CST: Character State Transformer for Object-Conditioned Human Motion Prediction

Kuan-Wei Tseng[1], Rei Kawakami[1], Satoshi Ikehata[2], and Ikuro Sato[1,3]

[1]Institute of Science Tokyo [2]National Institute of Informatics [3]Denso IT Laboratory

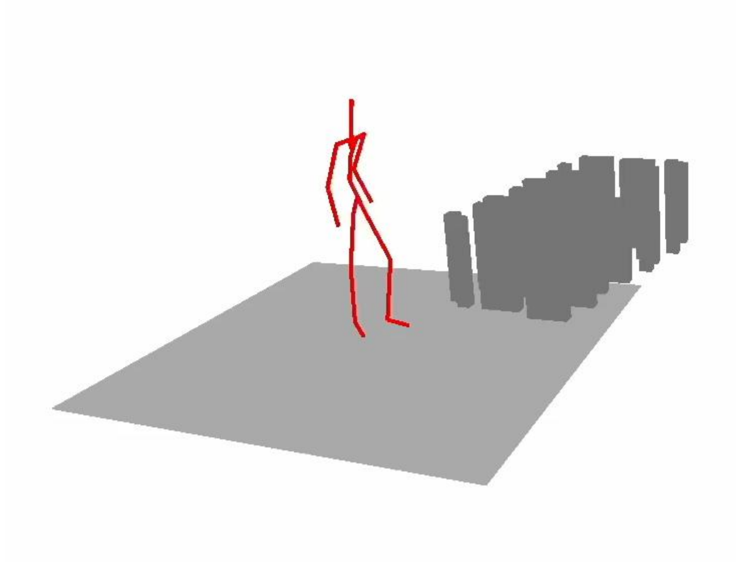
WACV 2025
TUCSON, ARIZONA • FEB 28 - MAR 4

CV4Smalls

Overview

Task

Human motion prediction aims to **generate the future human motions**, in form of the 3D poses of body joints, **given the historical human motion.**



Condition

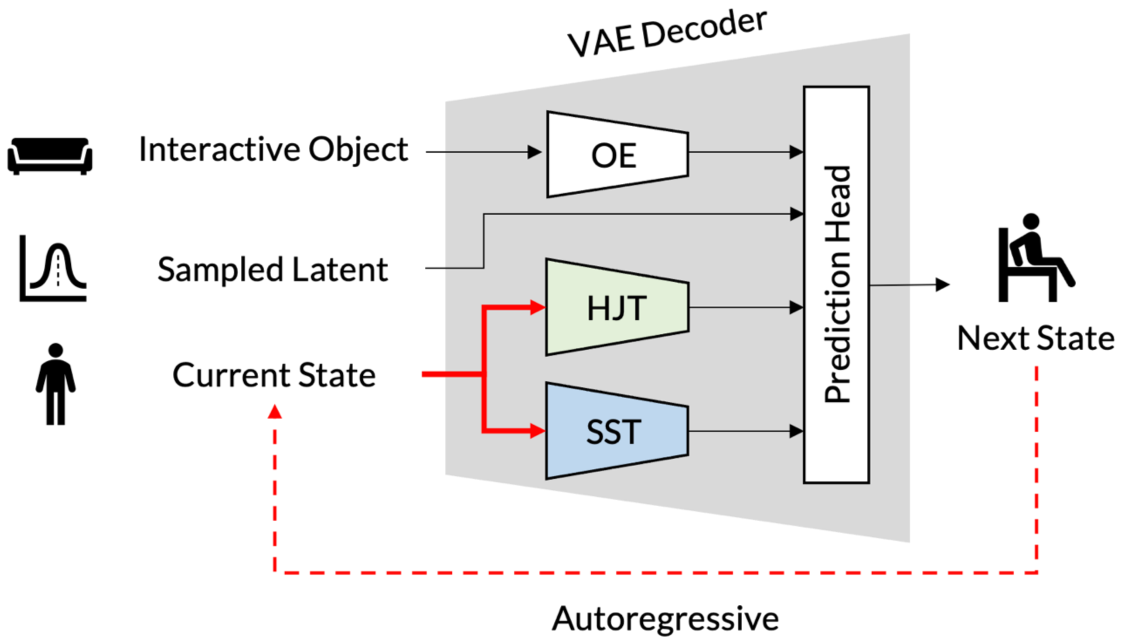
Interactable objects in the scene (e.g., chairs)

