## Introduction

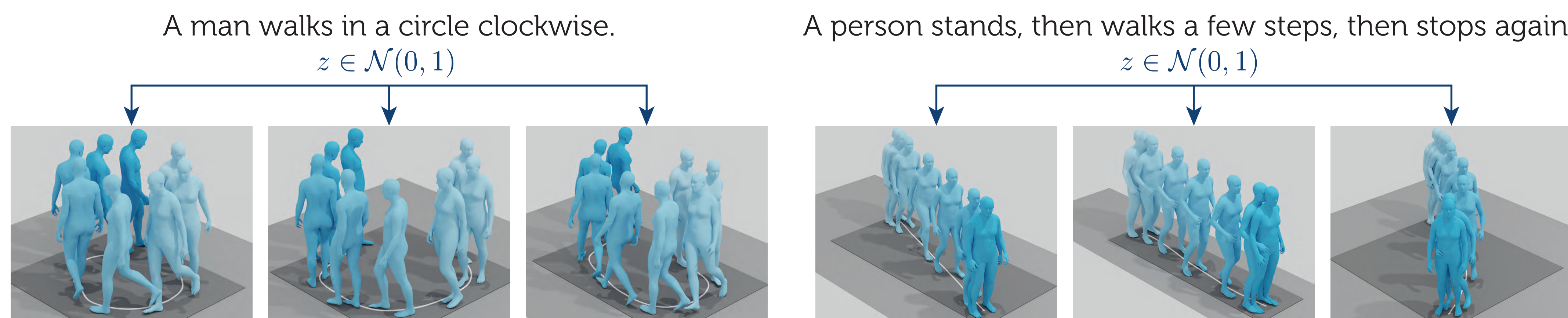
**Goal:** Given a textual description, the task is to generate **multiple** diverse 3D human motions

**Prior work:** Deterministic (generate only one motion), jittery, complex models and losses

**Solution:** Encode the text into a gaussian **distribution**, use a non-autoregressive Transformer

**Contributions:** ○ Novel cross-modal variational model that can produce **diverse** motions

- Achieves **state-of-the-art performance**
- Provide **extensive** ablation study



## TEMOS: Text-to-Motions

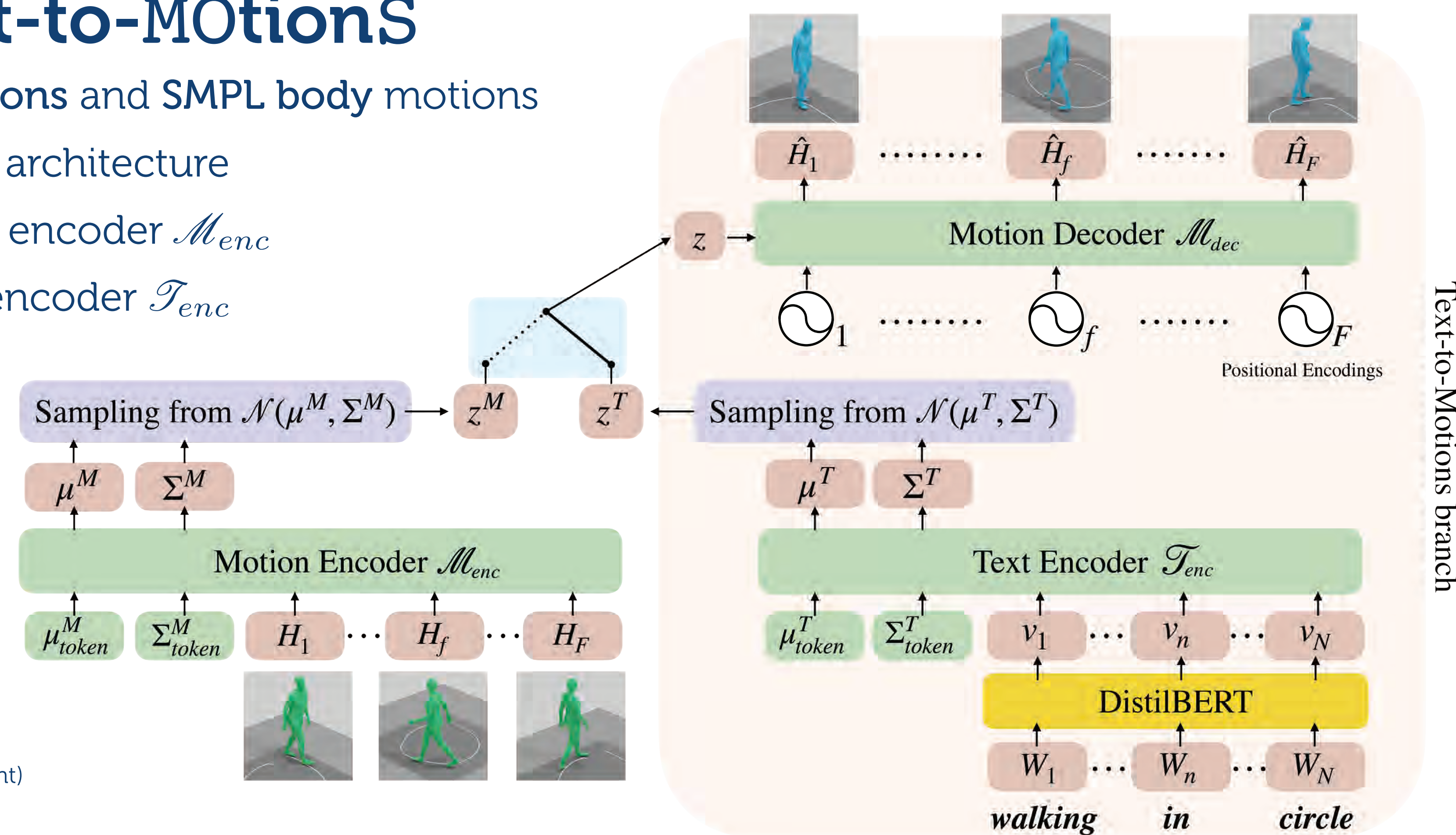
- Supports both **skeletons** and **SMPL** body motions
- Non auto-regressive architecture
- *Motion-level* motion encoder  $\mathcal{M}_{enc}$
- *Sentence-level* text encoder  $\mathcal{T}_{enc}$

