Digital Humans – Winter 24/25

Lecture 13_2 – Motion Synthesis with Diffusion

Prof. Dr. Gerard Pons-Moll

University of Tübingen / MPI-Informatics

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

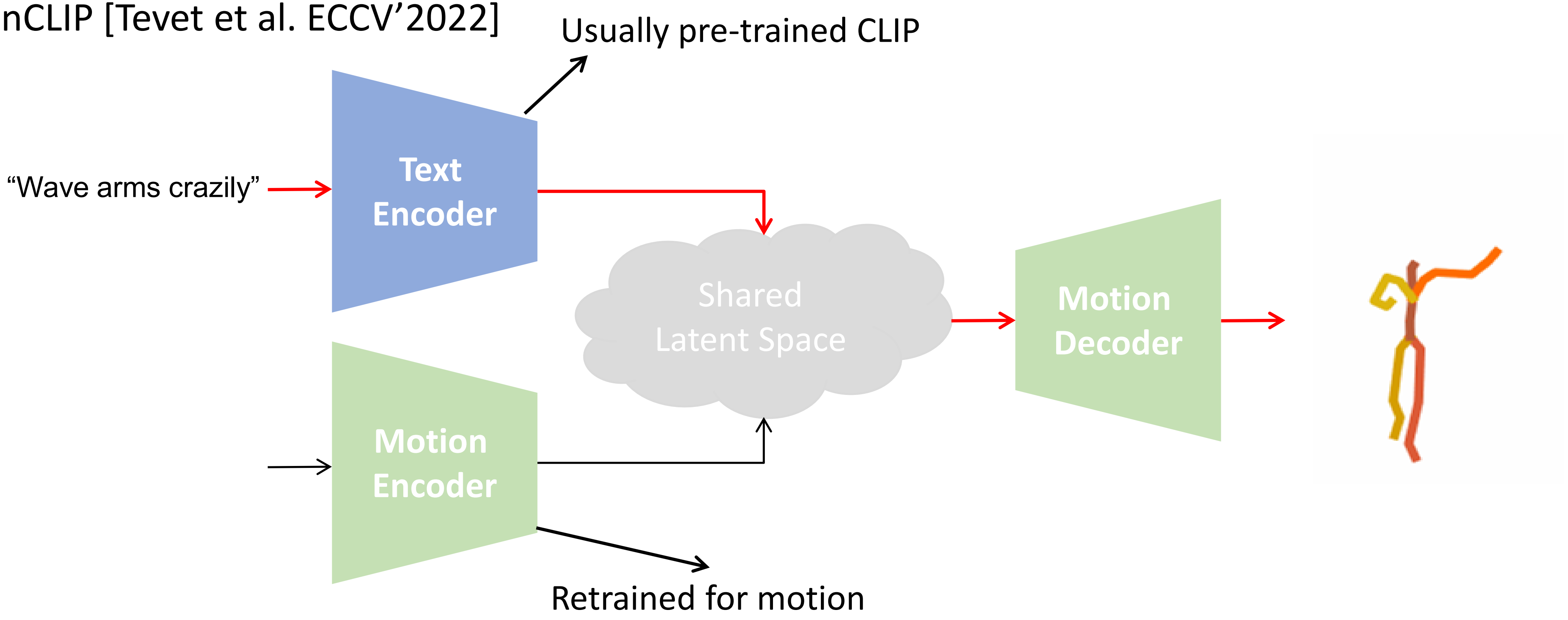
Main contents

- **Human motion diffusion model.**
  - **Text to motion generation.**
- Compositional motion generation with pretrained motion prior.
  - Long sequence.
  - Two-person interaction.
  - Trajectory control.
- Unified human motion synthesis and understanding with fine-grained semantics.
  - Bidirectional: text to motion and motion to text.
  - Hierarchical semantics: global and local text.

Classic way for text to motion generation

Text-to-motion using the VAE framework:

- TEMOS [Petrovich et al. ECCV'2022]
- T2M [Guo et al. CVPR'2022]
- MotionCLIP [Tevet et al. ECCV'2022]



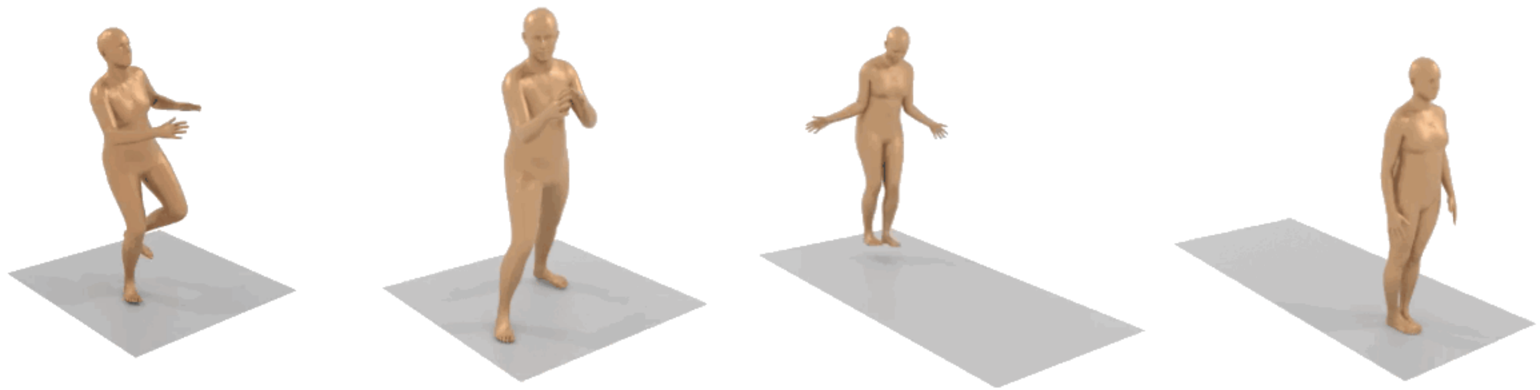
TEL AVIV UNIVERSITY

Human Motion Diffusion Model

Guy Tevet Sigal Raab Brian Gordon Yonatan Shafir

Daniel Cohen-Or Amit H. Bermano

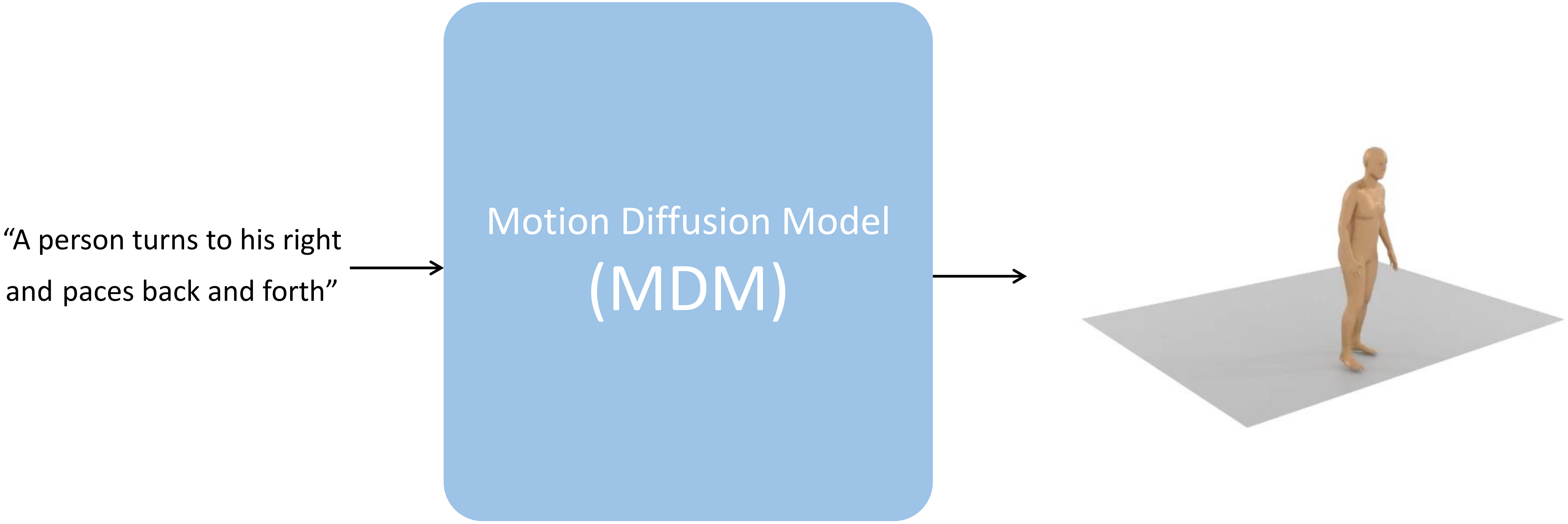
Tel Aviv University, Israel



"a man kicks with something or someone with his left leg." "A person punches in a manner consistent with martial arts." "A person is skipping rope." "a person walks backwards slowly."

MDM: A Human Motion Framework

Goal: given text, generate plausible human motion.



Tevet et al. MDM, ICLR’23.

MDM: A Human Motion Framework

High Quality

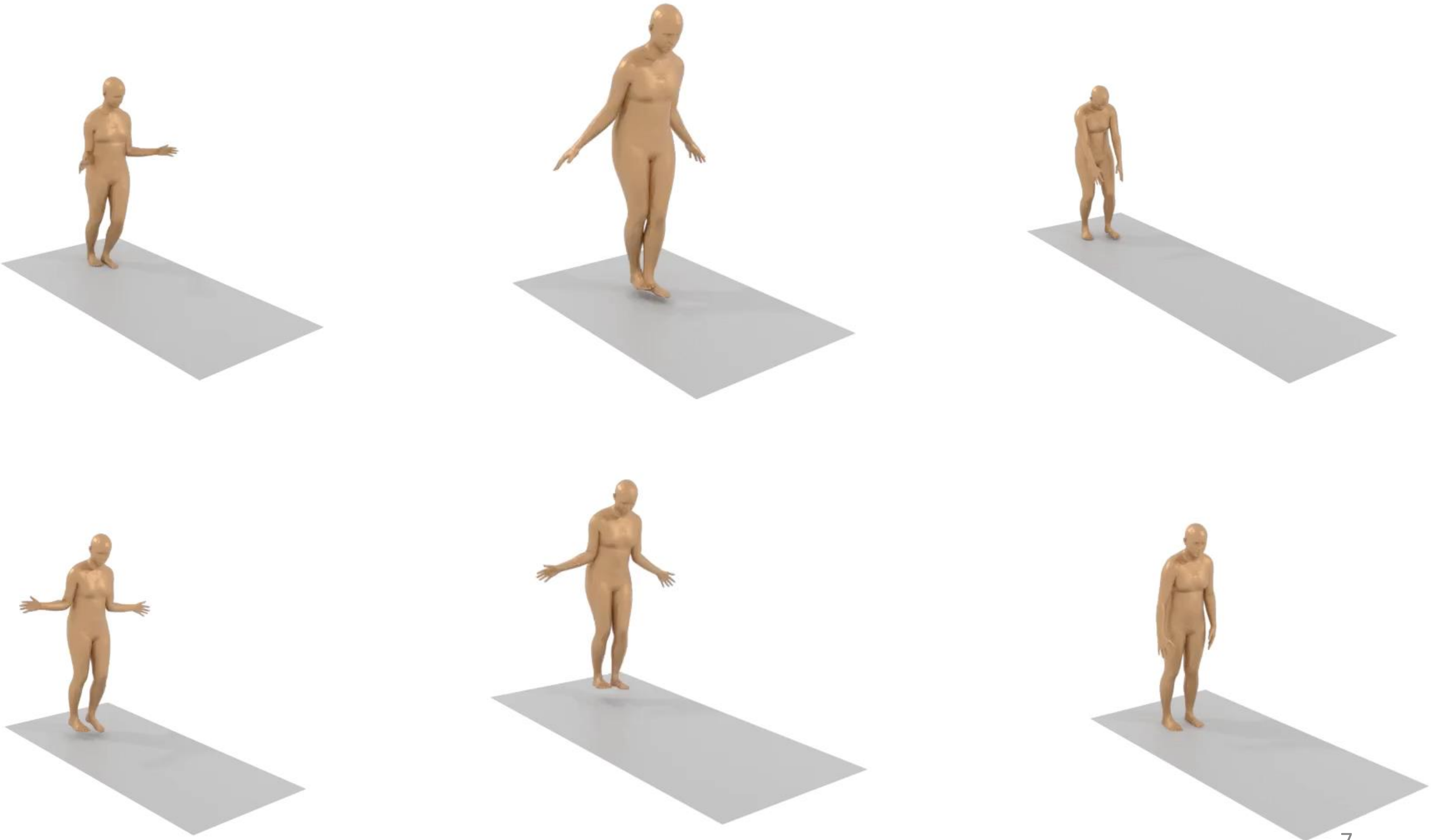
Global Position

Tevet et al. MDM, ICLR’23.

MDM: A Human Motion Framework

Variability

“A person is skipping rope.”

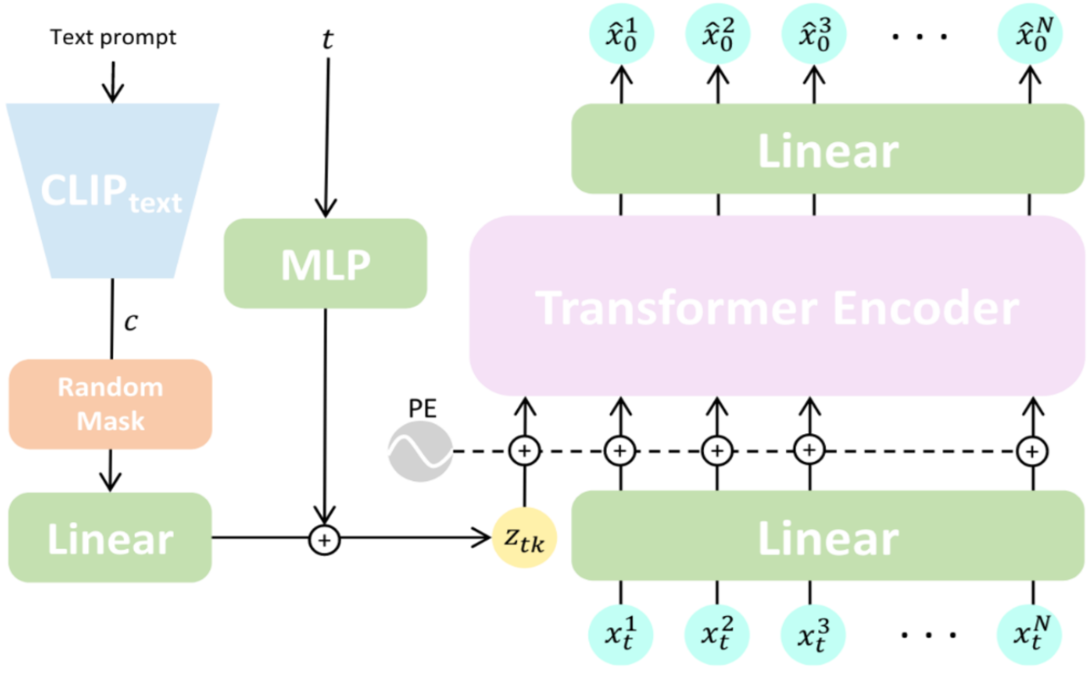


Tevet et al. MDM, ICLR’23.