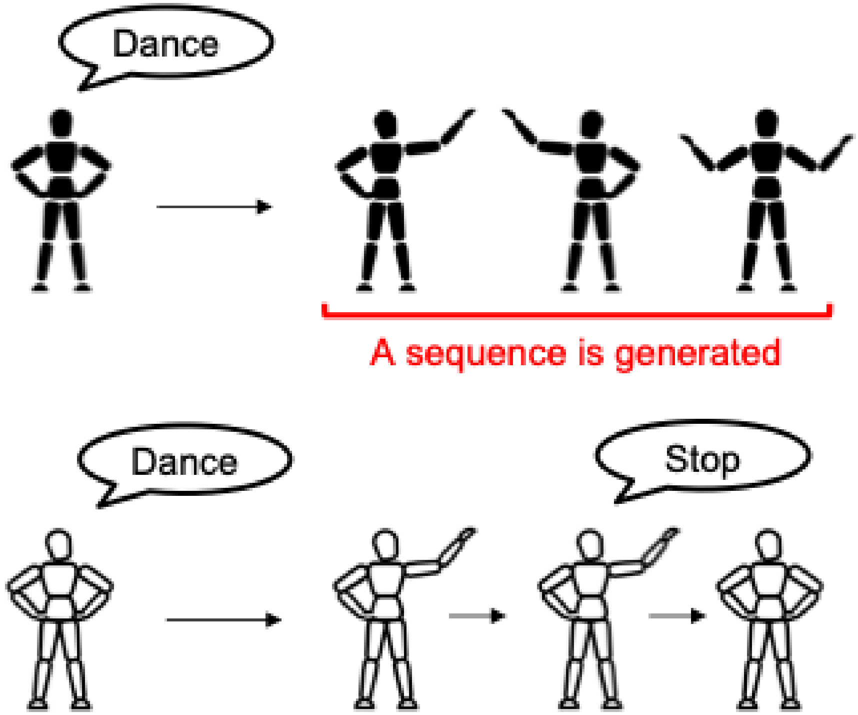
Difficulty

Scarce amount of homogeneous motion capture (MoCap) data

Auto-regressive human motion prediction

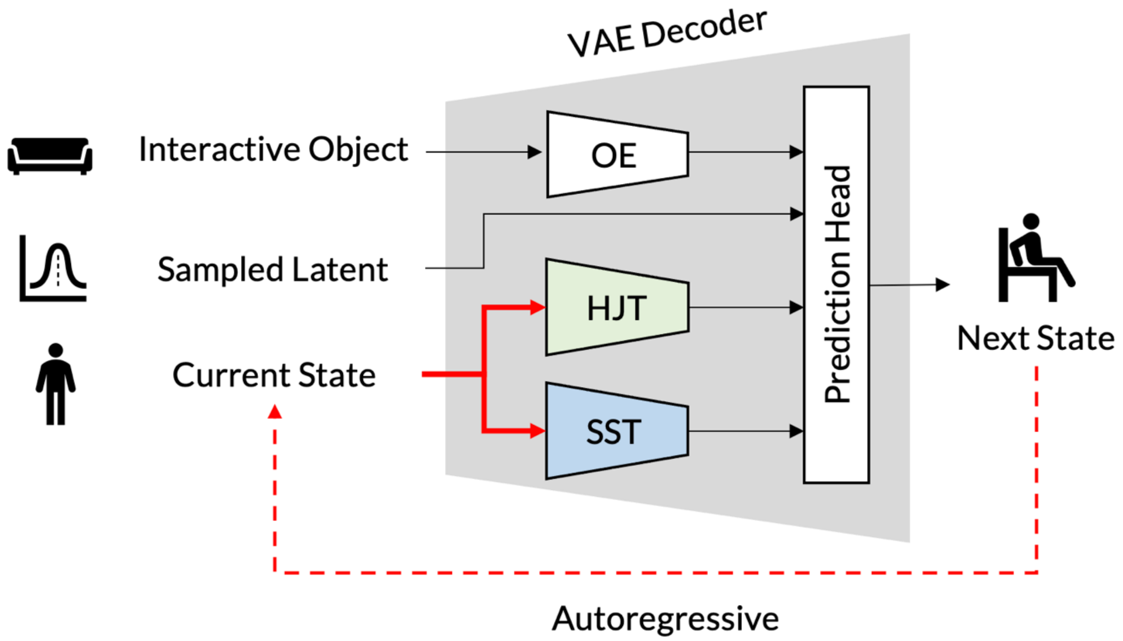
- We enhance the expressivity of auto-regressive model as it is important for **real-time applications** that includes dynamic inputs. For example, user keyboard input in the game controller.