### Training

- Reconstruction loss on the motion-to-motions branch (left)
- Reconstruction loss on the text-to-motions branch (right)
- Cross-modal losses to encourage a joint space between motion and text
- Gaussian priors to regularize the joint latent space

### Inference

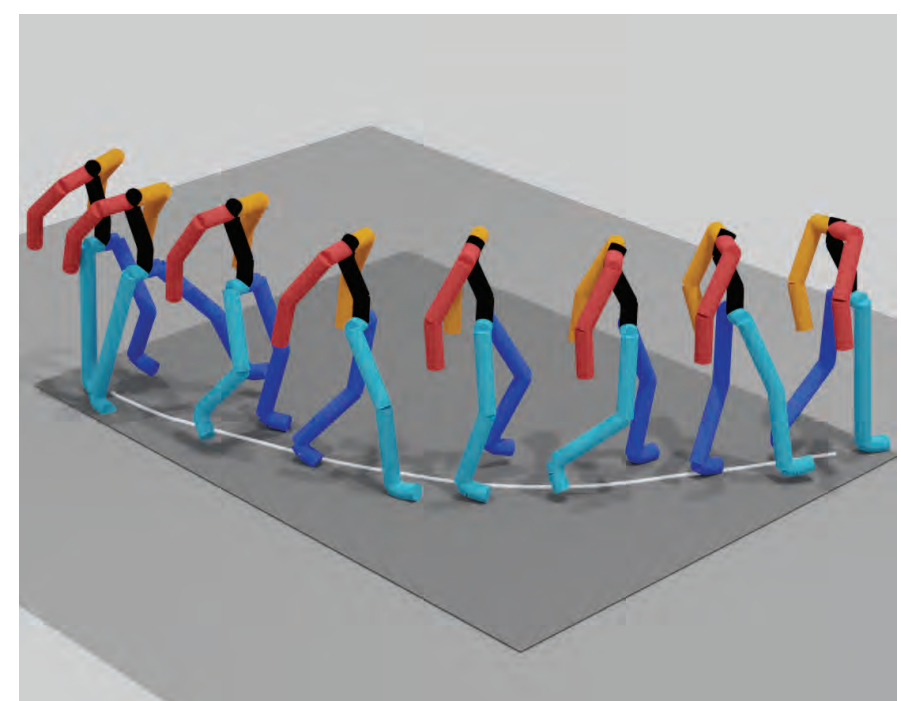
- Only the **text-to-motions** branch (right)
- Ability to sample **multiple** motions given a single textual description



