Synthesis of Compositional Animations from Textual Descriptions

Anindita Ghosh*[1,3], Noshaba Cheema[1,2,3], Cennet Oguz[1,3], Christian Theobalt[2,3], Philipp Slusallek[1,3]

[1]DFKI, [2]Max-Planck Institute for Informatics, [3]Saarland Informatics Campus

*Corresponding Author: anindita.ghosh@dfki.de

2021 ICCV OCTOBER 11-17 VIRTUAL

ABSTRACT

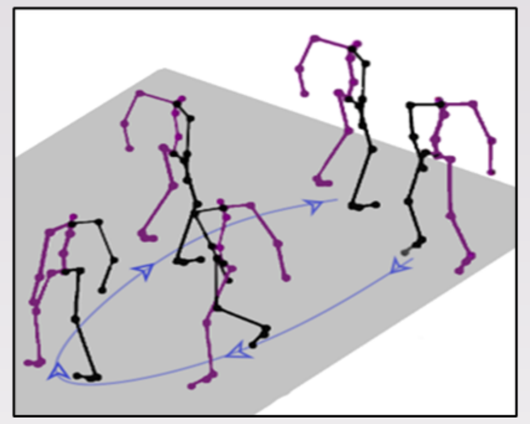
We present a learning-based method for generating animated 3D pose sequences depicting multiple sequential or superimposed actions provided in long, compositional sentences. We propose a hierarchical two-stream sequential model to explore a finer joint-level mapping between natural language sentences and 3D pose sequences corresponding to the given motion . We evaluate our proposed model on the KIT Motion-Language Dataset containing 3D pose data with human-annotated sentences. We show that our model advances the state-of-the-art on text-based motion synthesis in objective evaluations by a margin of 50%.

Input:

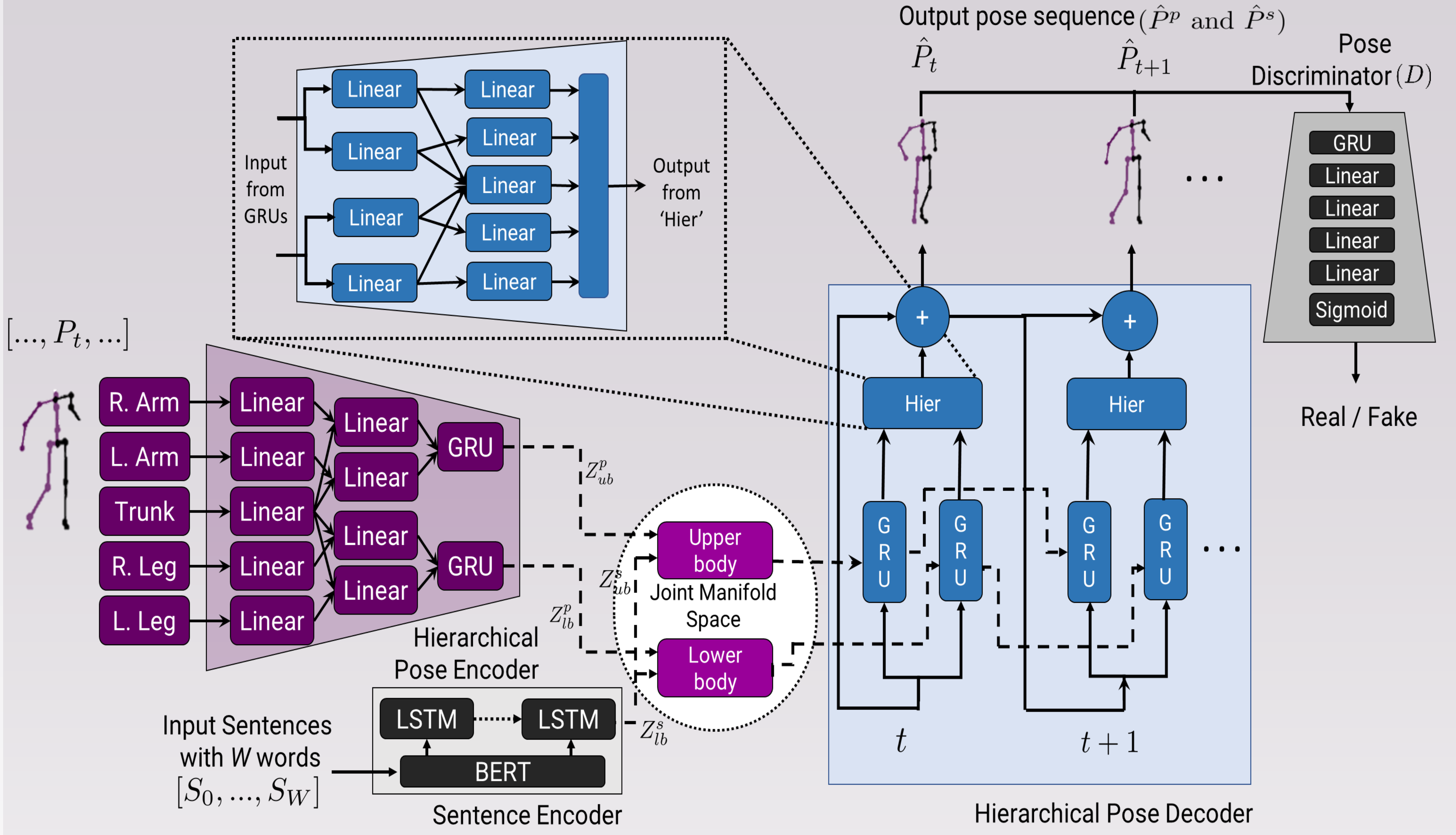
"A human walks in a clockwise circle."

