TEMOS: Generating diverse human motions from textual descriptions

Paper, video, PyTorch code, pretrained models available online

ECCV TEL AVIV 2022

https://mathis.petrovich.fr/temos

Mathis Petrovich[1,2] Michael J. Black[2] Gül Varol[1]

[1]LIGM, École des Ponts, Univ Gustave Eiffel, CNRS, France

[2]Max Planck Institute for Intelligent Systems, Tübingen, Germany

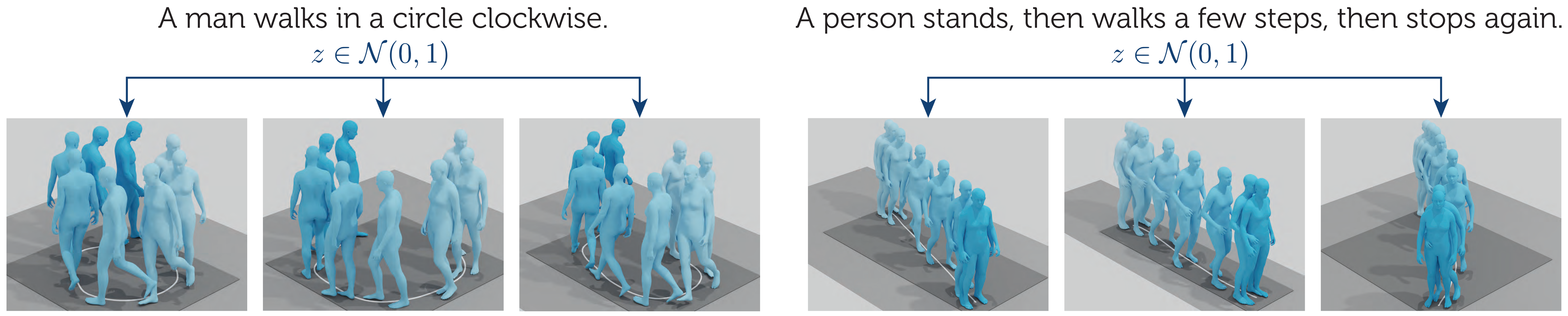
Introduction

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

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

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

Contributions:
- Novel cross-modal variational model that can produce **diverse** motions
- **State-of-the-art performance**
- **Extensive** ablation study



TEMOS: TExt-to-MOtionS

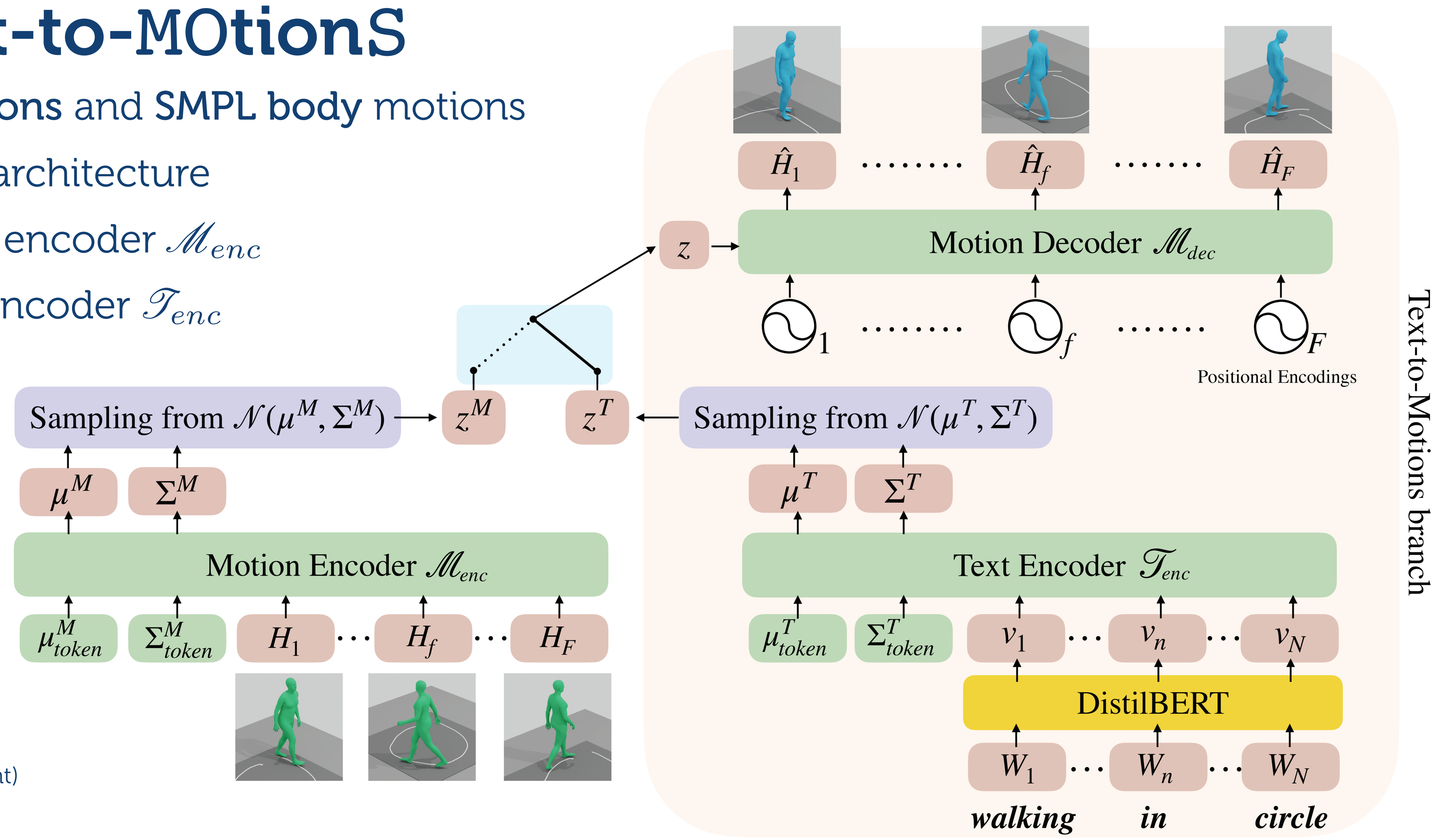
- Supports both **skeletons** and **SMPL body** motions
- Non-autoregressive architecture
- *Motion-level* motion encoder $\mathcal{M}_{enc}$
- *Sentence-level* text encoder $\mathcal{T}_{enc}$