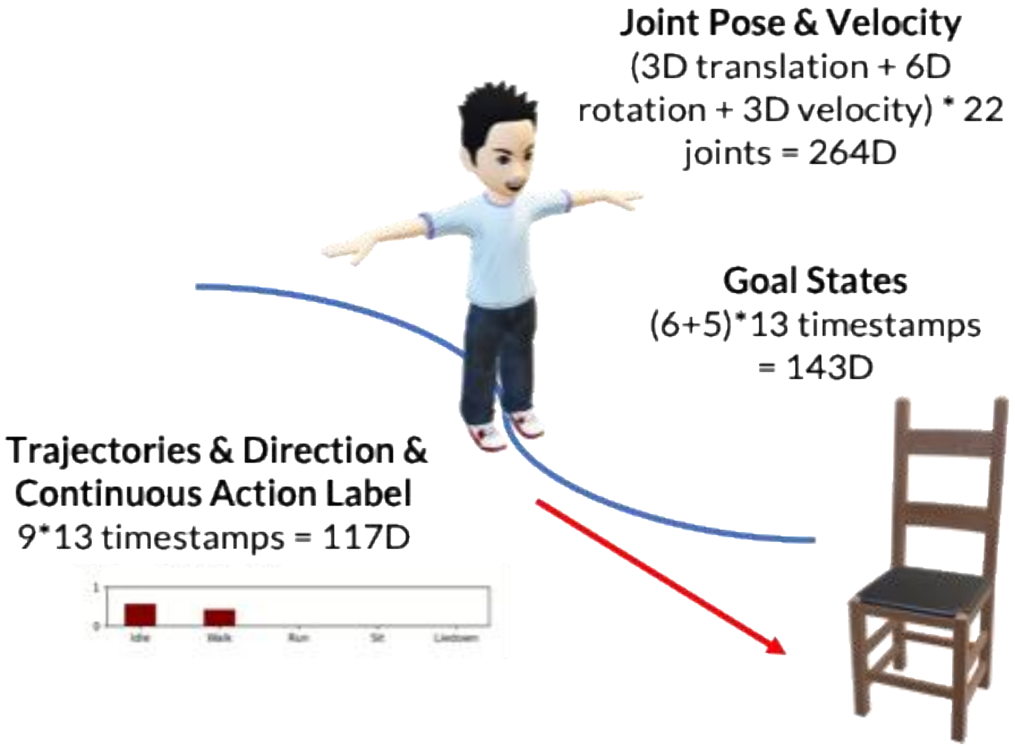




Challenges: Task specific data is limited

- High frame-rate motion capture data is required for training autoregressive motion prediction models. In addition, the data representation (i.e. state vector) are created only for a specific task.