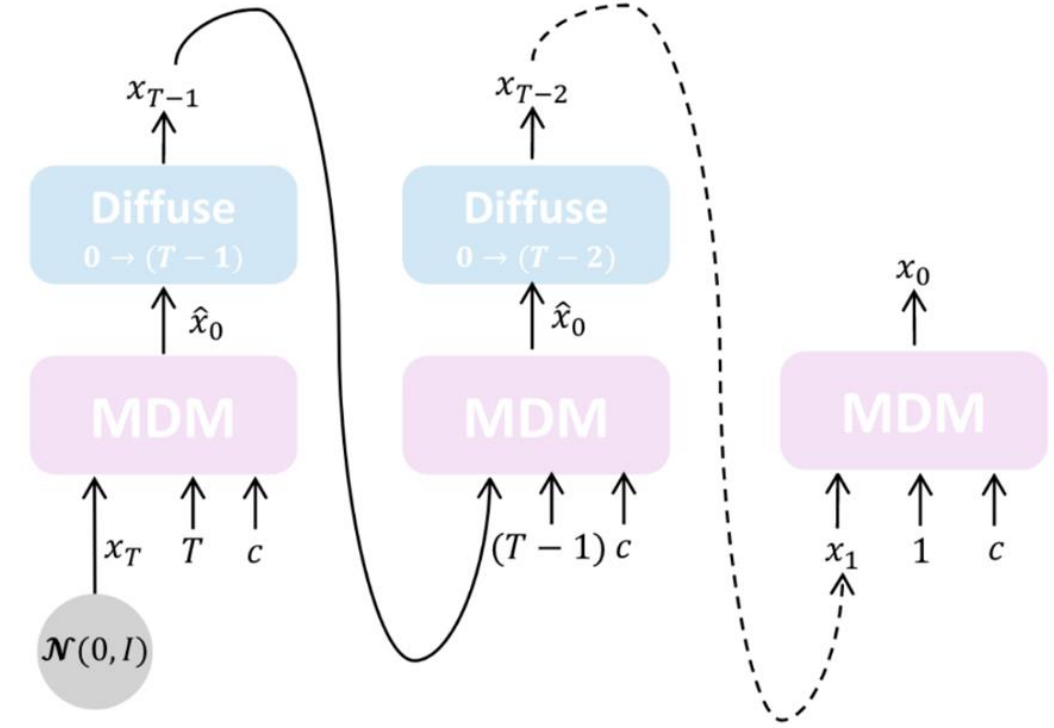
MDM framework

- CLIP text encoder + transformer based diffusion.
- Classifier-free guidance:
  - Training: randomly mask out text conditions.
  - Sampling: weighted combination of conditional and unconditional predictions.

MDM network architecture



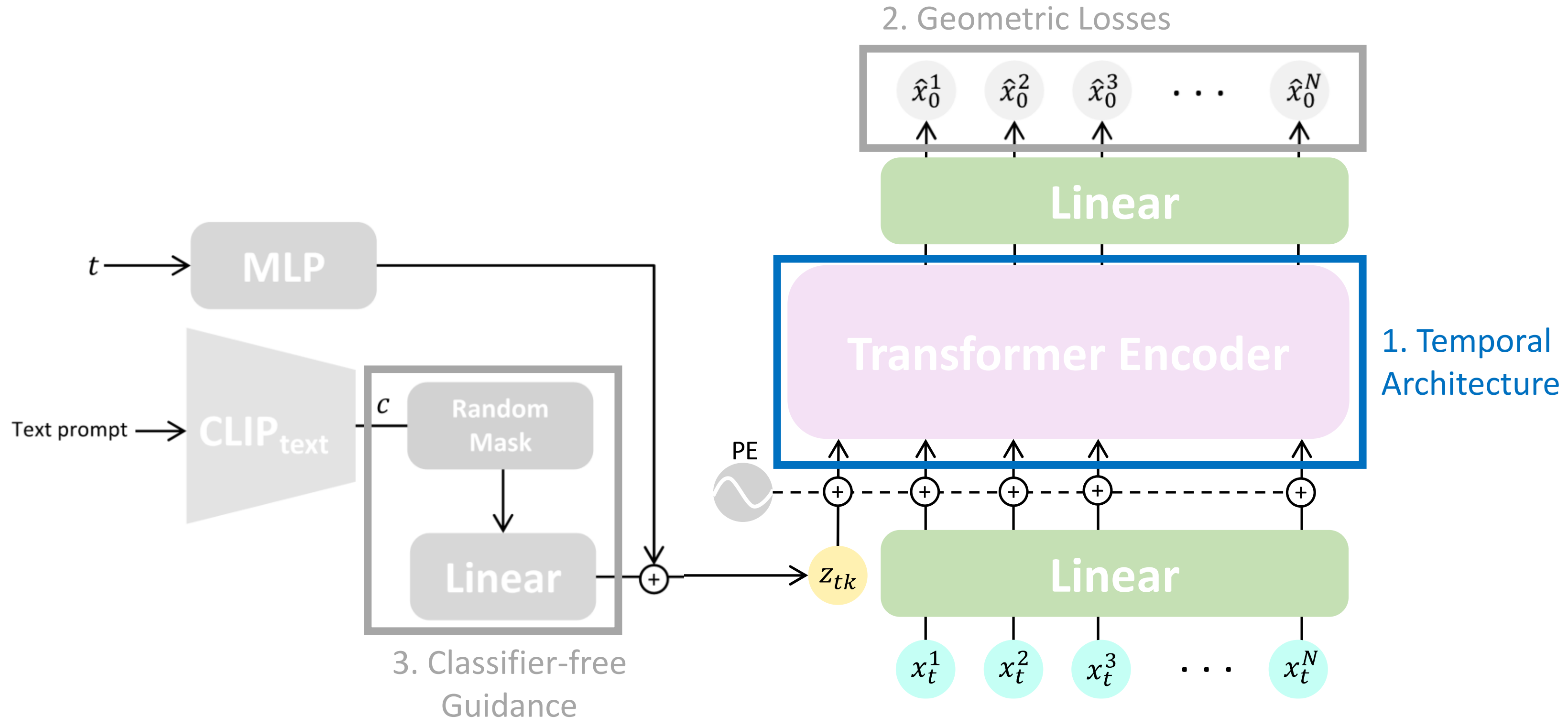
MDM reverse sampling

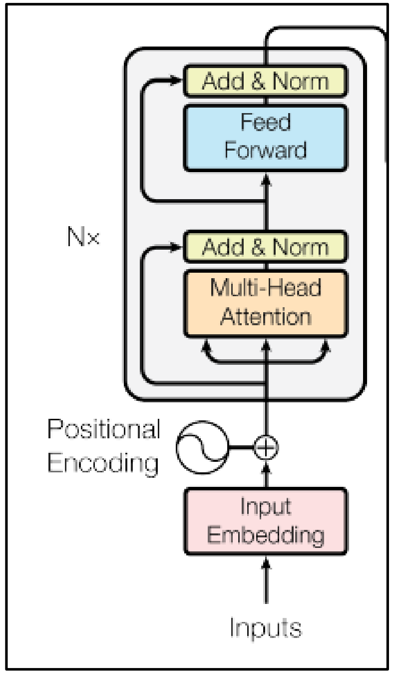


Tevet et al. MDM, ICLR'23.

MDM architecture in more detail

- 1. Temporal architecture: encode time information via transformer.
- 2. Geometric representation $x_0^i \in \mathbb{R}^{J \times D}$: joint rotations or positions.





Transformer encoder
[Vaswani et al. 2017]

Tevet et al. MDM, ICLR'23.

MDM training losses

- Diffusion loss on x_0 and geometric losses.

$$\mathcal{L} = \mathcal{L}_{\text{simple}} + \underbrace{\lambda_{\text{pos}}\mathcal{L}_{\text{pos}} + \lambda_{\text{vel}}\mathcal{L}_{\text{vel}} + \lambda_{\text{foot}}\mathcal{L}_{\text{foot}}}_{\text{Geometric loss terms}}$$

$$\mathcal{L}_{\text{simple}} = E_{x_0 \sim q(x_0|c), t\sim[1,T]}[\|x_0 - G(x_t, t, c)\|_2^2 \longrightarrow$$ Classic diffusion loss

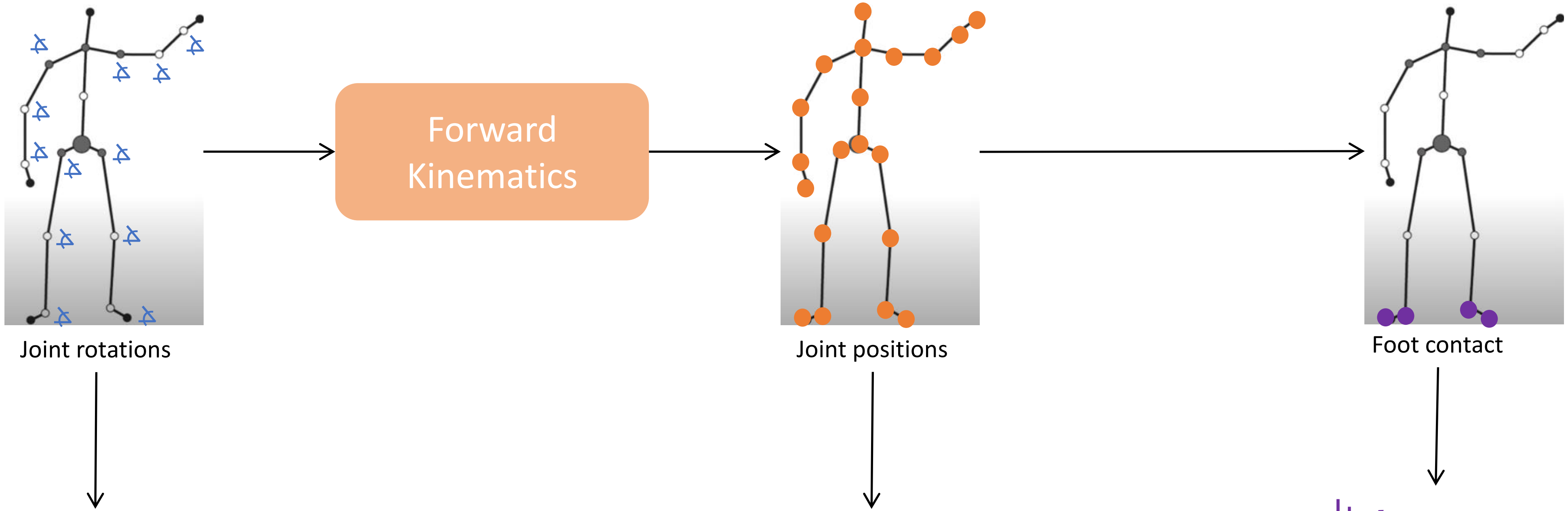
$$\mathcal{L}_{\text{pos}} = \frac{1}{N}\sum_{i=1}^{N} \|FK(x_0^i) - FK(\hat{x}_0^i)\|_2^2, \longrightarrow$$ Joint position after forward kinematics

$$\mathcal{L}_{\text{foot}} = \frac{1}{N-1}\sum_{i=1}^{N-1} \|(FK(\hat{x}_0^{i+1}) - FK(\hat{x}_0^i)) \cdot f_i\|_2^2, \longrightarrow$$ Foot contact $f_i \in \{0,1\}^J$

$$\mathcal{L}_{\text{vel}} = \frac{1}{N-1}\sum_{i=1}^{N-1} \|(x_0^{i+1} - x_0^i) - (\hat{x}_0^{i+1} - \hat{x}_0^i)\|_2^2 \longrightarrow$$ Velocity

Tevet et al. MDM, ICLR’23.

Geometric losses visualization



$$L_{recon} = \left\|\hat{X}_0 - X_0\right\|_2^2$$

$$L_{pos} = \left\|FK(\hat{X}_0) - FK(X_0)\right\|_2^2$$

$$L_{foot} = \left\|\frac{d}{dn}FK(\hat{X}_0) \cdot f_{mask}\right\|_2^2$$

Tevet et al. MDM, ICLR'23.

Geometric Losses - Results

MDM
- without geometric losses

Warm-up

Warm-up

Sit

Tevet et al. MDM, ICLR’23.

MDM sampling: classifier-free guidance

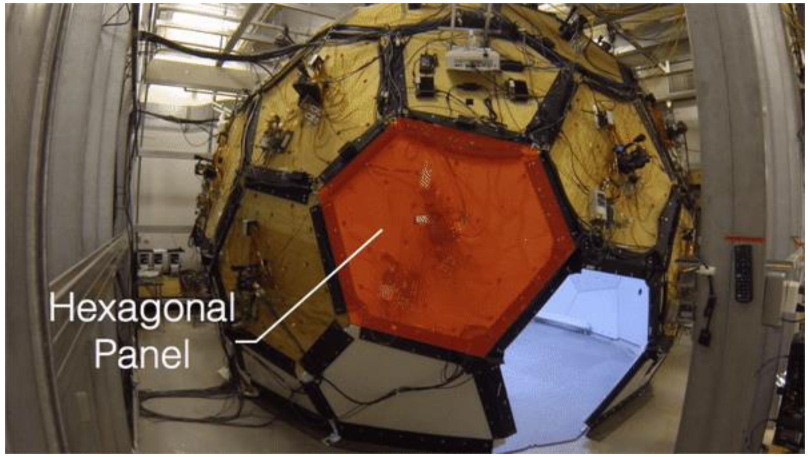
- Iterative reverse sampling with classifier-free guidance.
- Unconditional prediction: $\nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t) = -\frac{1}{\sqrt{1-\overline{\alpha}_t}} \epsilon_\theta(\mathbf{x}_t, t)$
- With condition y: $\nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t, y) = \nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t) + \nabla_{x_t} \log p(\mathrm{y}|\mathbf{x}_t)$
- Guidance:
  - Training time: randomly mask out condition y. MDM: $y =$ the CLIP latent.

$$
\begin{aligned}
\nabla_{\mathbf{x}_t} \log p(y|\mathbf{x}_t) &= \nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t|y) - \nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t) \\
&= -\frac{1}{\sqrt{1-\bar{\alpha}_t}} \Big(\underbrace{\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t, y)}_{\text{Conditional}} - \underbrace{\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)}_{\text{Unconditional}} \Big) \\
\bar{\boldsymbol{\epsilon}}_\theta(\mathbf{x}_t, t, y) &= \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t, y) - \sqrt{1-\bar{\alpha}_t}\, w \nabla_{\mathbf{x}_t} \log p(y|\mathbf{x}_t) \quad \text{Guidance scale w} \\
&= \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t, y) + w\big(\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t, y) - \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)\big) \\
&= (w+1)\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t, y) - w\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)
\end{aligned}
$$

Tevet et al. MDM, ICLR’23.

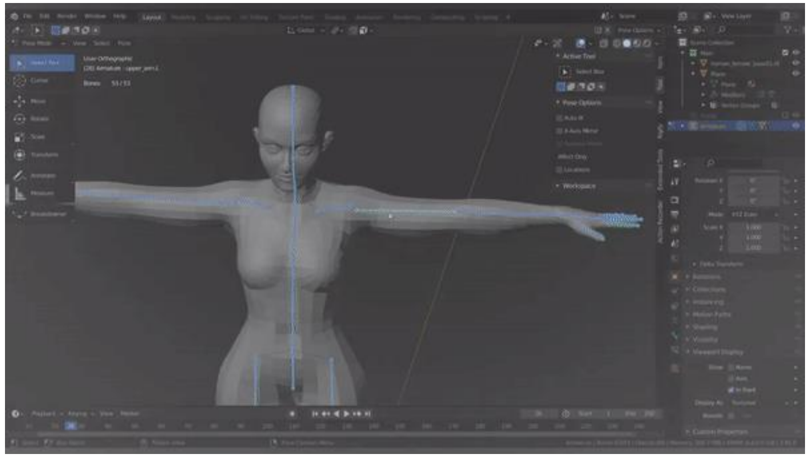
The Data: real capture or artist design.

