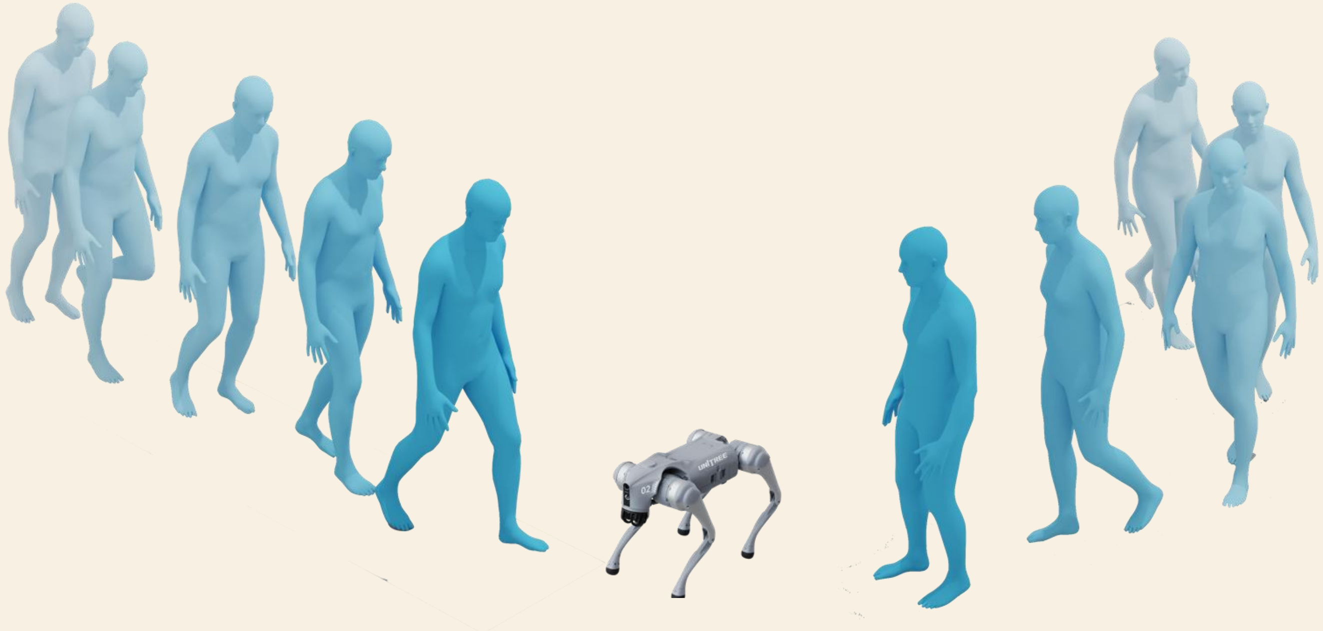
# Human Motion Generation: A Pathway to Safer, Physical Intelligence



Indro Spinelli

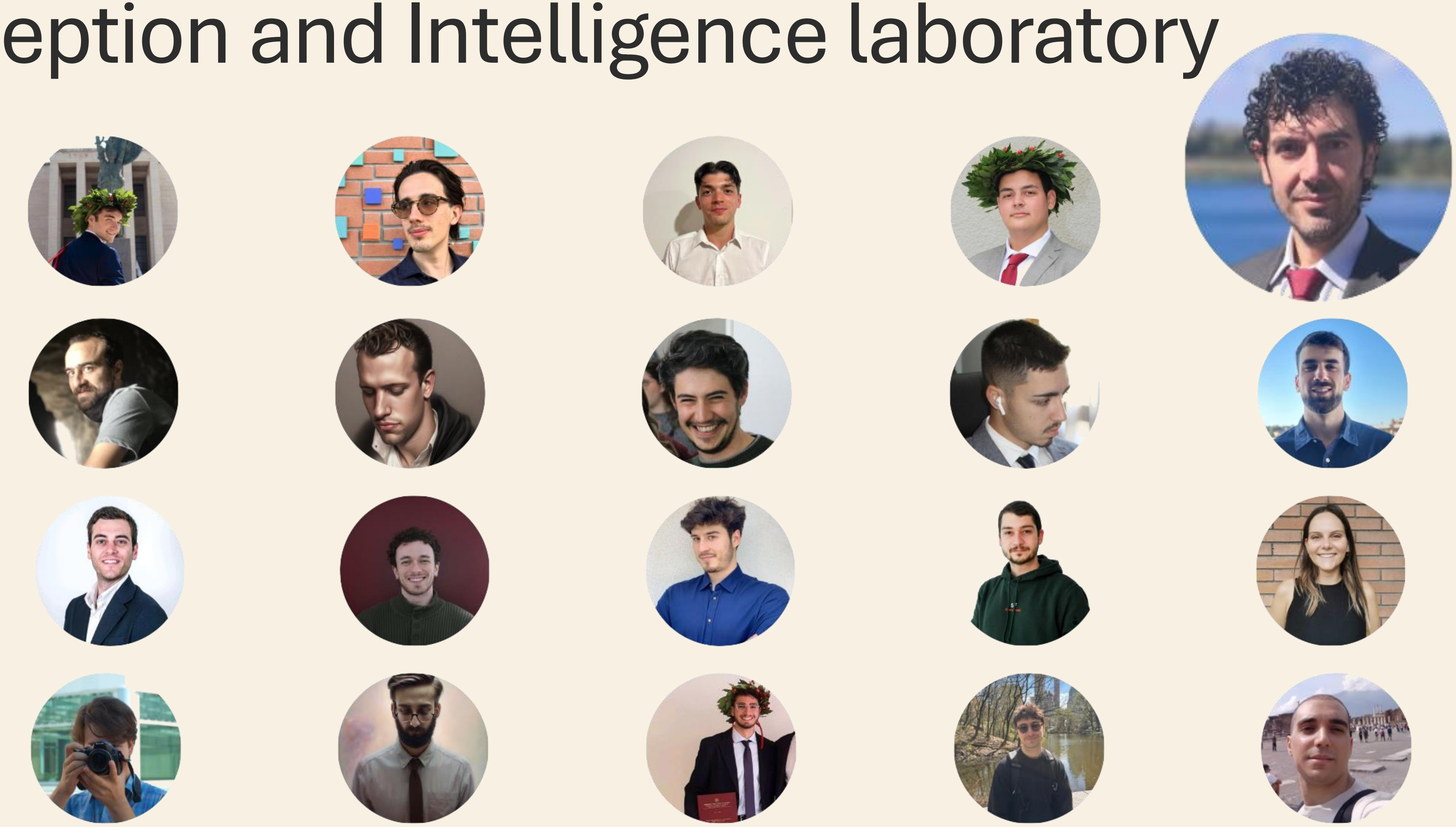
Assistant Professor

Sapienza University of Rome

PINlab

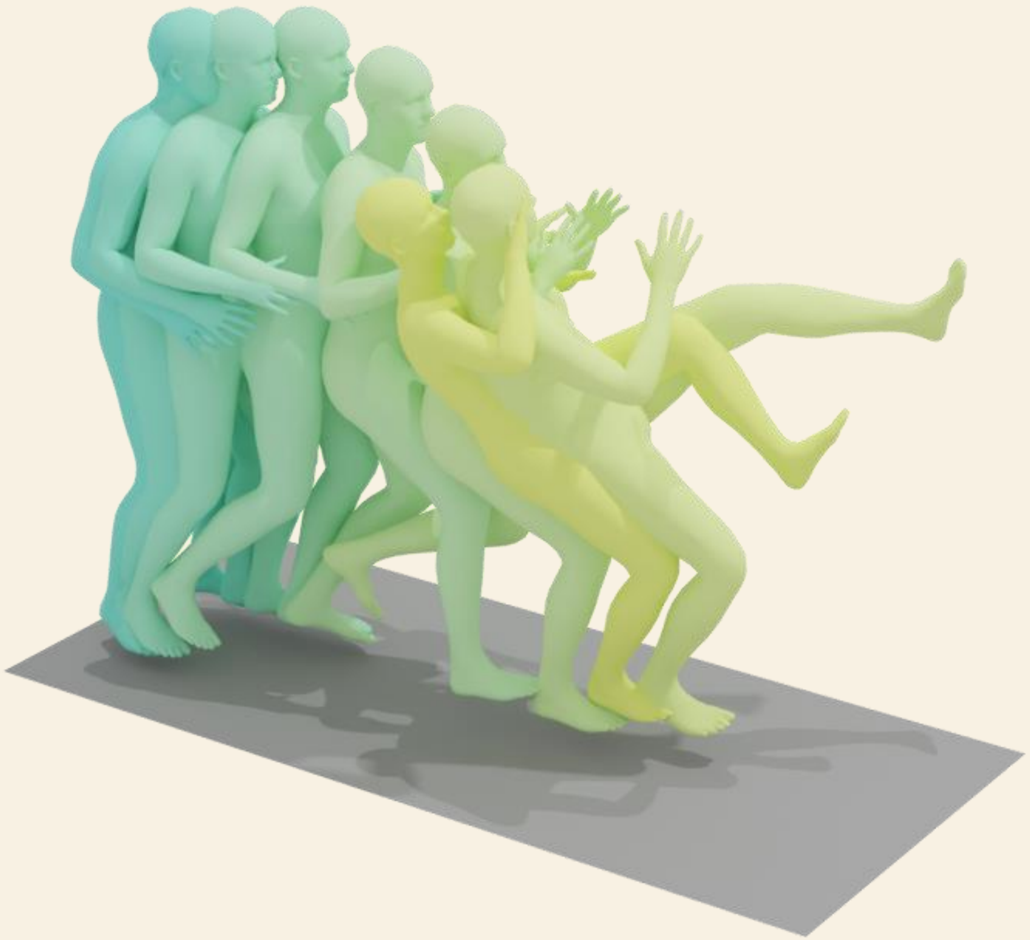
PERCEPTION AND INTELLIGENCE

# Perception and Intelligence laboratory

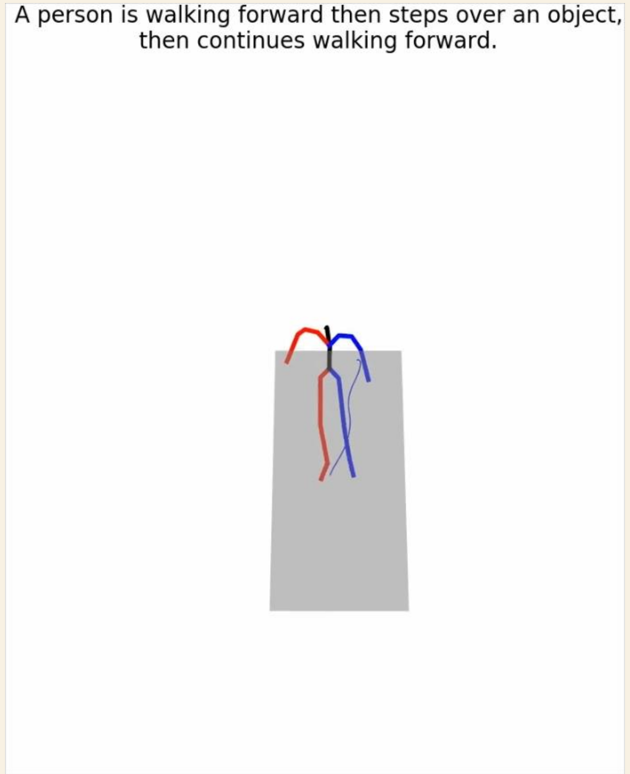


# Text to Motion

*“A person giving a single high kick”*



# A person is walking forward then steps over an objects, then continues walking forward



A person is walking forward then steps over an object, then continues walking forward.

Guo et al., **MoMask: Generative Masked Modeling of 3D Human Motions** CVPR 2024

# Today’s menu

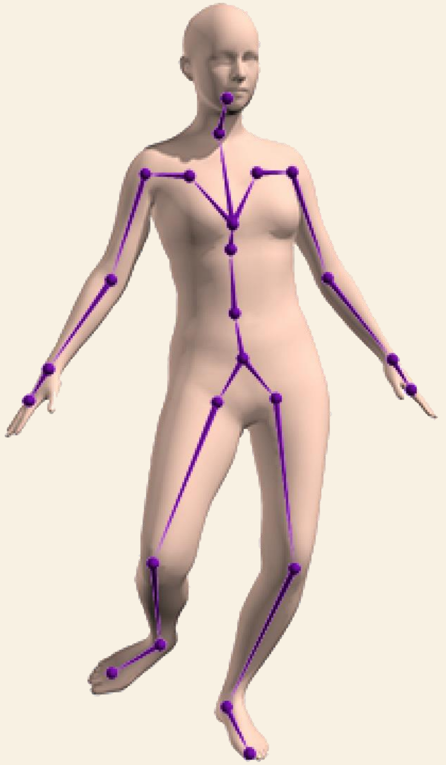
- Representing human motion in 3D
- Generative Model ZOO
- Control and conditioning (length, social factors, 3D scene)
- Human motion for robotic applications
- Human motion generation as gateway for video generation.

# SMPL: A Skinned Multi-Person Linear Model

Loper et al. SIGGRAPH Asia 2015

```
motion = torch.tensor(motion).float()


motion_params = {
        'root_orient': motion[:, :3],        # Global root orientation (3,)
        'pose_body': motion[:, 3:66],        # Body joint angles (63,)
        'trans': motion[:, 66:69],           # Global translation (3,)
        'betas': motion[:, 69:],             # Body shape parameters
        }
```

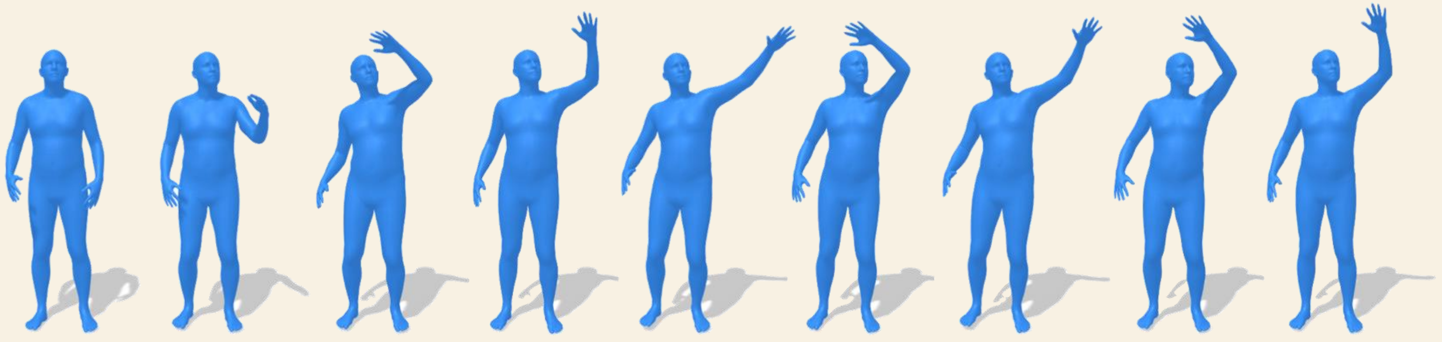


# Adding Text

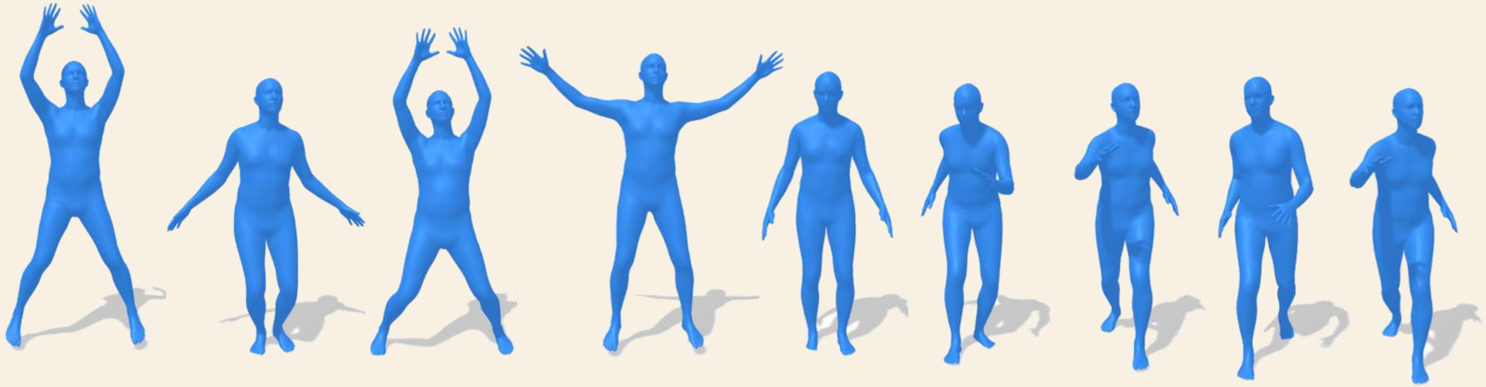
Guo et al. Generating Diverse and Natural 3D Human Motions From Text CVPR2022

HumanML3D is a 3D human motion-language dataset that originates from a combination of HumanAct12 and Amass dataset.
It covers a broad range of human actions:

- daily activities (e.g., 'walking', 'jumping’)
- sports (e.g., 'swimming', 'playing golf’)
- acrobatics (e.g., 'cartwheel’)
- artistry (e.g., 'dancing').



