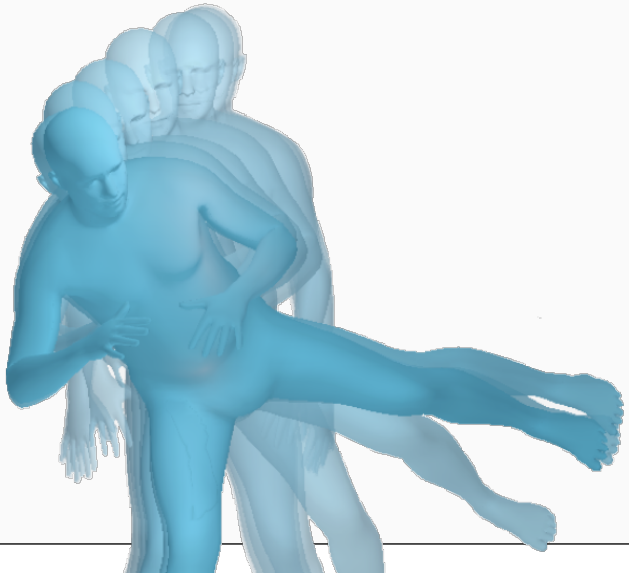
Action-Conditioned 3D Human Motion Synthesis with Transformer VAE

Mathis Petrovich[1], Michael J. Black[2], Gül Varol[1]
[1] LIGM, École des Ponts, Univ Gustave Eiffel, CNRS, France
[2] Max Planck Institute for Intelligent Systems, Tübingen, Germany

ICCV 2021

Guray Ozgur
University of Tübingen



Agenda

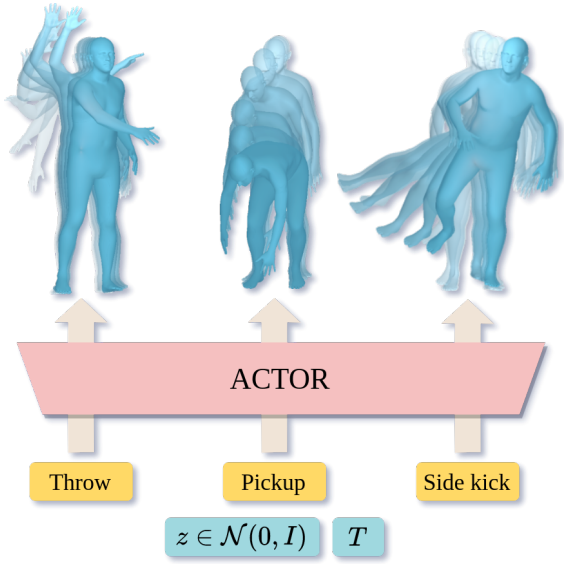
- Introduction
  - Goal and Motivation
  - Prior Work
- **ACTOR: AC**tion-Conditioned **T**ransf**OR**mer VAE
  - Method overview
  - Ablation study
  - Implementation details
- Results
  - Generating variable length sequences
  - Qualitative Results
  - Quantitative Results
  - Limitations
- Summary

Introduction

- Goal
- Motivation
- Prior Work

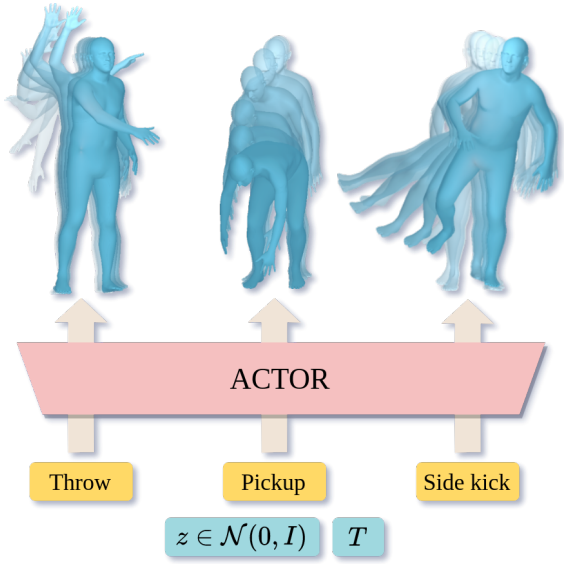
Goal

- **Generating synthetic** but **realistic** and **diverse** human motion sequence given an **action label**
- Learning from **noisy** 3D body poses estimated from **monocular** action recognition datasets



Motivation

- **Augmenting** existing MoCap **datasets**, which are expensive and limited in size
- **Serving as** additional **training data** for motion recognition
- A compact **action-aware latent space** for human motions



Petrovich et al. ICCV'21.