CMU Panoptic – Motion Capture [Joo et al. 2015]



https://www.youtube.com/watch?v=zQt6g-Jel7M&ab_channel=HanbyulJoo

Real motion capture



https://www.youtube.com/watch?v=antc20EFh6k&t=24s&ab_channel=kfiraberman

Artist design

The Data

HumanML3D [Guo et al. 2022]

- 15K examples
- ~7 sec each

( , “Wave arms crazily”)

( , “Right hand dribble”)

BABEL[Punnakkal et al. 2021]

MDM results: motion quality

Walk | Run | Jump

MDM

ACTOR
[Petrovich et al., 2021]

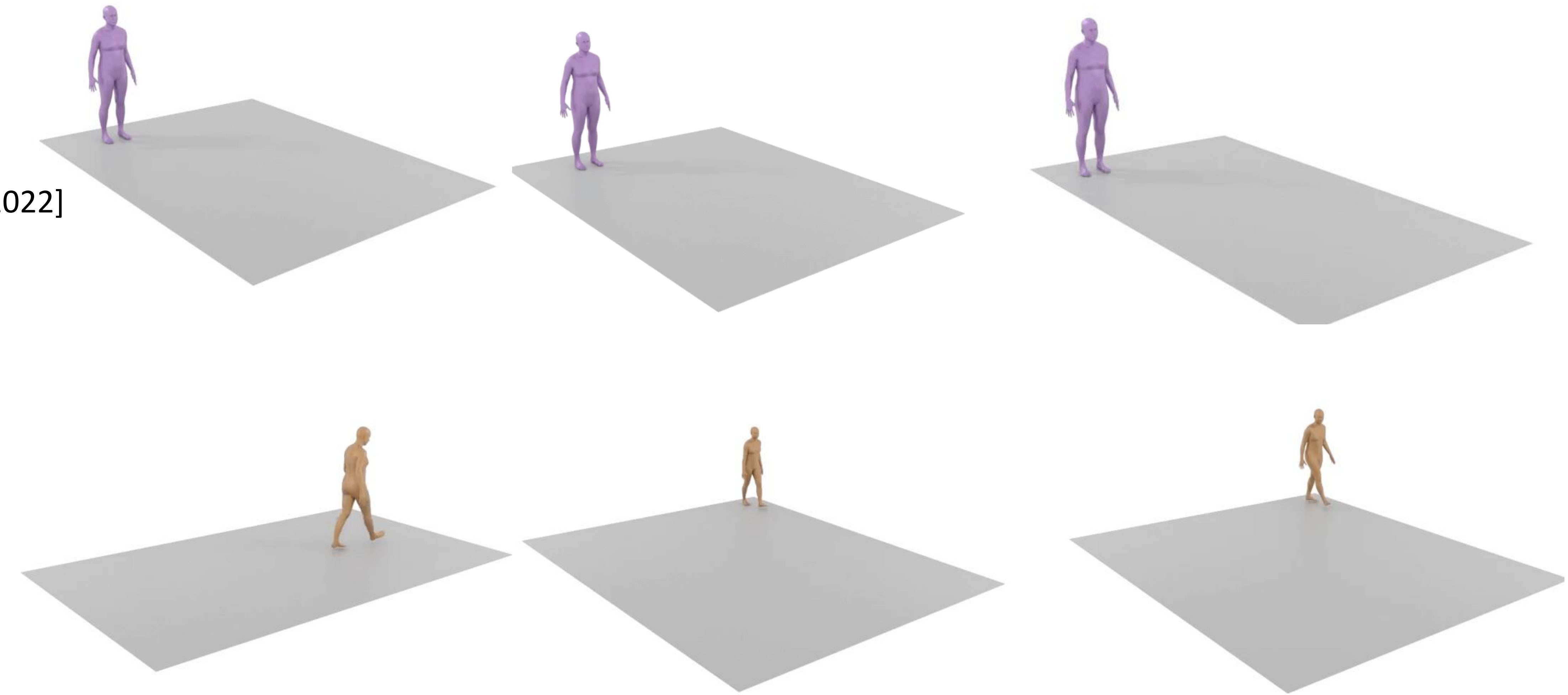
Tevet et al. MDM, ICLR’23.
Petrovich et al, ACTOR, ICCV’21

MDM results: motion diversity

“Person walking in an s shape”

T2M
[Guo et al. 2022]

MDM



Tevet et al. MDM, ICLR’23.
Guo et al, T2M, CVPR’22.

Main contents

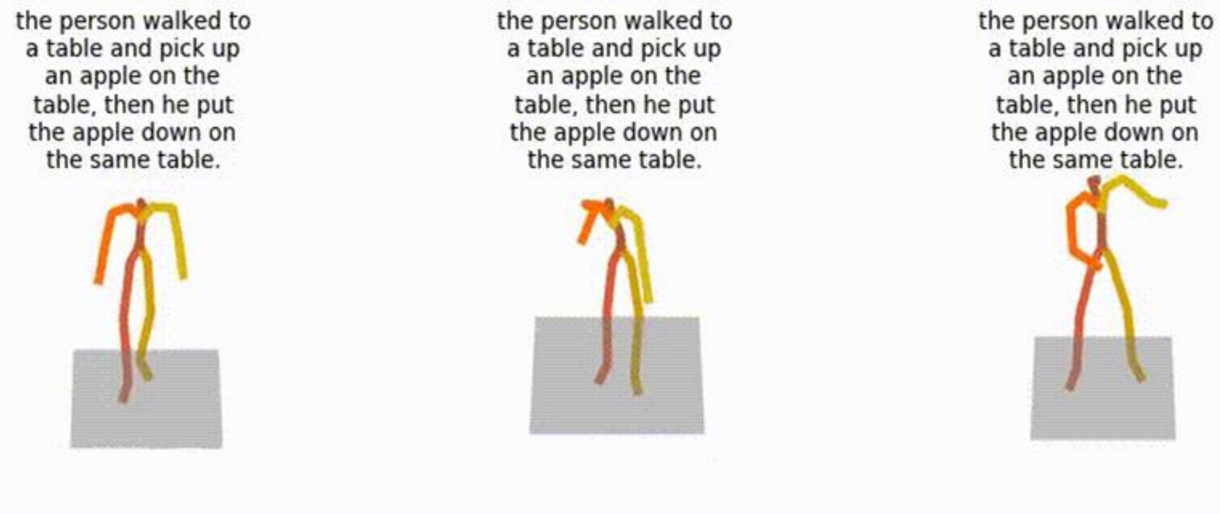
- Human motion diffusion model.
  - Text to motion generation.
- **Compositional motion generation with pretrained motion prior.**
  - **Long sequence.**
  - **Two-person interaction.**
  - **Trajectory control.**
- Unified human motion synthesis and understanding with fine-grained semantics.
  - Bidirectional: text to motion and motion to text.
  - Hierarchical semantics: global and local text.

MDM limitations

- Maximum generation length 196 frames (9.8 seconds).
- No condition on the location.
- Results not satisfying for human-object-interaction prompts.

No spatial control

Cannot generate complex long motion



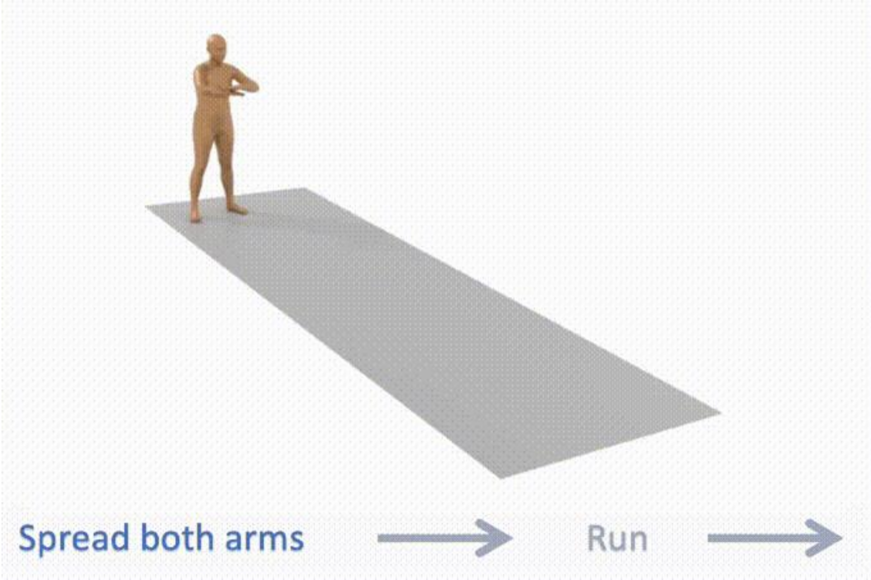
Tevet et al. MDM, ICLR’23.

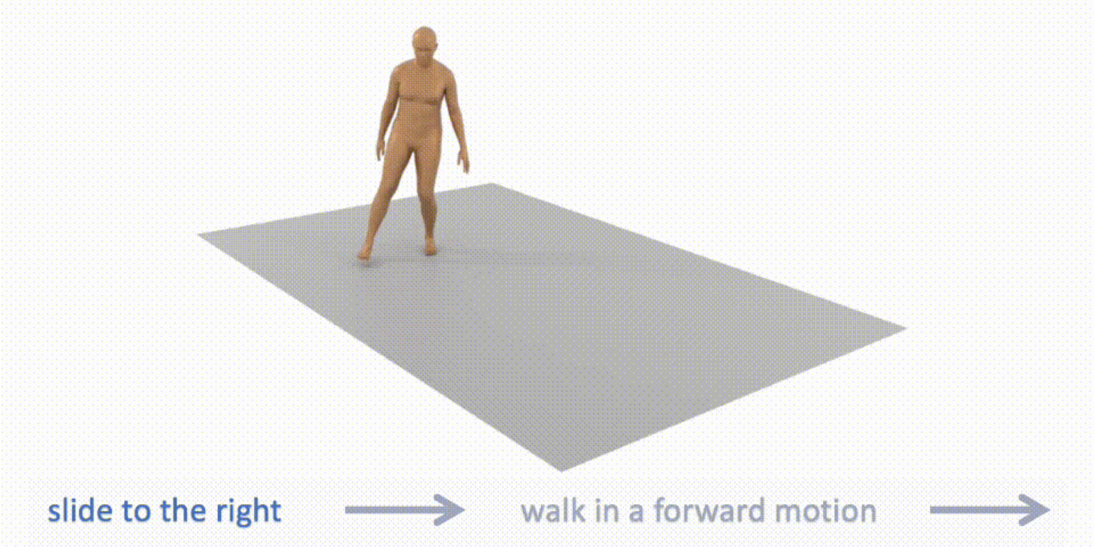
The data problem again

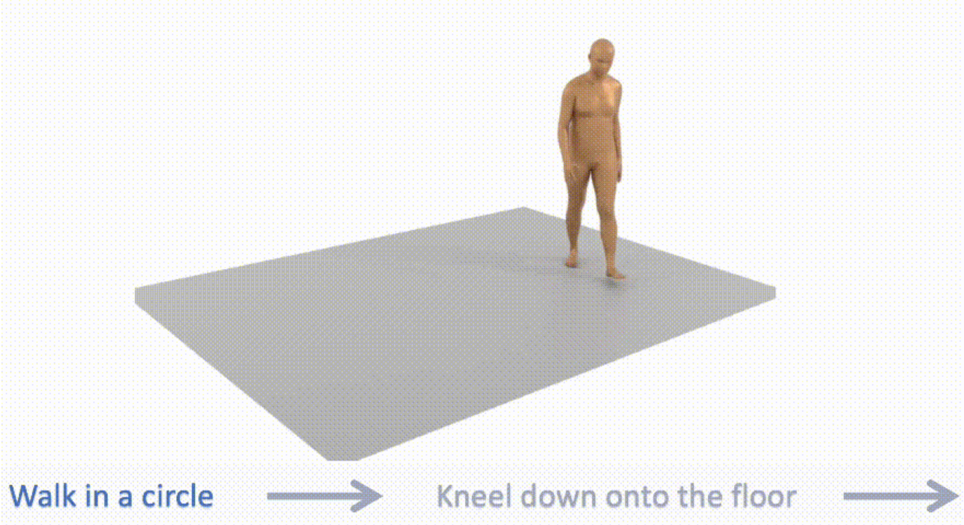
- Motion data is very expensive to obtain:
  - Real motion capture.
  - Designed by artists.
  - Training data for MDM: almost exclusively of short, single person sequences.
- Complex motions are compositional:
  - Very long motions.
  - Interaction between humans, or human and object.
  - Diverse control signals like text and spatial.
  - Can we use compose motions from pretrained motion priors?

PriorMDM: Human Motion Diffusion as a Generative Prior

Long Motions



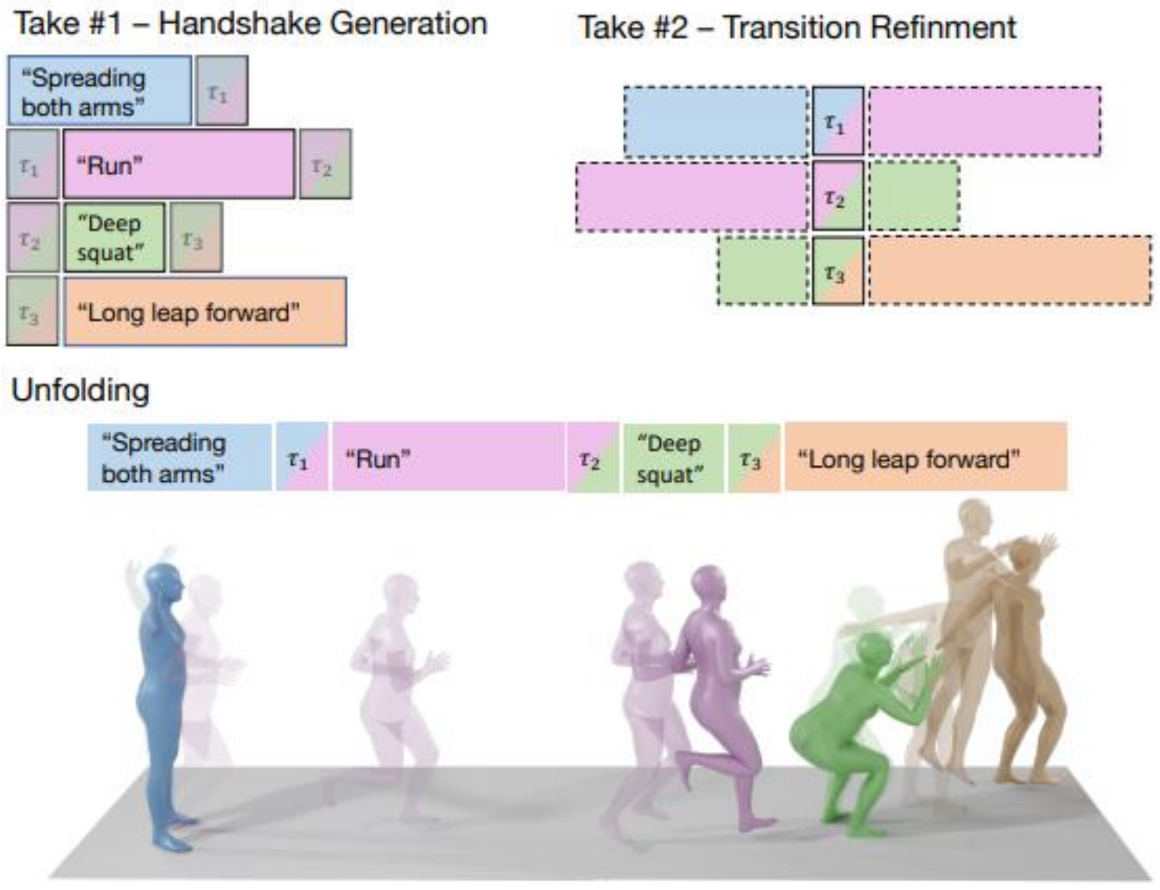




Shafir & Tevet et al. PriorMDM, ICLR’24.