1. The person is waves at someone with his left hand.
2. A person shakes an item with his left hand.
3. A person waves his left hand repeatedly above his head.



1. A person doing jumping jacks and then running on the spot.
2. A person is doing jumping jacks, then starts jogging in place.
3. A person does four jumping jacks then three front lunges.

# Generative Model ZOO

| | |
|---|---|
| Masked Modelling | Predict randomly masked tokens* |
| Auto Regressive | Predict next token given previous tokens |
| Diffusion | Reverse a noising process |

Tokens* A discrete representation of a single pose, or a small motion snippet.

# Masked Modelling

Inspired by BERT

The cat [MASK] on the [MASK]→ The cat sat on the mat

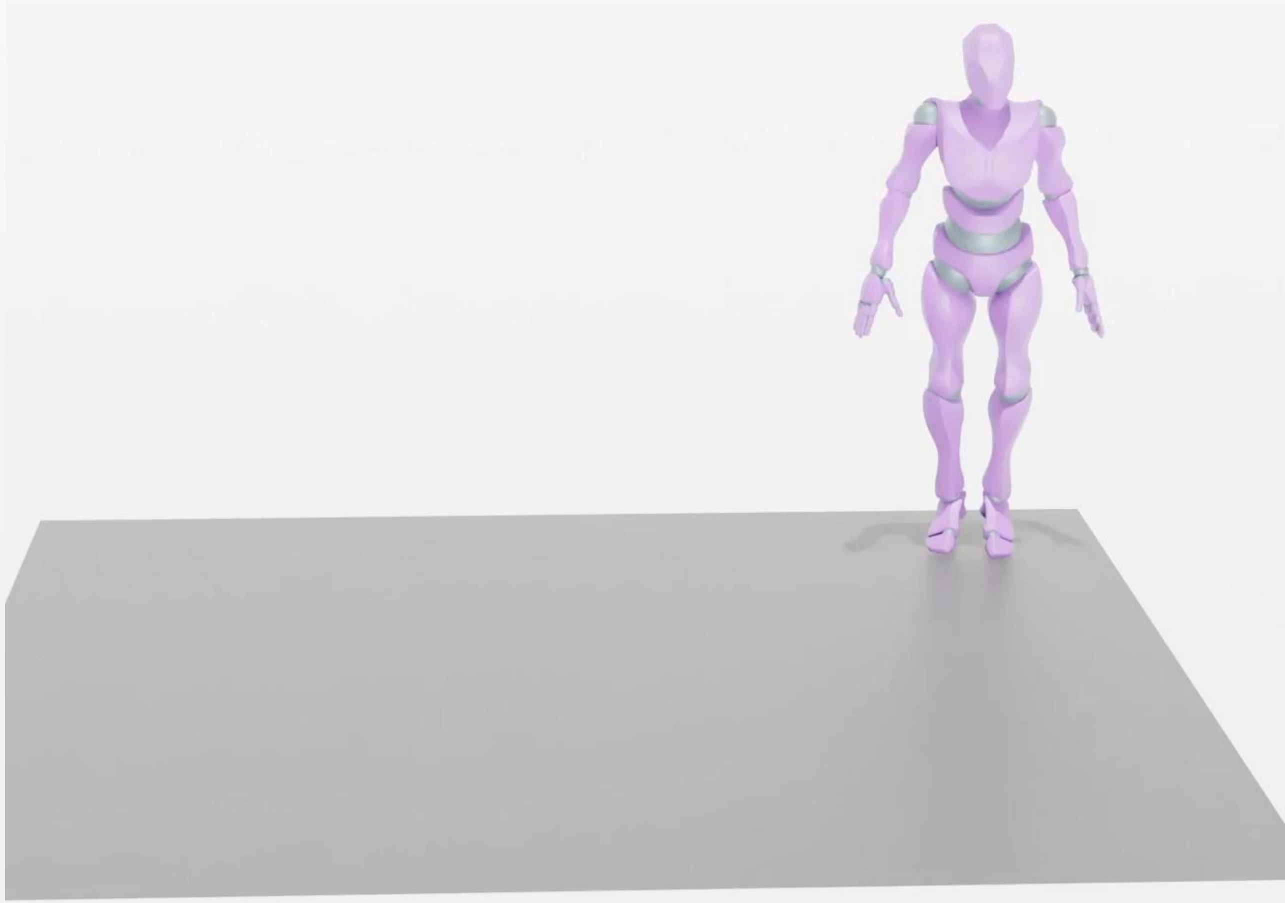
Guo et al., MoMask: Generative Masked Modeling of 3D Human Motions CVPR 2024

# Impainting

(Purple=Input, Cyan=Synthesis)

**"A person falls down and gets back up quickly.”**

**"A person is pushed."**



# Autoregressive

Notable example **GPT**

The cat sat on the → The cat sat on the mat

Left-to-right, one token at a time.

Jiang et al. MotionGPT: Human Motion as Foreign Language
NeurIPS2024

# Text-to-Motion

# Diffusion

Inspired by termodynamics, notable example StableDiffusion

The cat sat on the + ^%$ → saa → mat

Training: Add noise in steps → train model to denoise

Inference: Start from pure noise, reverse the noising process

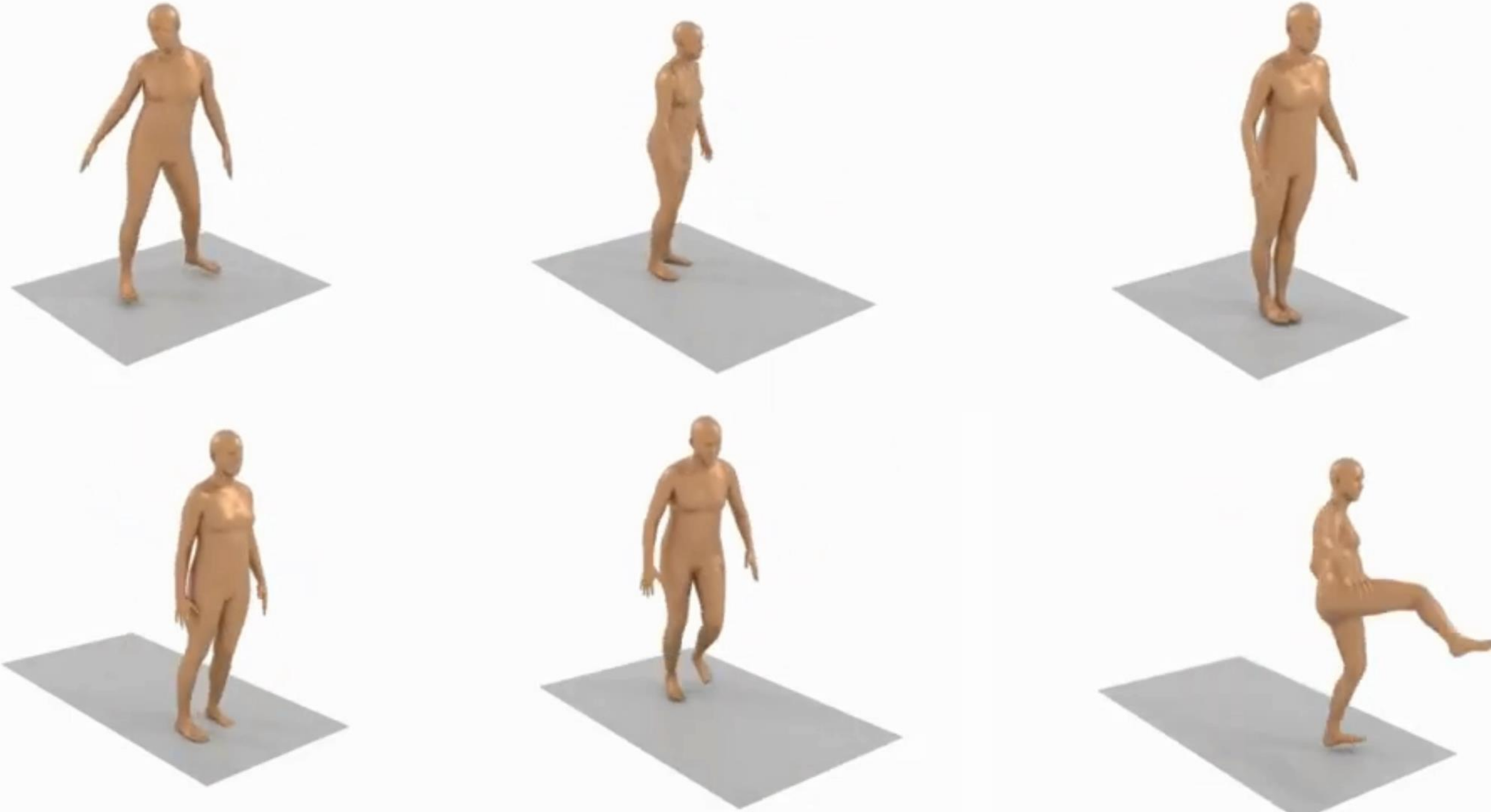
Tevet et al. **MDM: Human Motion Diffusion Model** ICLR2023

Chen et al **Executing your Commands via Motion Diffusion in Latent Space** CVPR2023

# Human Motion – a Many-to-many Problem

## Diversity

"A person kicks"



# Generation Length Control

State-of-the-art techniques for human behavior synthesis have limited control over the target sequence length.

Autoregressive: No global awareness of the target sequence length. The model just predicts the most likely next token. Stops when a special <EOS> (end-of-sequence) token is predicted

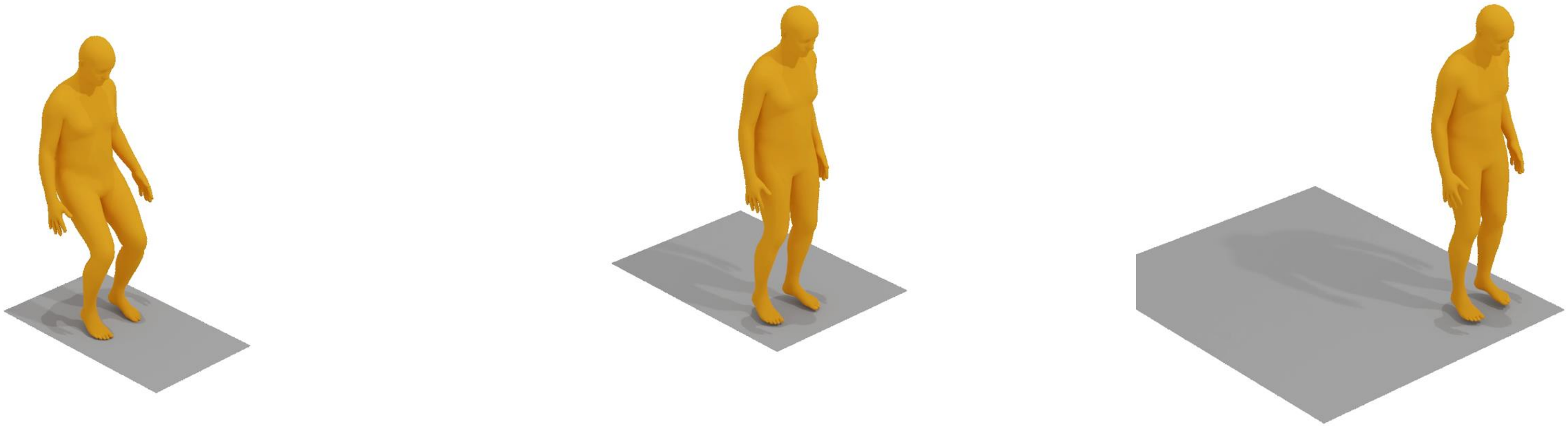
Masked and Diffusion Models: You always feed in a fixed-length sequence, e.g., 512 tokens. Model only predicts inside that window.

# Why is it important?

Suppose we want to generate a short kick motion: subsampling a longer one is insufficient, as it will not capture the intricate variations that occur when humans perform actions at different speeds.

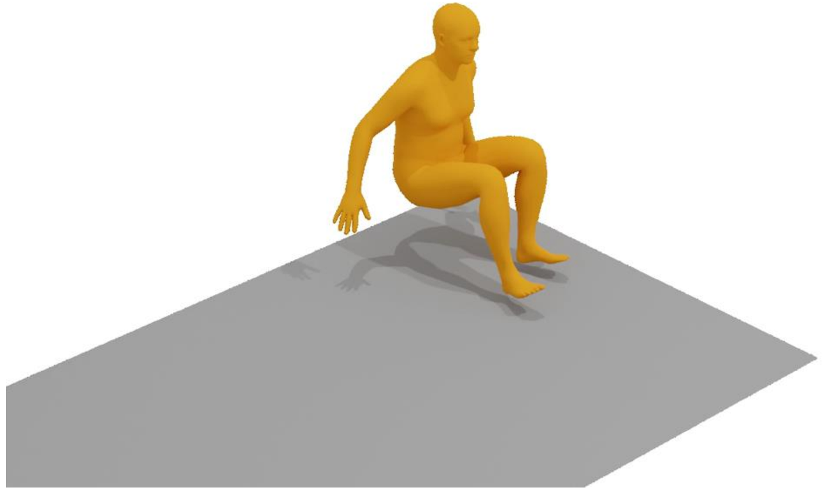
The embedding space should encode longer sequences using higher capacity representations, because they need more details to be generated, and shorter sequences with **less** capacity.