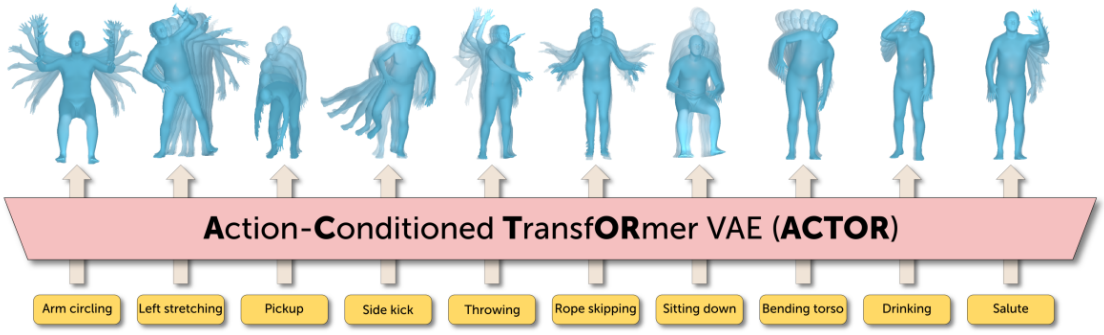
Prior Work on Motion Generation

- Human motion prediction (predicting poses from earlier poses)
- Unconstrained human motion synthesis (dominated by walking and running)
- **Conditioned human motion synthesis**
  - on Music (dance generation)
  - on Text
  - **on Action**
- Closest work

  Action2Motion, Guo et al. 20': An Autoregressive GRU-based (Gated Recurrent Unit) architecture

ACTOR: ACtion-Conditioned **T**ransf**OR**mer VAE

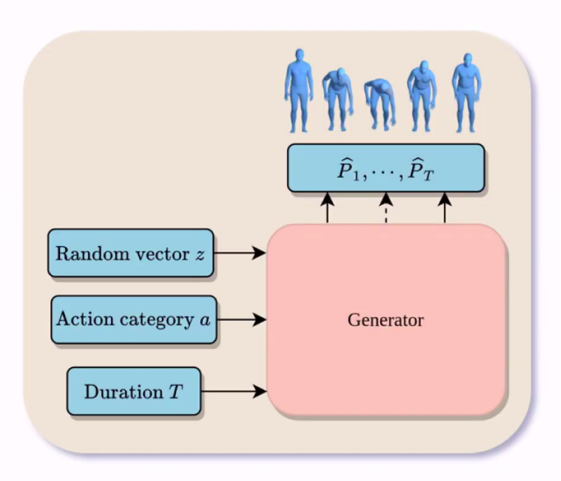
- Method overview
- Ablation study
- Implementation details



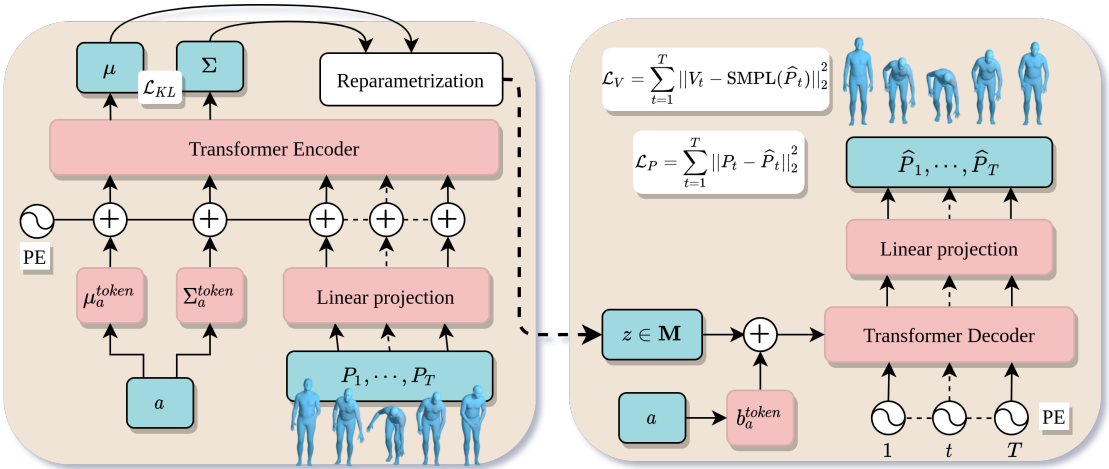
Petrovich et al. ICCV'21.

What is it?