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

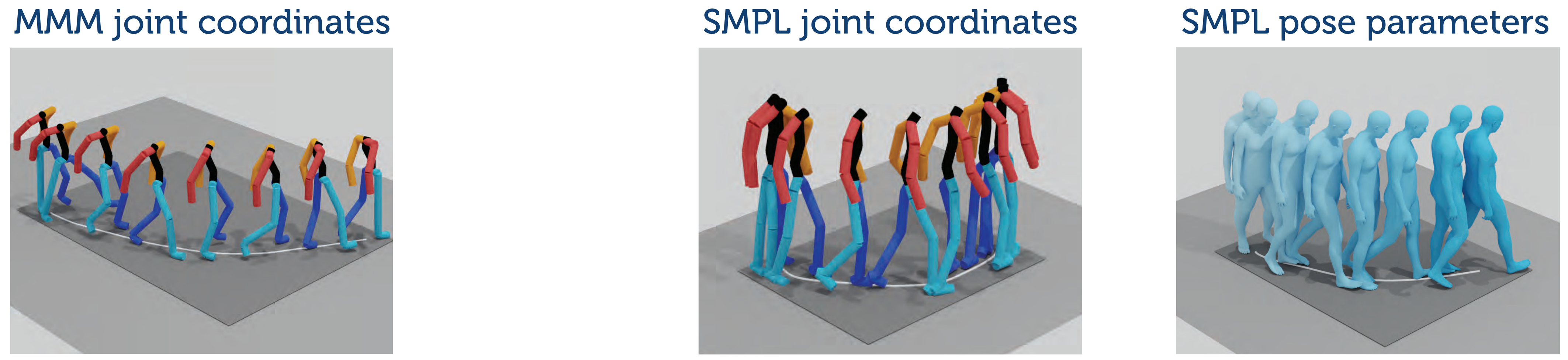
Inference

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

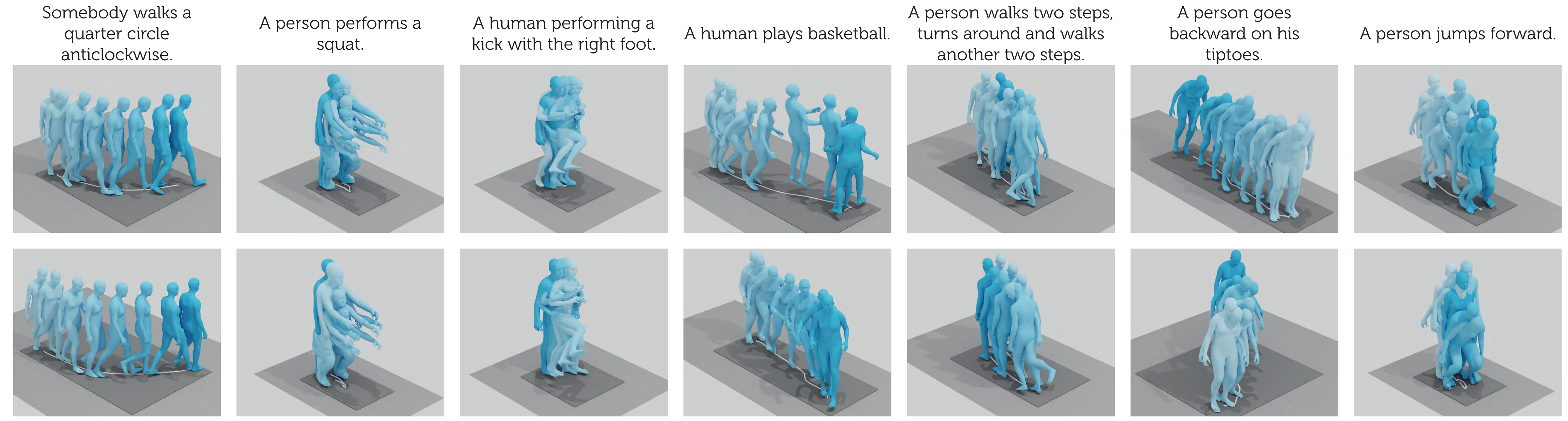


KIT Motion-Language Dataset

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



Qualitative results

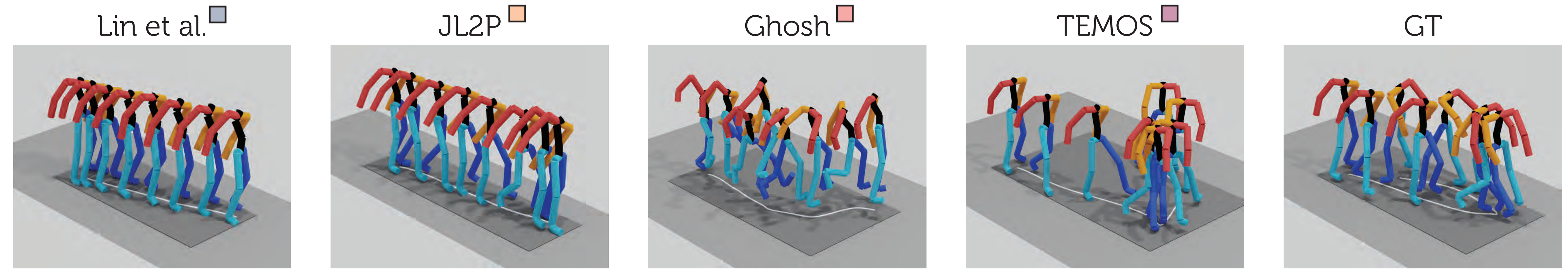


Ablation study: architecture and losses