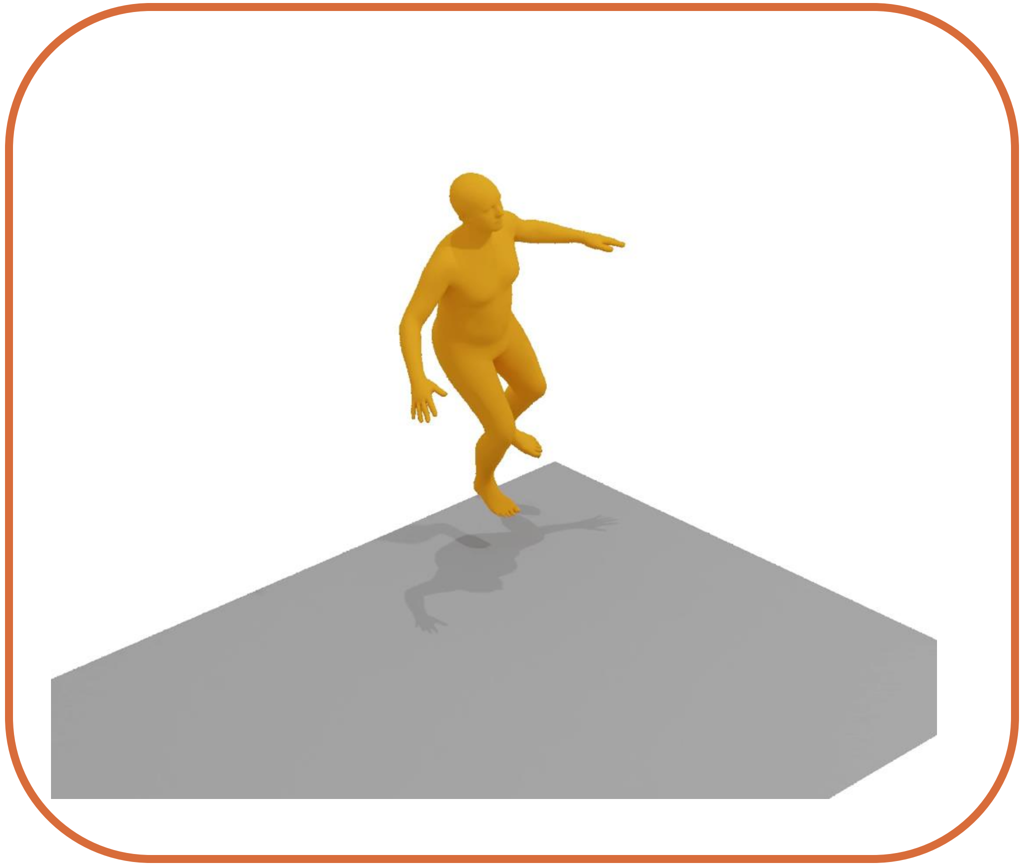
# Sampieri et al. Length-Aware Motion Syntesis via Latent Diffusion ECCV2024



# Blind Comparison - LADiff VS SoTA

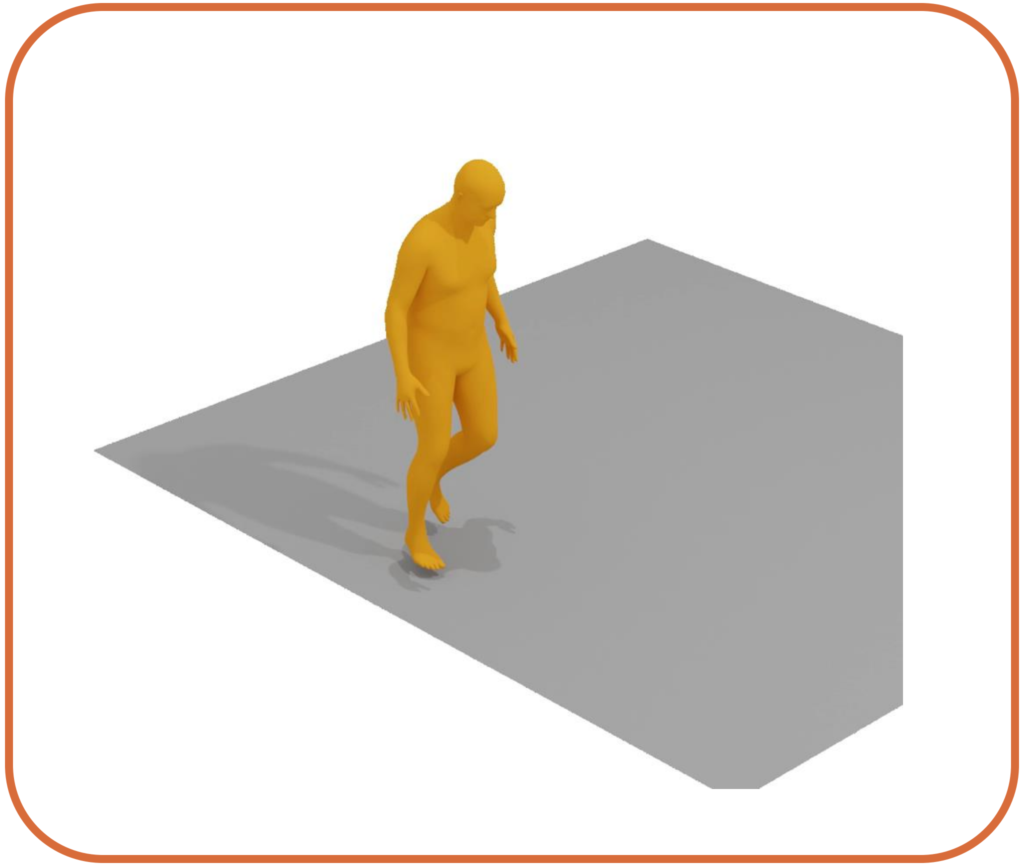
*“A person walks in a complete circle and then sits down.”*

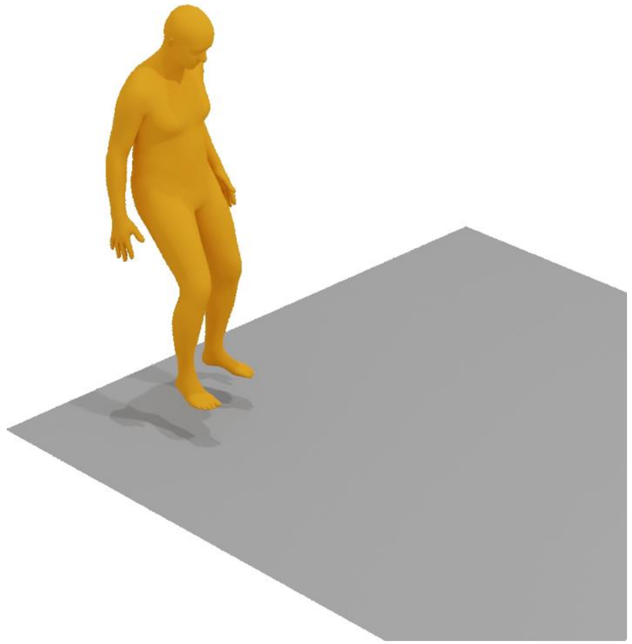




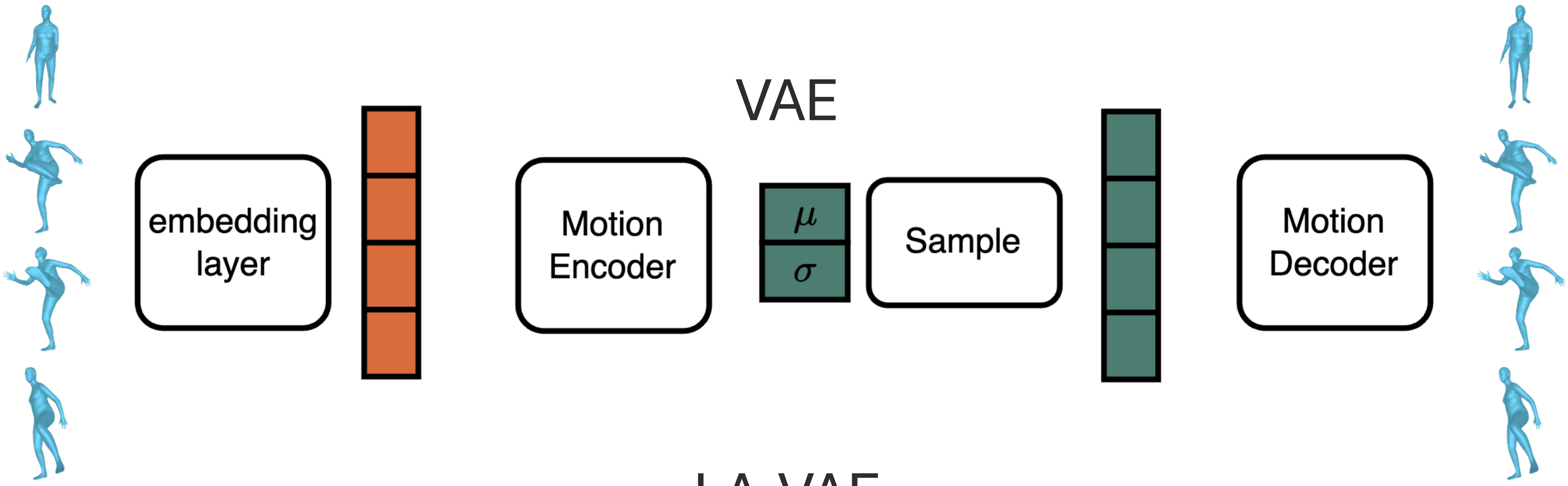
# Blind Comparison

*"A person walks in a complete circle and then sits down."*

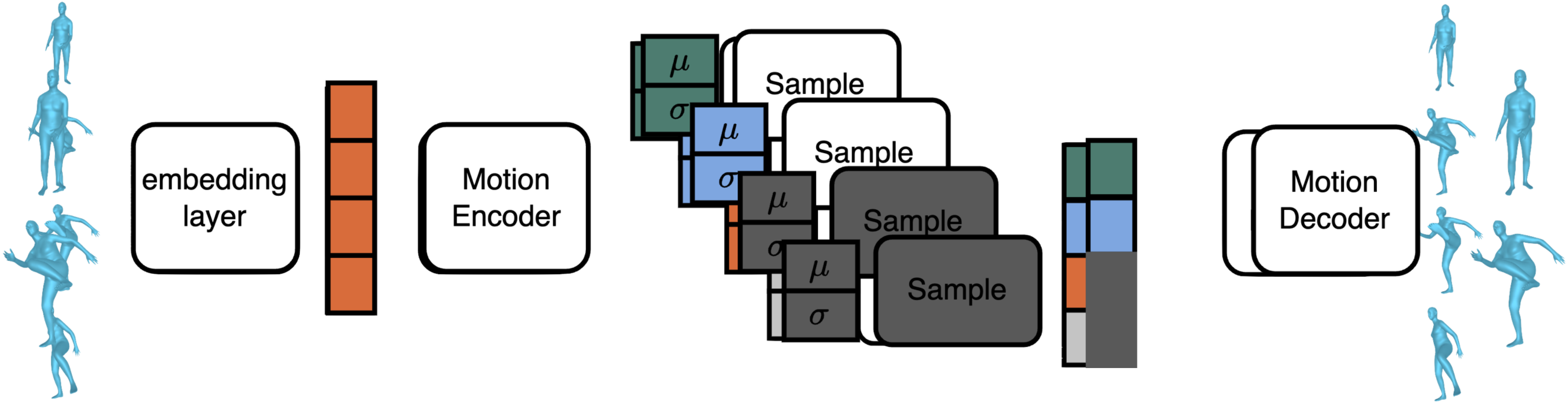




# VAE


embedding layer
Motion Encoder
$\mu$
$\sigma$
Sample
Motion Decoder

# LA-VAE

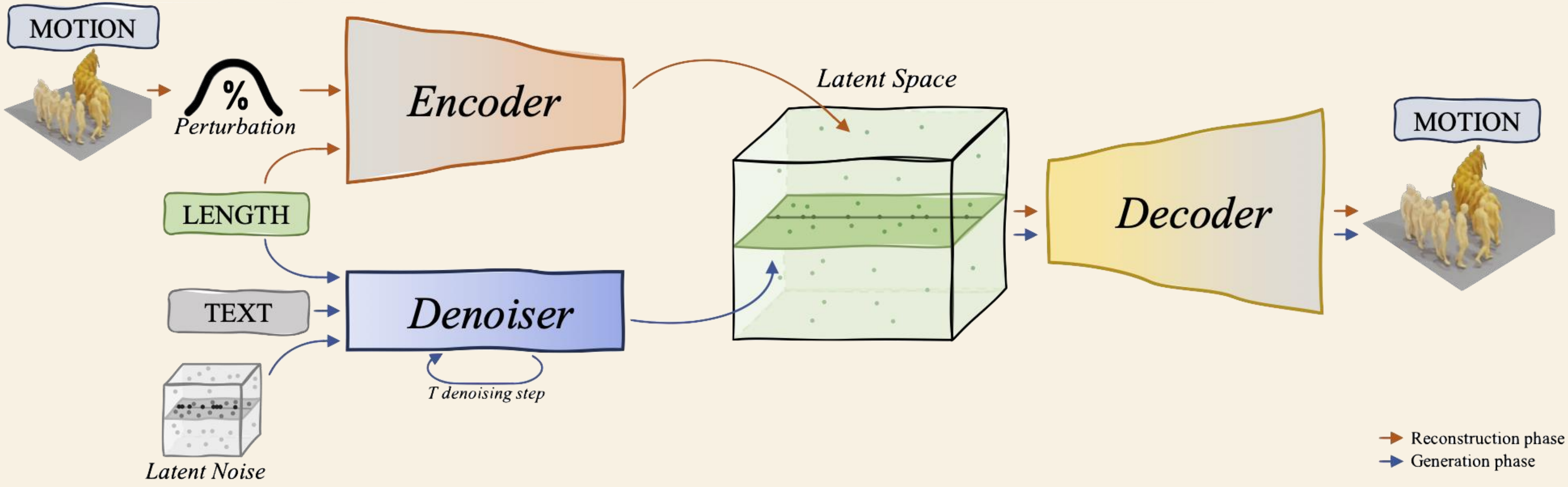

embedding layer
Motion Encoder
$\mu$
$\sigma$
Sample
$\mu$
$\sigma$
Sample
$\mu$
$\sigma$
Sample
$\mu$
$\sigma$
Sample
Motion Decoder

# Length-Aware VAE

- We split the full latent space into **K smaller subspaces**, each of size **D**.
  → Full space: $\mathbb{R}^{(K \times D)}$
  → Smallest subspace: $\mathbb{R}^{(1 \times D)}$

- As the motion gets longer, we gradually **unlock more subspaces**:
  → Number of active subspaces: $\mathbf{k = \lceil f / r \rceil}$
  → Where **f** = motion length (in frames), **r** = frames per subspace

- So, each motion **x** has a latent embedding $\mathbf{z} \in \mathbb{R}^{(k \times D),}$ depending on its length.



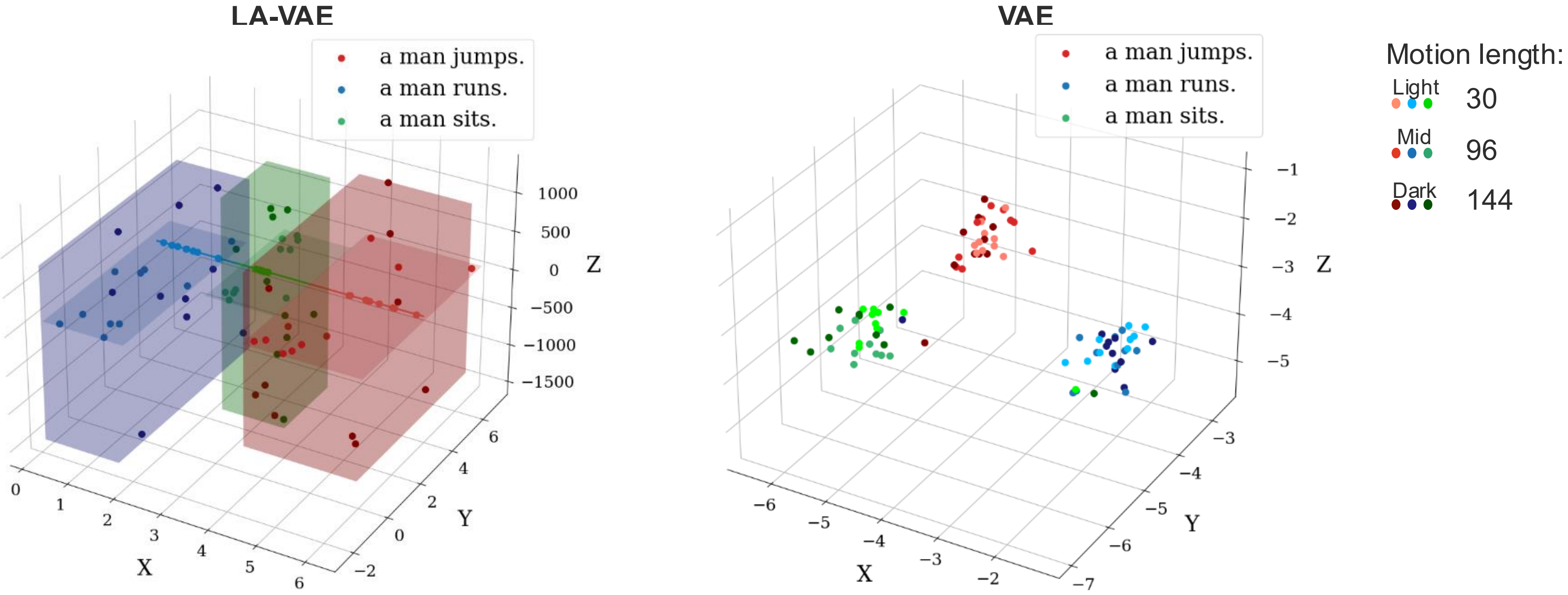
At inference time we select the right noise size for the intended generation length.