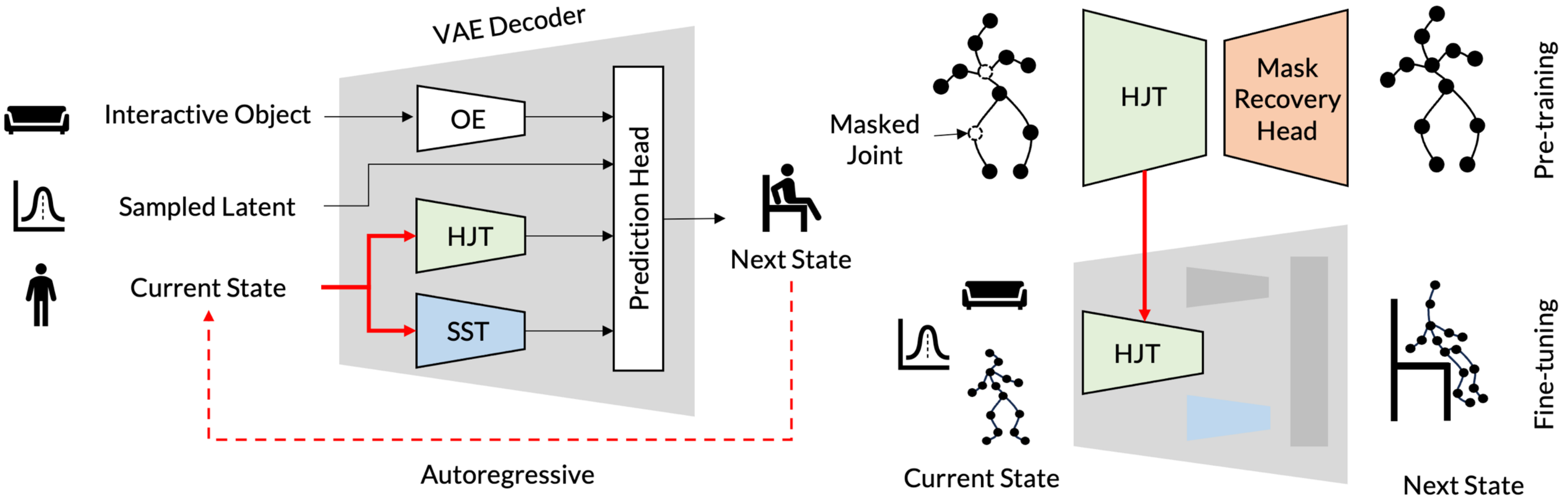

Joint Pose & Velocity
(3D translation + 6D rotation + 3D velocity) * 22 joints = 264D

Goal States
(6+5)*13 timestamps = 143D

Trajectories & Direction & Continuous Action Label
9*13 timestamps = 117D

Solution: Pre-trainable Transformer VAE

- To address the challenges posed by this limited data, we designed a transformer-based model to handle variable-length input that can be pretrained with static datasets in different representations.



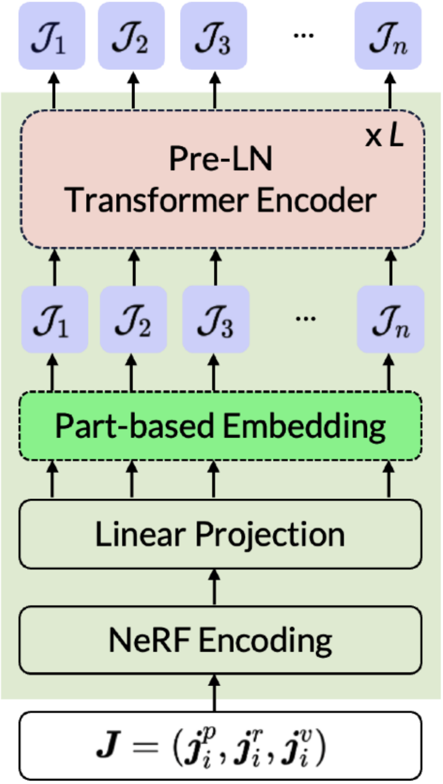
Task agnostic Human Joint Transformer (HJT)

Construction

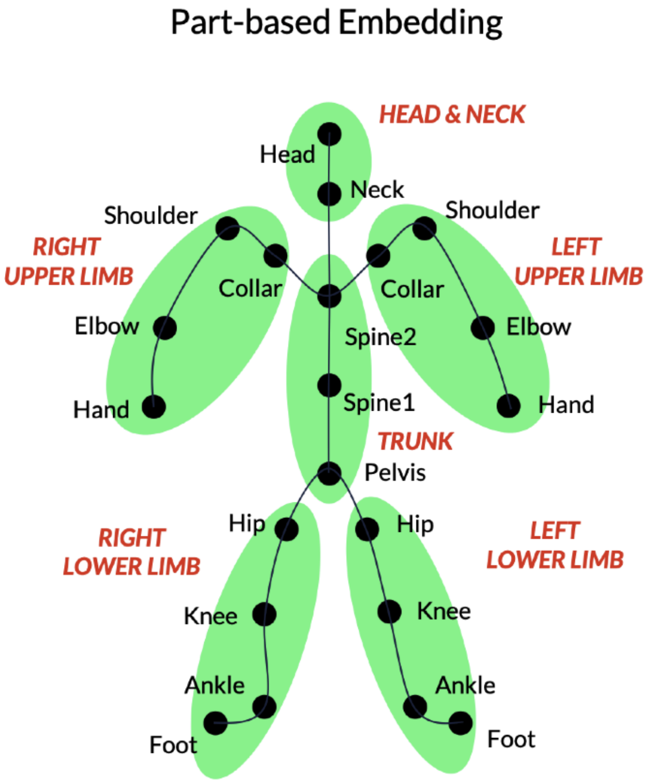
1. Input human joint state $\boldsymbol{J} = \{\boldsymbol{j}_i^p, \boldsymbol{j}_i^r, \boldsymbol{j}_i^v\}$ (position, rotation, and velocity)
2. Part-based embedding: Body part priors that is independent of number of joints.
3. Transformer: Exploit attention algorithm for feature extraction and correlation modeling