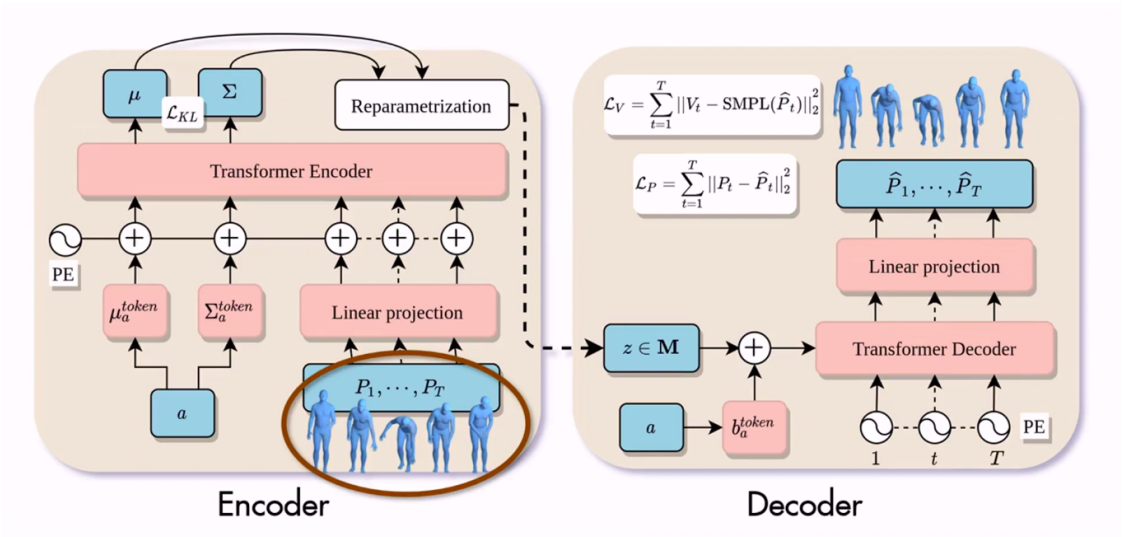
- Given
  - A single sequence-level latent vector
  - An action label
  - A specified duration
- Human motion synthesis in a single shot (non-autoregressive)
- Learnable tokens
- Variable length sequences with various body shapes
- Loss terms on rotations and vertices (SMPL)