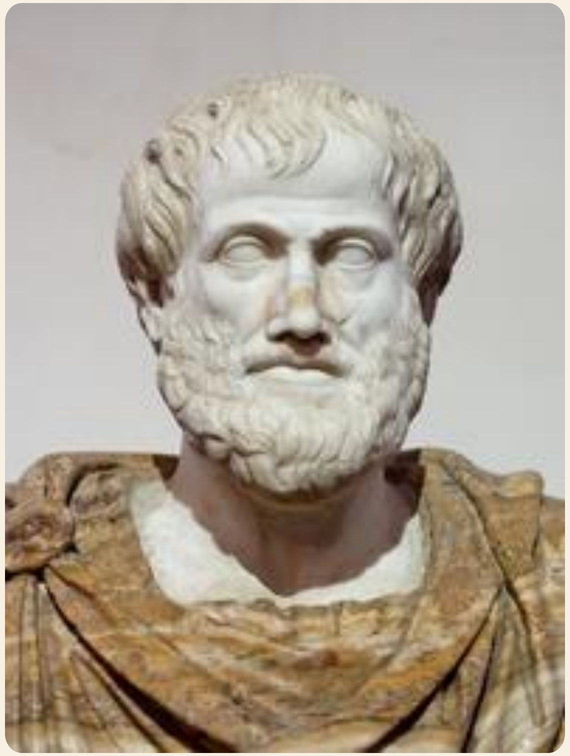
# Latent Space Structure

We compare the latent spaces using **t-SNE** for visualization.

Sequences of **30, 96, and 144 frames** are generated for **3 different actions**.

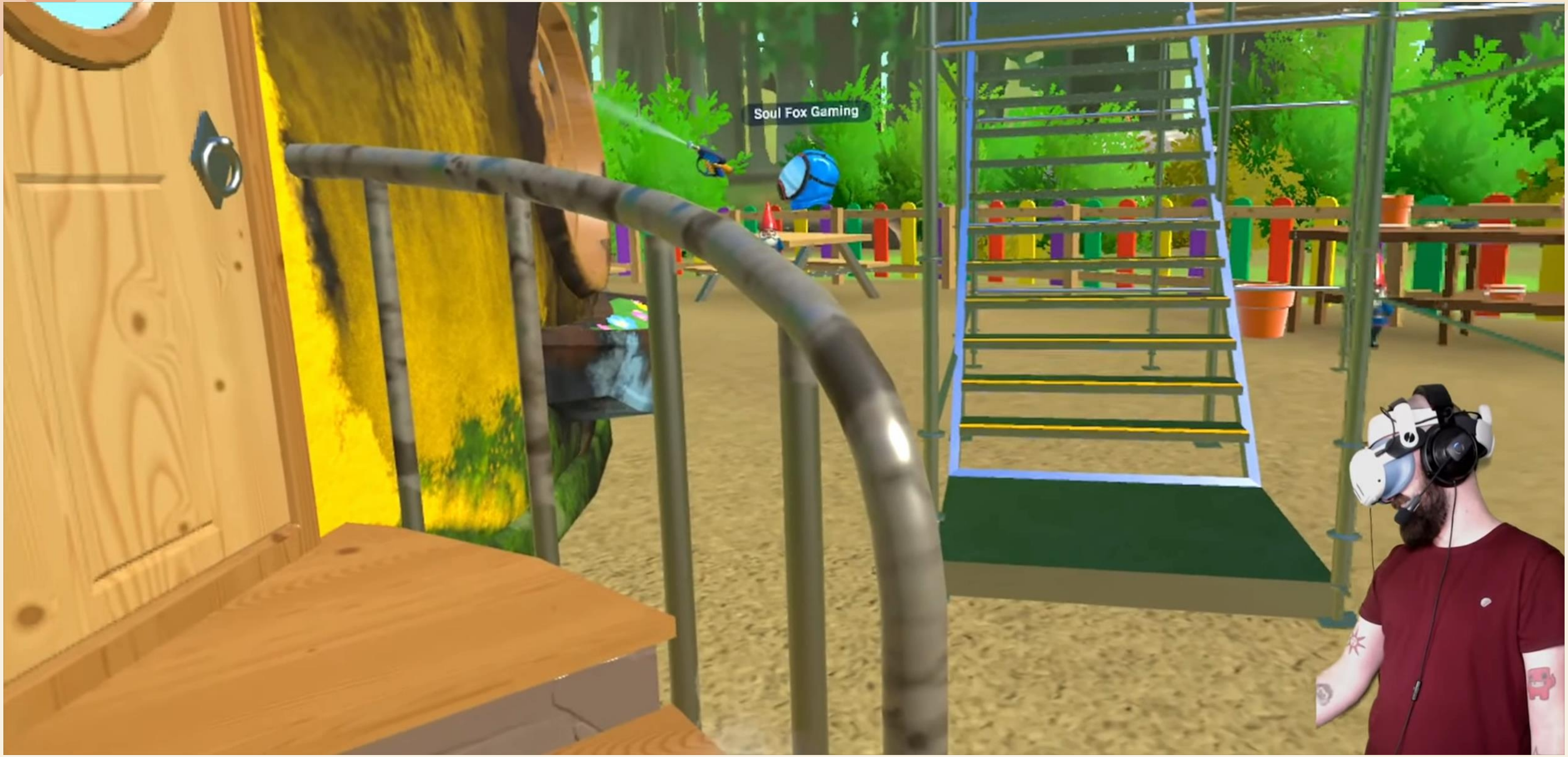
LA-VAE shows a more structured and length-aware representation than the standard VAE.





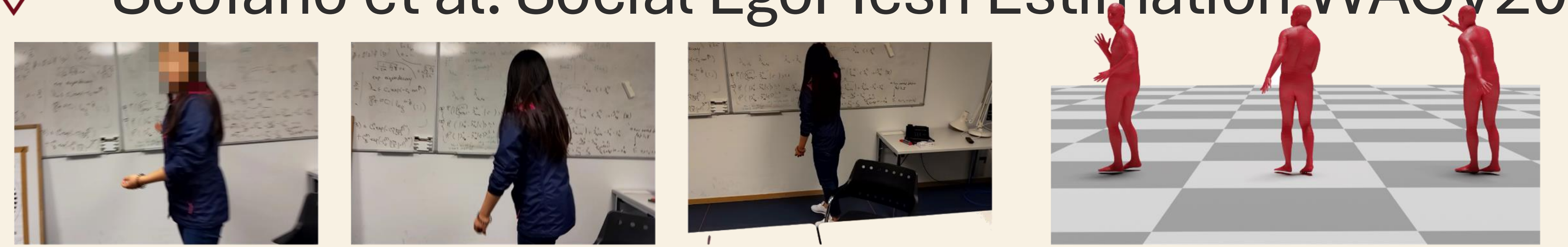
# Aristotle

“Man is by nature a social animal; an individual who is unsocial naturally and not accidentally is either beneath our notice or more than human.”



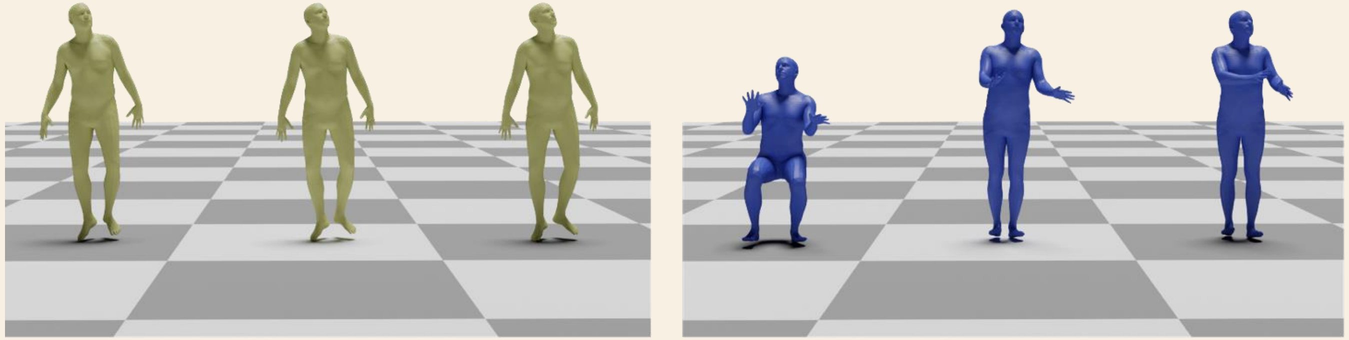
Powerwash simulator

# Scofano et al. Social EgoMesh Estimation WACV2025



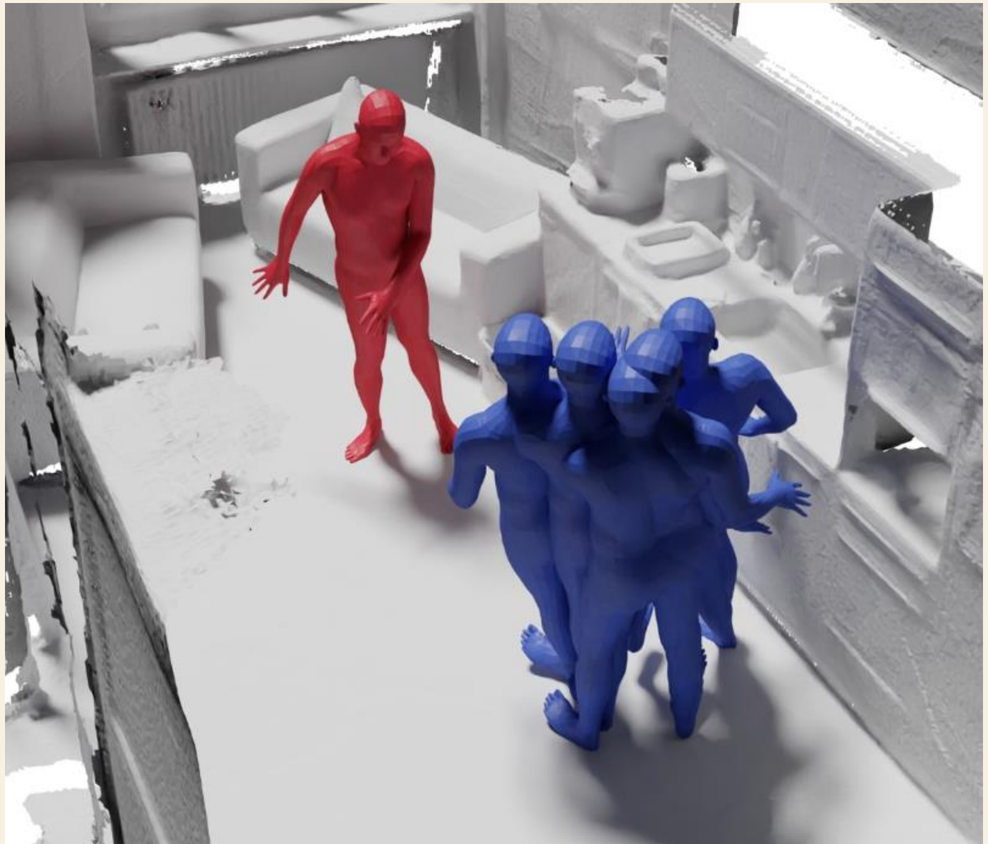
(a) Video stream of the wearer's egocentric view, not used in the model.

(b) Wearer's view of the interactee, extracted from the video stream.



(c) Prediction of the wearer's pose by EgoEgo.

(d) Prediction of the wearer's pose by SEE-ME.



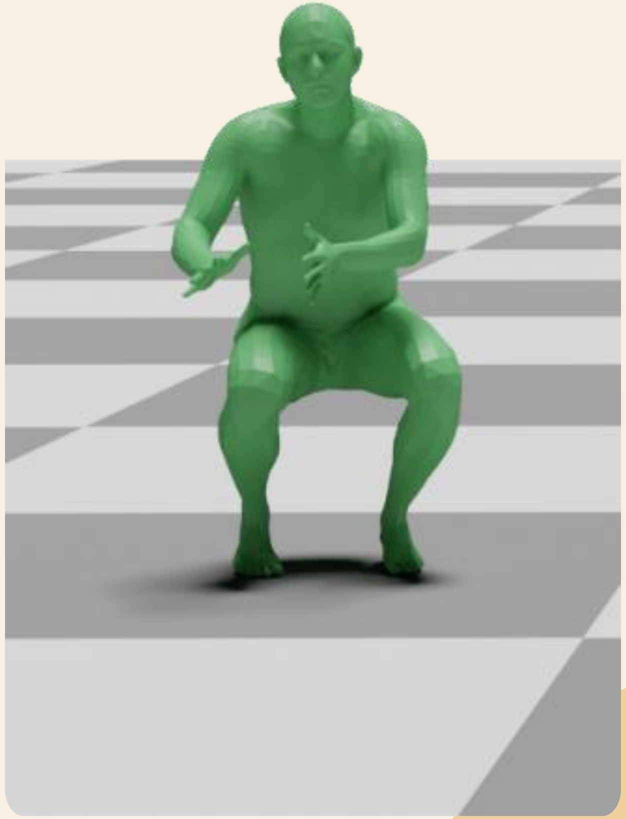
**Objective:** Estimate the 3D pose of the camera wearer in egocentric video sequences.

**Core Idea:** Integrate social interactions and scene understanding into the estimation process to enhance egocentric pose estimation accuracy.

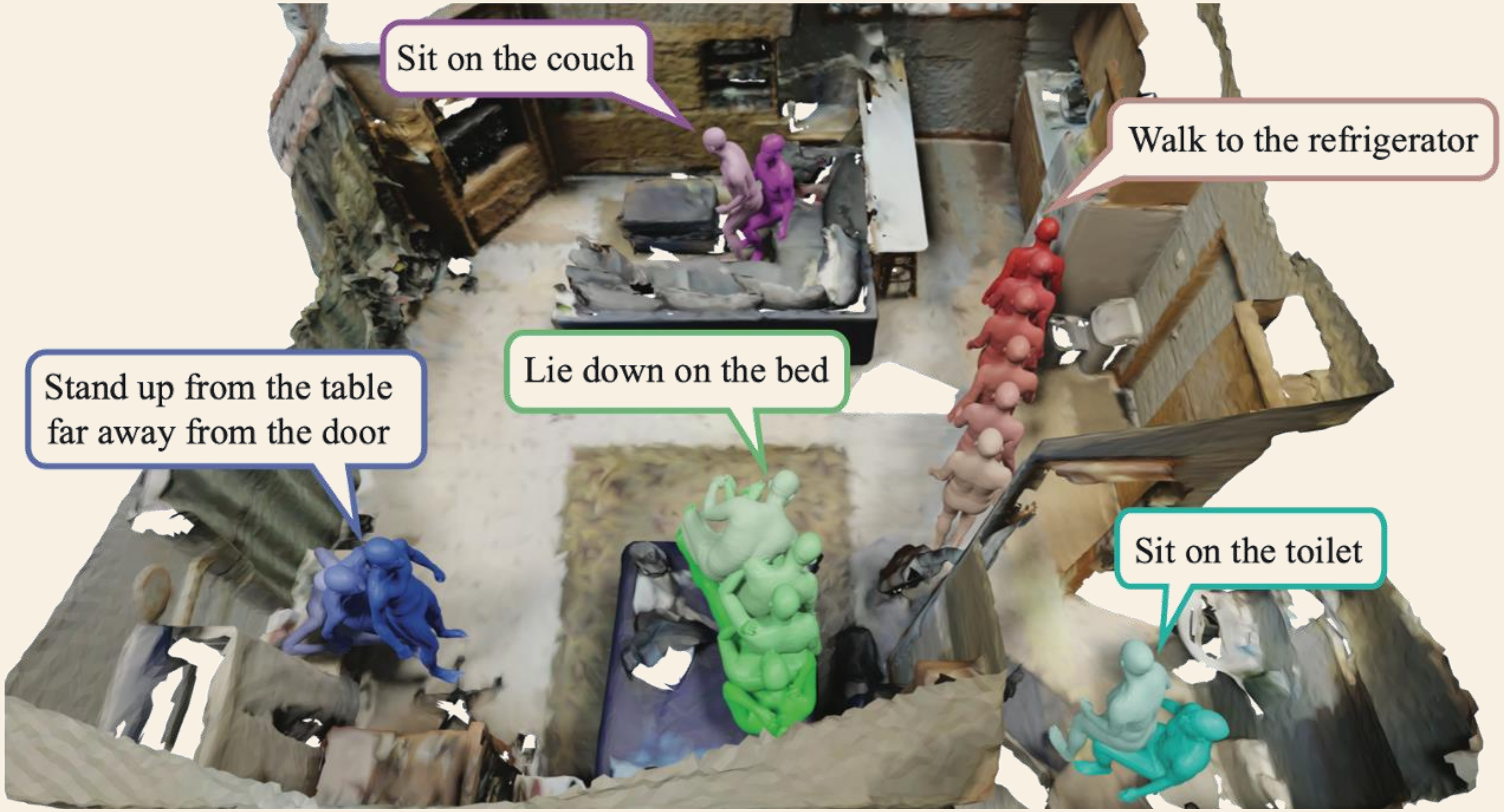
# Text-Motion-Scene

“ She asked me to stay and she told me to sit anywhere. But I looked around and I noticed there wasn’t a chair.”