Advantages

Task-independent, Variable-length



$\mathcal{J}_1$ $\mathcal{J}_2$ $\mathcal{J}_3$ ... $\mathcal{J}_n$
x L
Pre-LN Transformer Encoder
$\mathcal{J}_1$ $\mathcal{J}_2$ $\mathcal{J}_3$ ... $\mathcal{J}_n$
Part-based Embedding
Linear Projection
NeRF Encoding
$\boldsymbol{J} = (\boldsymbol{j}_i^p, \boldsymbol{j}_i^r, \boldsymbol{j}_i^v)$



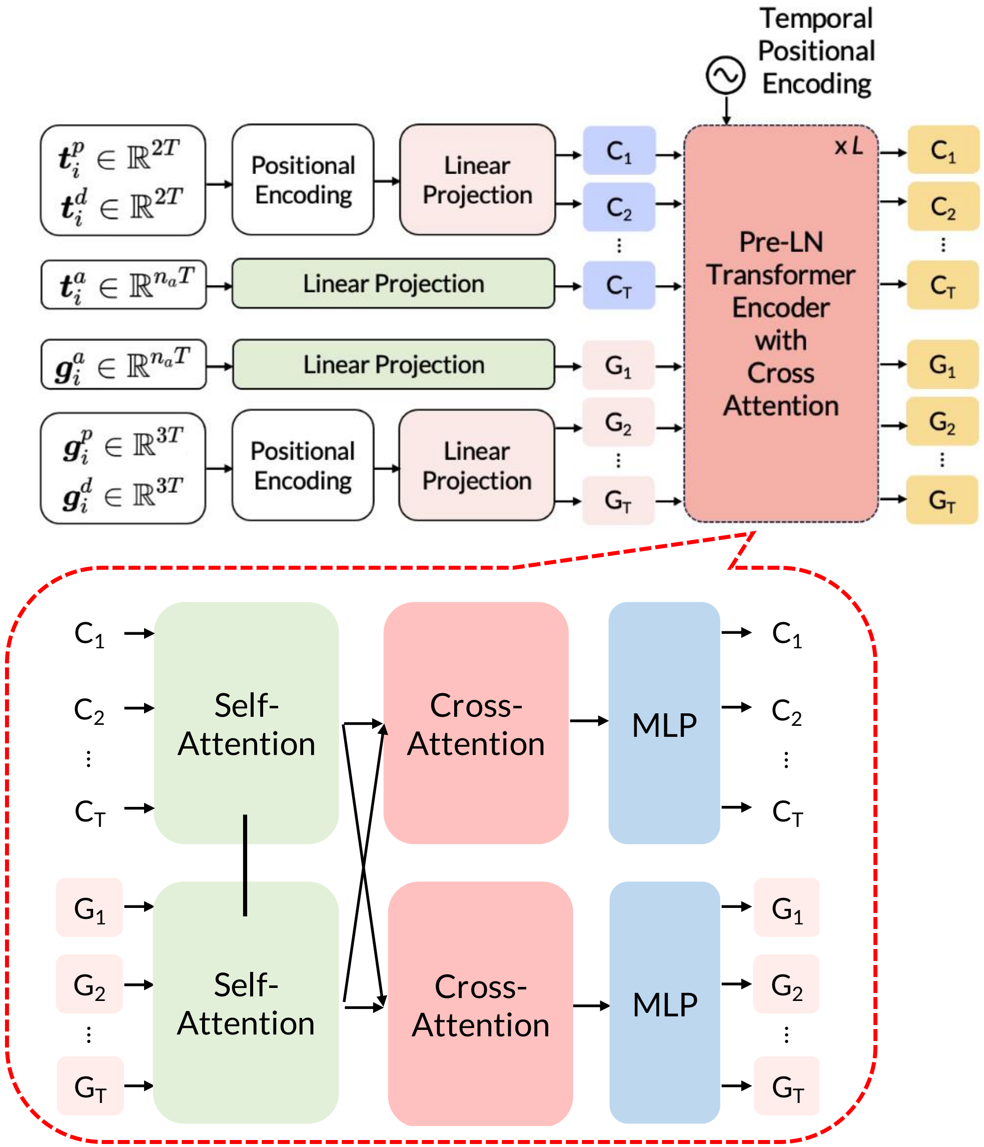
Task-specific Spatiotemporal State Transformer (SST)

Construction

1. Position and Direction are also mapped to high dimensional space by the same positional encoding.
2. Additional Linear Model is utilized to project the continuous action label.
3. The cross-attention layer models the correlation between trajectory and goal, allowing smooth transition between current state and goal state.

