ICCV23 PARIS

paper ● code ● data

PoseFix: Correcting 3D Human Poses with Natural Language

Ginger Delmas[1,2], Philippe Weinzaepfel[1], Francesc Moreno-Noguer[2], Grégory Rogez[1]

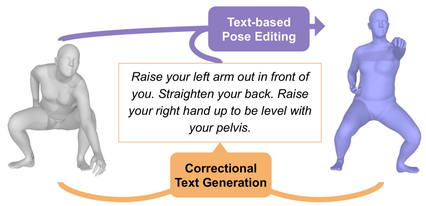
[1]NAVER LABS Europe [2]IRI CSIC-UPC

Institut de Robòtica i Informàtica Industrial

NAVER LABS Europe

Motivation & Contributions

Using text to learn fine-grained 3D body pose semantics.



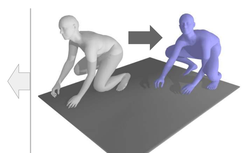
- **PoseFix dataset**: 3D pose pairs with modifying instructions
- **Text-based Pose Editing** model
- **Correctional Text Generative** model

The PoseFix dataset

135k pose pairs, 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.

PoseFix dataset = 3D *target* pose + 3D *source* pose + Human-written modifier + Several automatic modifiers (3 per sample)

135k pairs | 6k | 3x135k

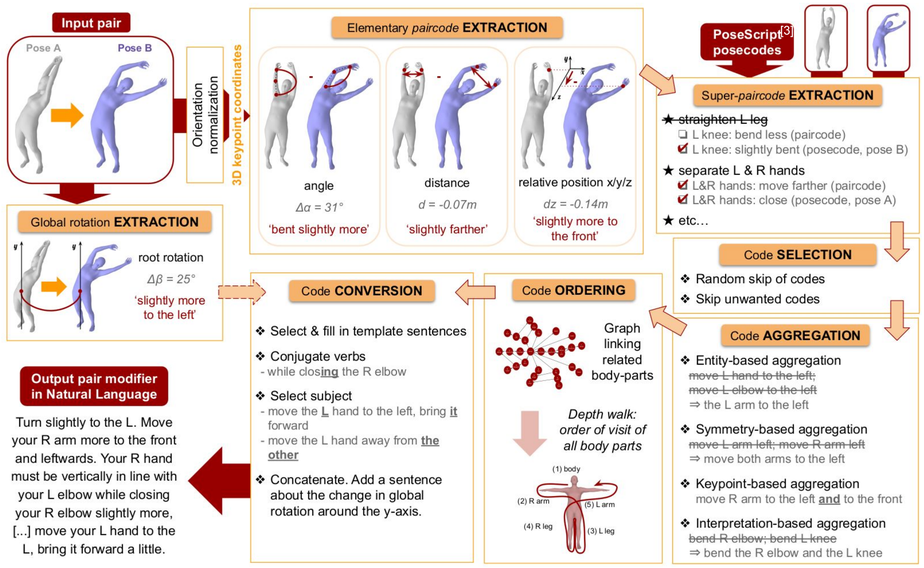
Diverse AMASS[1] poses

- **Similar enough** ⇒ get a modifier, not a description
- **Different enough** ⇒ get rich instructions

Collected on Amazon Mechanical Turk[2] (**AMT**)

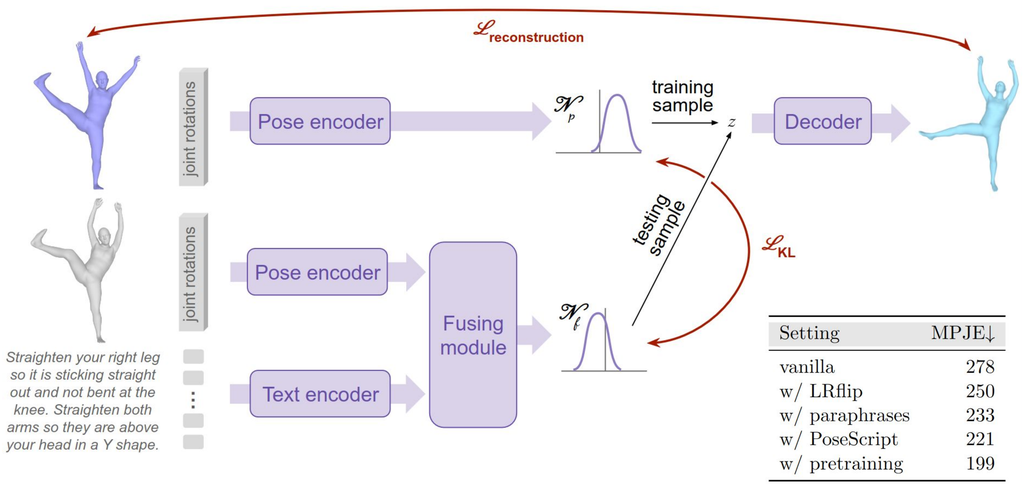
Automatically generated thanks to a randomized comparative pipeline

Get more training data at no cost: generate >10k modifiers in the time it takes to write 1!