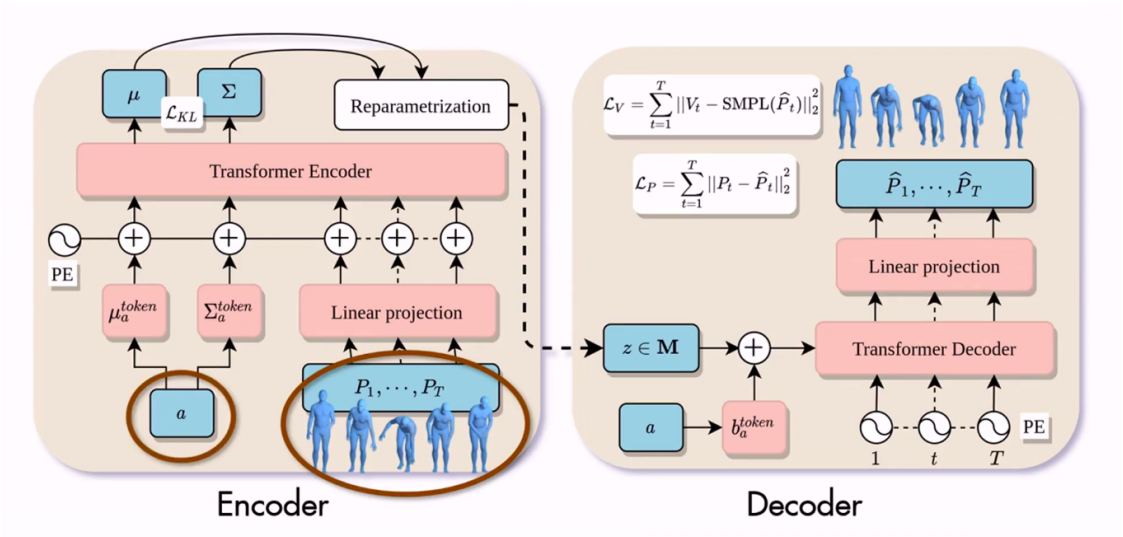
Encoder-Decoder



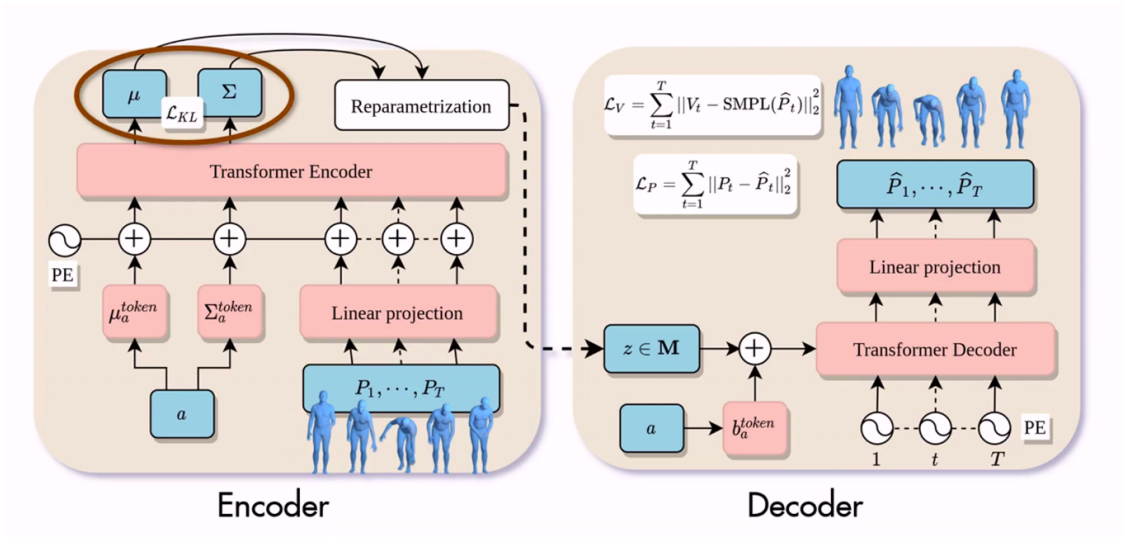
A sequence of poses of arbitrary length



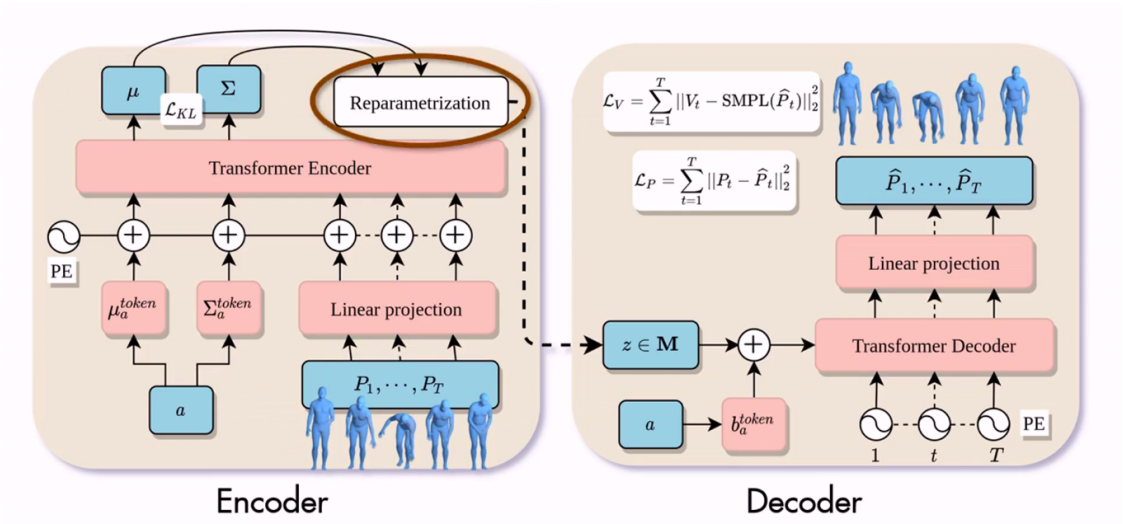
Action label



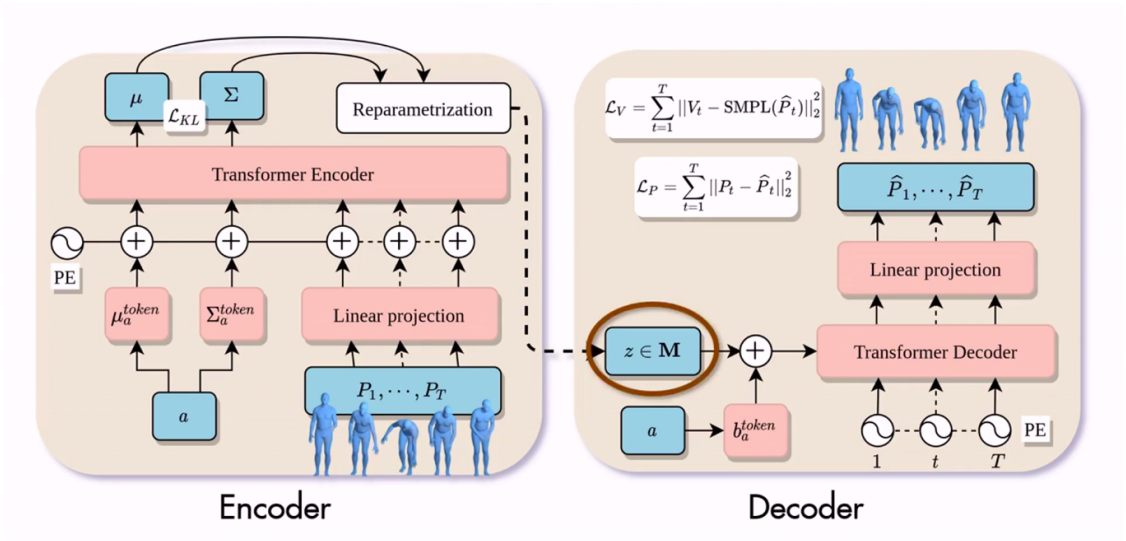
Distribution parameters of motion latent space