— Norwegian Wood, The Beatles



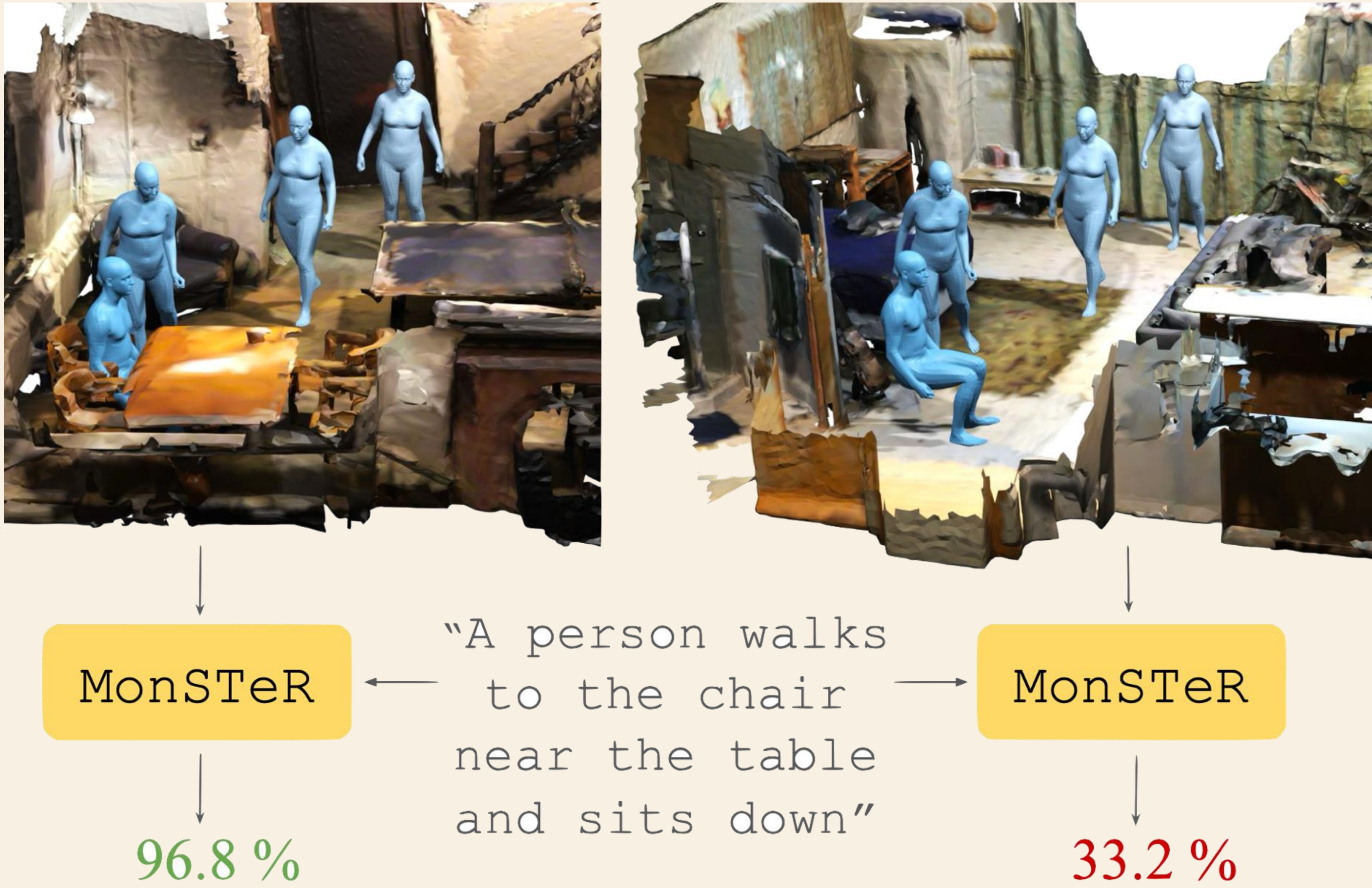
# Z. Wang et al. HUMANISE: Language-conditioned Human Motion Generation in 3D Scenes NeurIPS2022



# Collorone et al. MonSTeR : a Unified Model for Motion, Scene, Text Retrieval ICCV2025

Intention drives human movement in complex environments, but such movement can only happen if the surrounding context supports it.

Existing research has not yet provided tools to evaluate the alignment between skeletal movement (motion), intention (text), and the surrounding context (scene).



MonSTeR

"A person walks to the chair near the table and sits down"

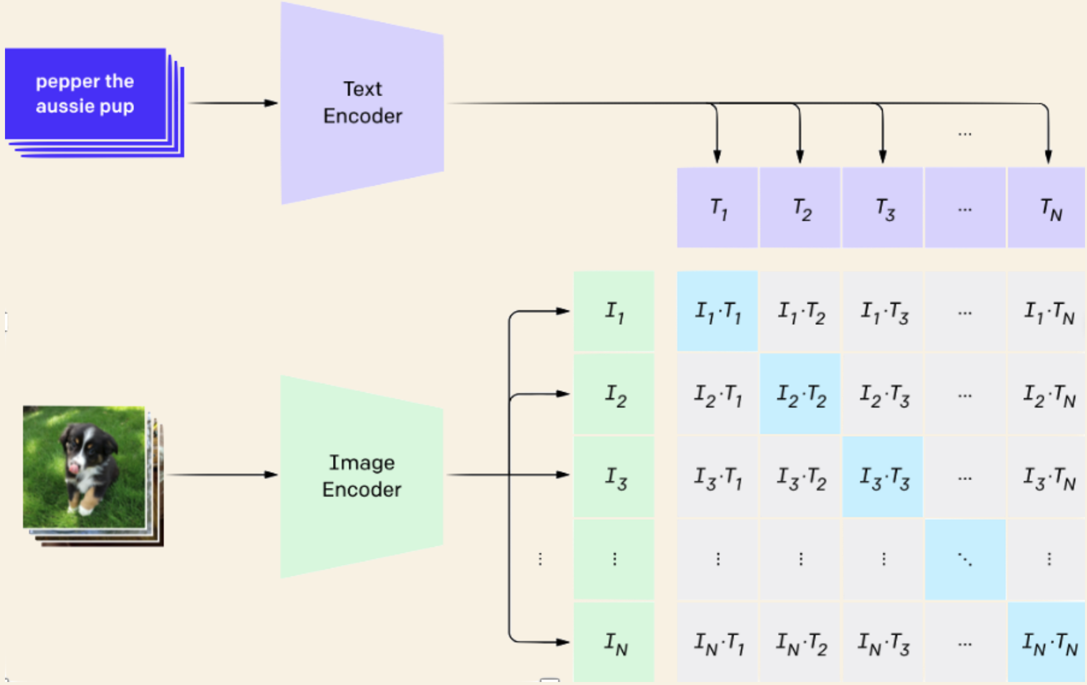
MonSTeR

96.8 %

33.2 %

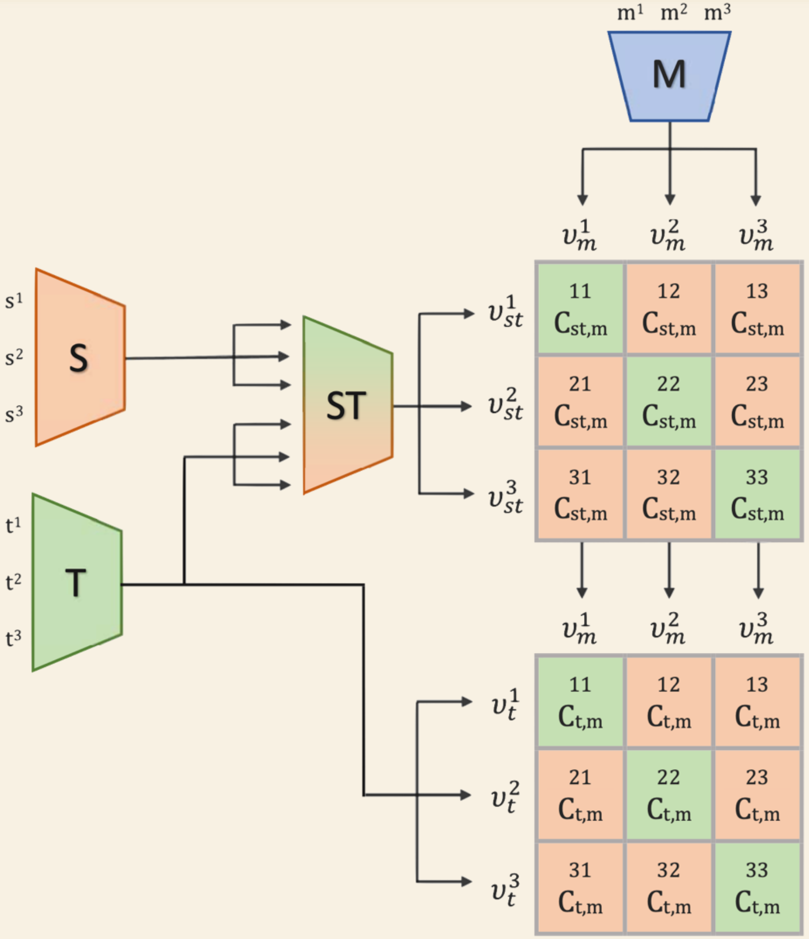
# Multimodal Alignment

Unimodal features (e.g., text, motion, scene) are treated as nodes, while bimodal (cross-modal) relationships (e.g., text-scene, motion-scene, motion-text) form the edges. Higher-order interactions are captured by aligning these bimodal embeddings with the remaining unimodal modality



|  | $T_1$ | $T_2$ | $T_3$ | … | $T_N$ |
|---|---|---|---|---|---|
| $I_1$ | $I_1 \cdot T_1$ | $I_1 \cdot T_2$ | $I_1 \cdot T_3$ | … | $I_1 \cdot T_N$ |
| $I_2$ | $I_2 \cdot T_1$ | $I_2 \cdot T_2$ | $I_2 \cdot T_3$ | … | $I_2 \cdot T_N$ |
| $I_3$ | $I_3 \cdot T_1$ | $I_3 \cdot T_2$ | $I_3 \cdot T_3$ | … | $I_3 \cdot T_N$ |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋱ | ⋮ |
| $I_N$ | $I_N \cdot T_1$ | $I_N \cdot T_2$ | $I_N \cdot T_3$ | … | $I_N \cdot T_N$ |

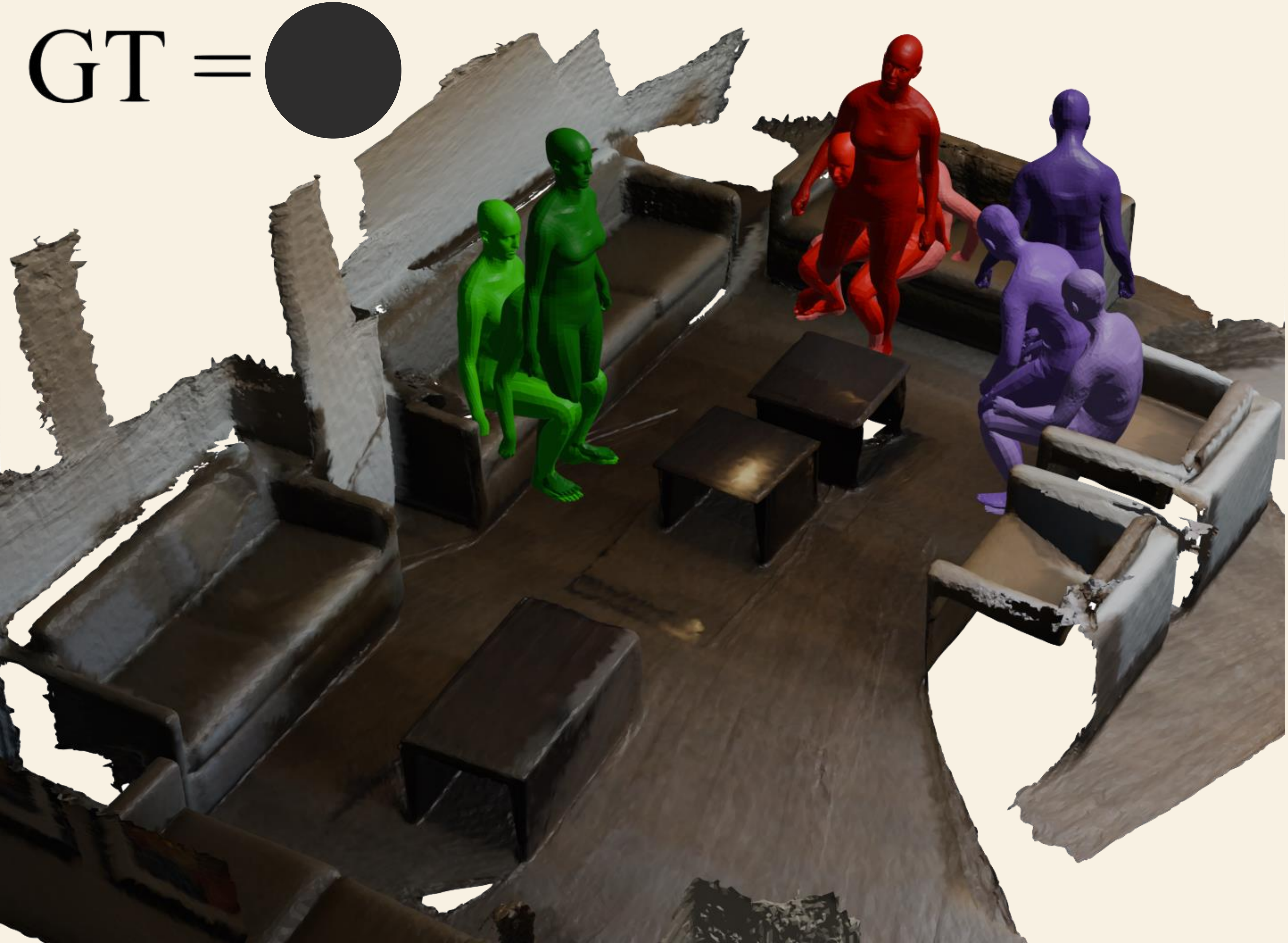
CLIP from OpenAI



|  | $v_m^1$ | $v_m^2$ | $v_m^3$ |
|---|---|---|---|
| $v_{st}^1$ | $C^{11}_{st,m}$ | $C^{12}_{st,m}$ | $C^{13}_{st,m}$ |
| $v_{st}^2$ | $C^{21}_{st,m}$ | $C^{22}_{st,m}$ | $C^{23}_{st,m}$ |
| $v_{st}^3$ | $C^{31}_{st,m}$ | $C^{32}_{st,m}$ | $C^{33}_{st,m}$ |

|  | $v_m^1$ | $v_m^2$ | $v_m^3$ |
|---|---|---|---|
| $v_t^1$ | $C^{11}_{t,m}$ | $C^{12}_{t,m}$ | $C^{13}_{t,m}$ |
| $v_t^2$ | $C^{21}_{t,m}$ | $C^{22}_{t,m}$ | $C^{23}_{t,m}$ |
| $v_t^3$ | $C^{31}_{t,m}$ | $C^{32}_{t,m}$ | $C^{33}_{t,m}$ |