Advantages

Effective spatial-temporal learning through SA/CA



Experiments set up: Quantitative Results

- Settings:
  - SAMP [1]: Baseline model. Trained on SAMP dataset for 100 epoch.
  - CST (from scratch): Proposed model. Trained on SAMP dataset for 100 epoch.
  - CST (pre-trained): Proposed model. With HJT Pre-trained on HumanAct12 dataset or CMU Motion Capture Dataset and finetuned on SAMP dataset.
- Evaluation Metrics:
  - Mean Per Joint Position/Rotation Error (MPJPE, MPJRE) on 1-step, 10-step, 30-step, and 60-step prediction results.
  - Fréchet distance (FD) on generated sequence (10 steps)

[1] Mohamed Hassan et al., Stochastic Scene-Aware Motion Prediction, in ICCV 2021.

Quantitative Results

- Reduced the number of parameters from ~17 M to ~12 M.
- CST trained directly on SAMP dataset also performs better than SAMP.
- CST pre-trained on HumanAct12/CMU improved more than 10% from the baseline SAMP.

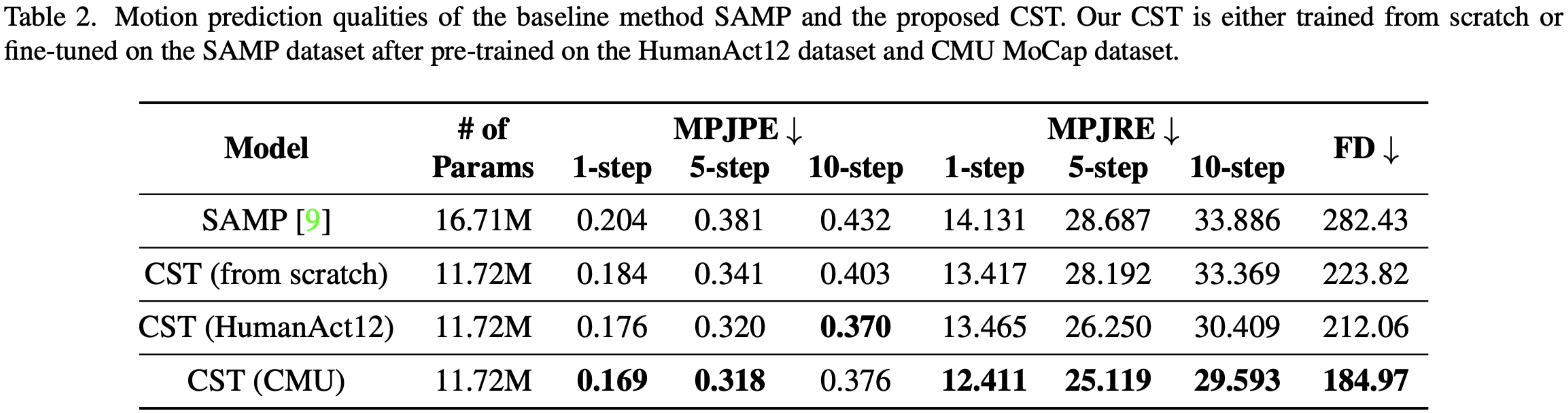
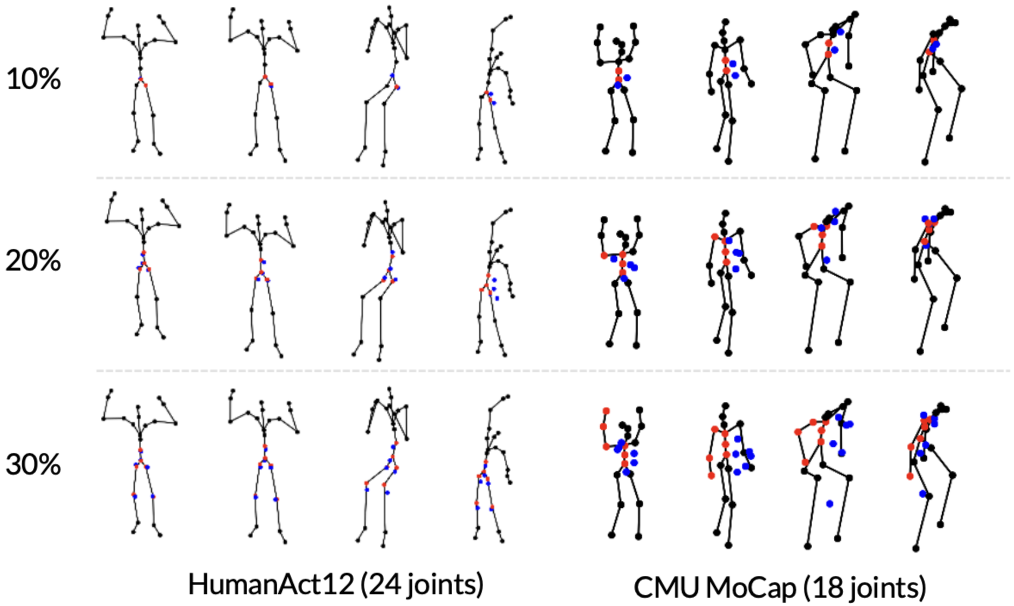


