PoseFix: Correcting 3D Human Poses with Natural Language

- Ginger Delmas, Philippe Weinzaepfel, Francesc Moreno-Noguer, Gregory Rogez
- NAVER LABS Europe

Prakash K Naikade

Problem

- Correctional Text Generative Model
 - **Given:** 3D Source pose and 3D Target pose
 - **Predict:** Correctional text generation

- Text-based Pose Editing Model
 - Given: 3D Target pose and Textual Prompt
 - Predict: 3D Target pose

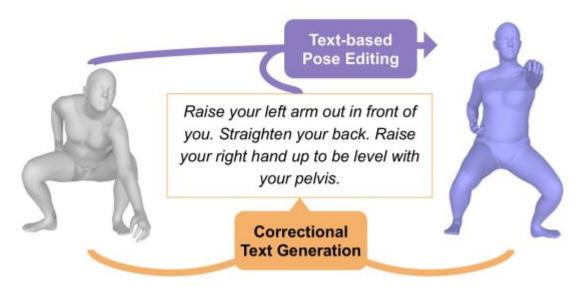
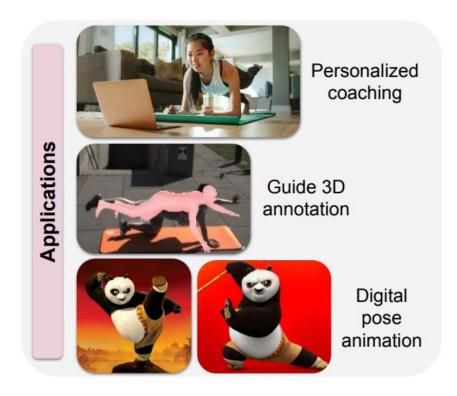


Fig 1: Problem description - Pose Fix

Motivation



- Sports Coaching
- Rehabilitation
- Labour Training
- Safety Instruction while handling dangerous machines
- Animations

Fig 2: Applications of PoseFix Models

Motivation



Join your hands in front of your chest.

Language-guided robot control

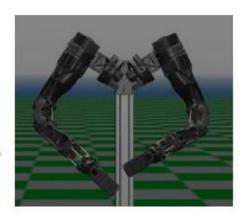


Fig 2a: Robot teaching application

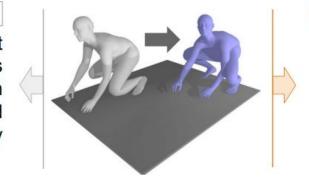
Key Contribution

- 1. **PoseFix dataset:** 3D pose pairs with modifying instructions
- 2. Text-based Pose Editing Model
- 3. Correctional Text Generative Model

- {pose A, pose B, text modifier}
- 135k pose pairs
- 3D human body poses were sampled from AMASS
- 6k human-written texts

Data collection on AMT

Stretch your thighs apart and project the knees forward so that they remain just along the elbows and then turn your face slightly to the left.



Automatic Comparative Pipeline

Your right thigh must be parallel to the floor while your right knee is bent to maximum, bring your right foot forward slightly, your right hand must be on the floor and your hands should be shoulder width apart.

Fig 3 - PoseFix Dataset

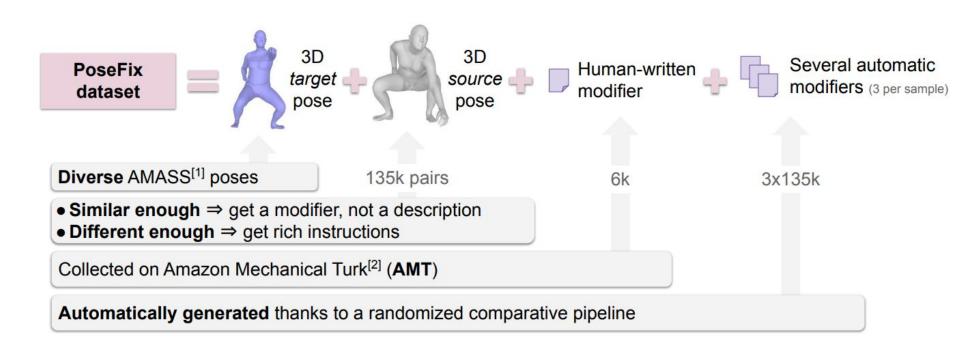


Fig 4 - PoseFix Dataset Generation

- Randomized comparative pipeline for automatic generation of modifiers:
 - more training data at no cost: generate >10k modifiers in the time it takes to write 1!