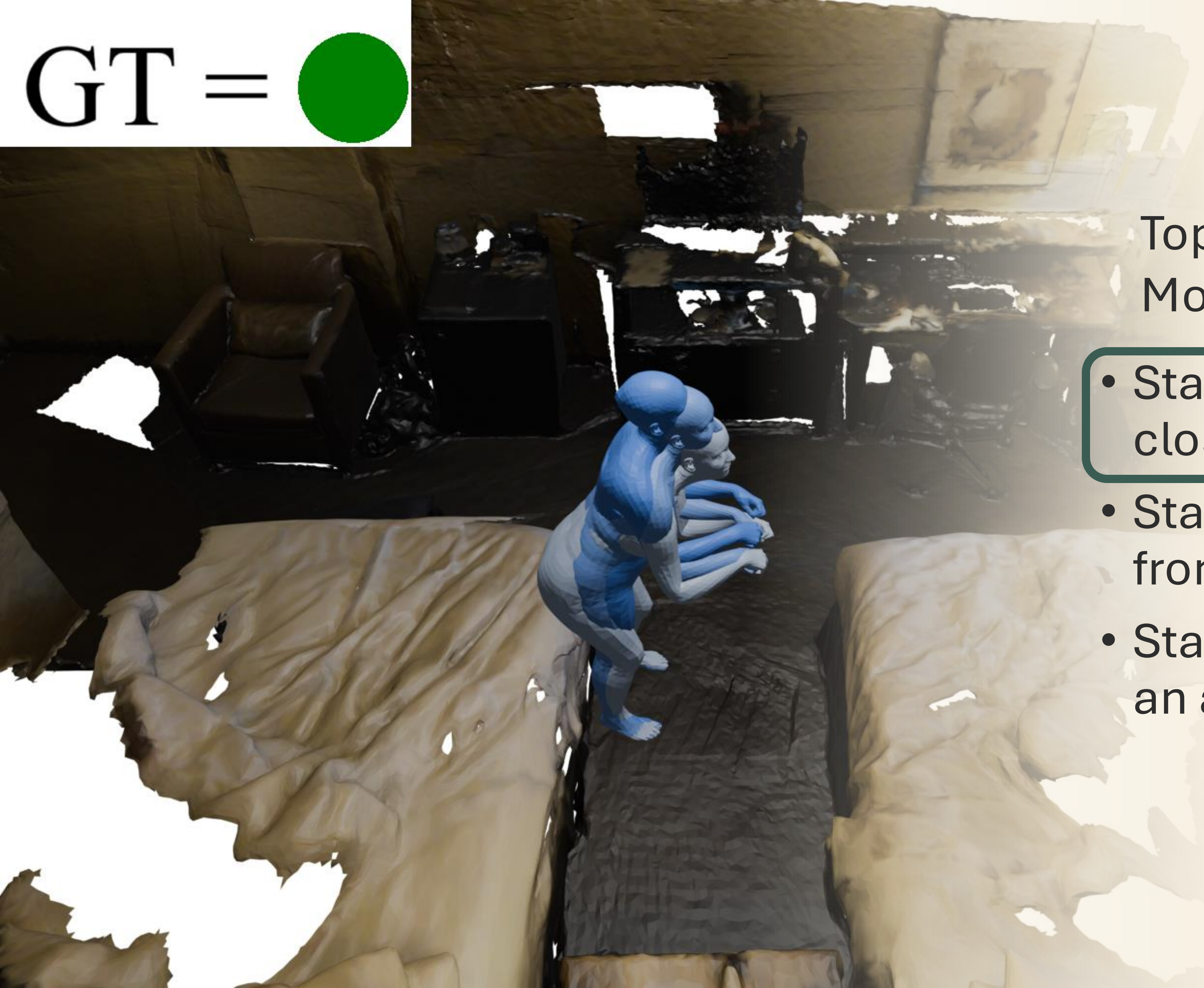
GT = ●



Top 3 motion retrieved by MonSTer:

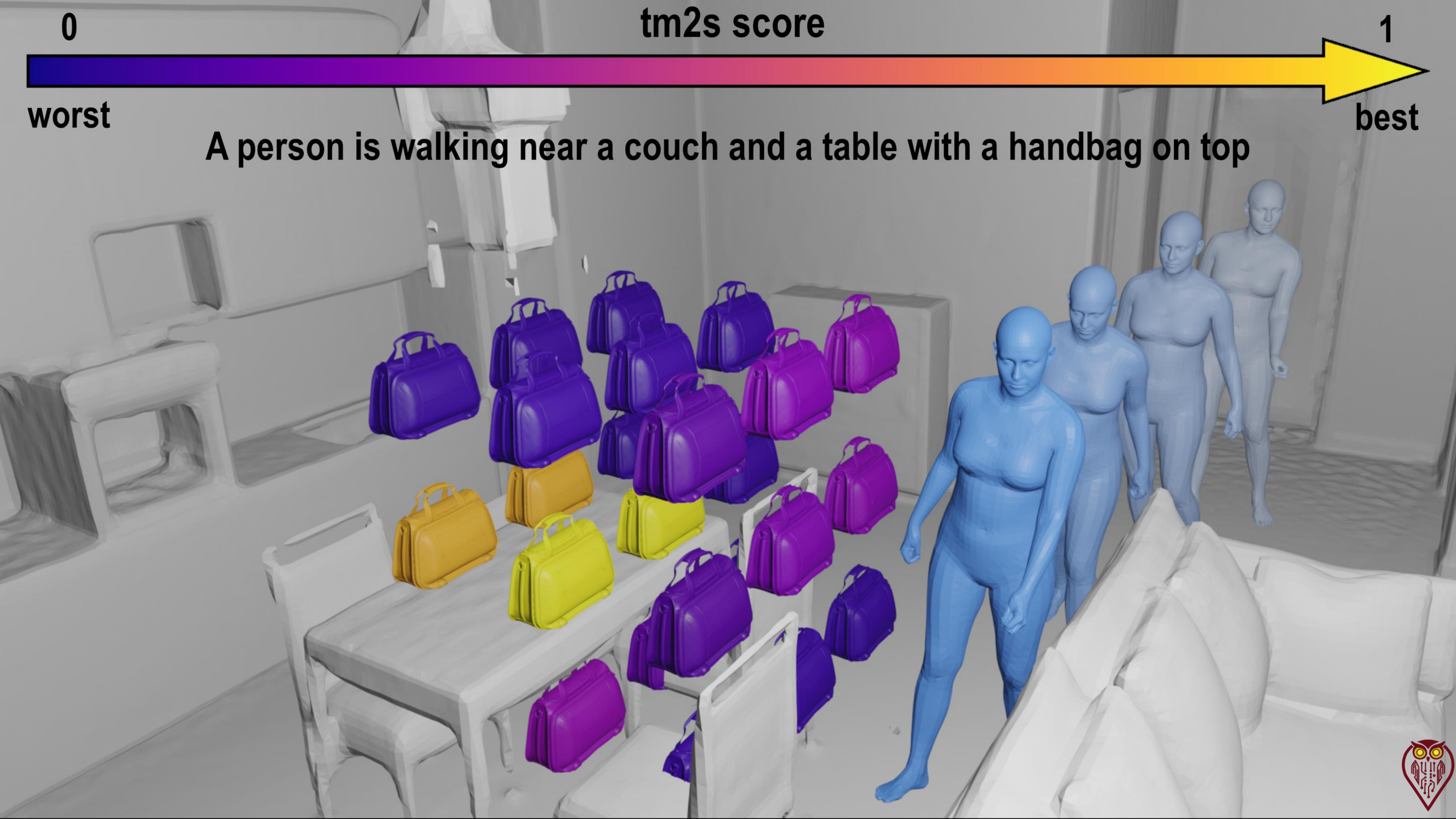
“Stand up from the sofa chair, which is near the door, a table and a wall.”

GT = 🟢



Top 3 text retrieved by MonSTer:

- Stand up from the bed close to the armchair.
- Stand up from the bed, far from the office chair.
- Stand up from the bed near an armchair and a suitcase.



tm2s score

0

1

worst

best

A person is walking near a couch and a table with a handbag on top

# FOLLOWING THE HUMAN THREAD IN SOCIAL NAVIGATION ICLR 2025

Two-stage reinforcement learning framework enabling robots to adapt to human movement.

Stage 1: learns a latent representation of social dynamics from fully observable human trajectories to condition robot navigation.

Stage 2: infers social dynamics solely from the robot's past, allowing real-time adaptation without privileged trajectory data.

Puig et al. Habitat 3.0: A Co-Habitat for Humans, Avatars and Robots ICLR 2024

# Performances…

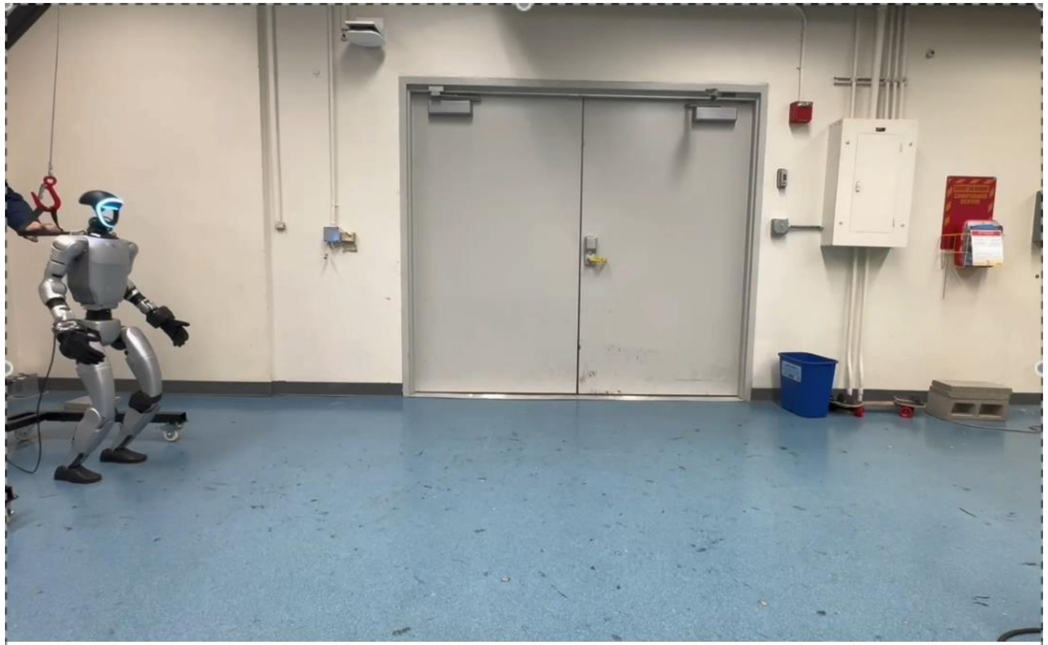
We showed state-of-the-art performance in finding and following humans:

- Finding Success (S): Increased from 76% to 91%. (+15%)
- Following Success (F): Increased from 0.29 to 0.39. (+10%)

# From text to motion to text to humanoid

Rise of end-to-end methods enabling humanoid robots to interpret and execute diverse whole-body motions directly from natural language commands.

Shao et al., **LangWBC: Lang**uage-directed Humanoid **W**hole **B**ody **C**ontrol via End-to-end Learning

