# Generating Human Interaction Motions in Scenes with Text Control

Hongwei Yi Justus Thies Michael J. Black Xue Bin Peng Davis Rempe

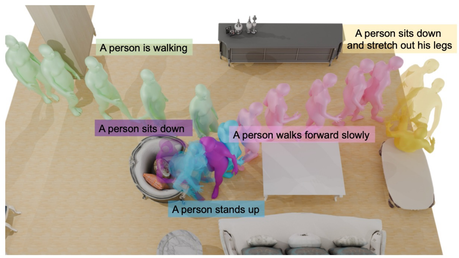
NVIDIA Max Planck Institute for Intelligent Systems, Tübingen, Germany
Technical University of Darmstadt Simon Fraser University

PERCEIVING SYSTEMS
MAX PLANCK INSTITUTE FOR INTELLIGENT SYSTEMS

3D Graphics & Vision
TECHNICAL UNIVERSITY OF DARMSTADT

## Goal

Generate a human motion with text control in 3D scenes.




## Limitation of Prior Work

1. Text to motion without scenes [1,2,3]
2. Scene-aware motion generation w./o. text control [4,5].

## Key Contribution

Fine-tuning an scene augmented model on a pre-trained text-to-motion diffusion model.

Motion Dataset with Text Annotation

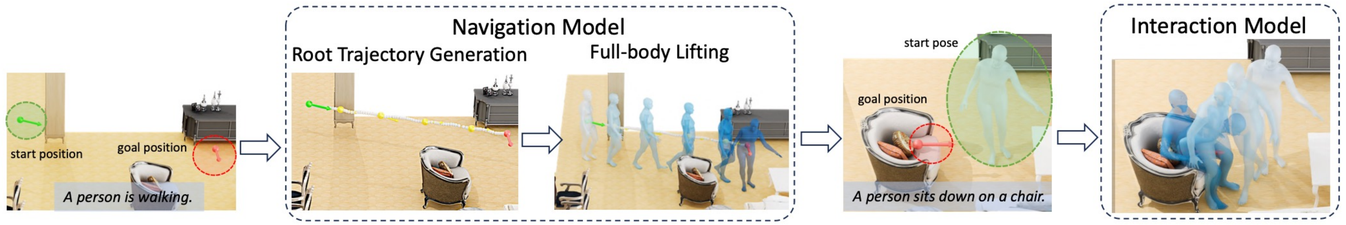
**Controllable Scene-agnostic** Pretrained Text-to-Motion Model

**Finetune**

3D Scenes or Objects

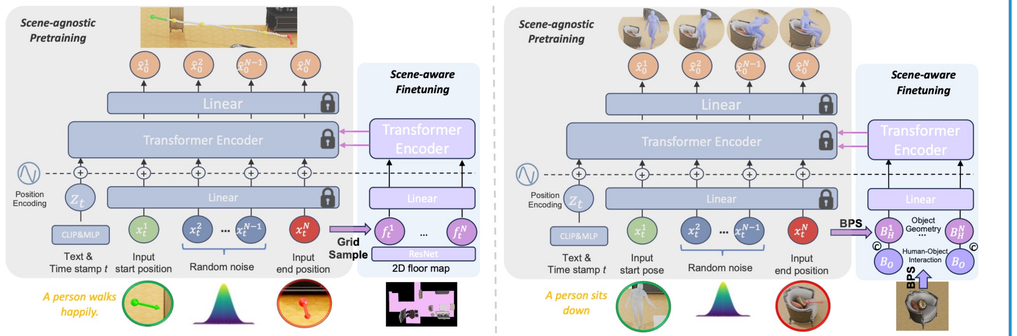
**Scene-aware** Finetuning

## Pipeline



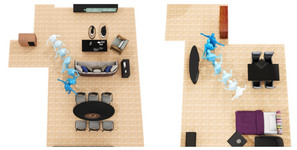

Navigation Input | Generated Navigation Motion | Interaction Input | Generated Interaction

## Architecture

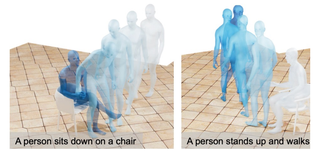



a) Root trajectory model | b) Interaction model

## Data Creation

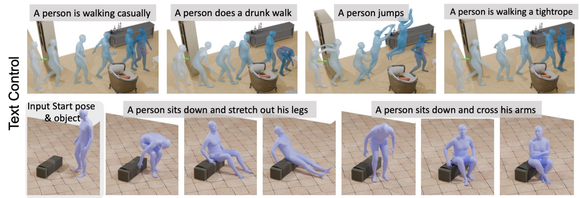


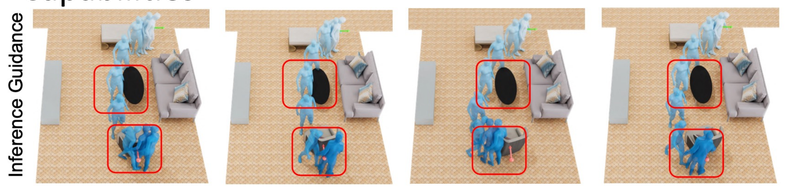
a) Loco-3D-Front: locomotion in different rooms



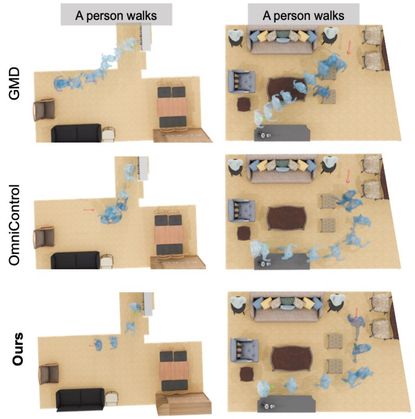

b) Interaction with different objects and text descr

## Capabilities

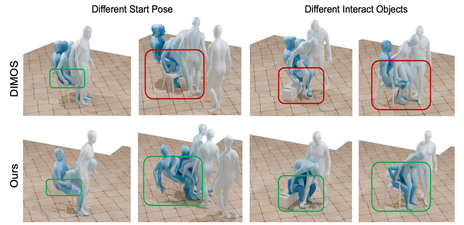



## Locomotion Results




## Interaction Results




## Reference

[1] Tevet et.al., Human motion diffusion model. ECCV20.
[2] Shafir et.al., Human motion diffusion as a generative prior. ICLR24.
[3] Xie et.al., Omnicontrol: Control any joint at any time for human motion generation. ICLR24.
[4] Hassan et.al., Stochastic sceneaware motion prediction. ICCV21.
[5] Zhao et.al, Synthesizing diverse hu- man motions in 3d indoor scenes. ICCV23.

https://research.nvidia.com/labs/toronto-ai/tesmo/