Deep Compositional Phase Diffusion for Long Motion Sequence Generation

Ho Yin Au, Jie Chen, Junkun Jiang, Jingyu Xiang
Department of Computer Science, Hong Kong Baptist University

NEURAL INFORMATION PROCESSING SYSTEMS

香港浸會大學
HONG KONG BAPTIST UNIVERSITY
DEPARTMENT OF COMPUTER SCIENCE
計算機科學系

Intro: Long-term Compositional Generation

- Human Motion Generation
  - Short duration motion clip of a single semantics
- Long-term Compositional Generation
  - Multiple **sequentially connected clips**, each with single semantics.
    - Each clip is aligned to the **semantic** condition
    - **Transitions** between clips are smooth and natural
  - Existing models <u>struggle to synthesize</u> natural and seamless transitions

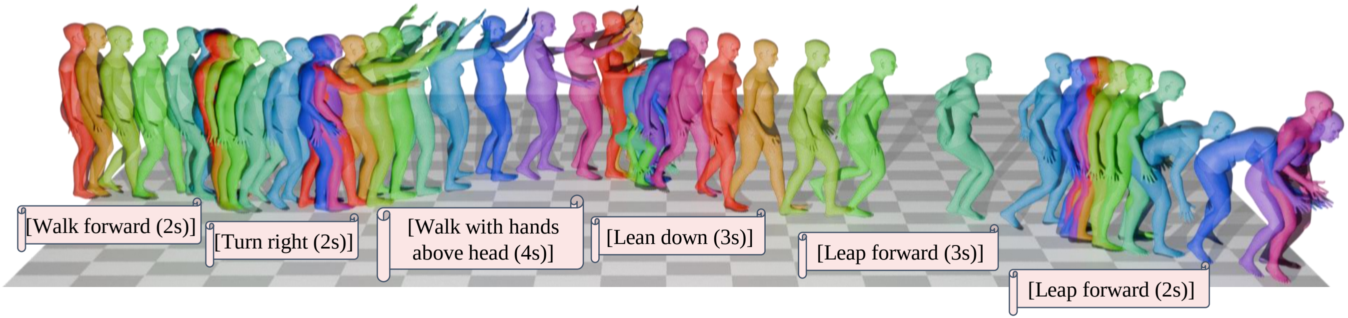


Illustration of Long-term Compositional Generation

Related Works

- TEACH (3DV 2022)
  - Autoregressively synthesizes motion clips, having minor discontinuities between clips
  - Blending transitions are generated using spherical linear interpolation (SLERP)
- Limitation
  - Blending transitions often appear unrealistic, reducing overall motion realism

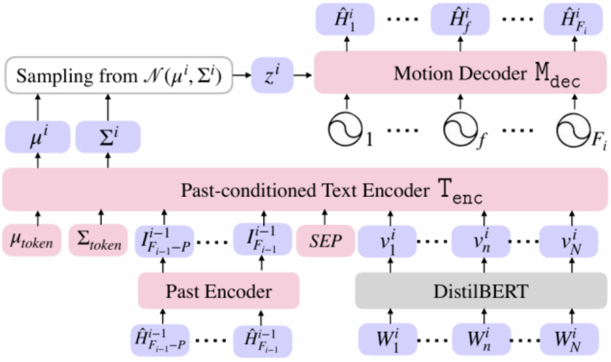
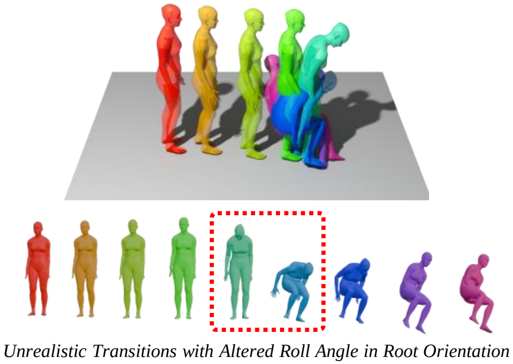


Illustration of TEACH's autoregressive variational encoder-decoder architecture



Result visualization of [walk(2.4s), sit down(3.6s)]

- Athanasiou et al., Teach: Temporal action composition for 3d humans. In 3DV, pp. 414–423, 2022.