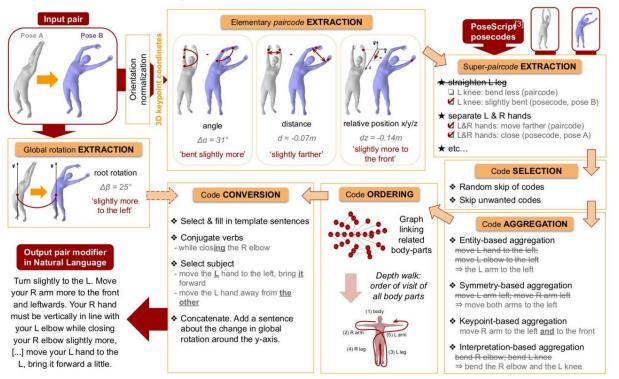
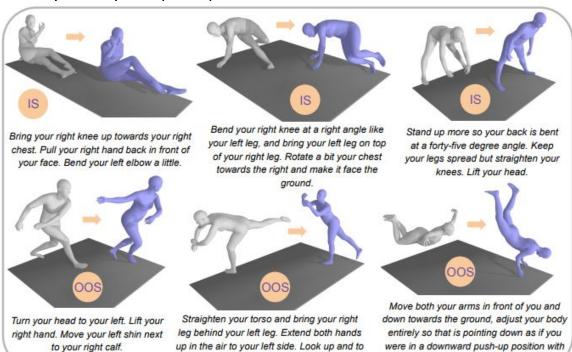


Fig 5 - PoseFix Dataset Generation - Randomized comparative pipeline

- In-sequence pairs (IS)
- Out-of-sequence pairs (OOS)



vour left.

Automatically generated

Your right thigh must be parallel to the floor while your right knee is bent to maximum, bring your right foot forward slightly, your right hand must be on the floor and your hands should be shoulder width apart.

Automatically generated

Bring the right arm, the left leg and the left hand up, bring the right hand slightly to the left and move both hands backward a little, stretch the left arm, the left thigh must be parallel to the ground. Bring the left and a bit to the back.

Fig 6 - Examples of pose pairs and their annotated modifier in PoseFix

your legs elevated towards the sky.



Left: Data presented to the annotators. The slider makes it possible to look at the poses under different viewpoints.

Right: word cloud of the PoseFix annotations.

Fig 7 - Annotators

Text-based Pose Editing

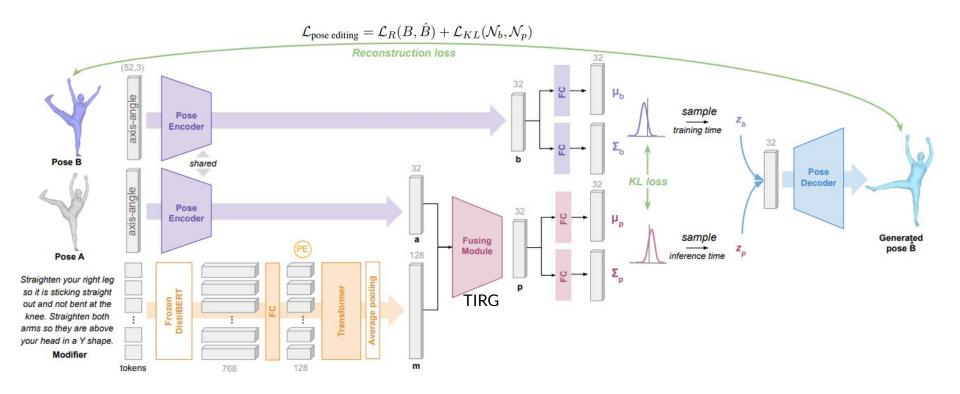


Fig 8: Overview of our text-based pose editing baseline

Text-based Pose Editing

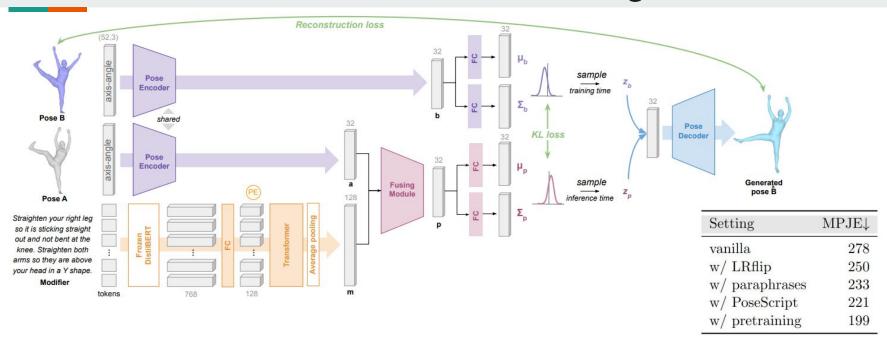


Fig 8: Overview & Evaluation -The top part represents a standard VAE, where poses are encoded into a Gaussian distribution. At training time, a latent variable is sampled and decoded into a pose to learn pose reconstruction. The bottom left part represents the conditioning: the text is encoded using a frozen DistilBERT with a small transformer on top. It is combined with source pose features in the fusion module, from which we predict a Gaussian distribution. A KL loss ensures the alignment of the distributions from the standard VAE and the conditioning. At test time, we sample from the latter to predict the target pose.