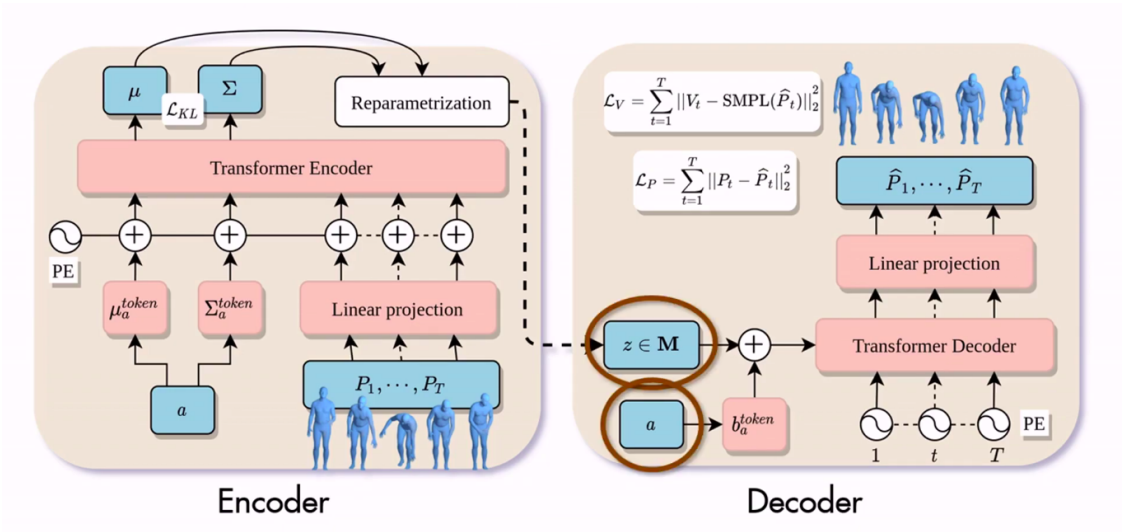
Reparametrization trick of VAEs



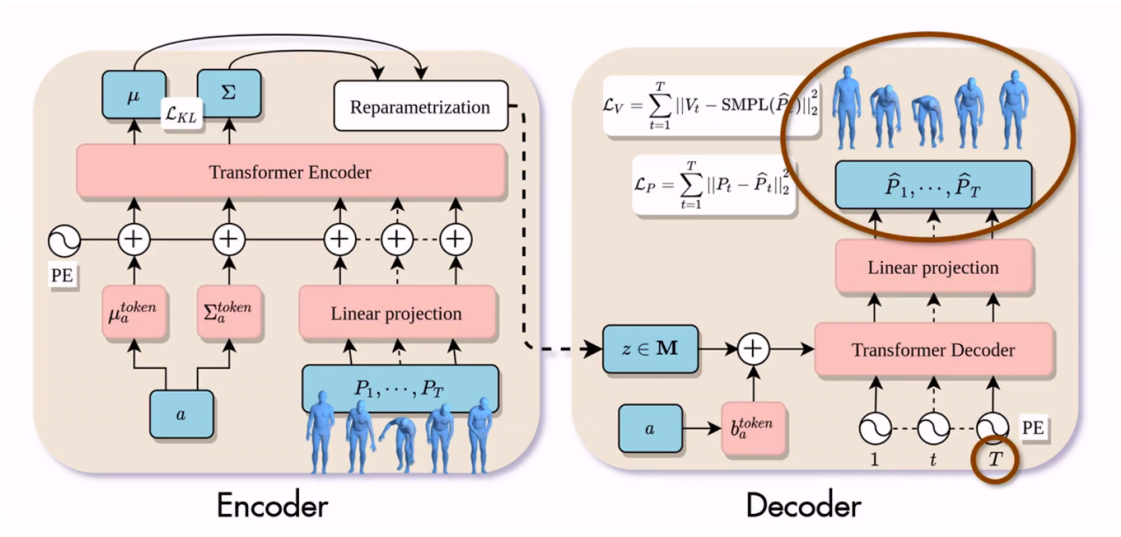
Sampling a latent vector



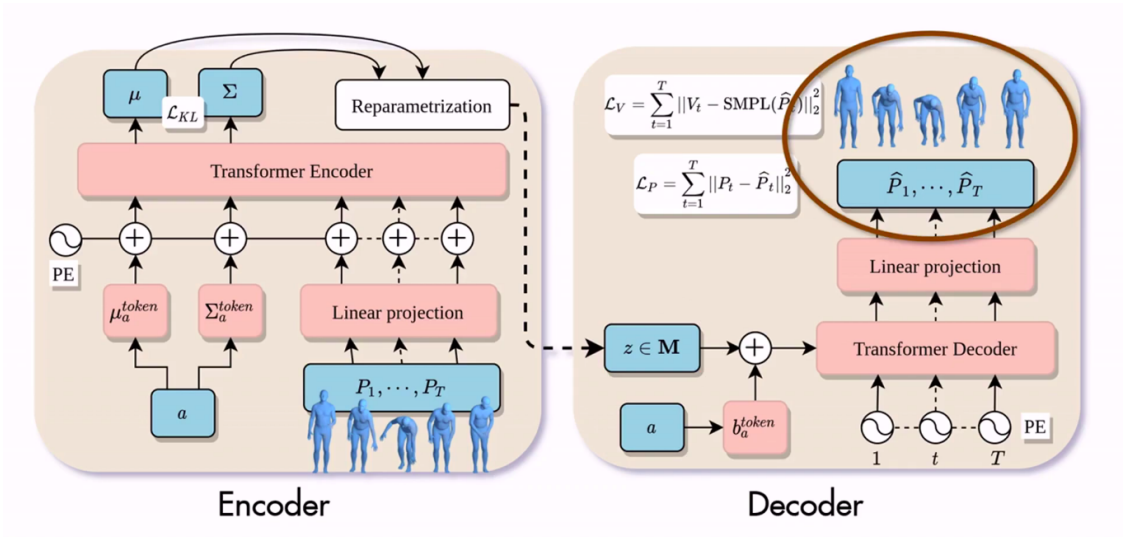
Conditioning on an action label



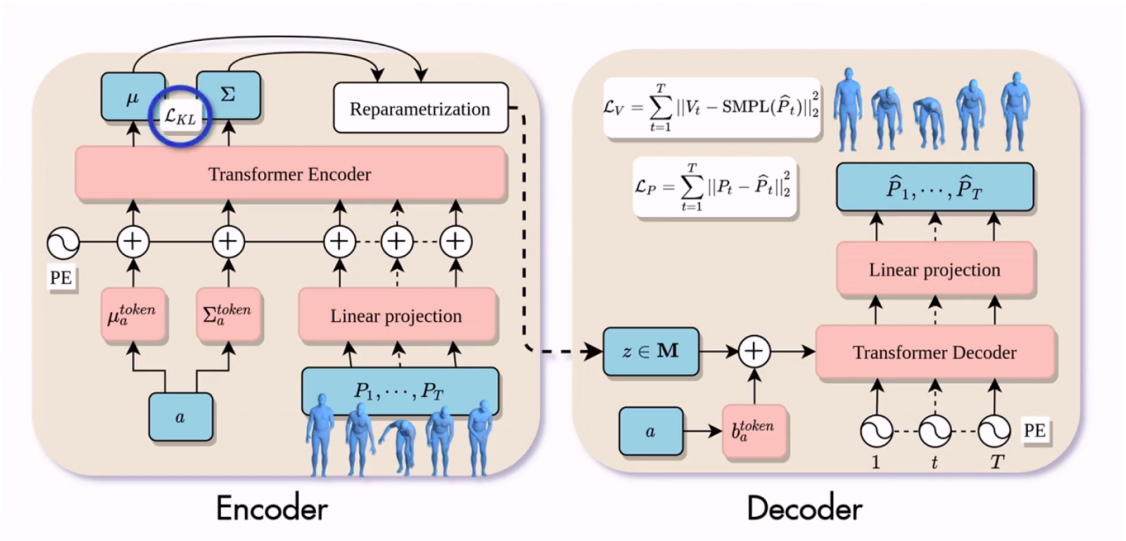
Given duration



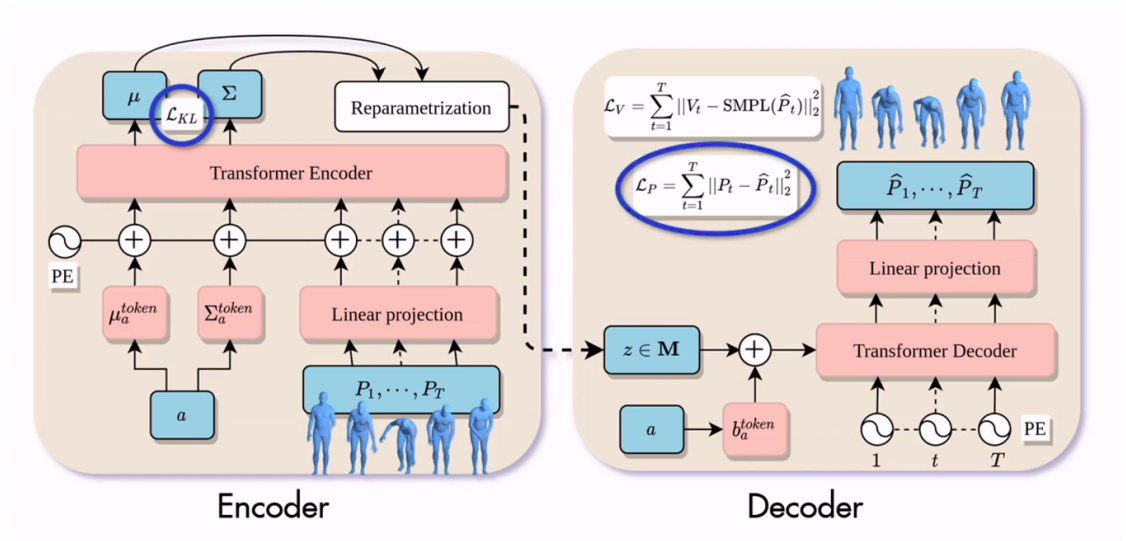