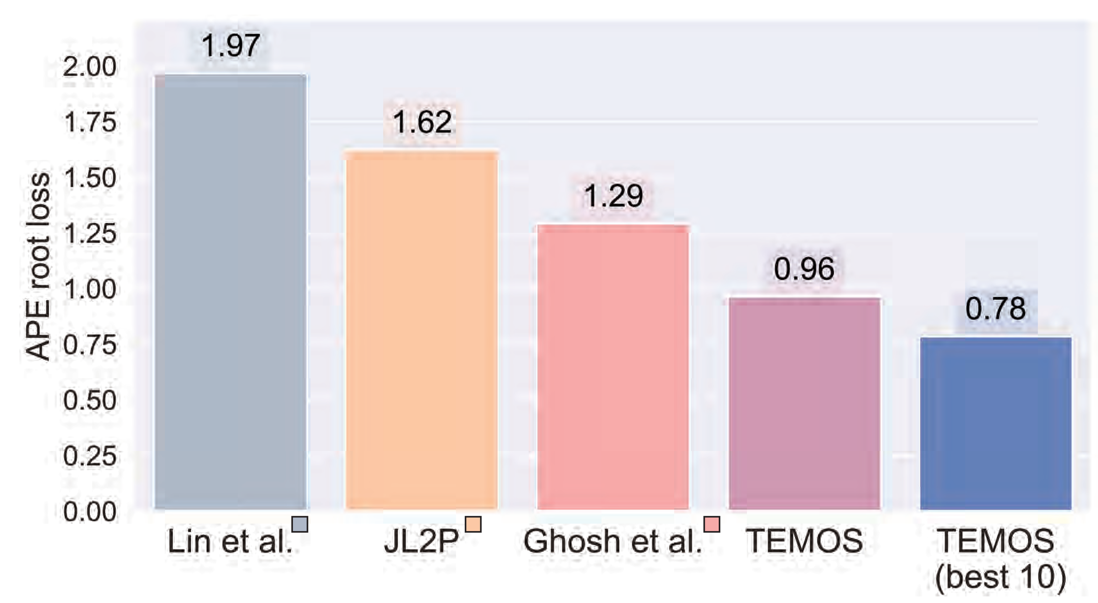
Arch.	$\mathcal{L}_{KL}$	$\mathcal{L}_E$	Average Positional Error ↓ root joint	glob. traj.	mean loc.	mean glob.	Average Variance Error ↓ root joint	glob. traj.	mean loc.	mean glob.
GRU	$KL(\phi^T,\phi^M)+KL(\phi^M,\phi^T)+KL(\phi^T,\psi)+KL(\phi^M,\psi)$	✓	1.443	1.433	0.105	1.451	0.600	0.599	0.007	0.601