Related Works

- PriorMDM (ICLR 2024)
  - Synthesizes all motion clips in parallel, resulting in significant discontinuities between clips
  - Blending transitions are generated using the human motion diffusion model (MDM)
- Limitation
  - Substantial discontinuities between clips, and generated transitions fail to smooth them

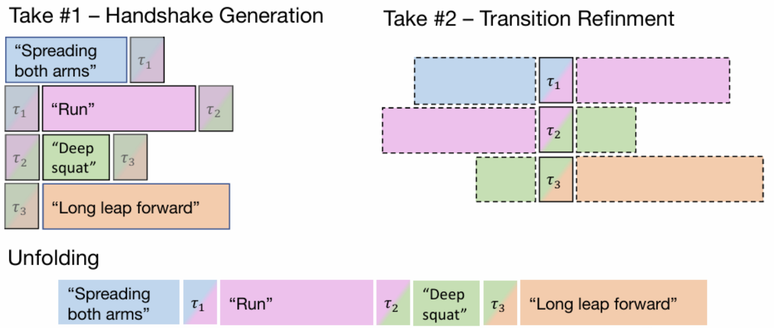
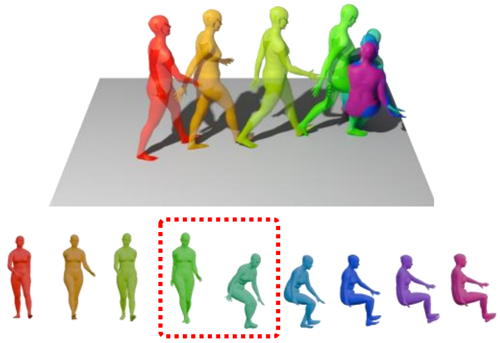


Illustration of priorMDM's DoubleTake algorithm



Result visualization of [walk(2.4s), sit down(3.6s)]

- Shafir et al., Human motion diffusion as a generative prior, ICLR, 2024.

Related Works

- DeepPhase (SIGGRAPH 2022)
  - Encode motion into the periodic latent space using fixed-length convolutions and FFT
  - Excels at motion extrapolation and inbetweening, effectively capturing **motion dynamics**
- Limitation
  - Results in variable numbers of phase latents, more difficult to learn text-motion alignment.

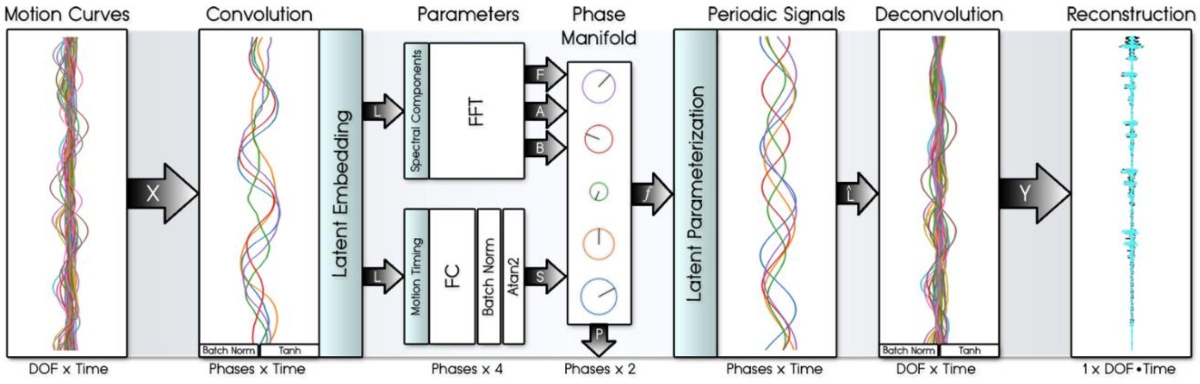


Illustration of DeepPhase's periodic autoencoder architecture

- Starke et al., Deepphase: Periodic autoencoders for learning motion phase manifolds. ACM Transactions on Graphics, 2022

Framework Core Idea

- Integrates **motion dynamics from adjacent segments** into diffusion process
  - Minimizes discontinuity between consecutive motion segments
  - Utilizes these dynamics to generate realistic blending transitions
- Employs **phase mixing** to jointly integrate semantic and transitional conditions