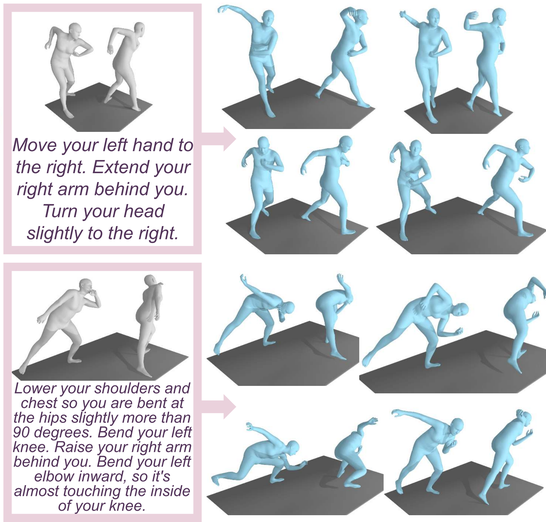


Text-based Pose Editing

⇒ **Conditional VAE**[4]

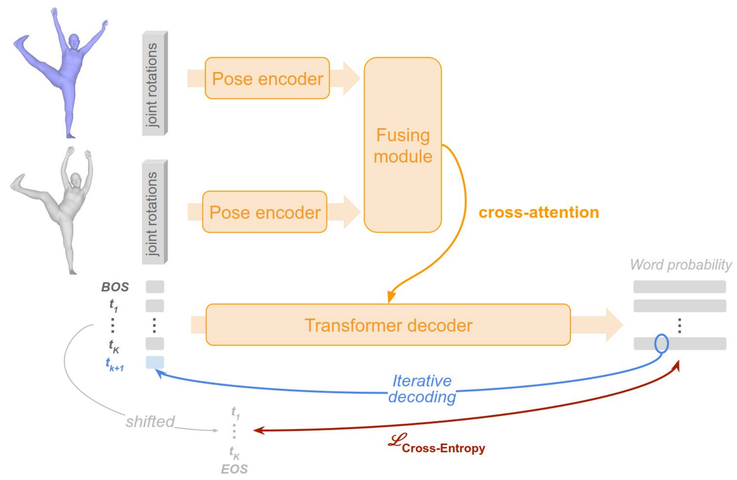


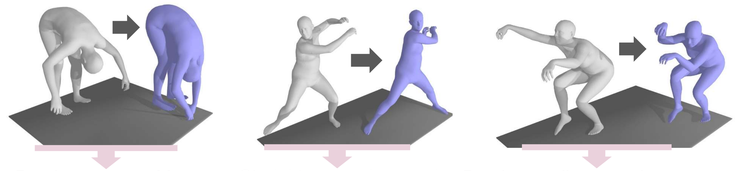
Setting	MPJE↓
vanilla	278
w/ LRflip	250
w/ paraphrases	233
w/ PoseScript	221
w/ pretraining	199



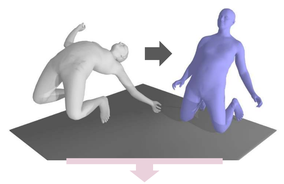
Correctional Text Generation

⇒**Auto-regressive model**

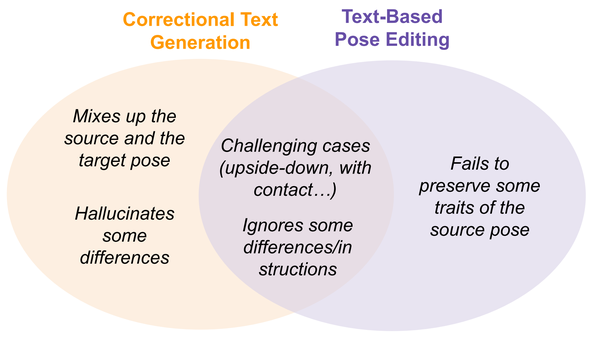




Setting	R@1-precision↑
random text	3.1
original text	62.7
vanilla	6.8
with pretraining	58.4
+ LRflip	60.7

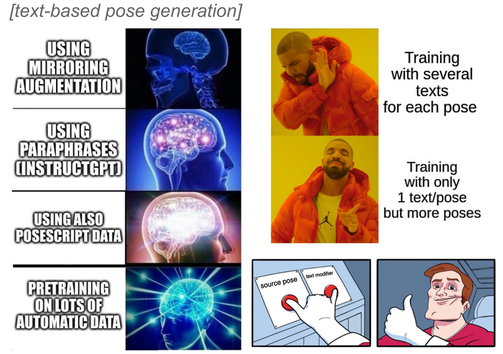


Challenges & limitations



Some conclusions

[text-based pose generation]



References

[1] AMASS: Archive of Motion Capture as Surface Shapes, Mahmood et al., ICCV 2019

[2] https://www.mturk.com/

[3] PoseScript: : 3D Human Poses from Natural Language, Delmas et al., ECCV 2022

[4] Auto-encoding variational bayes, Kingma and Welling, ICLR, 2014

[5] Training language models to follow instructions with human feedback, Ouyang et al., NeurIPS 2022