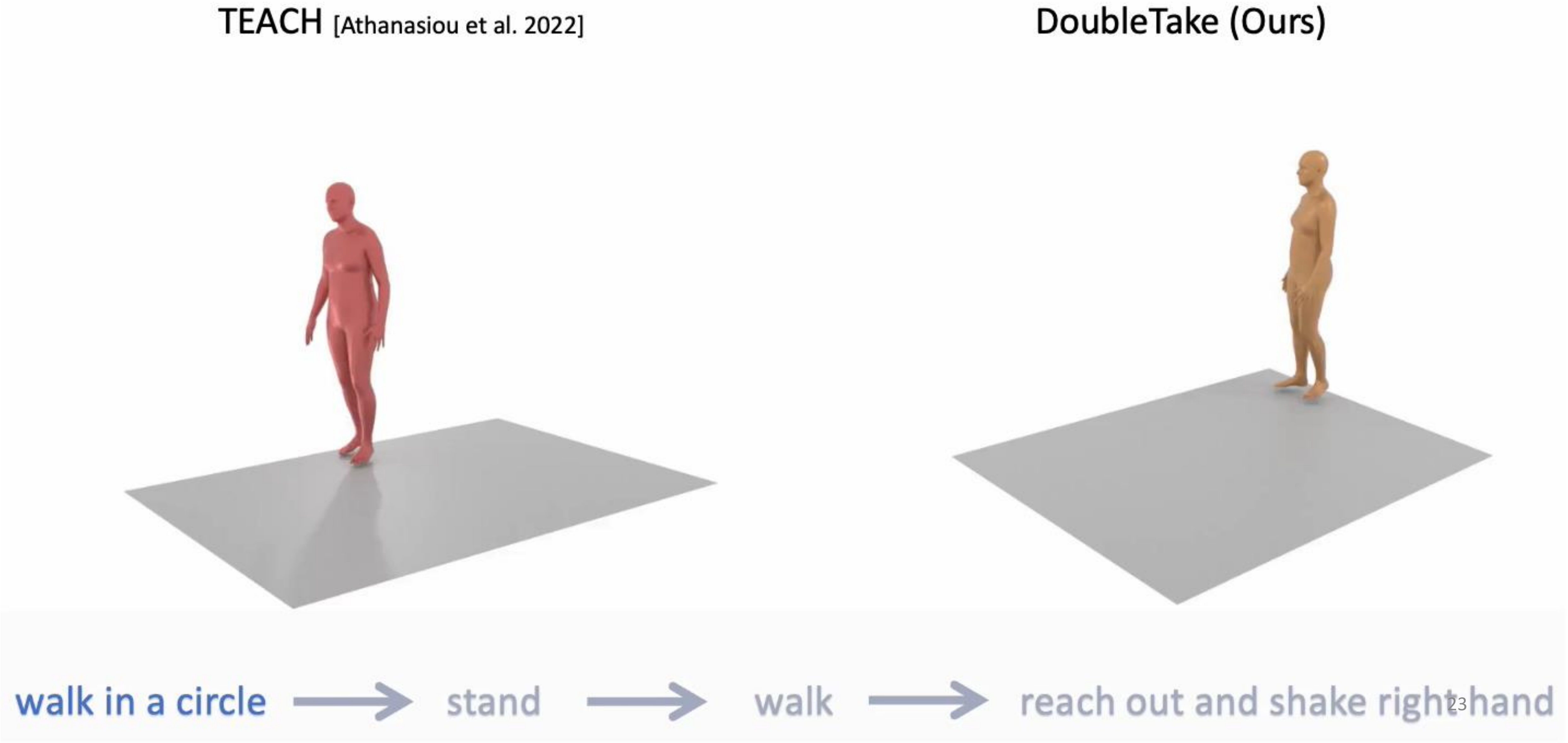
DoubleTake: composition in temporal domain

- Independent denoising + mixed diffusion in transition regions.
- Each temporal window is conditioned on different text.
- Communicate between segments via handshake (~1s/window):
  - Inside the handshake τ: average of previous suffix and current prefix.
- Second take: add noise to handshake part and denoise back.



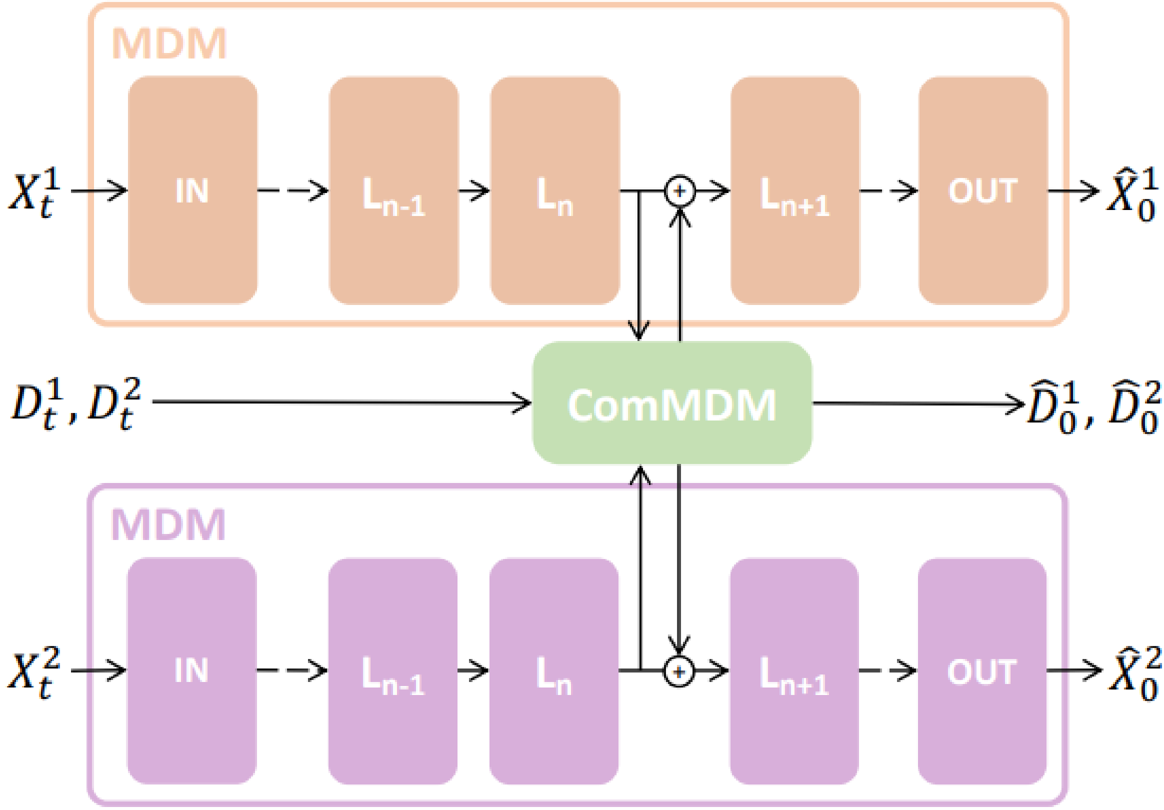
Shafir & Tevet et al. PriorMDM, ICLR'24.

DoubleTake results



Two person motion generation

- Compositional: keep individual models and learn the interaction.
- ComMDM: learns to correct the output from individual models.
- Few-shot learning: trained on 55 two-person motion sequences.

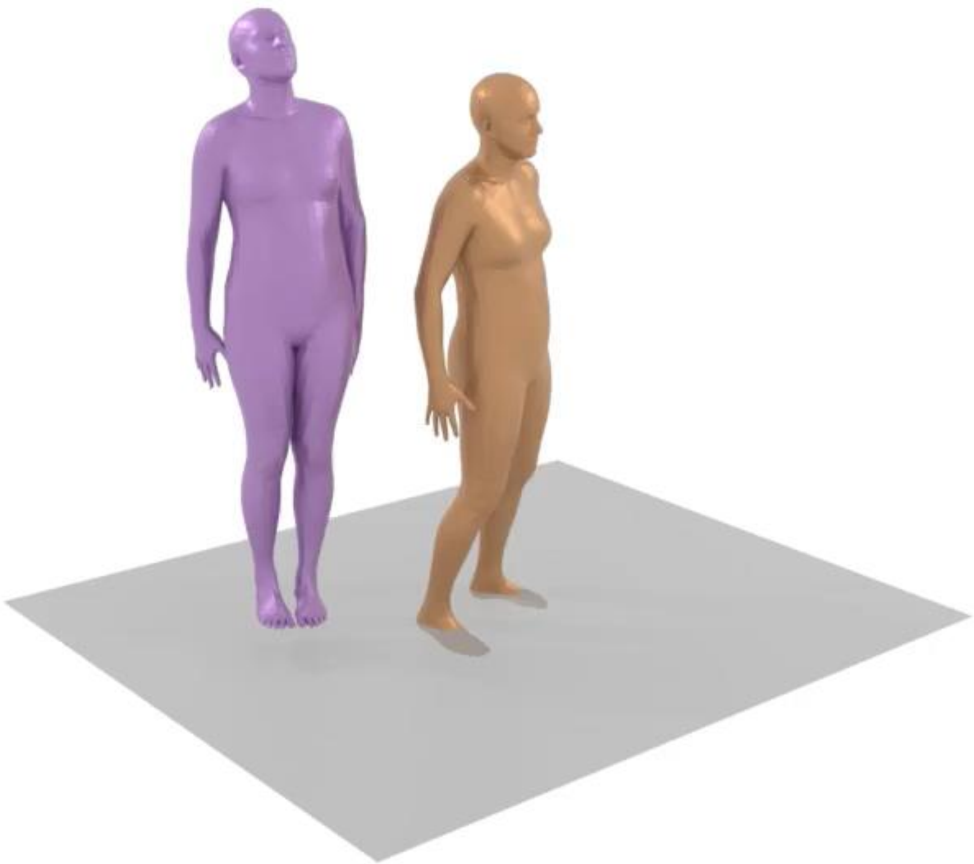


Shafir & Tevet et al. PriorMDM, ICLR’24.

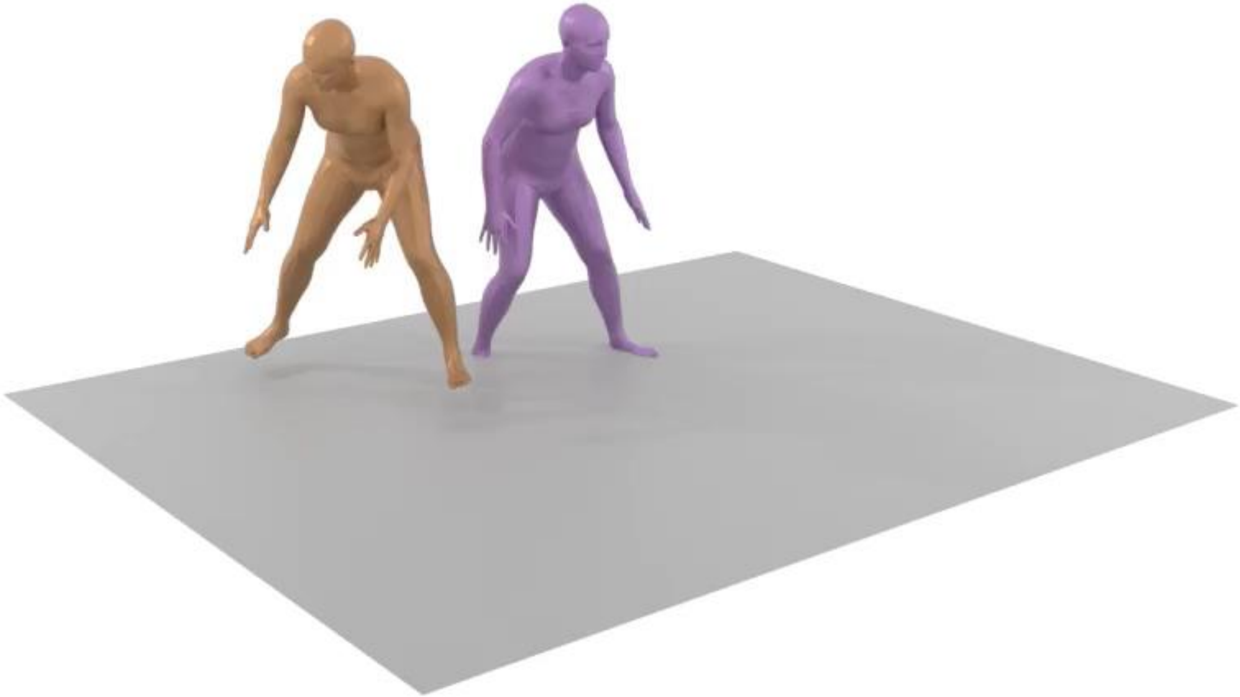
Shafir & Tevet et al. PriorMDM, ICLR'24.

ComMDM for two person interaction

"A Capoeira practice. One is kicking and the other is avoiding the kick."



"The two people are playing basketball, one with the ball the other is defending."



MDM fine tuned for trajectory control

- Goal: control the motion with given trajectory.
- Training: fine tune the model to denoise only free body parts.
- Sampling: set hard constraint on the given trajectory.

Algorithm 1 Fine-tuning method

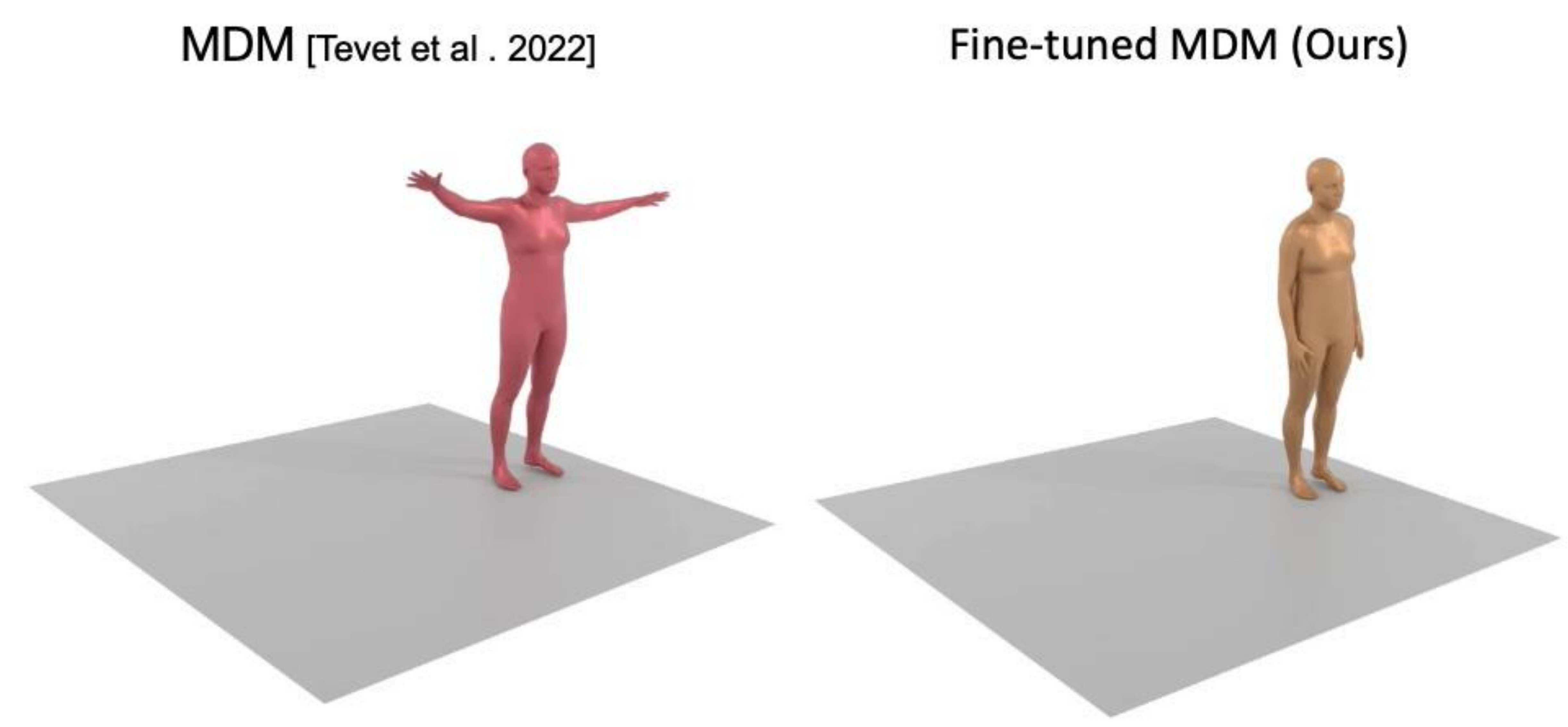
repeat
  $x_0 \sim q(x_0)$
  $t \sim$ Uniform$(\{1, \ldots, T\})$
  $\epsilon \sim \mathcal{N}(0, I)$
  $\epsilon[trajectory] = 0$ ▷ **Our addition**
  Take gradient descent step on:
    $\nabla_\theta \| x_0 - \epsilon_\theta(\sqrt{\bar{\alpha}_t} x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, t) \|$
until Converged

Algorithm 2 Sampling method

$x_0^{(T)} = 0$
for $t = T, \ldots, 0$ **do**
  $x_0^{(t)}[trajectory] =$given trajectory ▷
Original in-painting
  $\epsilon \sim \mathcal{N}(0, I)$
  $\epsilon[trajectory] = 0$ ▷ **Our addition**
  $x_0^{(t-1)} = \epsilon_\theta(\sqrt{\bar{\alpha}_t} x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, t)$
end for

Shafir & Tevet et al. PriorMDM, ICLR’24.