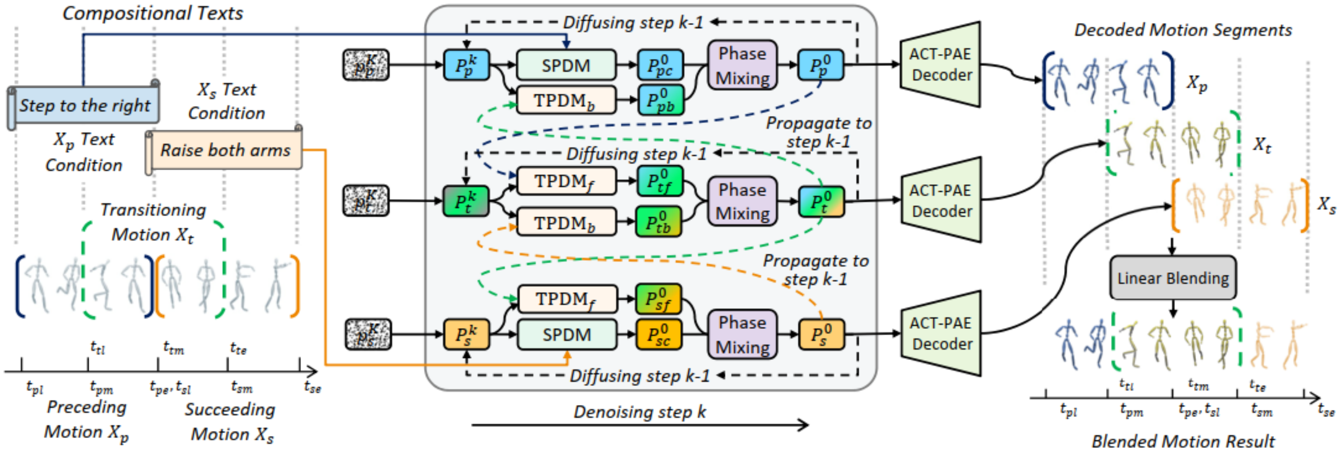


Illustration of our Compositional Motion Generation pipeline

Framework Overview

- Three Components:
  - ACT-PAE: Encode human motion sequence into **phase latent space**
  - SPDM: Incorporates **semantic information** into the diffusion process
  - TPDM: Integrates **adjacent motion dynamics** into the diffusion process

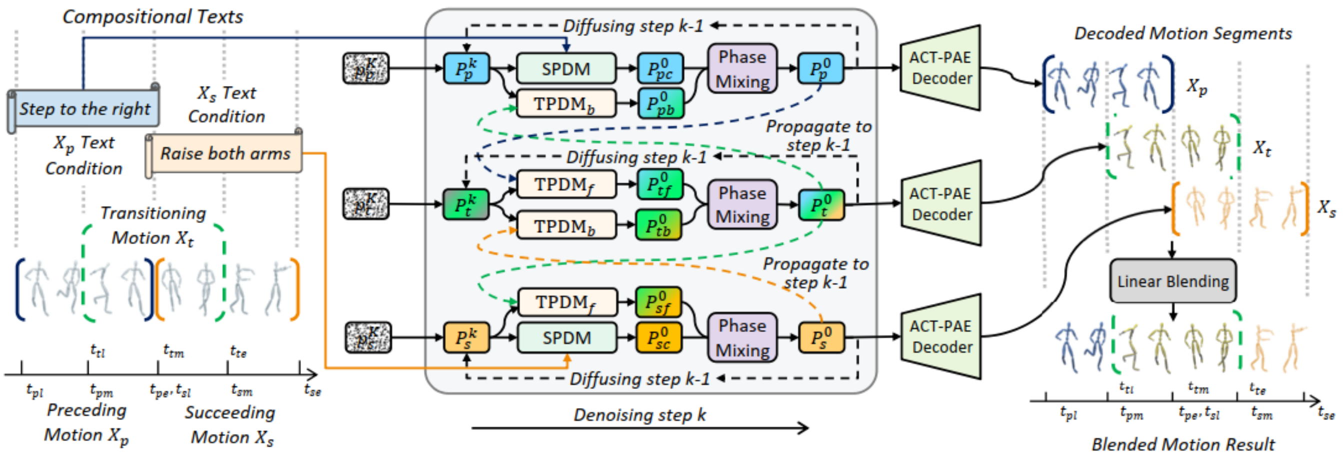
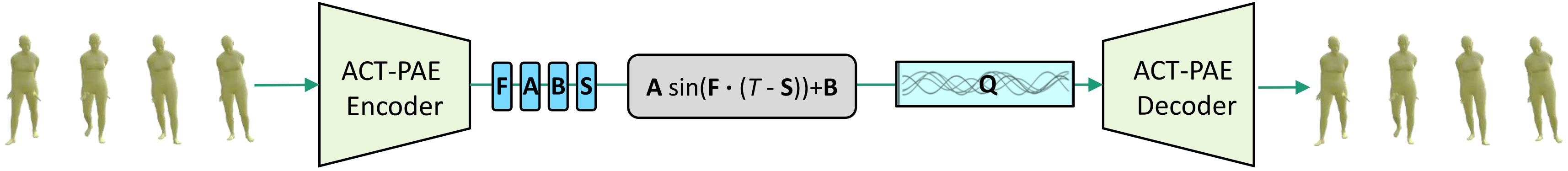