→

Output:



NETWORK ARCHITECTURE AND TRAINING



Objective. Minimize the Pose Prediction loss (L_R), the Embedding Similarity loss (L_E), Velocity Reconstruction loss (L_V) and Adversarial loss (L_D, L_G) :

$$L_R = \mathcal{L}(\hat{P}^s, P) + \mathcal{L}(\hat{P}^p, P)$$

$$L_E = \mathcal{L}(Z^p_{ub}, Z^s_{ub}) + \mathcal{L}(Z^p_{lb}, Z^s_{lb})$$

$$L_V = \mathcal{L}(\hat{P}_{vel}, P_{vel})$$

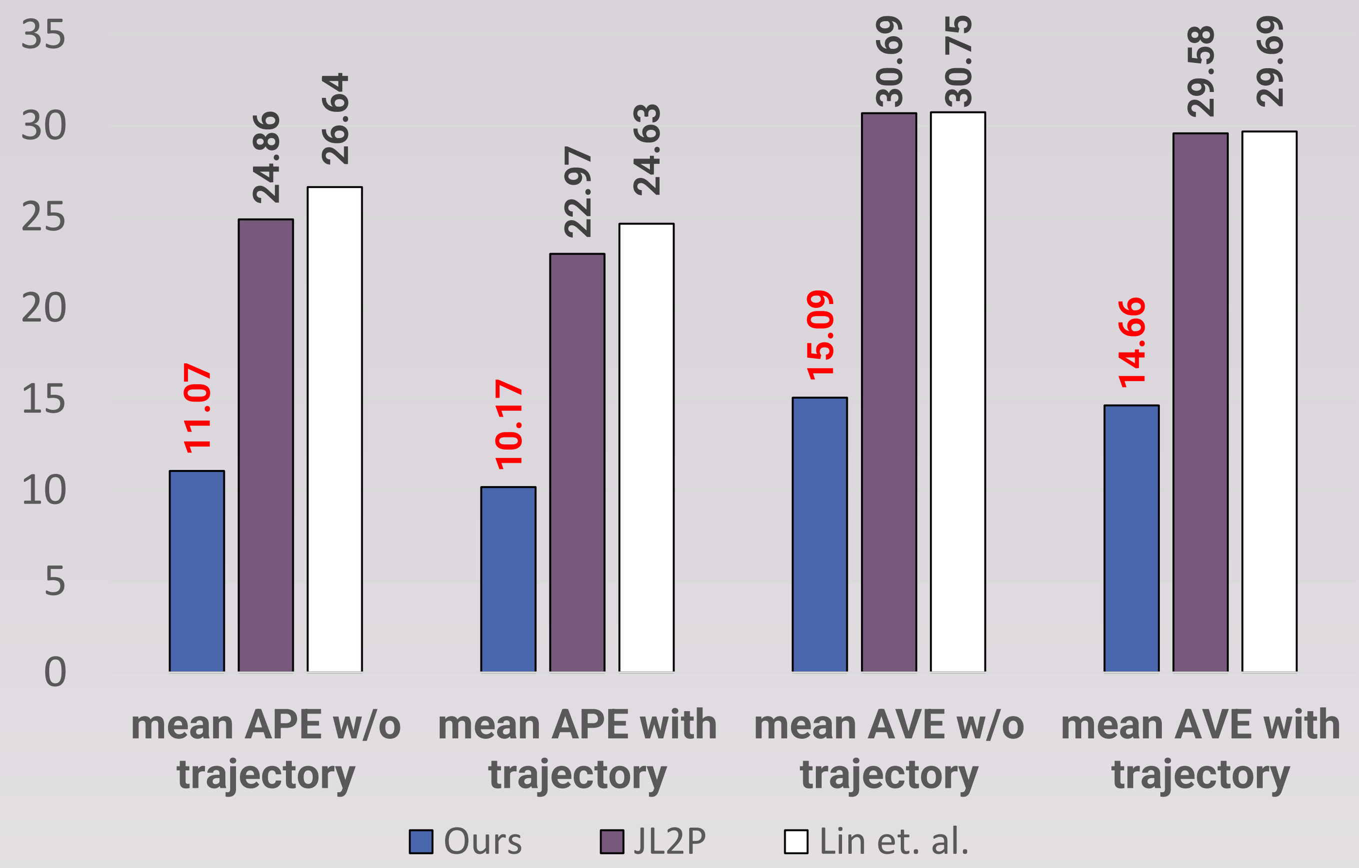
$$L_G = \mathcal{L}_2(D(\hat{P}), 0) + \mathcal{L}_2(D(P), 1)$$