MDM fine tuned for trajectory control

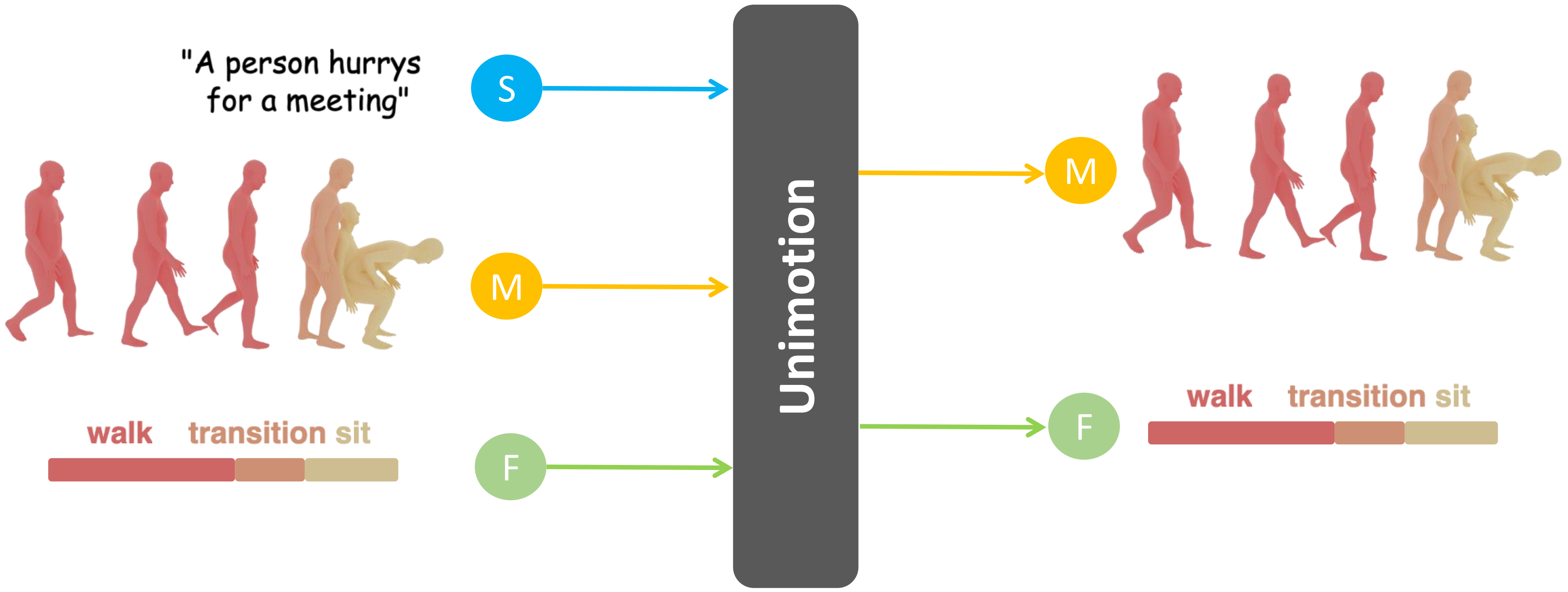


Shafir & Tevet et al. PriorMDM, ICLR’24.

Main contents

- Human motion diffusion model.
  - Text to motion generation.
- Compositional motion generation with pretrained motion prior.
  - Long sequence.
  - Two-person interaction.
  - Trajectory control.
- **Unified human motion synthesis and understanding with fine-grained semantics.**
  - **Bidirectional: text to motion and motion to text.**
  - **Hierarchical semantics: global and local text.**

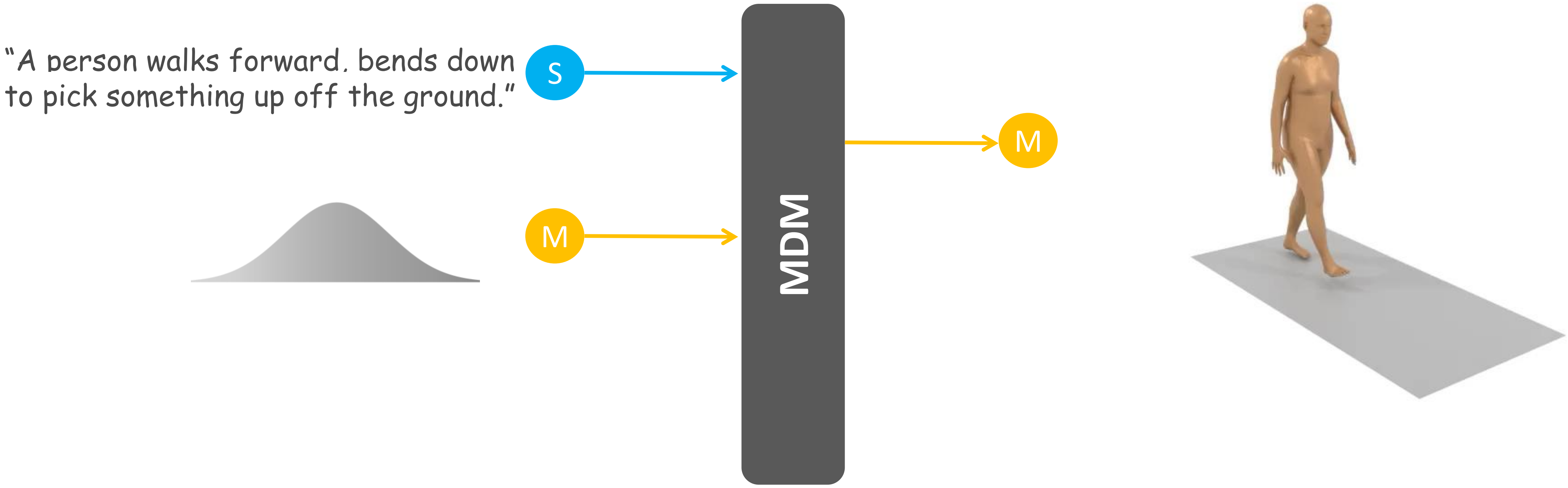
Goal: Unify **motion tasks** into a single model



Li et al. UniMotion, 3DV'25

Multi-tasks

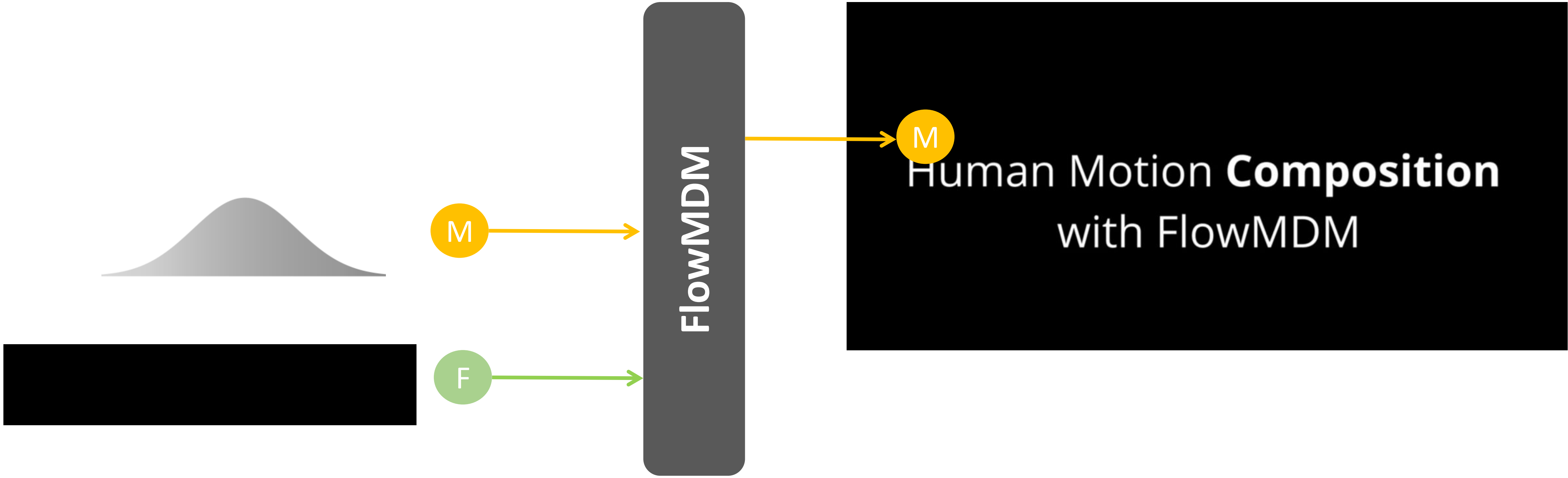
- Sequence-level Text-to-Motion Generation



MDM: Tevet et al, ICLR' 23

Multi-tasks

- Sequence-level Text-to-Motion Generation
- Frame-level Text-to-Motion Generation

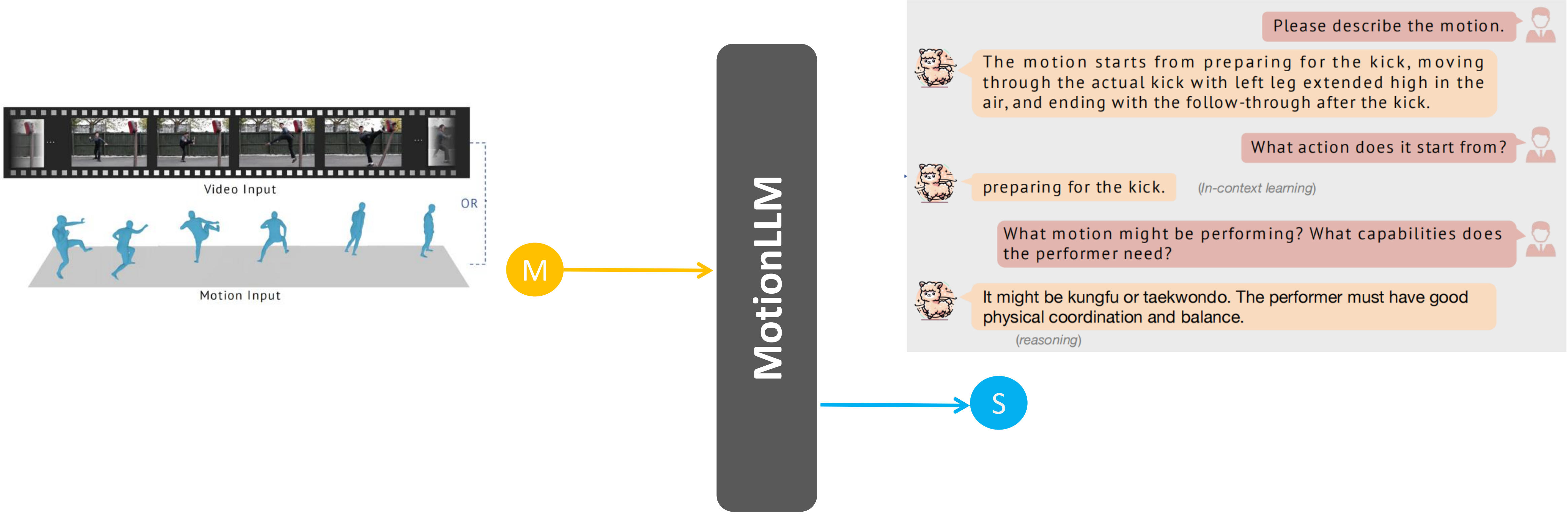


FlowMDM: Barquero et al, CVPR’ 24

Li et al. UniMotion, 3DV’25

Multi-tasks

- Sequence-level Text-to-Motion Generation
- Frame-level Text-to-Motion Generation
- Motion-to-Text Understanding

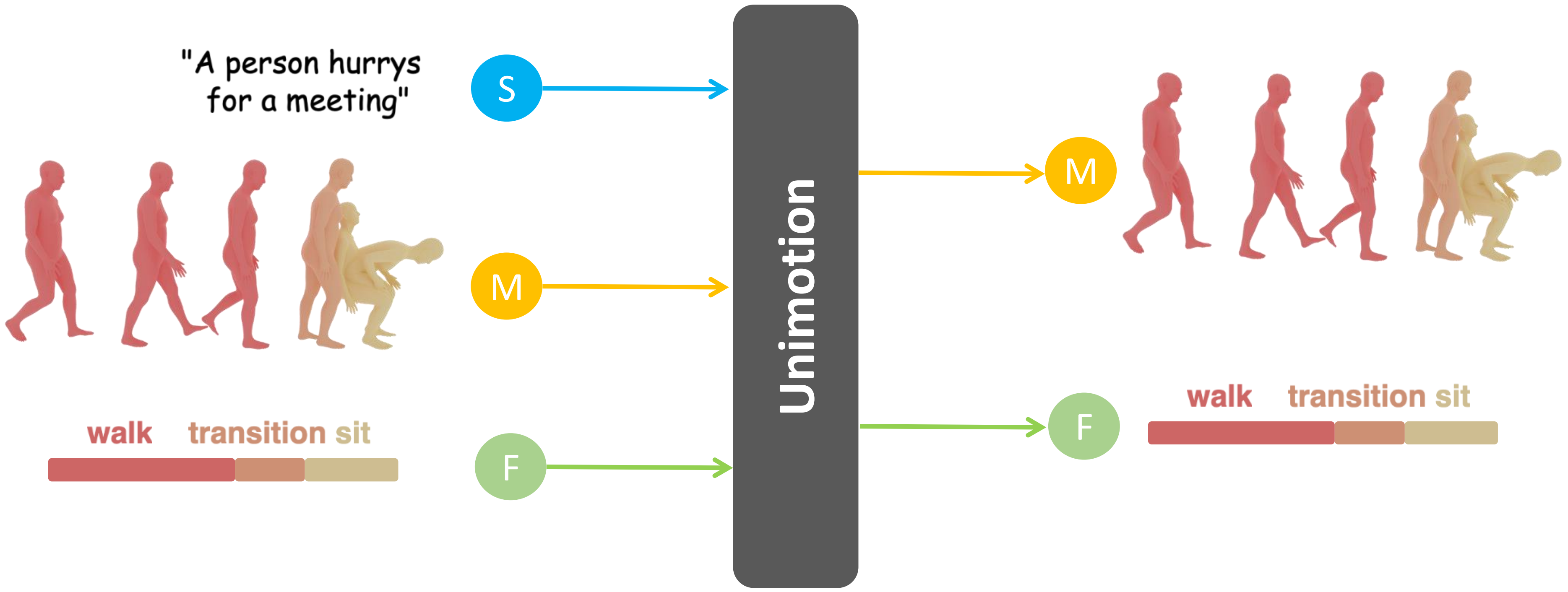


MotionLLM: ArXiv’24

Li et al. UniMotion, 3DV’25

Goal: Unify **motion tasks** into a single model

Our model unifies all these tasks into a single model allowing for **multi-model input**.