Illustration of our Compositional Motion Generation pipeline

ACT-PAE

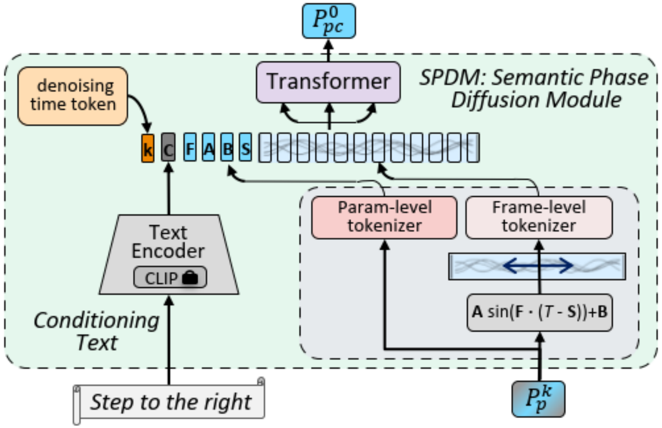
- **A**ction **C**en**T**ric **P**eriodic **A**uto**E**ncoder (ACT-PAE):
  - Encode motion $\boldsymbol{X} \in \mathbb{R}^{\mathrm{N \times E}}$ into four phase latents $\boldsymbol{F}, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{S} \in \mathbb{R}^{\mathrm{Q}}$
  - Reparameterizes latents into periodic signal $\boldsymbol{Q} \in \mathbb{R}^{\mathrm{N \times Q}}$ using: $\boldsymbol{Q} = \boldsymbol{A} \sin(\boldsymbol{F} \cdot (T - \boldsymbol{S})) + \boldsymbol{B}$
  - Decode $\boldsymbol{Q}$ back to reconstructed motion $\widehat{\boldsymbol{X}}$
- Advantage over Traditional PAE:
  - Avoiding fixed-window motion slicing, enabling the capture of unified and semantic meaningful motion dynamics in the motion segment as a **cohesive unit**



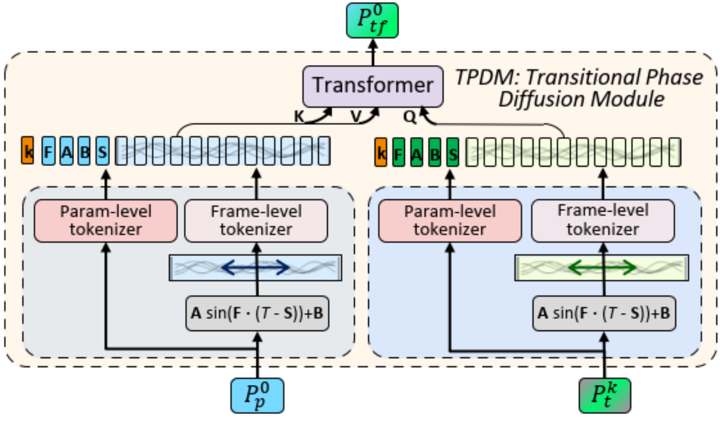
Detail of the Action centric Periodic AutoEncoder (ACT-PAE)

SPDM and TPDM

- **S**emantic **P**hase **D**iffusion **M**odule (SPDM):
  - Denoise phase latents $\boldsymbol{P}_{\mathrm{p}}^{\mathrm{k}}$ utilizing semantic condition **C**
  - Inputs include both phase latents $\boldsymbol{P}_{\mathrm{p}}^{\mathrm{k}}= [\boldsymbol{F}, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{S}]$ and periodic signal $\boldsymbol{Q} = \boldsymbol{A} \sin(\boldsymbol{F} \cdot (T - \boldsymbol{S}))+\boldsymbol{B}$
- **T**ransitional **P**hase **D**iffusion **M**odule (TPDM):
  - Functions similarly to SPDM, but uses adjacent phase parameters $\boldsymbol{P}_{\mathrm{p}}^{0}$ as input
- Advantage over Latent Diffusion Model:
  - The periodic signal $\boldsymbol{Q}$ captures **spatial-temporal context** within the phase latents $\boldsymbol{P}$



Detail of the Semantic Phase Diffusion Module (SPDM)