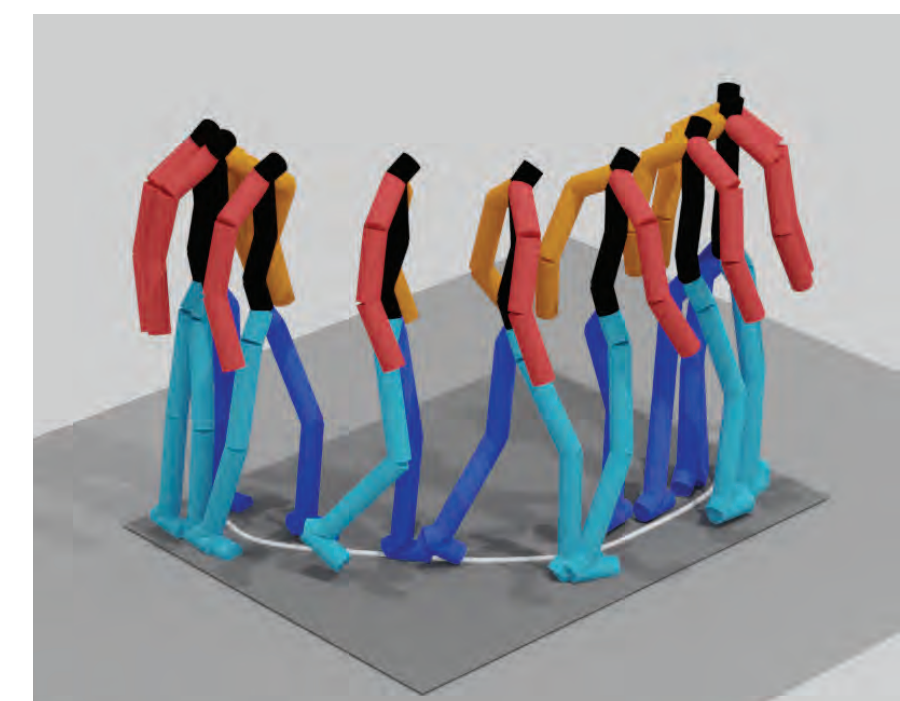
## KIT Motion-Language Dataset

- 3911 motions sequences with 6353 sequence level sentence annotations
- Processed with MMM framework, as in prior work
- Processed with SMPL (correspondance with AMASS)