$$L_D = \mathcal{L}_2(D(\hat{P}), 1)$$

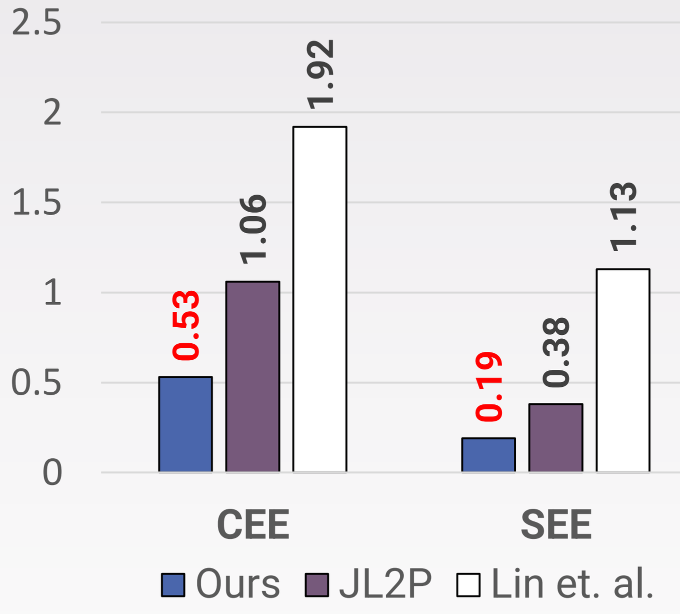
$\mathcal{L}$: Smooth L_1 loss

$\mathcal{L}_2$: Binary Cross Entropy loss

QUANTITATIVE RESULTS



Our method shows more than 50% improvement on the mean Average Positional Error (APE) and the mean Average Variance Error (AVE) of joint positions over the state-of-the-art methods of JL2P and Lin et al.