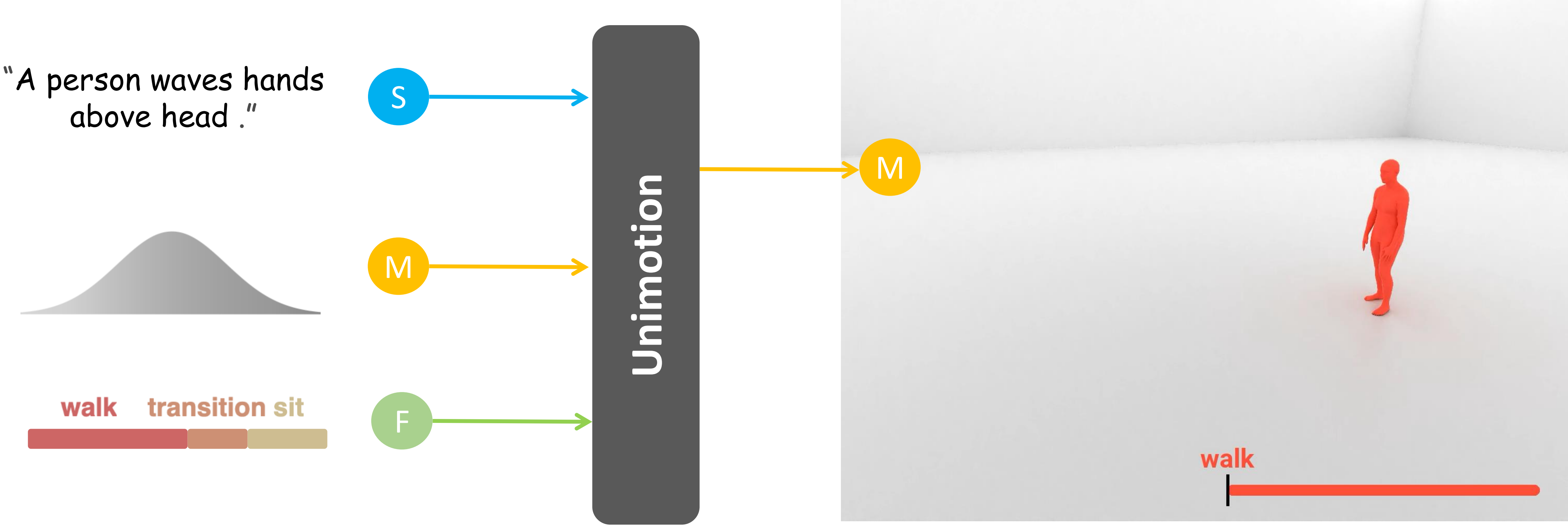


Additionally, its flexibility allows for novel tasks, not yet performed by prior works.

Li et al. UniMotion, 3DV'25

Novel tasks

- Hierarchical Text-to-Motion Generation

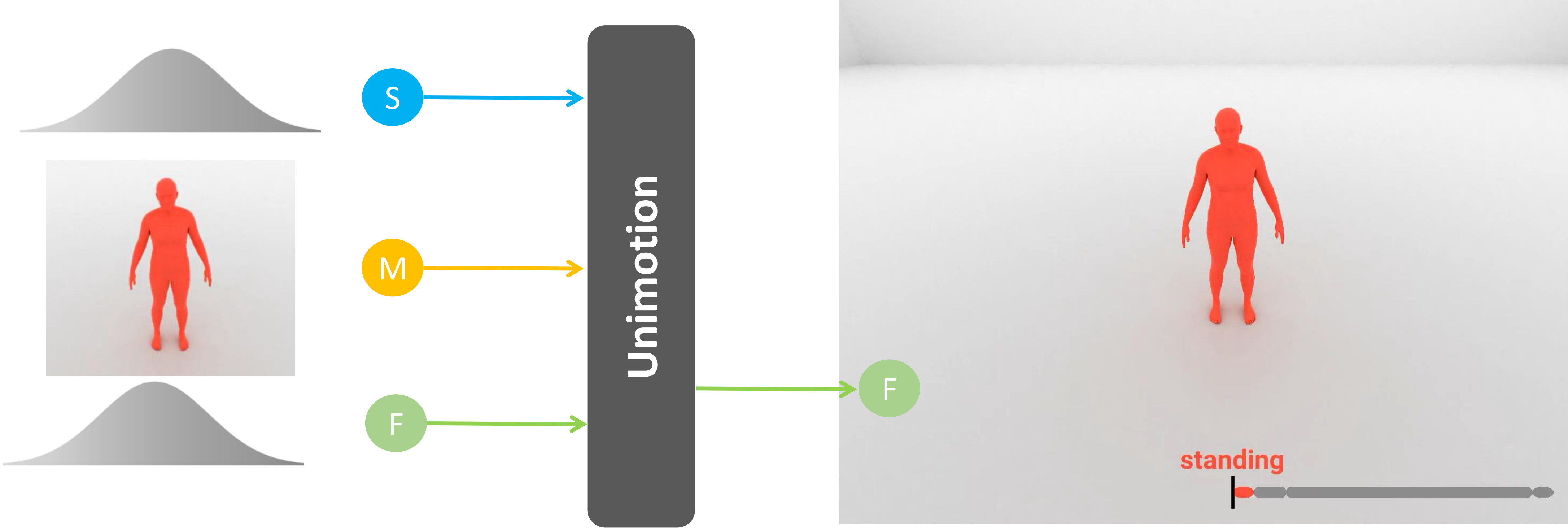


Li et al. UniMotion, 3DV'25

Novel tasks

- Hierarchical Text-to-Motion Generation
- Motion-to-Text Understanding

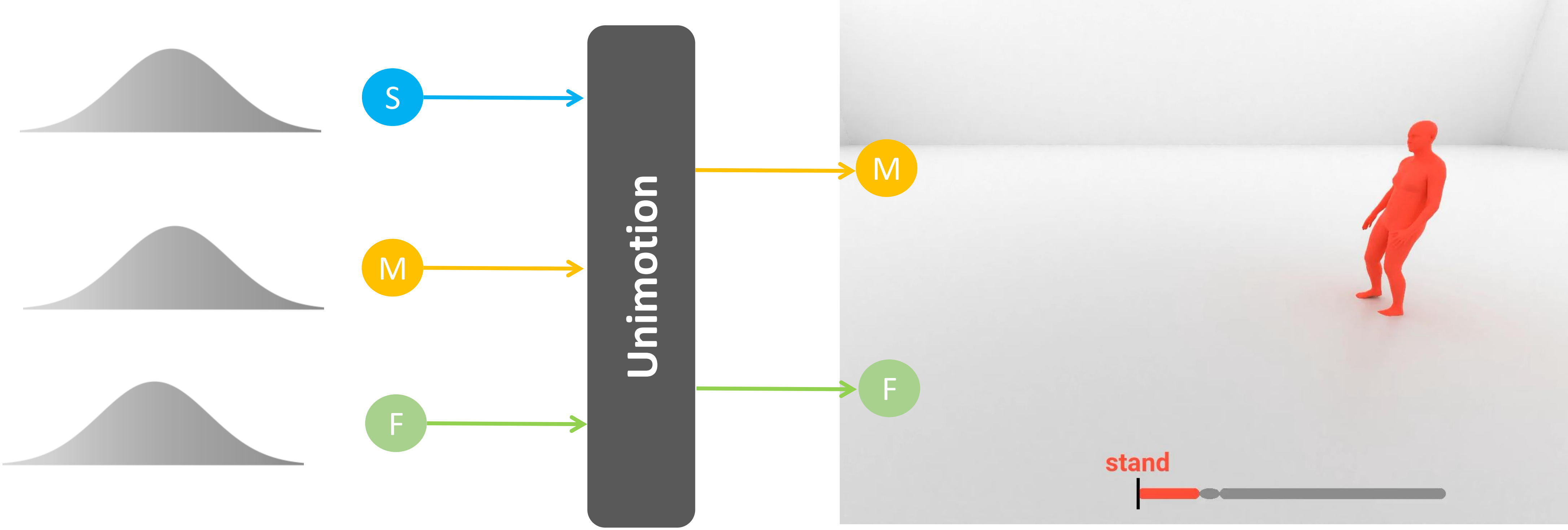
Fine-grained with temporal alignment!



Li et al. UniMotion, 3DV'25

Novel tasks

- Hierarchical Text-to-Motion Generation
- Motion-to-Text Understanding
- Unconditional Joint Generation

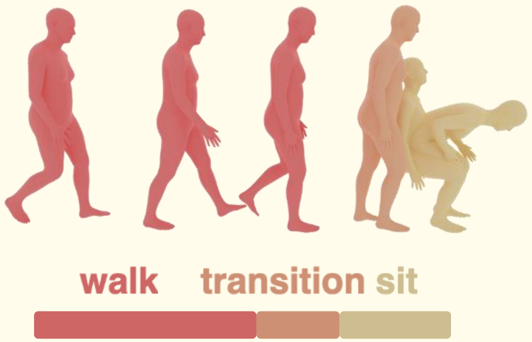


Method

Input

Noised (Motion, Frame-level text) + Seq-level text

Clean motion



Frame-level text

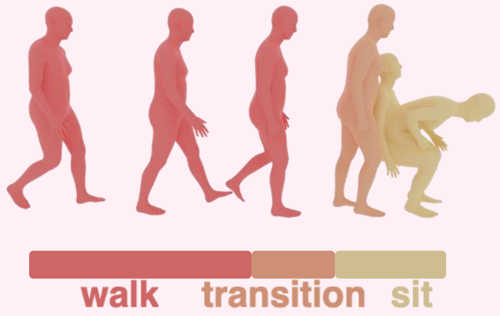
Seq-level text

"A person hurrys for a meeting"