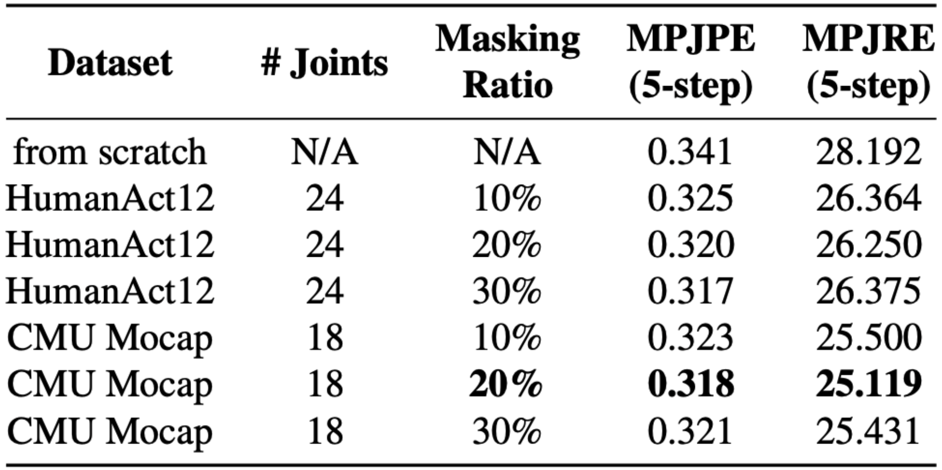
Table 2. Motion prediction qualities of the baseline method SAMP and the proposed CST. Our CST is either trained from scratch or fine-tuned on the SAMP dataset after pre-trained on the HumanAct12 dataset and CMU MoCap dataset.

Model	# of Params	MPJPE ↓ 1-step	MPJPE ↓ 5-step	MPJPE ↓ 10-step	MPJRE ↓ 1-step	MPJRE ↓ 5-step	MPJRE ↓ 10-step	FD ↓
SAMP [9]	16.71M	0.204	0.381	0.432	14.131	28.687	33.886	282.43
CST (from scratch)	11.72M	0.184	0.341	0.403	13.417	28.192	33.369	223.82
CST (HumanAct12)	11.72M	0.176	0.320	**0.370**	13.465	26.250	30.409	212.06
CST (CMU)	11.72M	**0.169**	**0.318**	0.376	**12.411**	**25.119**	**29.593**	**184.97**

Human Joint Transformer – Pretraining

- Task agnostic Human Joint Transformer (HJT) can be pre-trained to improve performance on human object interaction tasks.
- We adopt the **masked skeleton reconstruction** as the training scheme for HJT pre-training.
- Based on our observations, when the masking ratio is excessively high, training may collapse to an average pose and become stuck in a local minimum.