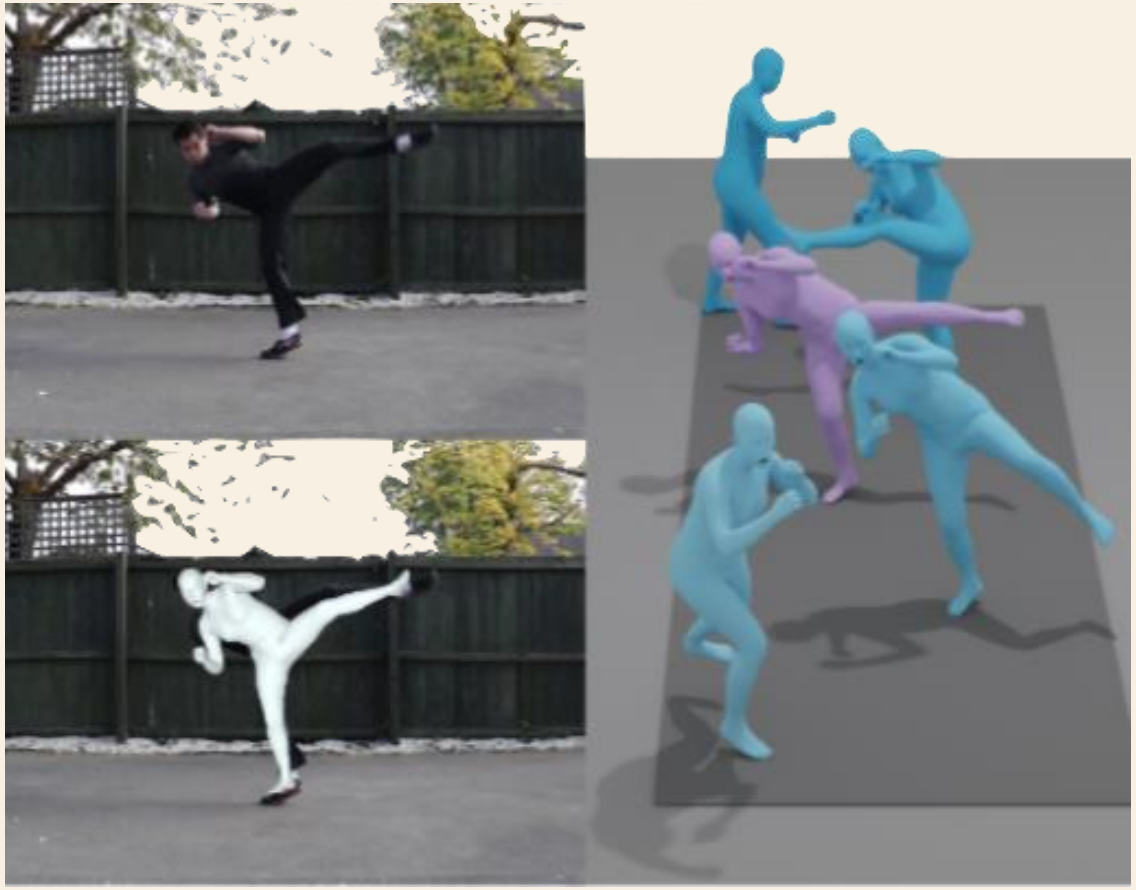


*A person is moving forward briskly*

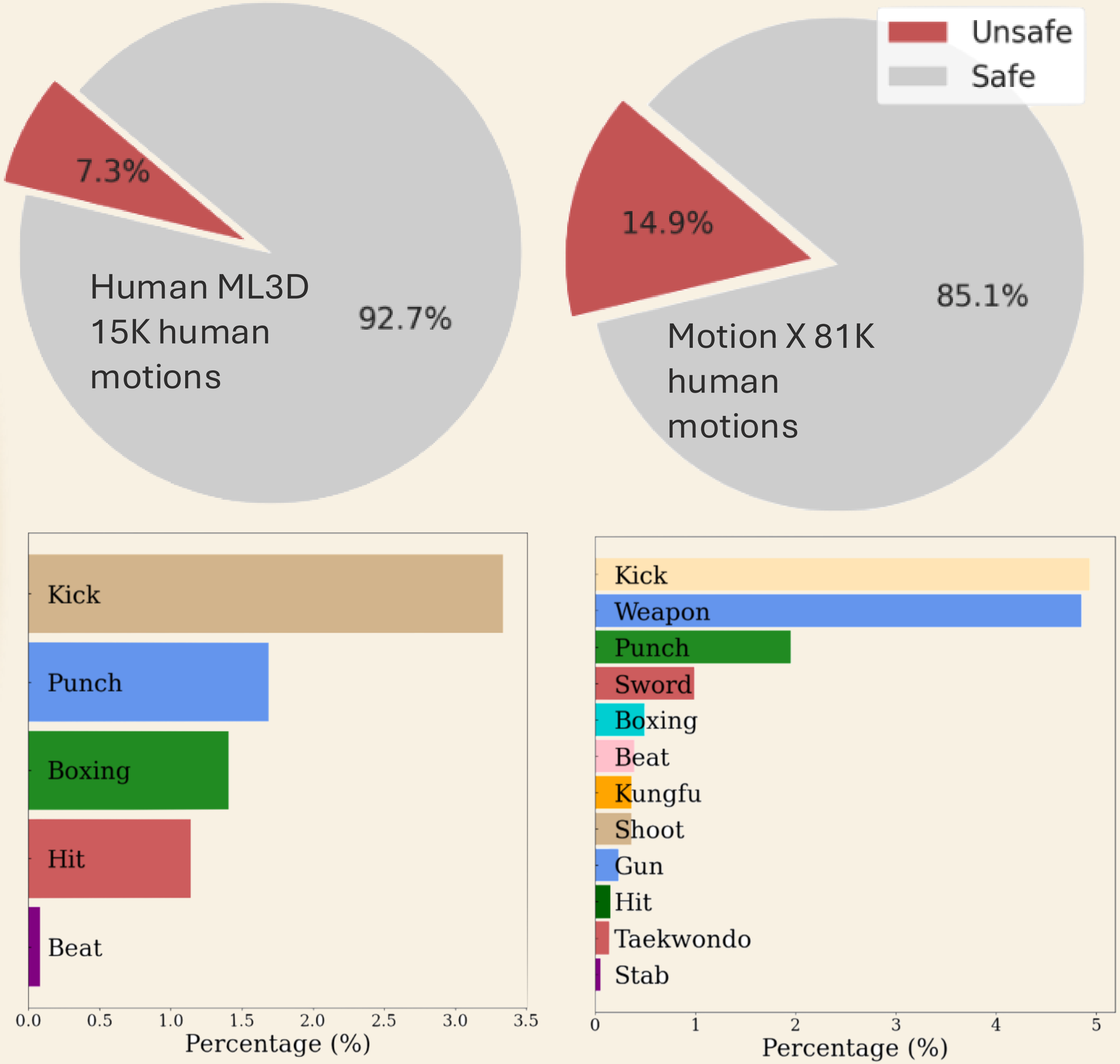
Wave -> Run -> Wave (Side View)

# Challenges in Web-Scraped Human Motion Data

Scraping the web for human motion data can lead to inclusion of violent content.



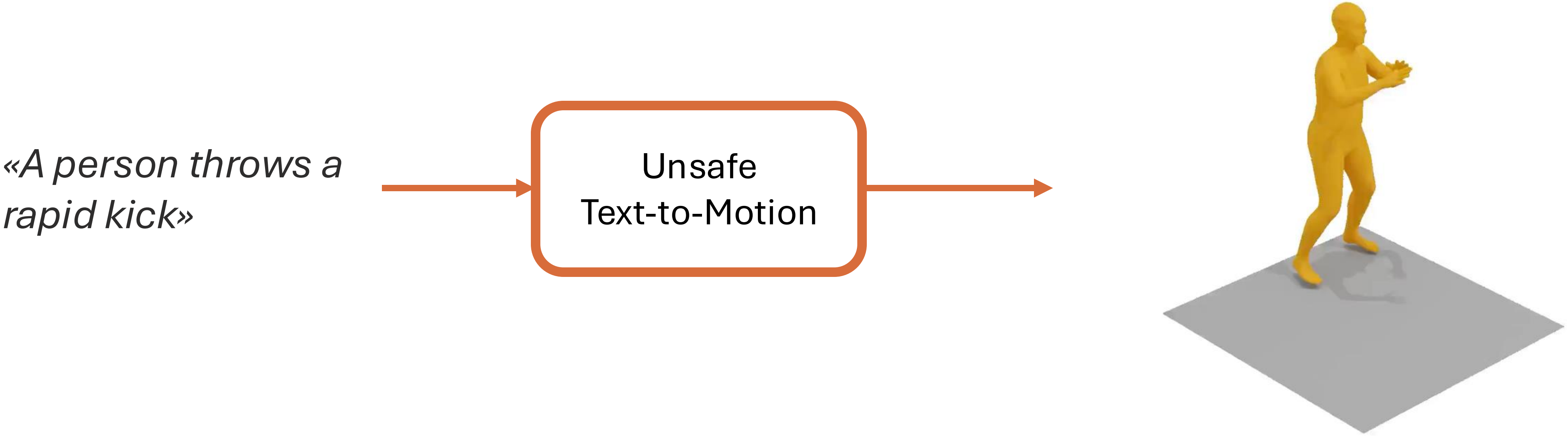
# Violent motion in today's 'small' datasets.



# Goal

Considering a human motion generative model.

We want to make it incapable of creating violent motions.



*«A person throws a rapid kick»*

Unsafe Text-to-Motion

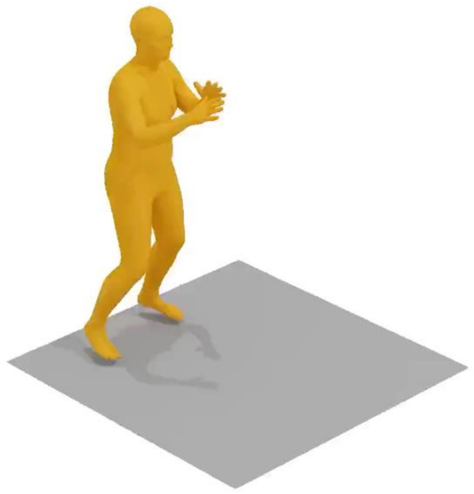
# Goal

Considering a human motion generative model.

We want to make it incapable of creating violent motions like this:

*«A person throws a rapid kick»*

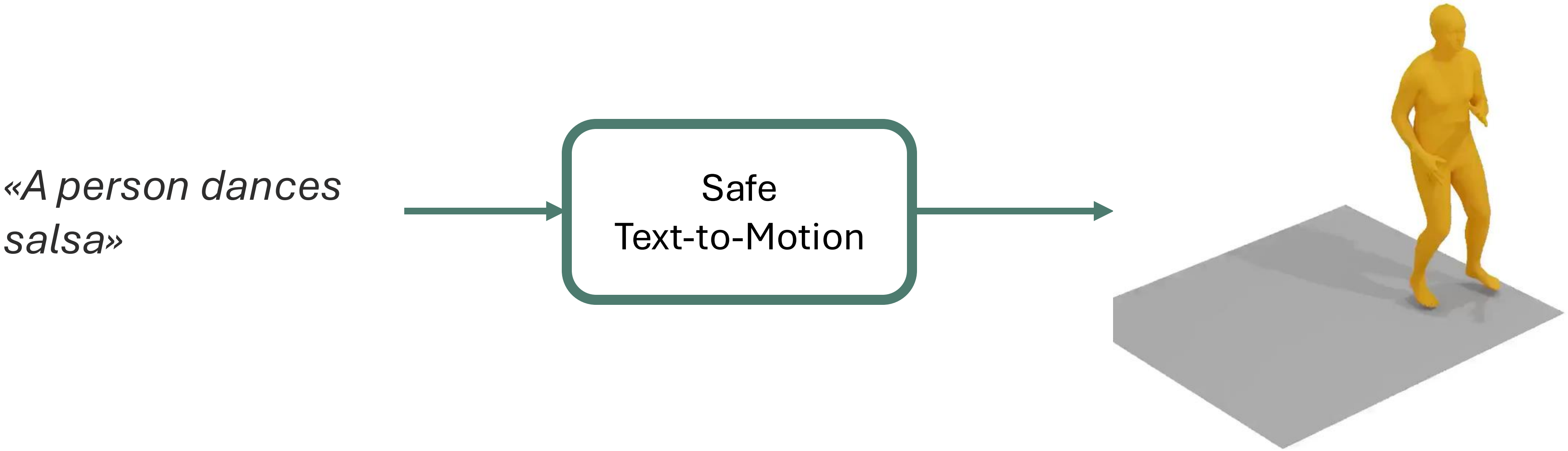
Safe Text-to-Motion



# Goal

Considering a human motion generative model.

We want to make it incapable of creating violent motions … while preserving safe motion generation.



*«A person dances salsa»*

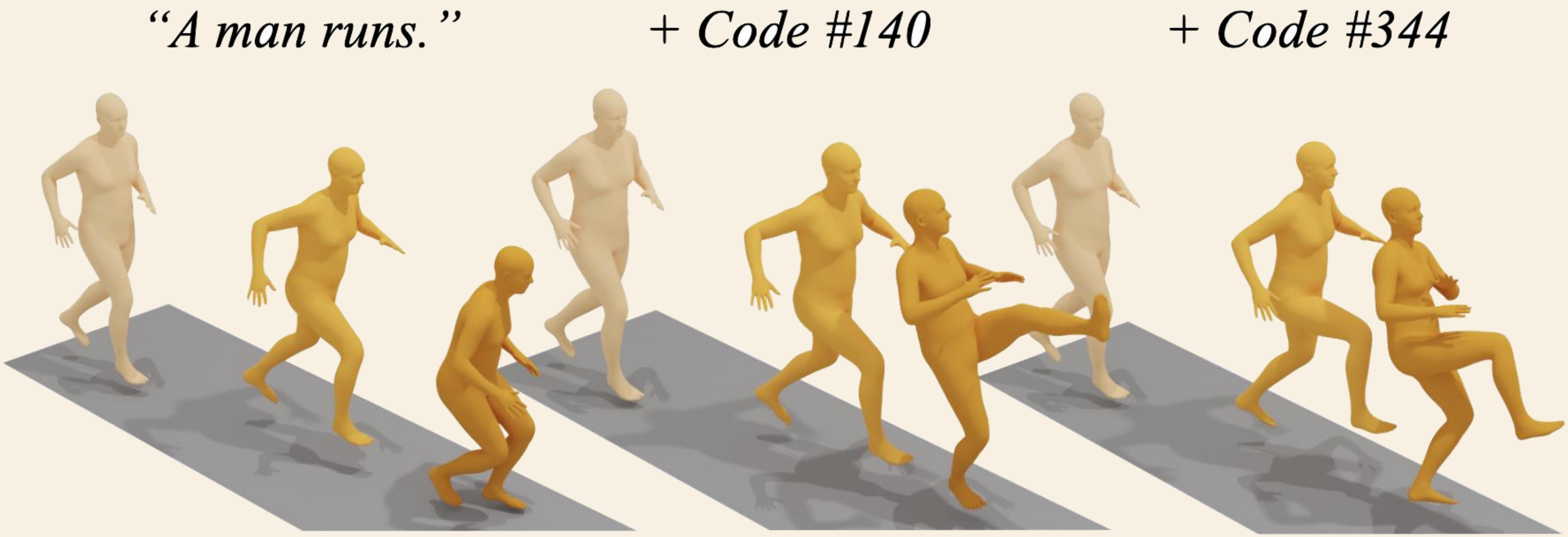
Safe Text-to-Motion

# De Matteis et al. Human Motion Unlearning

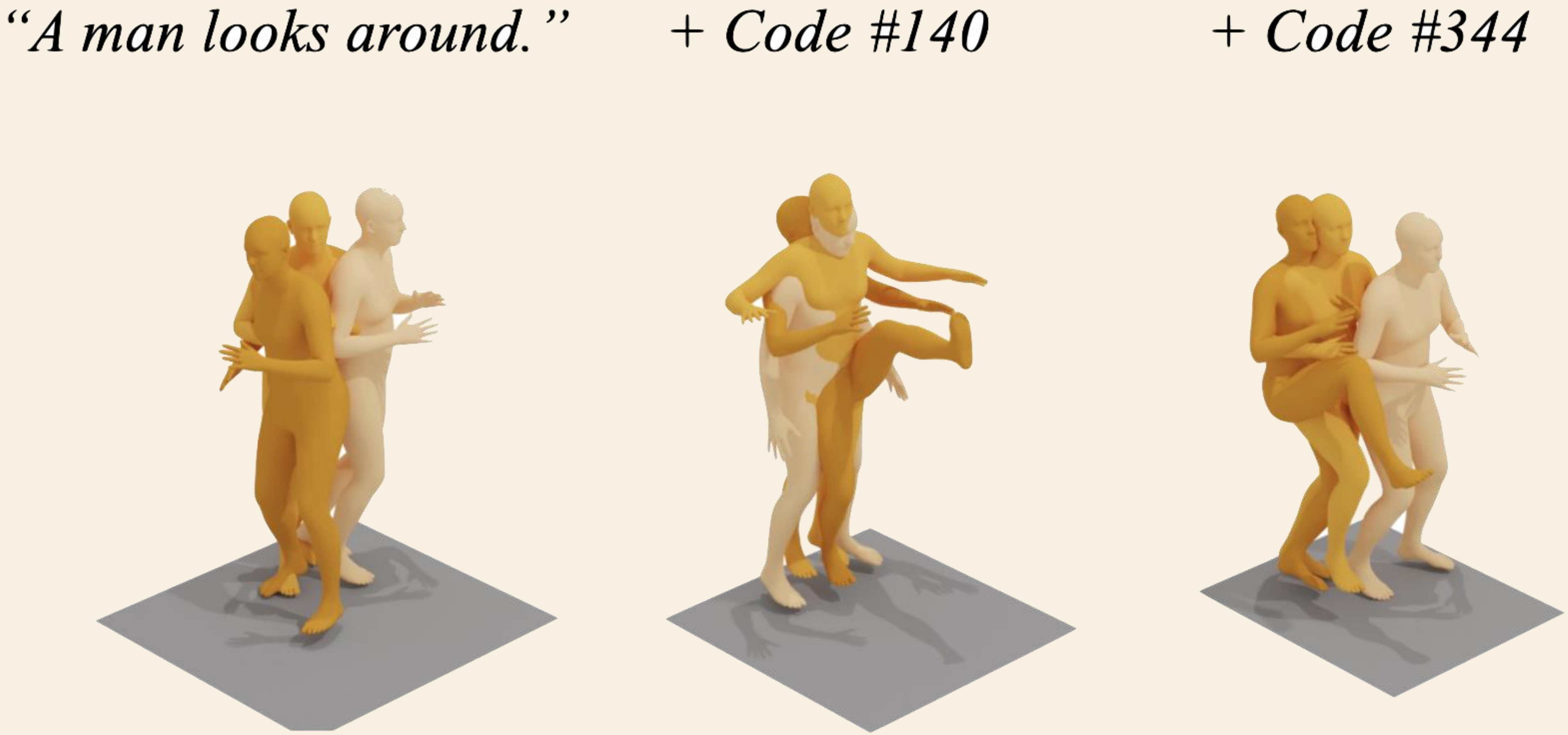
1) Take state of the art Text-to-Motion models (MoMask)
2) Take state of the art Unlearning algorithms
    1) UCE Gandikota et al. Unified Concept Editing in Diffusion Models WACV 2024
    2) RECE Gong et al. Reliable and efficient concept erasure of text-to-image diffusion models ECCV2024
3) Build a benchmark and propose a solution crafted for T2M.

For point 3, we focused on creating an unlearning algorithm that tackles unlearning in VQ-VAE models.