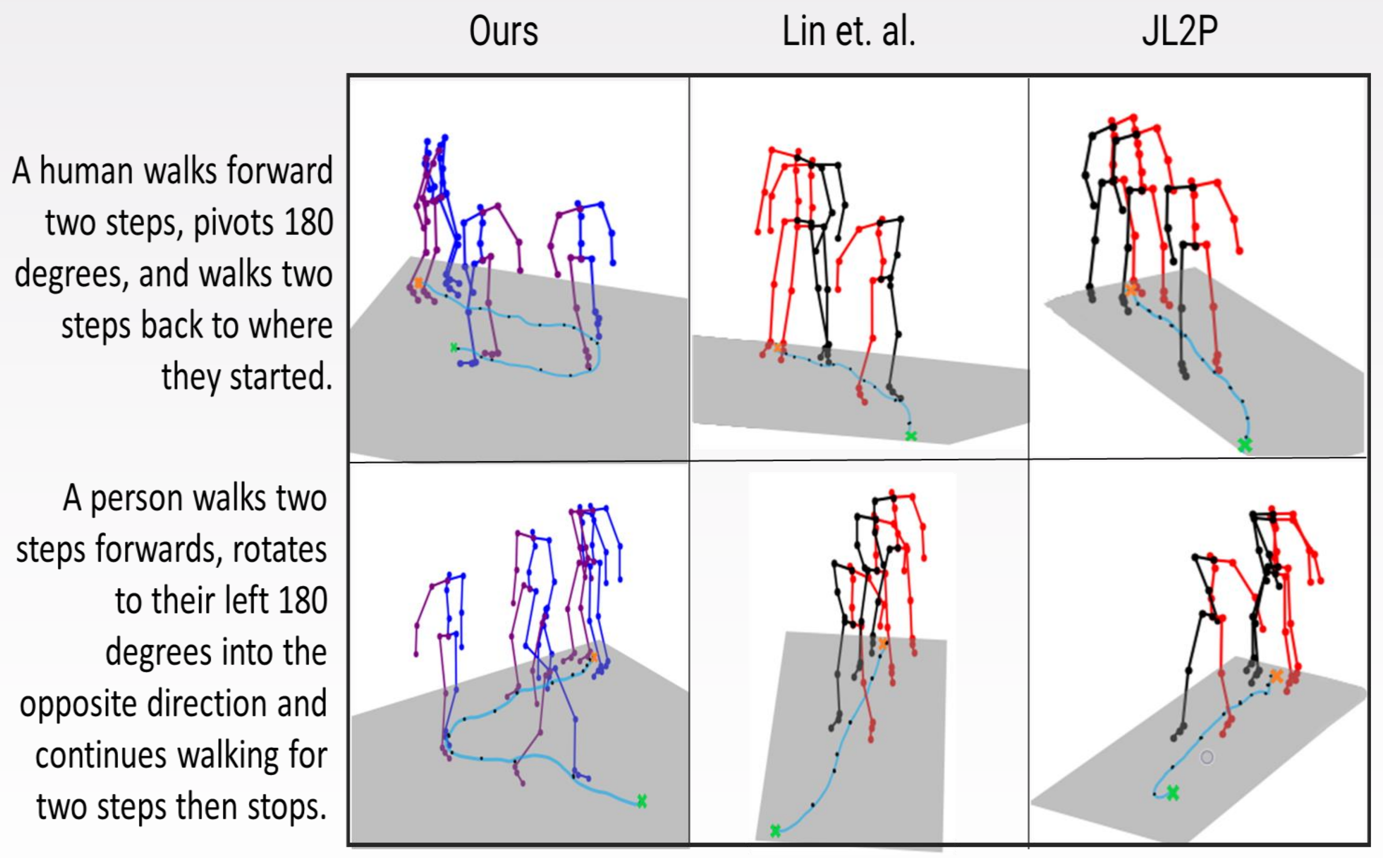


The joint embedding space learned by our method can correlate poses and corresponding sentences better than JL2P and Lin et al. as seen from the Content Encoding Error (CEE) and the Style Encoding Error (SEE) metrics.