Text-based Pose Editing

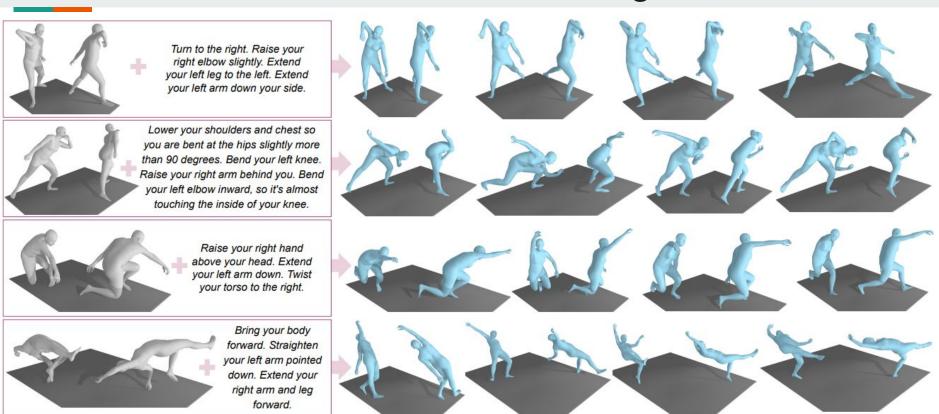


Fig 9: Generated poses for the text-based pose editing task

Correctional Text Generation

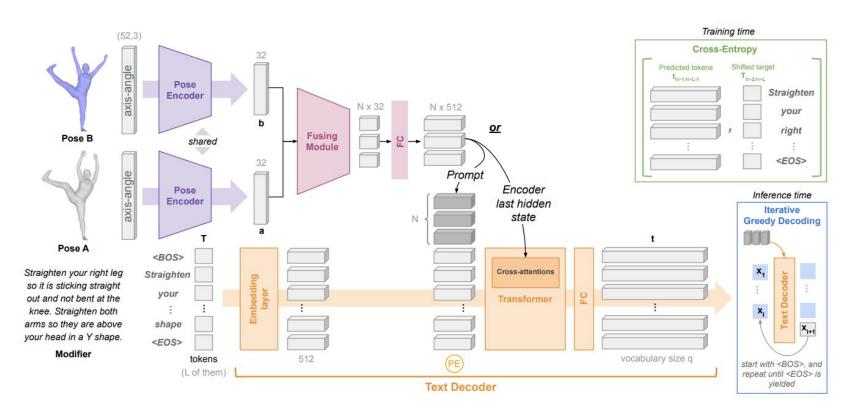


Fig 10: Overview of our baseline for correctional text generation

Correctional Text Generation

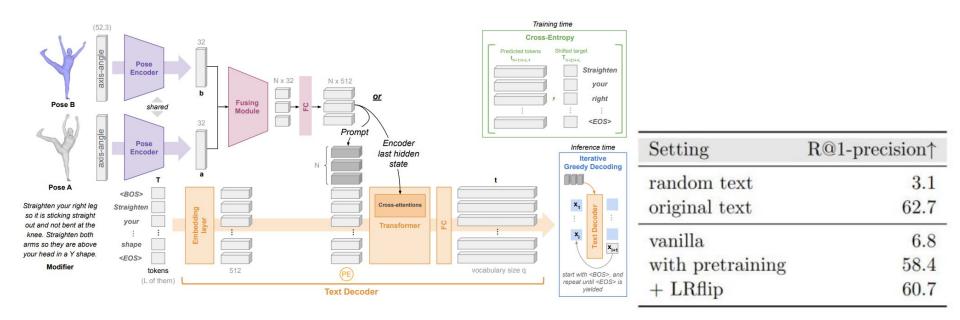


Fig 11: Overview & Evaluation: The bottom part represents a standard auto-regressive transformer model: the next word is predicted from the previously generated tokens. The decoder outputs a distribution of probabilities over the vocabulary for each token. The top part represents the conditioning on the pose pair: the two pose embeddings are fused together into a set of "pose tokens", further used for conditioning via prompting or via cross-attentions in the transformer. At inference, the modifier is generated iteratively using the greedy approach.

Correctional Text Generation



Bend over more. Move your right arm down. Move your left arm to the right.

Move your right leg to the left and bend your right knee slightly. Move your left leg back and to the right. Turn your head to the right. Move your right arm to the right and bend your right elbow. Move your left hand to the right.

Bend your elbows and move your hands closer to each other. Turn your head to the right.

Straighten your legs and lean back. Lower your arms to your chest.

Raise your left leg and extend it out to the side.
Turn your head to the left.

Lean back and to the left. Lower your arms to your sides. Turn your head to the right.

Fig 12: Generated correctional texts for PoseFix pose pairs

Challenges & Limitations

Correctional Text Generation Text-Based Pose Editing

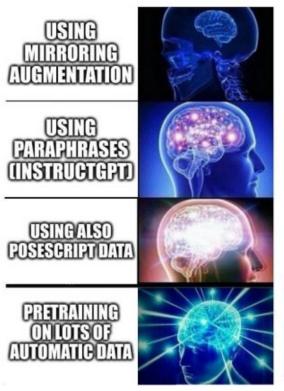
Mixes up the source and the target pose

Hallucinates some differences Challenging cases (upside-down, with contact...)

Ignores some differences/in structions

Fails to preserve some traits of the source pose

Conclusions





Training with several texts for each pose

Training with only 1 text/pose but more poses





Thank You!

Questions Please!