# Codebooks contain disentangled concepts and can be identified



# Does our assumption hold?



*"A man looks around."*

*+ Code #140*

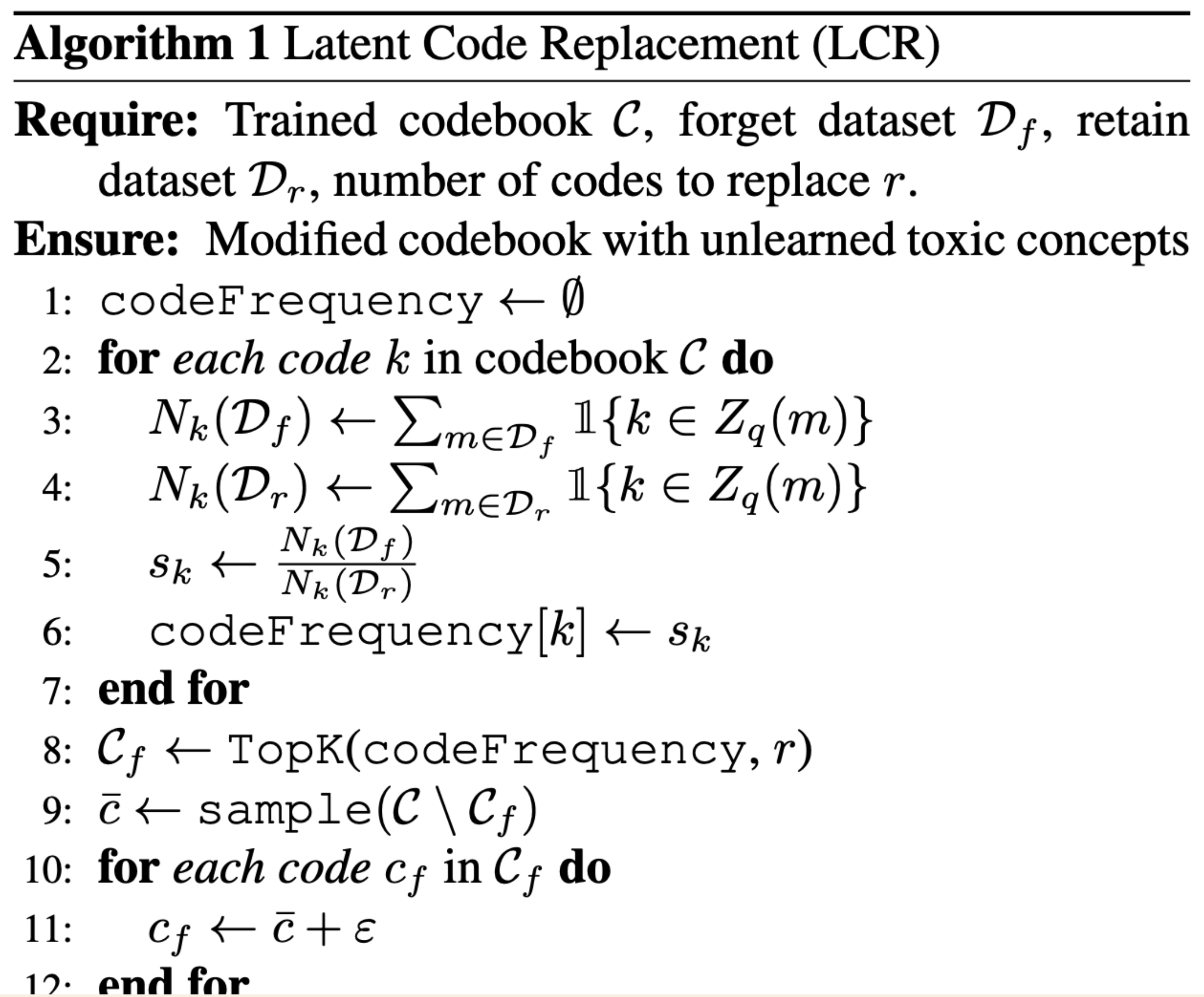
*+ Code #344*

# LCR

The Forget set contains all the violent motions present in our training data.

The Retain set is the training data without the violent motions.

We overwrite the violent codebooks to some other value which is close by but safe.



**Algorithm 1** Latent Code Replacement (LCR)

**Require:** Trained codebook $\mathcal{C}$, forget dataset $\mathcal{D}_f$, retain dataset $\mathcal{D}_r$, number of codes to replace $r$.

**Ensure:** Modified codebook with unlearned toxic concepts

```
codeFrequency $\leftarrow \emptyset$
for each code $k$ in codebook $\mathcal{C}$ do
    $N_k(\mathcal{D}_f) \leftarrow \sum_{m \in \mathcal{D}_f} \mathbb{1}\{k \in Z_q(m)\}$
    $N_k(\mathcal{D}_r) \leftarrow \sum_{m \in \mathcal{D}_r} \mathbb{1}\{k \in Z_q(m)\}$
    $s_k \leftarrow \frac{N_k(\mathcal{D}_f)}{N_k(\mathcal{D}_r)}$
    codeFrequency$[k] \leftarrow s_k$
end for
$\mathcal{C}_f \leftarrow$ TopK(codeFrequency, $r$)
$\bar{c} \leftarrow$ sample($\mathcal{C} \setminus \mathcal{C}_f$)
for each code $c_f$ in $\mathcal{C}_f$ do
    $c_f \leftarrow \bar{c} + \varepsilon$
end for
```

# From text to motion to text to video

- Both condition generative models on **textual descriptions**
- Both involve **sequence generation** with rich temporal dynamics
- Multimodal Alignment (text ↔ dynamics)
- Use of Diffusion, Transformers, Contrastive Learning
- Evaluation metrics: Diversity, Realism, Semantic Accuracy

| Task | Input | Output | Domain |
|---|---|---|---|
| **Text-to-Motion** | Sentence | 3D joint trajectory | Skeleton space |
| **Text-to-Video** | Sentence | RGB video | Pixel space |

# Challenges of video generation unlearning

No access to training data. It is impossible to access the Retain and Forget set (training dataset unknown).

Massive diffusion models (OpenSora), not based on the VQ-VAE. It is impossible to associate toxic concepts to codebooks.

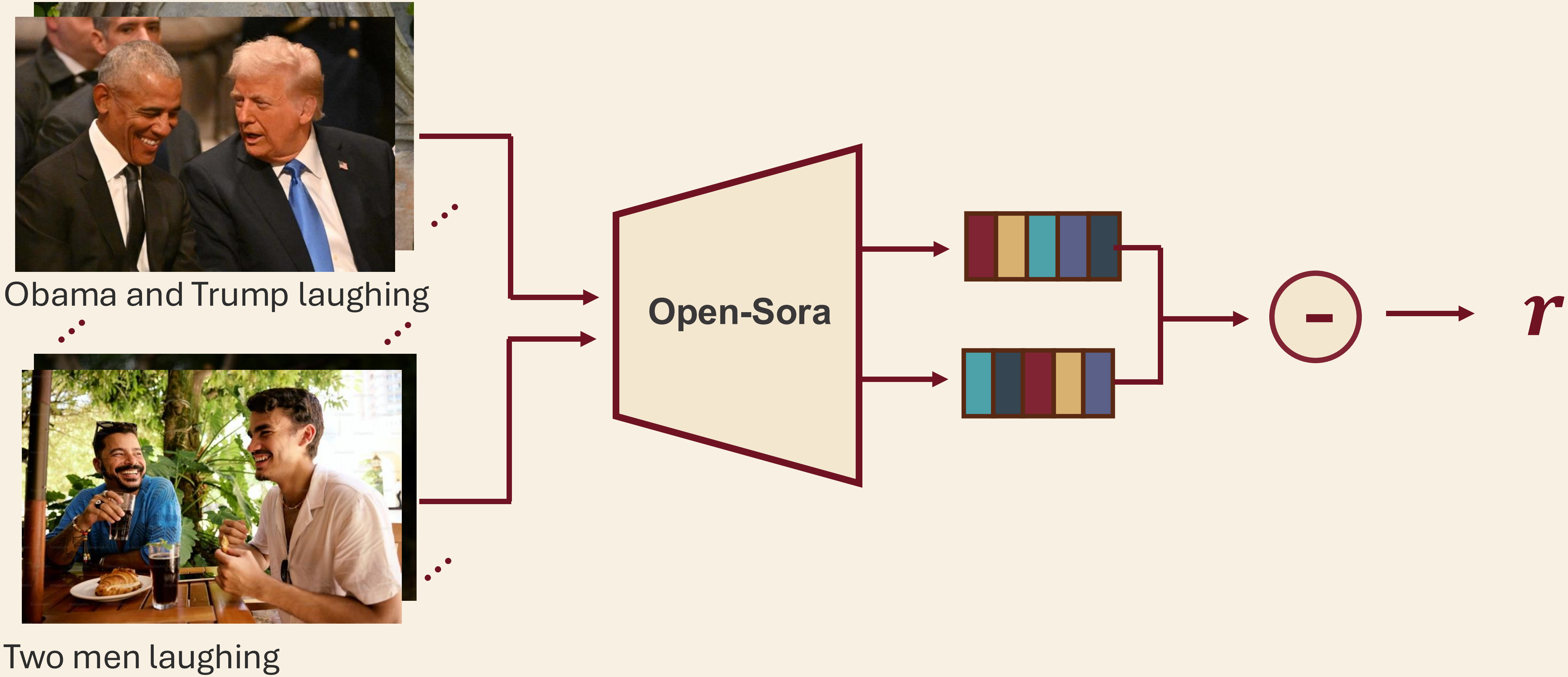
# Facchiano et al. Video Unlearning via Low-Rank Refusal Vectors (Under Review)

Build on the fly retain and forget set of paired text-video:

| | |
|---|---|
| Naked woman bathed in warm sunlight | Woman bathed in warm sunlight |
| Naked man with windswept hair by the sea | Man with windswept hair by the sea |
| Naked woman taking a quiet mirror moment | Woman taking a quiet mirror moment |
| Naked man smiling under tropical blooms | Man smiling under tropical blooms |
| Naked woman shielding her eyes at the shore | Woman shielding her eyes at the shore |

The videos are not shown for their sensitivity.

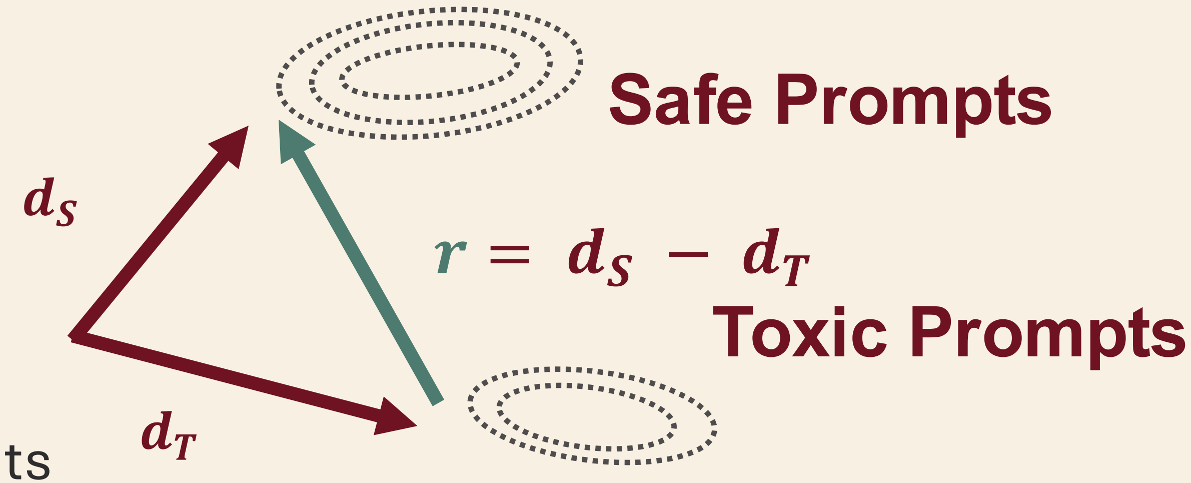
# Extraction of the concept "famous" person

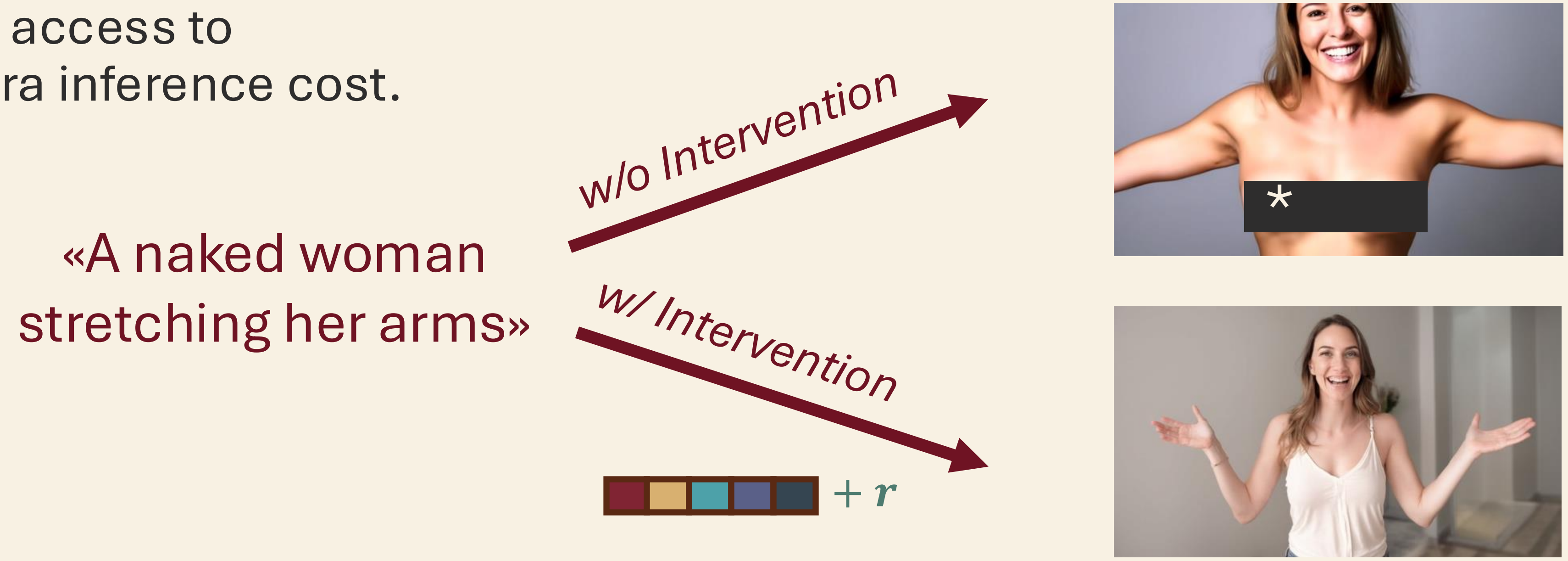


Zeng et al Open-sora: Democratizing efficient video production for all

# Steering

This vector is embedded in the model weights to precisely suppress unwanted concepts like nudity, violence, or copyrighted material, without needing retraining or access to original data, and with no extra inference cost.





# Copyright

A close-up of a Rolex watch under sunlight. 24 FPS. For motion score.

# Privacy

Angela Merkel pointing 24 FPS. For motion score.

# Human Motion Generation: A Pathway to Safer, Physical Intelligence

🧠 Bridging language and physical intelligence

🤖 Enabling embodied agents that understand us

🛡 Ensuring safety and trust in generative models

Buzzwords I'm currently obsessed with:

- ICL for Robotics — using large language models directly for robot behavior and decision-making.
- Safety & Robustness — making sure LLMs and ICL setups don't break in the real world.

# Thank you!

Length-Aware Motion Synthesis via Latent Diffusion
*A Sampieri, A Palma, I Spinelli, F Galasso* ***ECCV*** **2024**

Social EgoMesh Estimation
*L Scofano, A Sampieri, E De Matteis, I Spinelli, F Galasso* ***WACV*** **2025**

MonSTeR : a Unified Model for Motion, Scene, Text Retrieval
*L Collorone, M Gioia, M Pappa, P Leoni, G Ficarra, O Litany, I Spinelli, F Galasso* **ICCV 2025**

Following the Human Thread in Social Navigation
*L Scofano, A Sampieri, T Campari, V Sacco, I Spinelli, L Ballan, F Galasso* ***ICLR*** **2025**

Human Motion Unlearning
*E De Matteis, M Migliarini, A Sampieri, I Spinelli, F Galasso* **Preprint**

Video Unlearning via Low-Rank Refusal Vector
*S Facchiano, S Saravalle, M Migliarini, E De Matteis, A Sampieri, A Pilzer, E Rodolà, I Spinelli, L Franco, F Galasso* **Preprint**