Detail of the Transitional Phase Diffusion Module (TPDM)

Compositional Phase Diffusion Pipelines

- Pipeline for **Compositional Motion Generation**:
  - Denoise phase latent using SPDM (for $\boldsymbol{P}_\mathrm{c}^0$) and two TPDMs (for $\boldsymbol{P}_f^0$ and $\boldsymbol{P}_\mathrm{b}^0$)
  - Perform **phase mixing** on the denoised outputs: $\boldsymbol{P}^0 = \mathrm{r}\frac{\boldsymbol{P}_f^0+\boldsymbol{P}_\mathrm{b}^0}{2} + (1-\mathrm{r})\boldsymbol{P}_\mathrm{c}^0$
  - Diffuse $\boldsymbol{P}^0$ to the step k − 1 or decode via ACT-PAE to generate $\boldsymbol{X}_\mathrm{p}, \boldsymbol{X}_\mathrm{t}, \boldsymbol{X}_\mathrm{s}$
  - **Linear blend** the transition segment $\boldsymbol{X}_\mathrm{t}$ into the overlap region between $\boldsymbol{X}_\mathrm{p}$ and $\boldsymbol{X}_\mathrm{s}$

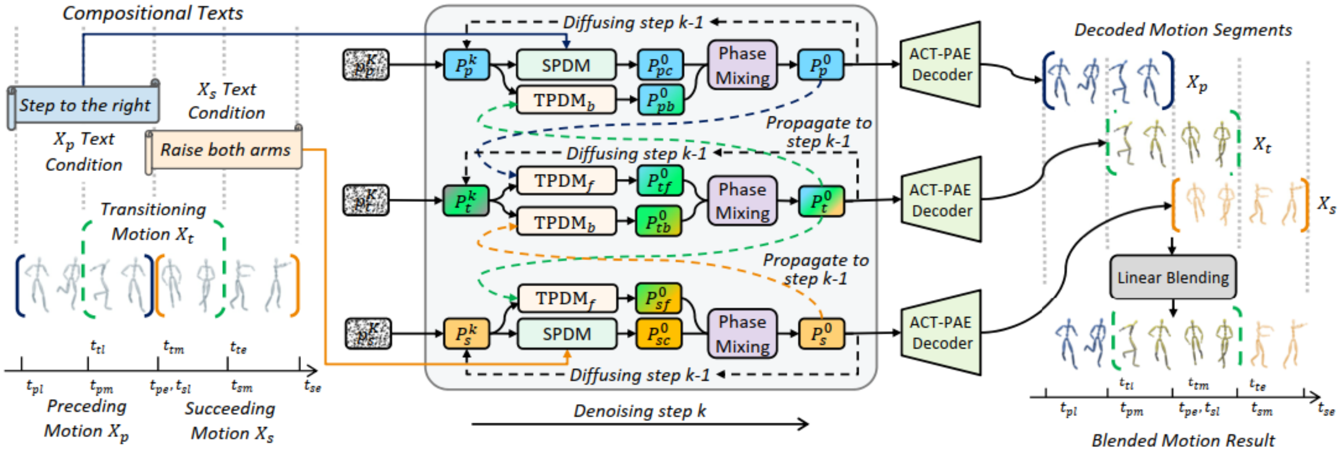


Illustration of our Compositional Motion Generation pipeline

Compositional Phase Diffusion Pipelines

- Pipeline for **Motion Inbetweening**:
  - Encode user-provided motion segment into phase latents ($\boldsymbol{P}_p^0$ and $\boldsymbol{P}_s^0$) using ACT-PAE encoder
  - Denoise the inbetweening motion $\boldsymbol{X}_\mathrm{i}$ and the transitioning motions $\boldsymbol{X}_{\mathrm{t}_1}$ and $\boldsymbol{X}_{\mathrm{t}_2}$
  - **Linear blend** the transition segment $\boldsymbol{X}_{\mathrm{t}_1}$ and $\boldsymbol{X}_{\mathrm{t}_2}$ into the overlapping regions of $\boldsymbol{X}_\mathrm{p}$, $\boldsymbol{X}_\mathrm{i}$, $\boldsymbol{X}_\mathrm{s}$
  - Note that $\boldsymbol{X}_\mathrm{i}$ can be further conditioned with text input by incorporating an optional SPDM