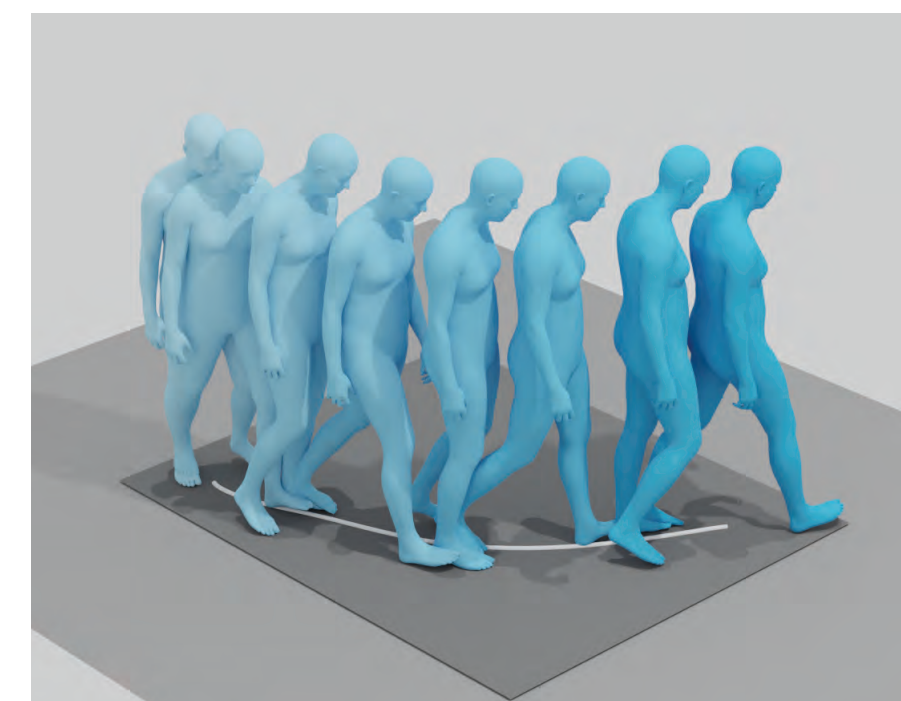
MMM joint coordinates



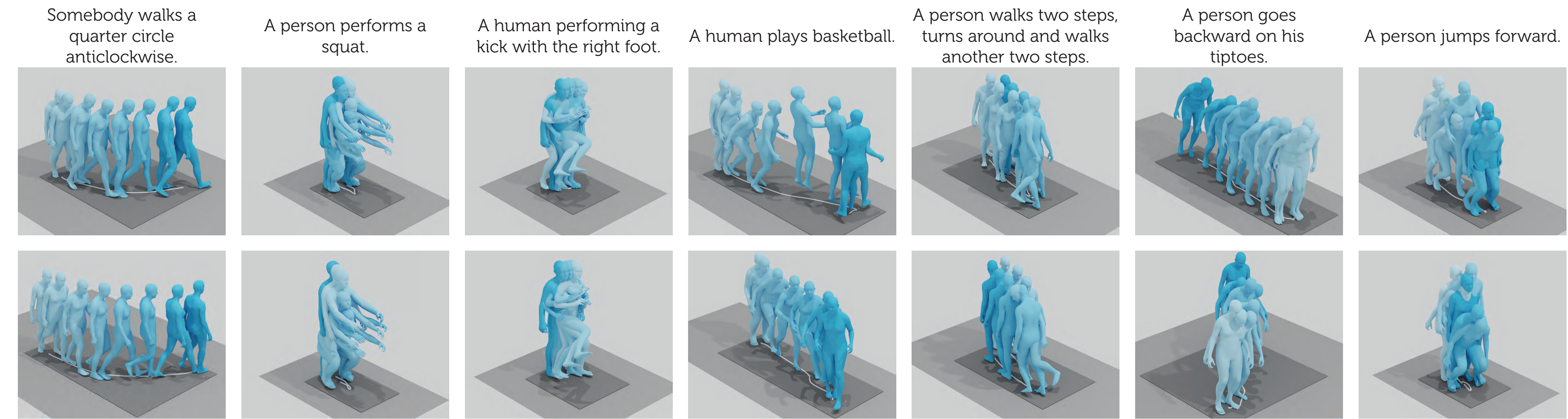
SMPL joint coordinates



SMPL pose parameters



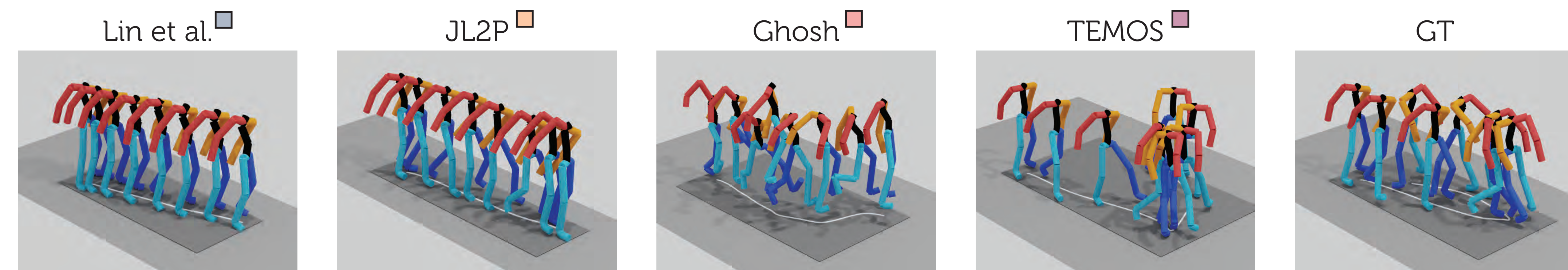
## Qualitative results



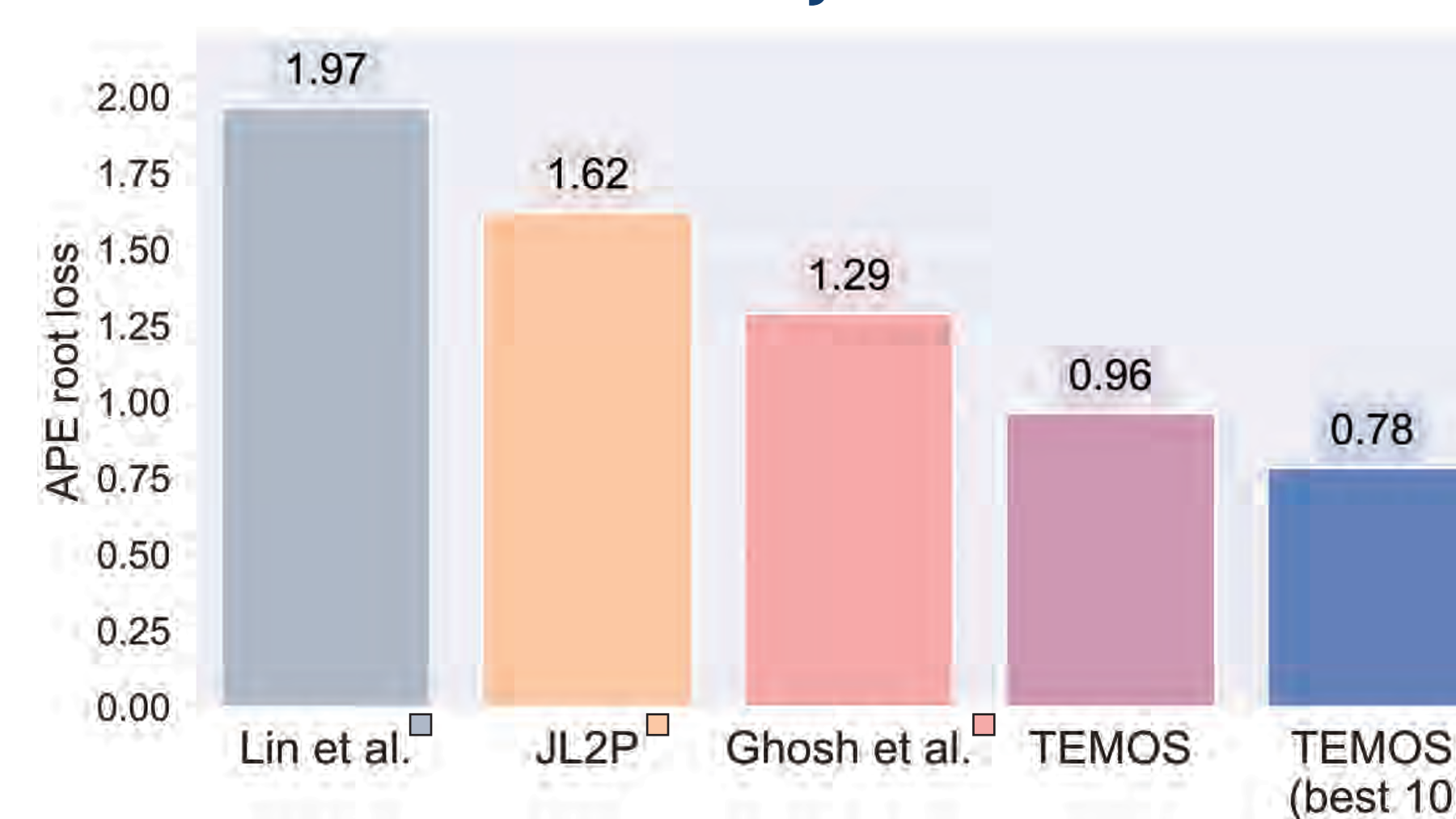
## Ablation study: architecture and losses

Arch.	$\mathcal{L}_{KL}$	$\mathcal{L}_E$	Average Positional Error ↓				Average Variance Error ↓			
			root joint	glob. traj.	mean loc.	mean glob.	root joint	glob. traj.	mean loc.	mean glob.
GRU	$KL(\phi^T, \phi^M) + KL(\phi^M, \phi^T) + KL(\phi^T, \psi) + KL(\phi^M, \psi)$	✓	1.443	1.433	0.105	1.451	0.600	0.599	0.007	0.601
Transf.	$KL(\phi^T, \psi)$ w/out $\mathcal{M}_{enc}$	✗	1.178	1.168	0.106	1.189	0.506	0.505	0.006	0.508
Transf.	$KL(\phi^T, \phi^M) + KL(\phi^M, \phi^T) + KL(\phi^T, \psi) + KL(\phi^M, \psi)$	✗	1.091	1.083	0.107	1.104	0.449	0.448	0.005	0.451
Transf.	$KL(\phi^T, \psi) + KL(\phi^M, \psi)$ w/out cross-modal KL losses	✗	1.080	1.071	0.107	1.095	0.453	0.452	0.005	0.456
Transf.	$KL(\phi^T, \psi) + KL(\phi^M, \psi)$ w/out cross-modal KL losses	✓	0.993	0.983	0.105	1.006	0.461	0.460	0.005	0.463
Transf.	$KL(\phi^T, \phi^M) + KL(\phi^M, \phi^T)$ w/out Gaussian priors	✓	1.049	1.039	0.108	1.065	0.472	0.471	0.005	0.475
Transf.	$KL(\phi^T, \phi^M) + KL(\phi^M, \phi^T) + KL(\phi^T, \psi) + KL(\phi^M, \psi)$	✓	<b>0.963</b>	<b>0.955</b>	<b>0.104</b>	<b>0.976</b>	<b>0.445</b>	<b>0.445</b>	0.005	<b>0.448</b>