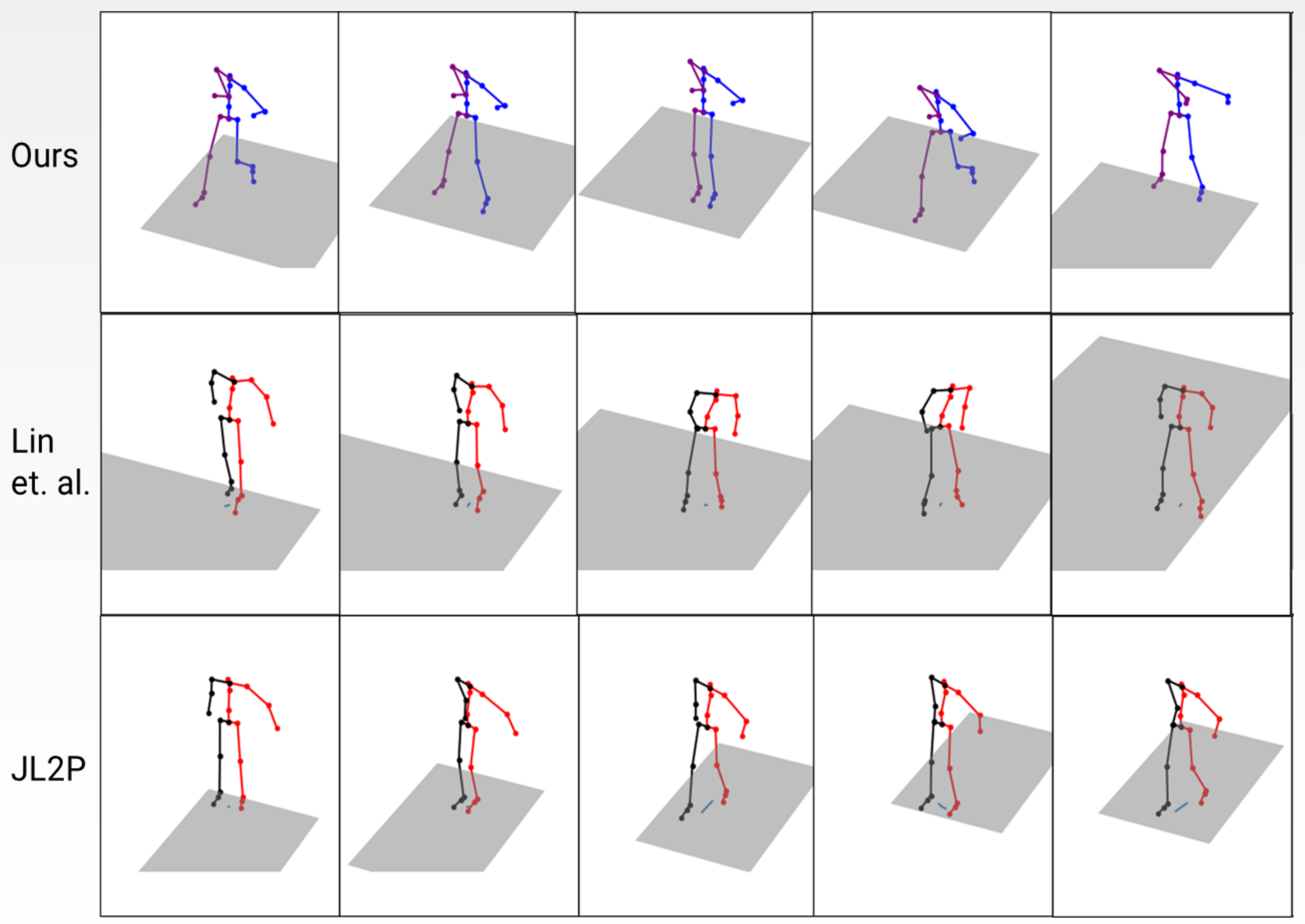
APPROACH

- We introduce a hierarchical joint embedding space that learns the embeddings of pose and language simultaneously.
- We separate our intermediate pose embeddings hierarchically to limb embeddings such that our model learns features from the different components of the body.
- We have a two-stream sequential network to separately learn the upper and the lower body movements and focus on the end joints of the body.
- We use contextualized BERT embeddings with handpicked word feature embeddings to improve text understanding.
- We further use additional loss terms and a pose discriminator to further improve the plausibility of the synthesized motion.

QUALITATIVE RESULTS



Our method accurately synthesizes a trajectory that matches the semantics of a given sentence (top) and waltz dance motion (right) compared to the benchmark methods of Lin et al. and JL2P.



A human performs the steps of a waltz dance while it is holding its hands like it is leading a partner with its hands.

This research is funded by the BMBF grants XAINES (01|W20005) and IMPRESS (01|S20076), EU Horizon 2020 grant Carousel+ (101017779) and an IMPRS-CS Fellowship. Computational resources provided by the BMWi grants 01MK20004D and 01MD19001B.

CODE AND RESOURCES:
https://github.com/anindita127/Complextext2animation