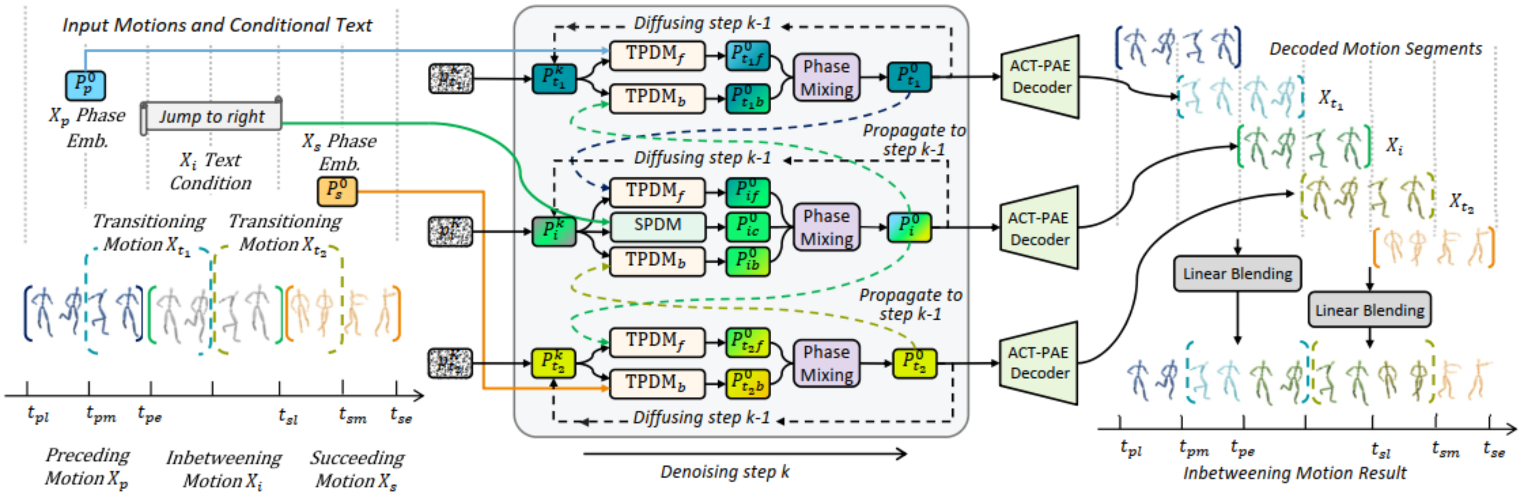


Illustration of our Motion Inbetweening pipeline

Compositional Phase Diffusion Pipelines

- Advantages of the pipelines:
  - **Coherent Transitions**:
    Bidirectional TPDMs progressively propagate phase information throughout the sequence, ensuring smoother and more coherent motion generation
  - **Scalable and Flexible**:
    Pipeline supports an arbitrary number of segments by rearranging SPDM and TPDM modules, with parallel processing for efficient generation of long motion sequences.