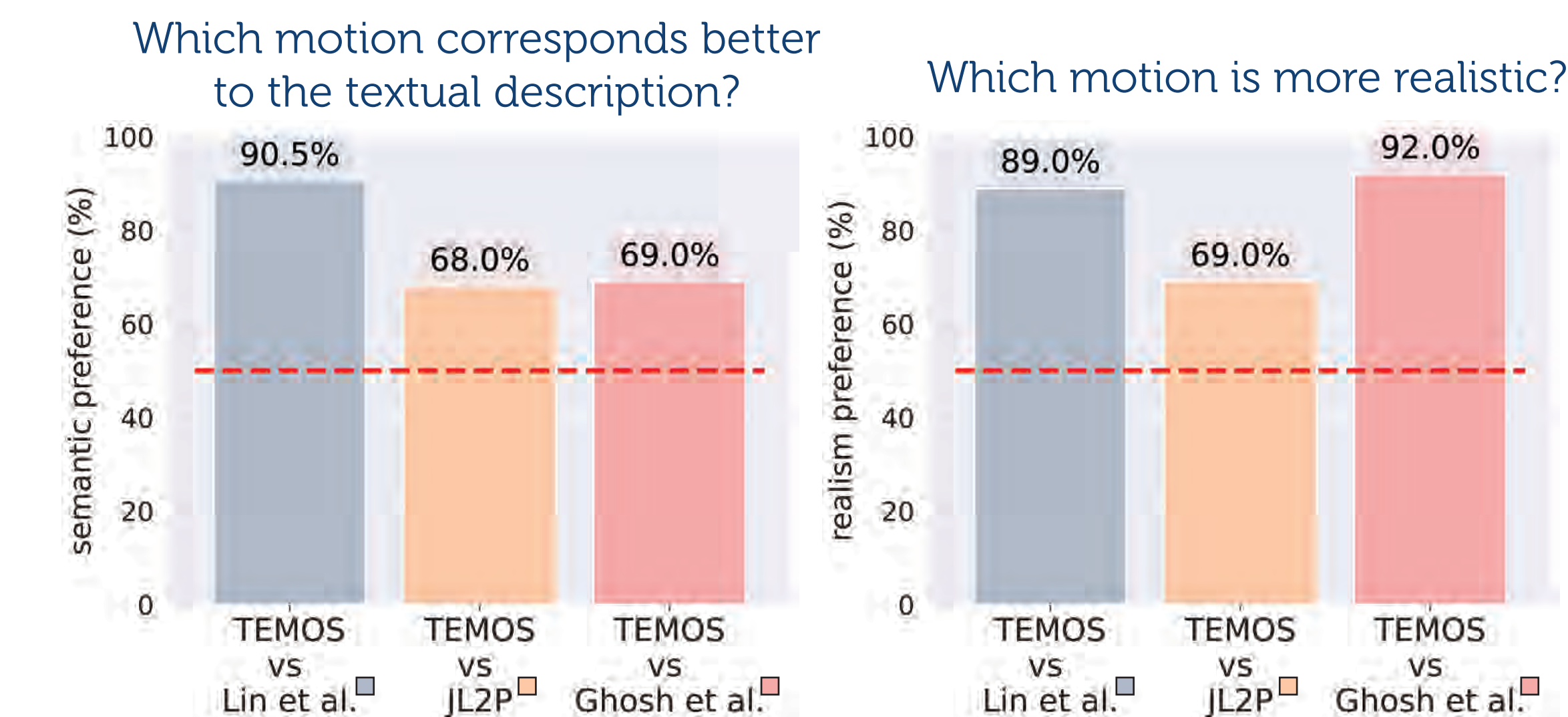
## Comparison with previous work



APE root joint error



Human studies



## References

- Lin et al. Generating animated videos of human activities from natural language descriptions (NeurIPS Workshop 2018)
- Ahuja et al. Language2Pose: Natural language grounded pose forecasting (3DV 2019)
- Ghosh et al. Synthesis of compositional animations from textual descriptions (ICCV 2021)