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.	$KL(\phi^T,\phi^M)+KL(\phi^M,\phi^T)+KL(\phi^T,\psi)+KL(\phi^M,\psi)$	✓	**0.963**	**0.955**	**0.104**	**0.976**	**0.445**	**0.445**	0.005	**0.448**

Comparison with previous work

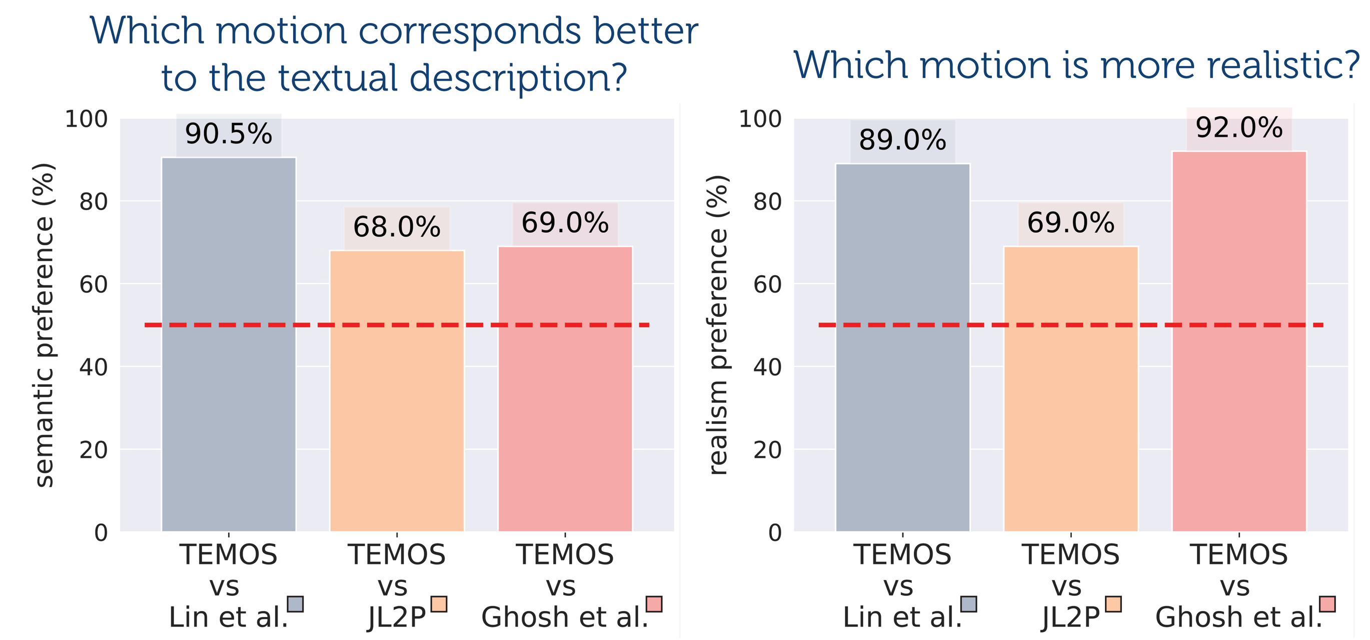
A person walks two steps, turns around and walks another two steps.



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)