Decoder generates human motion



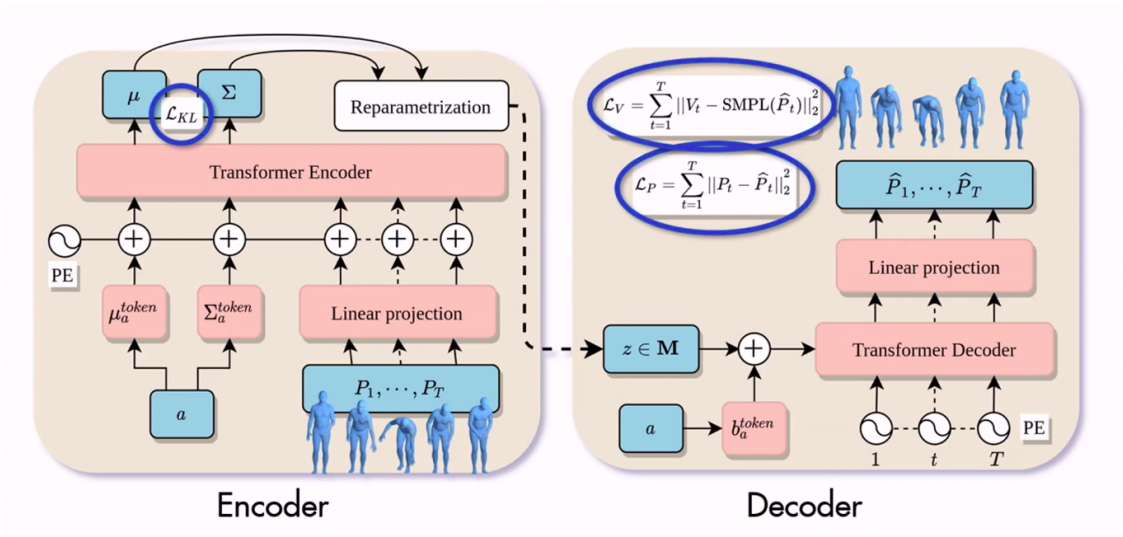
KL loss regularize the latent space



1^{st} Reconstruction loss: L_2 loss on SMPL poses



2^{nd} Reconstruction loss: L_2 loss on vertices



Training Data: Estimates of SMPL poses (NOISY)

NTU13 (13 actions)

- RGB-D dataset, subset of NTU-120
- SMPL poses estimated with VIBE

UESTC (40 actions)

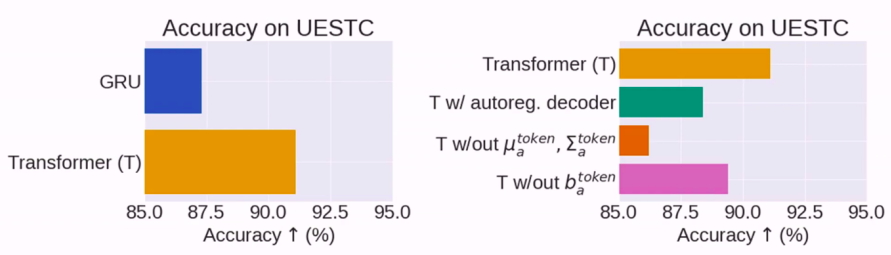
- RGB-D dataset
- SMPL poses estimated with VIBE

HumanAct12 (12 actions)

- RGB-D + polarization images, subset of PHSPDataset
- SMPL poses estimated

NTU13: Liu et al. 18', UESTC: Ji et al. 18', Action2Motion: Guo et al. 20', SMPL: Loper 15', VIBE: Kocabas et al. 20'