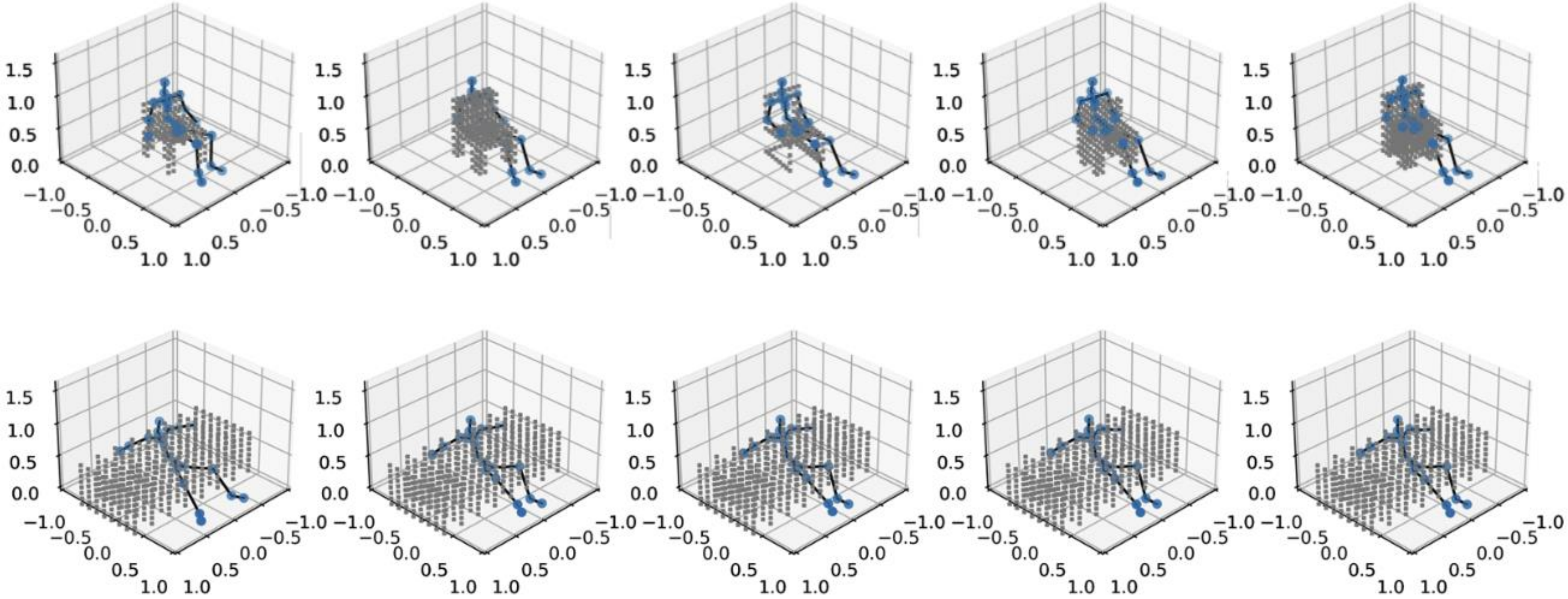




Dataset	# Joints	Masking Ratio	MPJPE (5-step)	MPJRE (5-step)
from scratch	N/A	N/A	0.341	28.192
HumanAct12	24	10%	0.325	26.364
HumanAct12	24	20%	0.320	26.250
HumanAct12	24	30%	0.317	26.375
CMU Mocap	18	10%	0.323	25.500
CMU Mocap	18	**20%**	**0.318**	**25.119**
CMU Mocap	18	30%	0.321	25.431

Qualitative Results

- We visualize the motion prediction results of the proposed CST along with the interactive object (e.g. chairs). Note that the penetration artifact is the consequence of imperfect fitting that exists in the training data.



Qualitative Results: Video

Summary

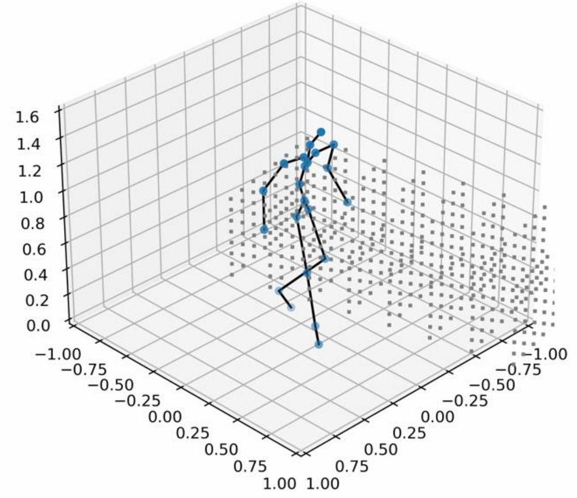
Task

Autoregressive human motion predictor conditioned on interactive object.

Method

Character State Transformer (CST)

1. **HJT/SST**: Applicable to different data format and scenarios so that we can introduce **pre-training** for better learning of the human motion.
2. **Positional Embedding**: Learnable Part-based positional encoding to provide priors on human structure.



Evaluation

CST pre-trained on HumanAct12 improved more than 10% from the baseline SAMP.

Recognition and Learning Algorithm Laboratory

Institute of SCIENCE TOKYO

NII
大学共同利用機関法人 情報・システム研究機構
国立情報学研究所
National Institute of Informatics

DENSO IT LAB

Thank you very much!

For more information, please visit our project page

https://kuan-wei-tseng.github.io/cst/

WACV 2025
TUCSON, ARIZONA • FEB 28 - MAR 4

CV4Smalls