Model

Output

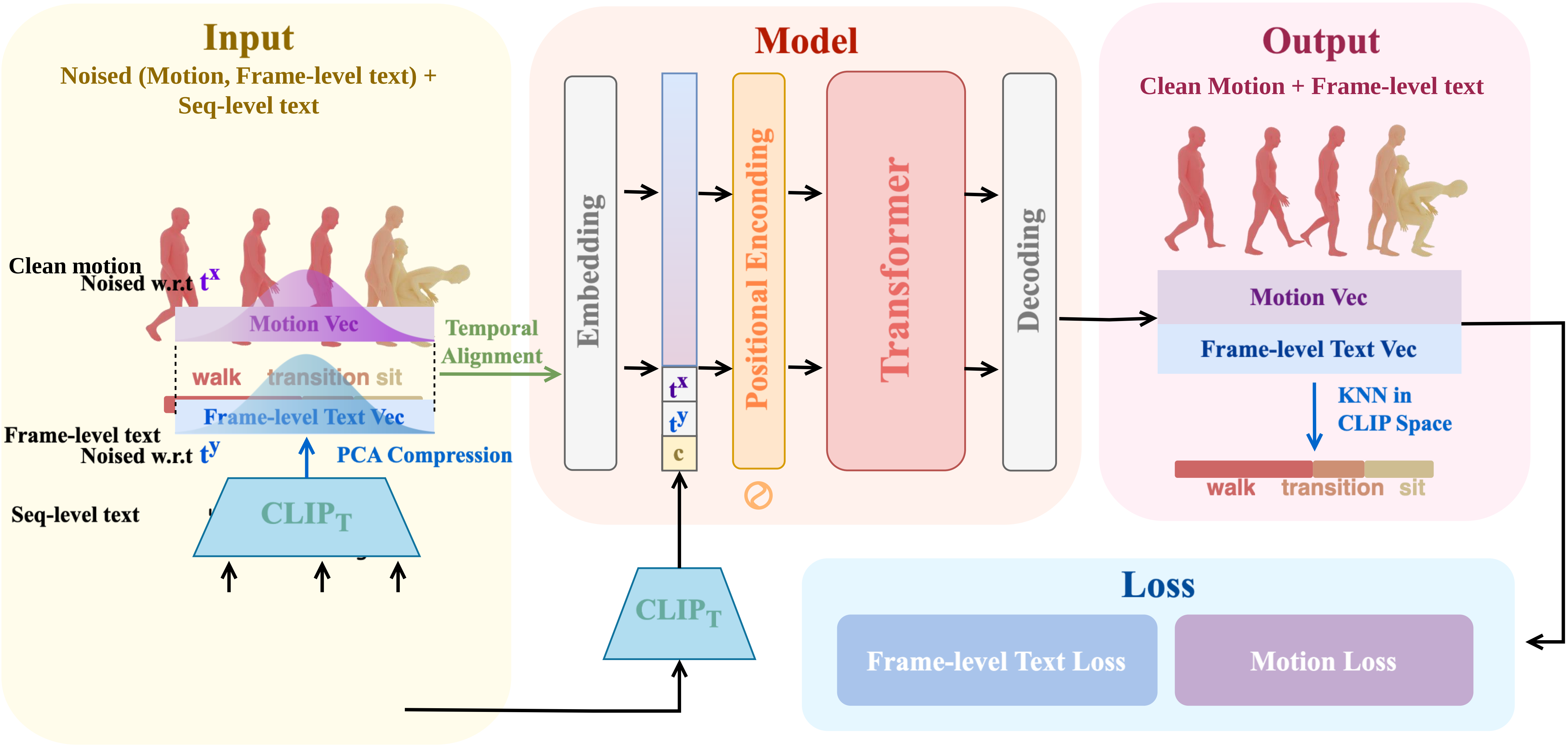
Clean Motion + Frame-level text



Loss

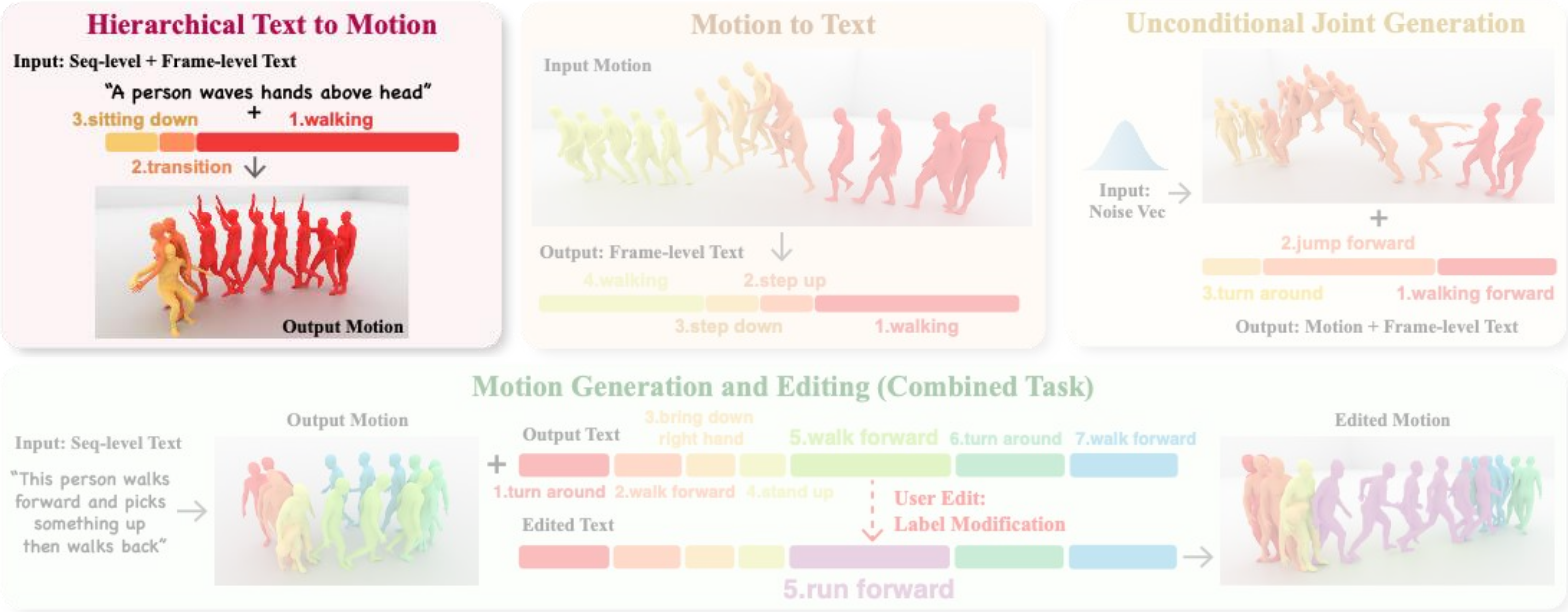
Li et al. UniMotion, 3DV’25

Method



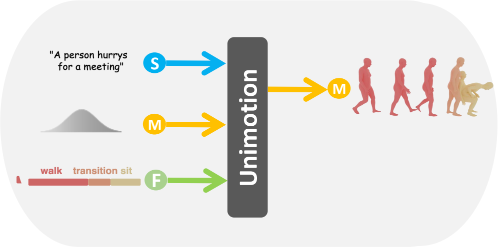
Li et al. UniMotion, 3DV'25

Multi-tasks



Li et al. UniMotion, 3DV'25

Frame-level Text-to-Motion Generation



Our Method achieves flexible hierarchical control and stronger correspondence

FlowMDM
- X Sequence-level control
- ✓ Frame-level control

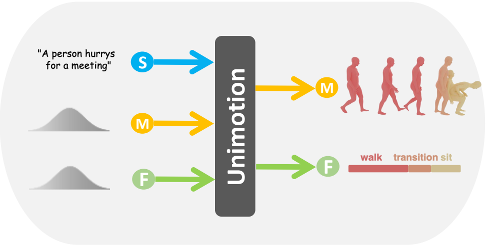
STMC
- X Sequence-level control
- ✓ Frame-level control

Ours
- ✓ **Sequence-level control**
- ✓ **Frame-level control**

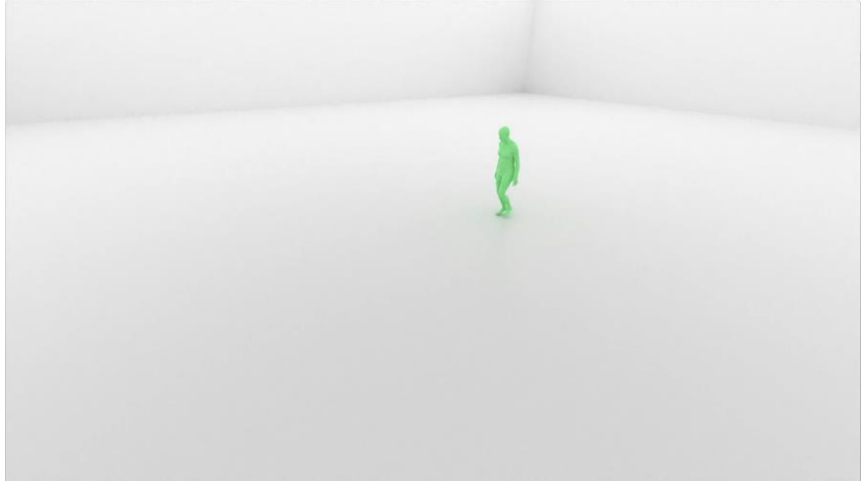
Li et al. UniMotion, 3DV'25
Barquero et al. FlowMDM, CVPR'24
Petrovich et al. STMC, CVPRW24.

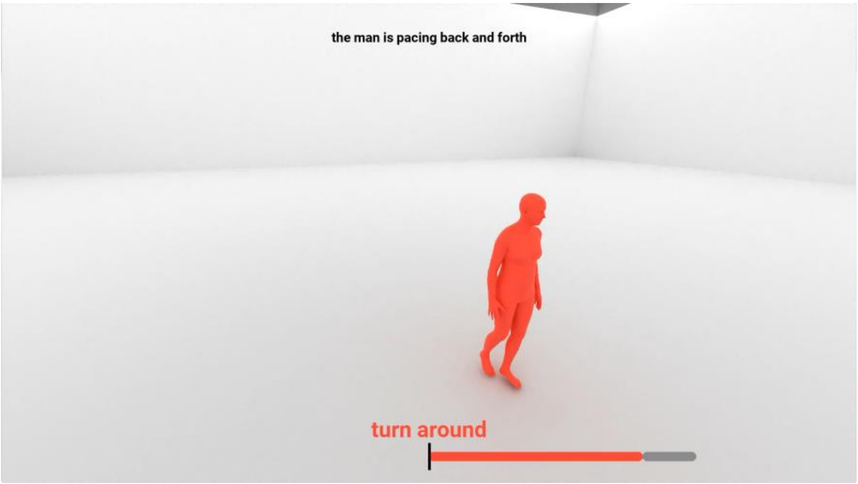
Sequence-level Text-to-Motion Generation

Our Method achieves motion synthesis and motion understanding at the same time



The man is pacing back and forth





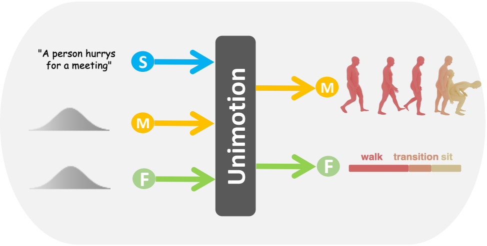
MDM
- ✓ Motion generation
- X Motion understanding

Ours
- ✓ **Motion generation**
- ✓ **Motion understanding**

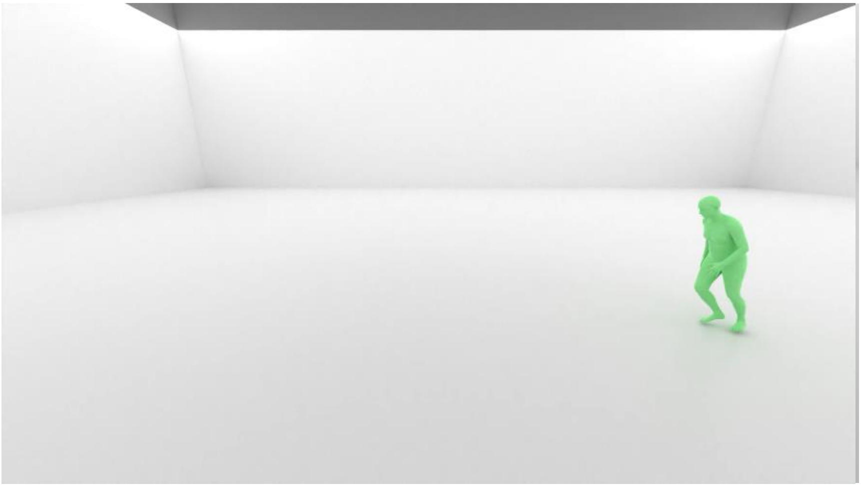
Li et al. UniMotion, 3DV'25
Tevet et al. MDM, ICLR'23.

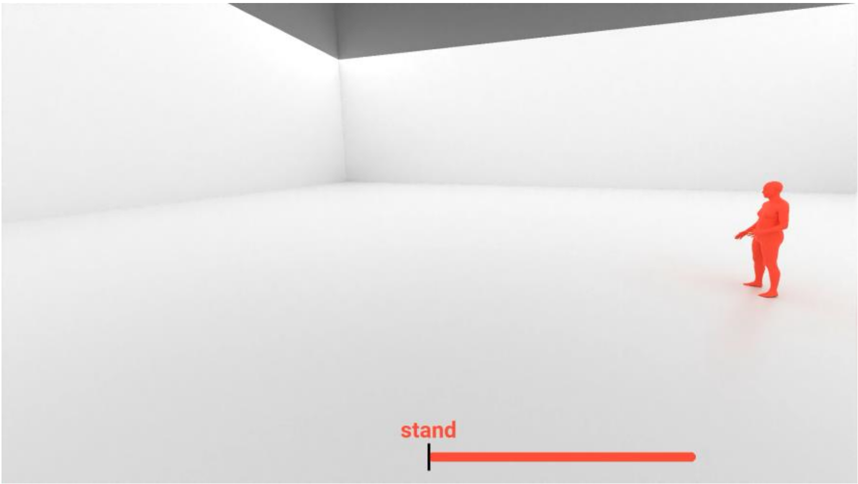
Sequence-level Text-to-Motion Generation

Our Method achieves motion synthesis and motion understanding at the same time



a person sprinting ahead, and then slowing down



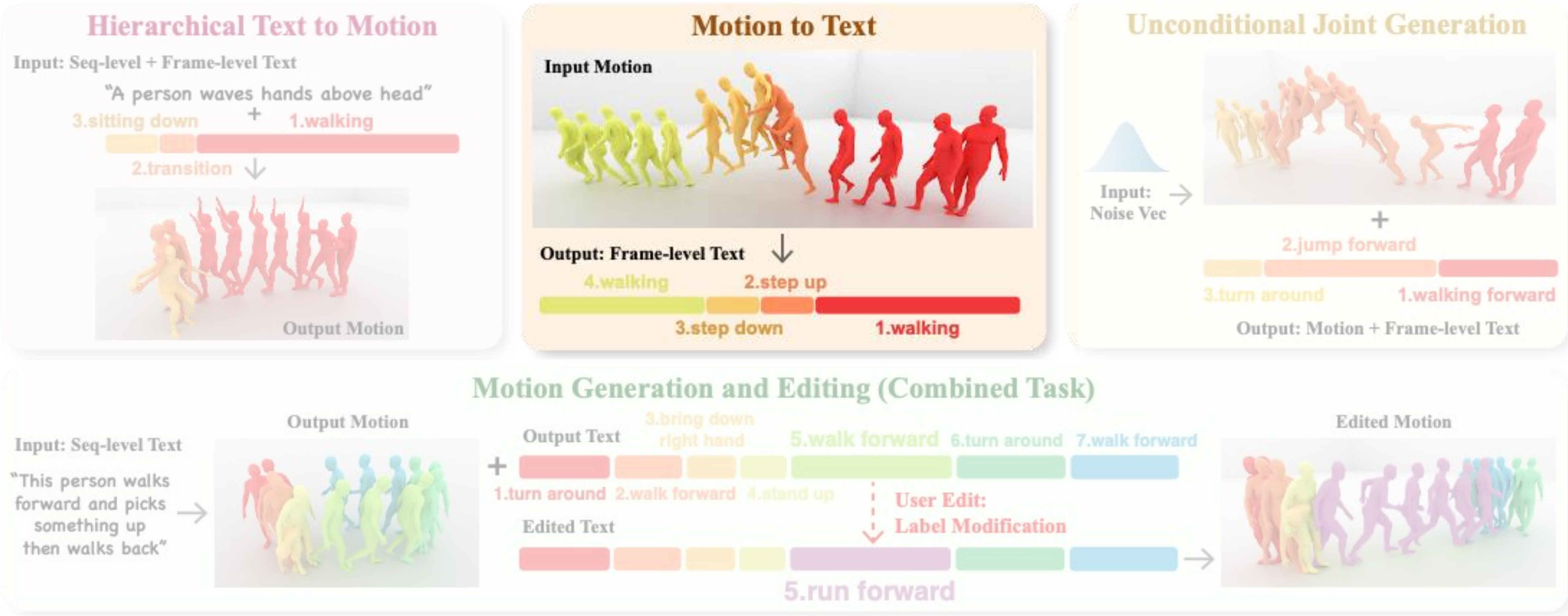


MDM
- ✓ Motion generation
- x Motion understanding

Ours
- ✓ **Motion generation**
- ✓ **Motion understanding**

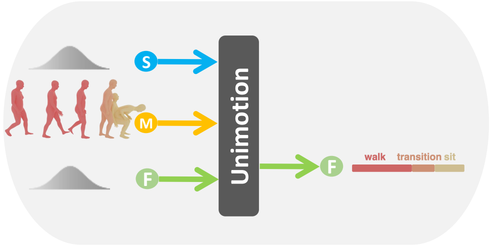
Li et al. UniMotion, 3DV'25
Tevet et al. MDM, ICLR'23.