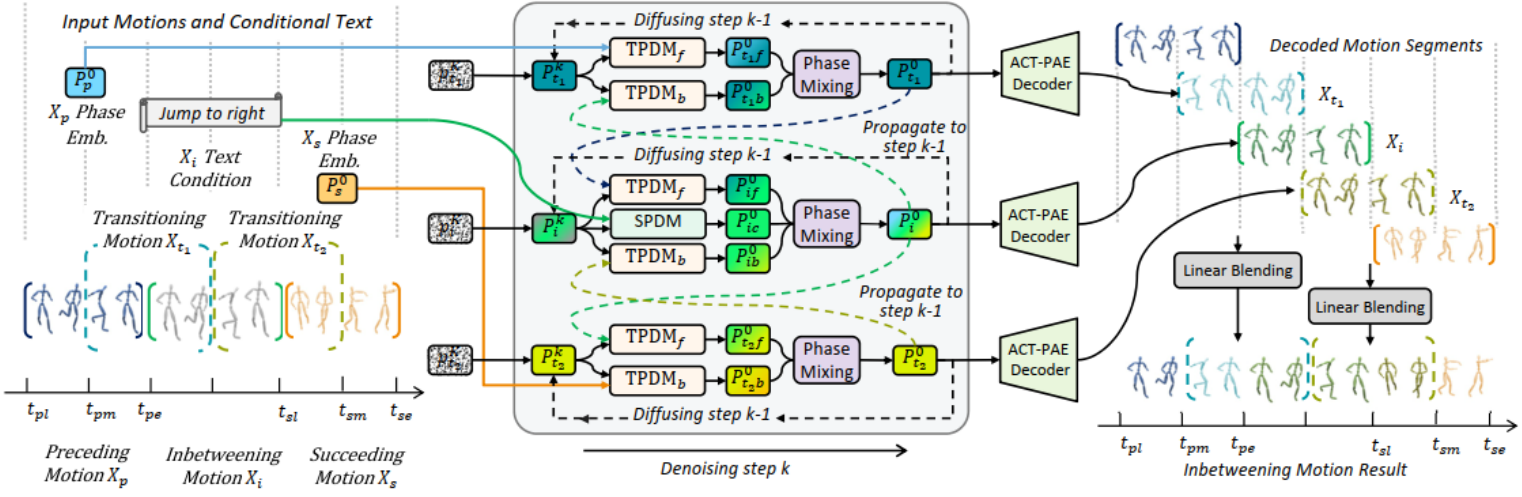
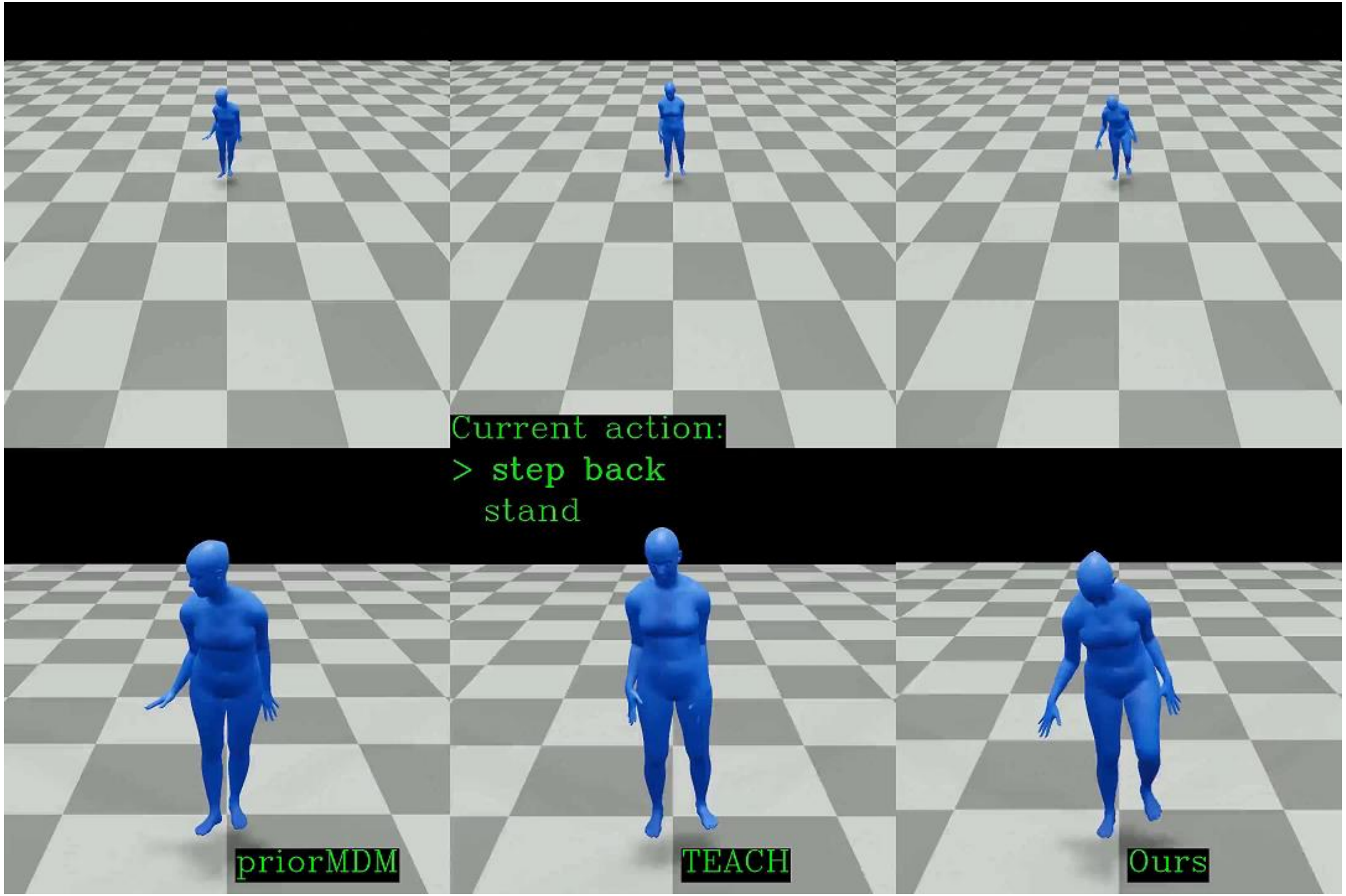
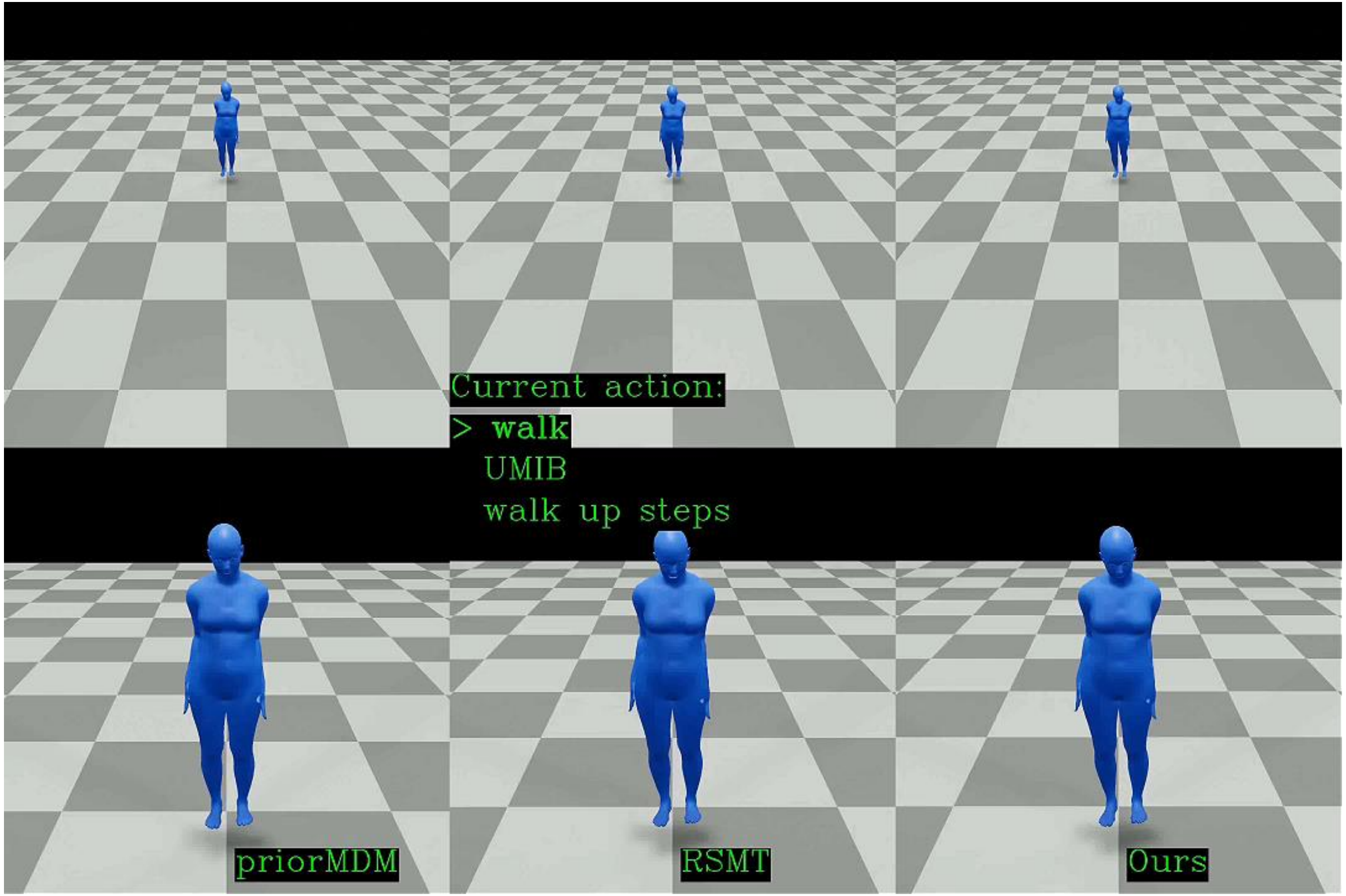


Illustration of our Motion Inbetweening pipeline

Result: Compositional Motion Generation



Result: Motion Inbetweening



NEURAL INFORMATION PROCESSING SYSTEMS

Summary

- Operating within a unified phase latent space facilitates alignment of transitional dynamics, enables smoother transitions between motion clips
- Scalable and efficient framework supports diverse motion generation tasks and allows parallel processing of arbitrary number of motion segments

Github Code

arXiv Paper