Multi-tasks

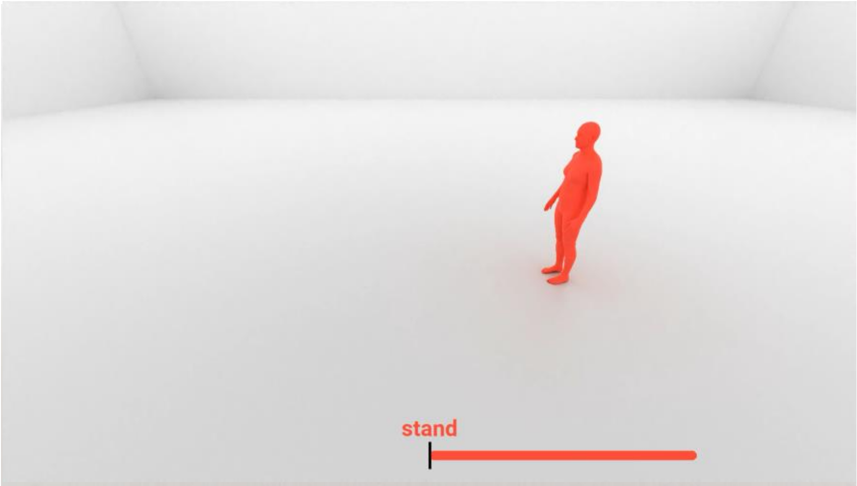


Li et al. UniMotion, 3DV'25

Fine-grained Motion-to-text Understanding



Our Method achieves motion understanding with fine-grained temporal awareness

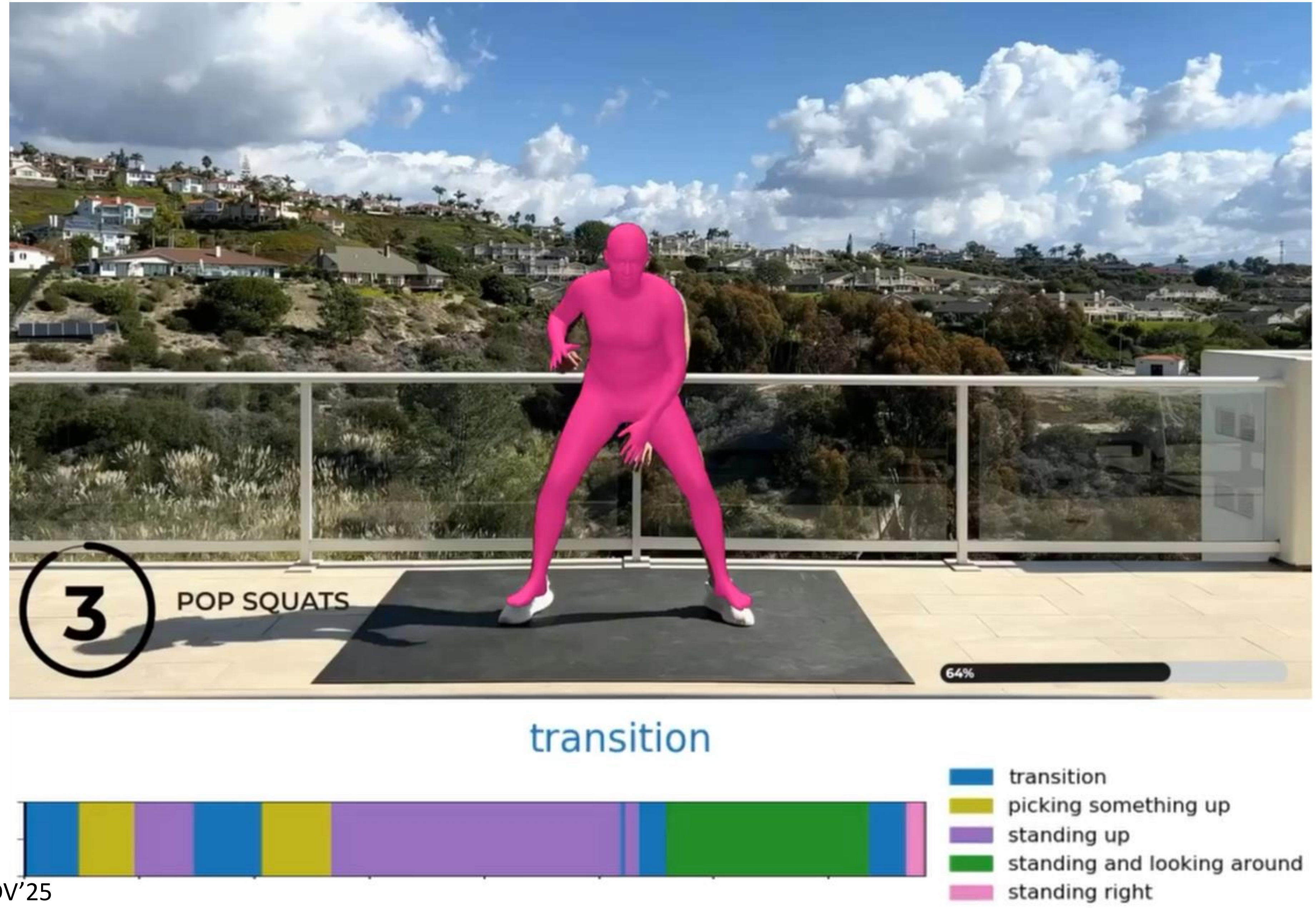


Ours

✓ **Motion generation**
✓ **Motion understanding**

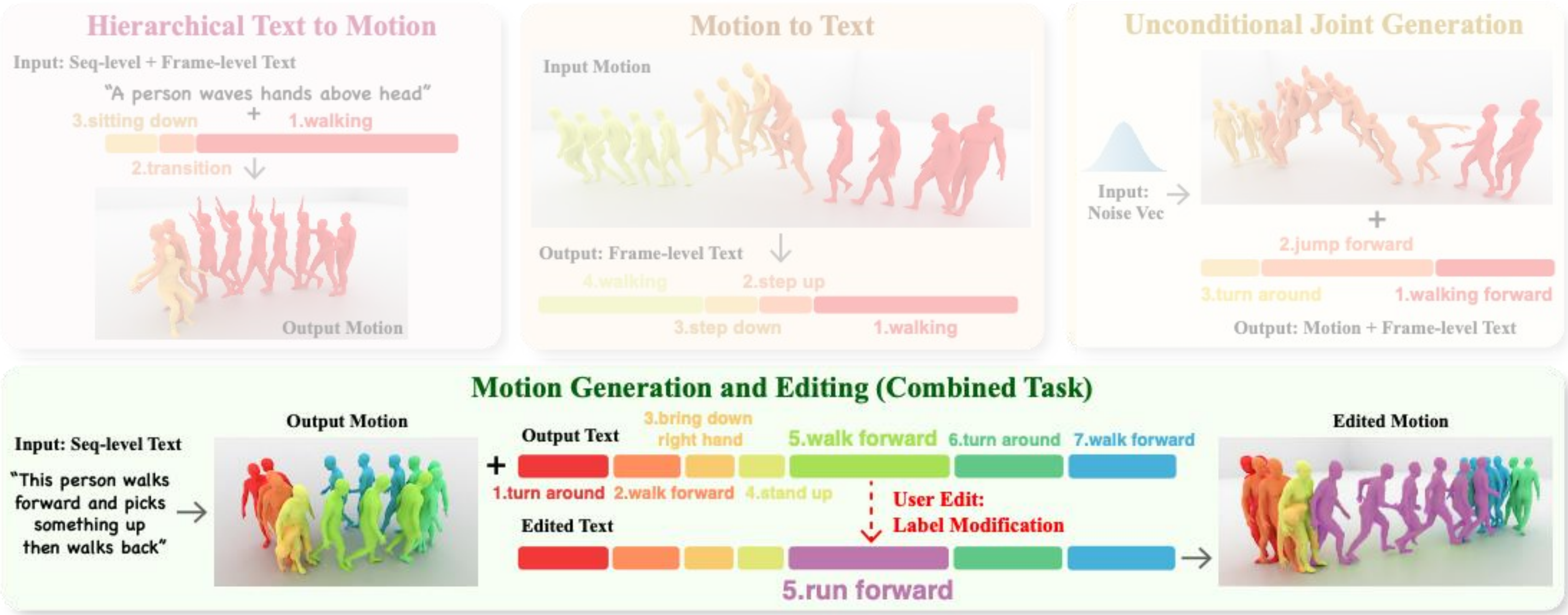
Li et al. UniMotion, 3DV'25

2D-video annotation



Li et al. UniMotion, 3DV'25

Multi-tasks



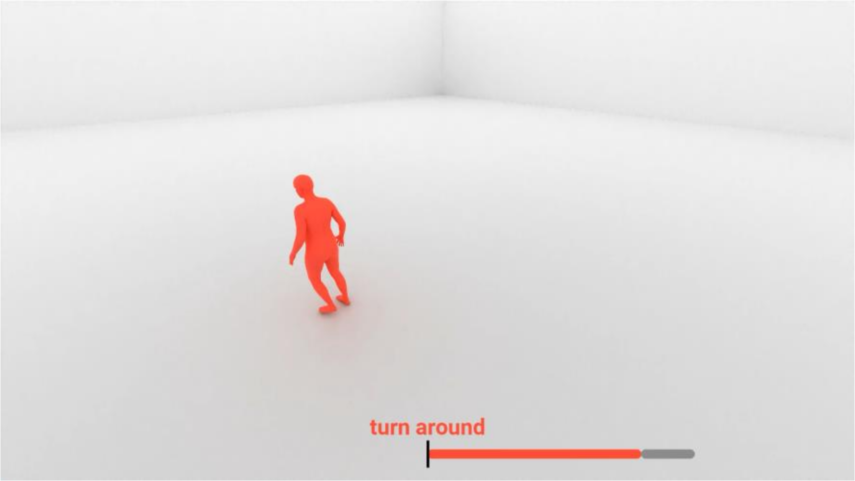
Li et al. UniMotion, 3DV'25

Motion generation and Editing

User input: "The person walks forward and picks something up then walks back."

Model output:

1.turn around | 2.walk forward | 3.bring down right hand | 4.stand up | 5.walk forward | 6.turn around | 5.walk forward



Li et al. UniMotion, 3DV'25

Motion generation and Editing

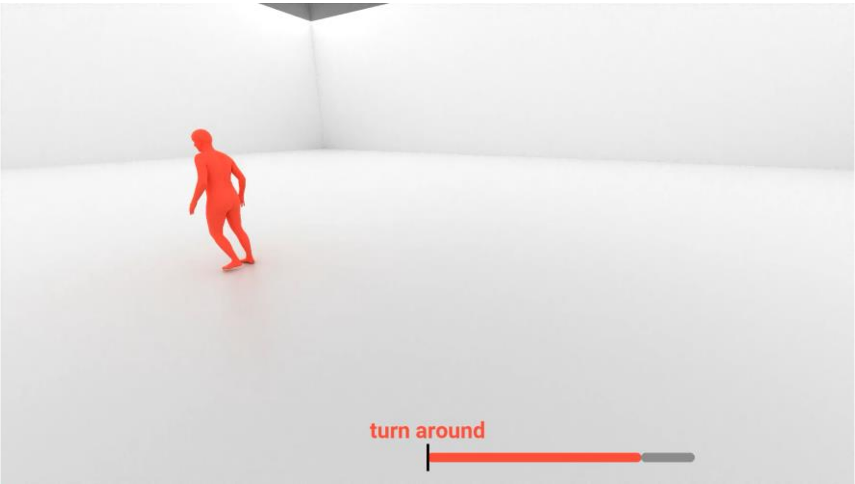
User input: “The person walks forward and picks something up then walks back.”

Model output:

1.turn around | 2.walk forward | 3.bring down right hand | 4.stand up | 5.walk forward | 6.turn around | 5.walk forward

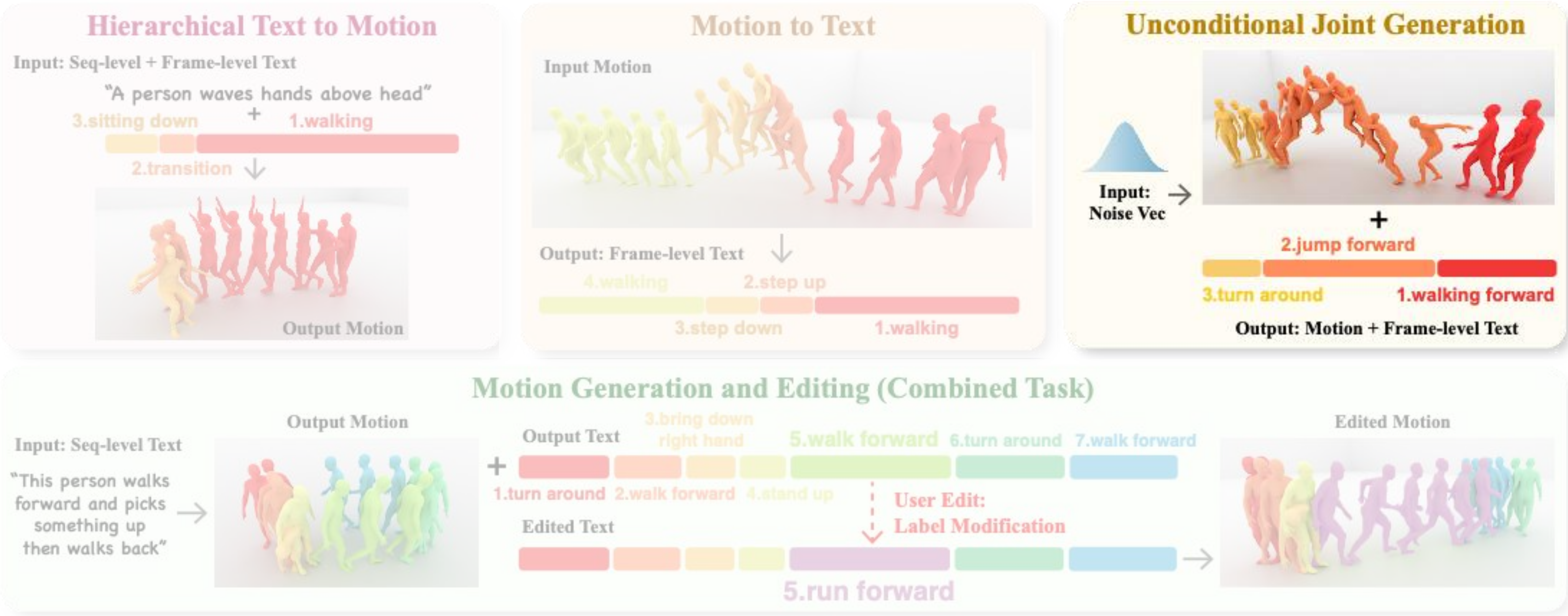
User edit:

1.turn around | 2.walk forward | 3.bring down right hand | 4.stand up | **5.jog forward** | 6.turn around | 5.walk forward

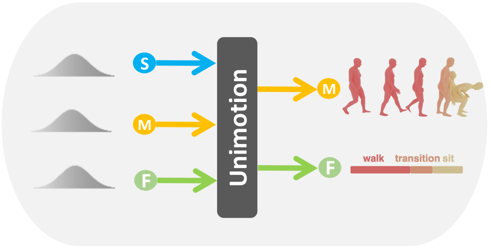


Li et al. UniMotion, 3DV’25

Multi-tasks

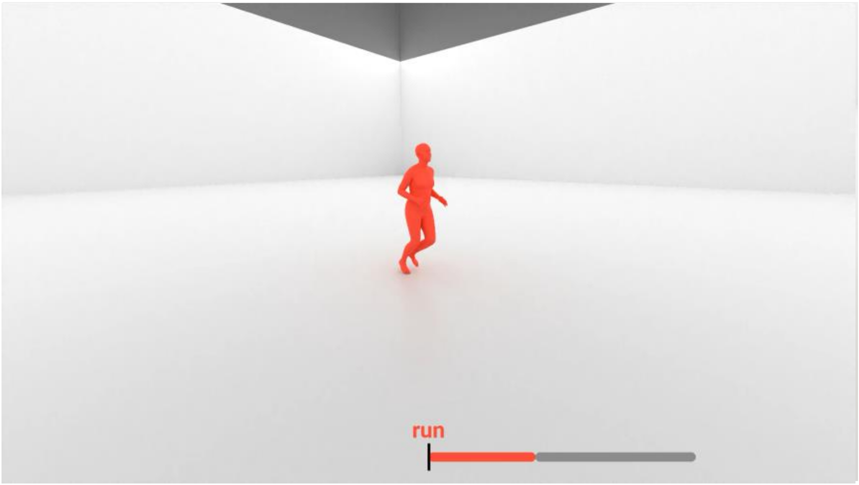


Li et al. UniMotion, 3DV'25

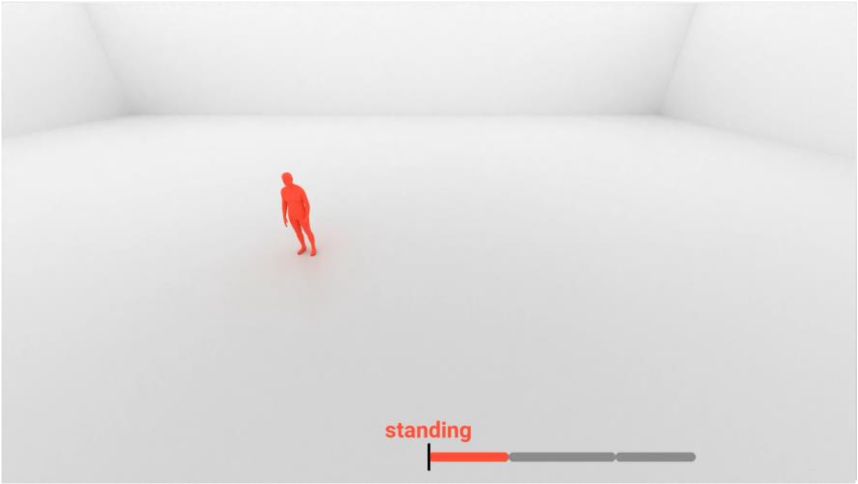


Unconditional Joint Generation

Our method is the first to do unconditional joint generation



Ours



Ours

✓ Motion generation
✓ Motion understanding

Li et al. UniMotion, 3DV'25

Take home messages

- Diffusion is also powerful for motion generation, but requires geometry constrain to train a good model.

- Pretrained motion prior is useful to generate complex compositional motions.

- A unified motion helps text and motion in both directions.

Slide credits

- Thanks to Guy Tevet for kindly providing the slides for diffusion based human motion models.

Thank you!

"A person is standing and waving goodbye!"