Ablation study



NTU13: Liu et al. 18', UESTC: Ji et al. 18', Action2Motion: Guo et al. 20', SMPL: Loper 15', VIBE: Kocabas et al. 20'

Implementation details

- **Architectural details.**
  - embedding dimensionality to 256
  - the number of layers to 8
  - the number of heads in multi-head attention to 4
  - the dropout rate to 0.1
  - the dimension of the intermediate feedforward network to 1024
  - Gaussian Linear Error Units (GELU)
- **Runtime.** Training takes 24 hours for 2K epochs on NTU, 19h hours for 5K epochs on HumanAct12, and 33 hours for 1K epochs on UESTC on a single Tesla V100 GPU, using 4GB GPU memory with batch size 20.

Additional experiments

- Weight of the KL loss

$$L = L_V + L_P + \lambda_{KL} L_{KL}$$

  $\lambda_{KL} = 1e-5$

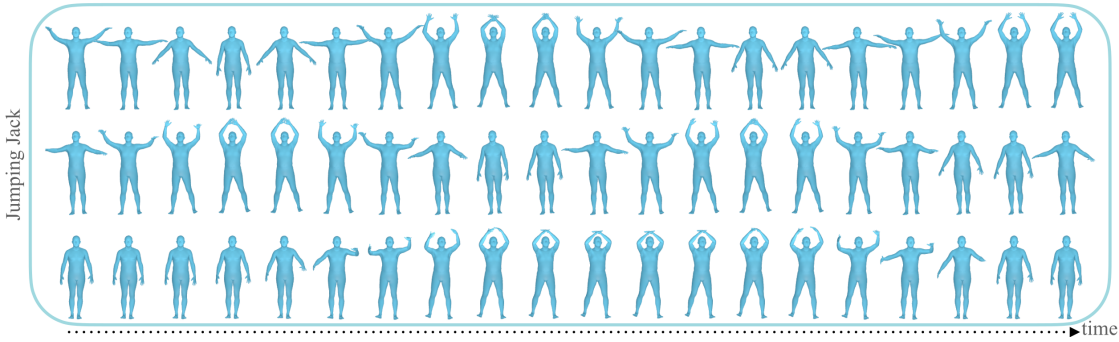
- Influence of the batch size
  10, **20**, 30, 40
- Number of Transformer layers
  2, 4, 6, **8**
- SMPL pose parameter representation
  quaternions, rotation matrices and **6D continuous representations**

NTU13: Liu et al. 18', UESTC: Ji et al. 18', Action2Motion: Guo et al. 20', SMPL: Loper 15', VIBE: Kocabas et al. 20'

Results

- Qualitative Results
- Generating variable length sequences
- Quantitative Results
- Limitations

Qualitative Results (Realistic, Diverse, Smooth)



Petrovich et al. ICCV'21.

Qualitative Results (Realistic, Diverse, Smooth)

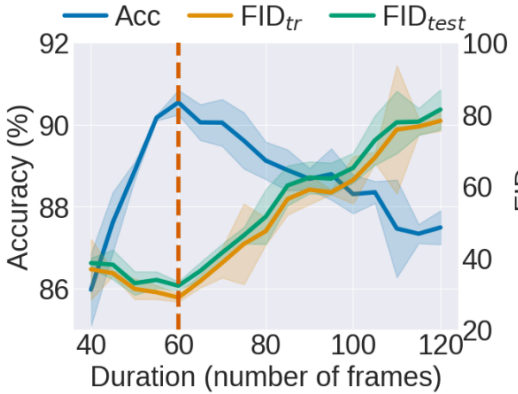
Metrics

- **FID:** Fréchet inception distance
  - measures similarity between the distributions
- **Acc.:** Action Recognition Accuracy
- **Div.:** Diversity
- **Multimod.:** Multi-modality
  - measures per-action diversity

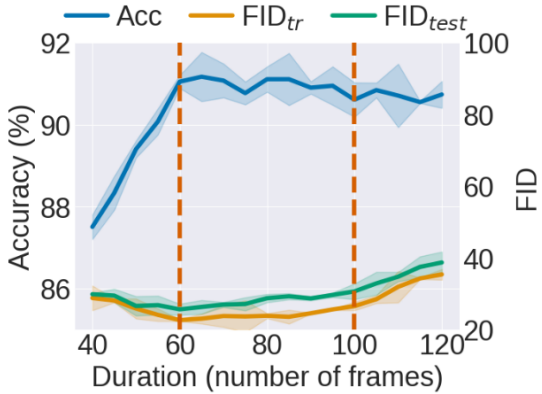
Generating variable length sequences

- Model capability is evaluated on UESTC with
  - fixed size (60 frames)
  - variable size (60-100 frames)
- Performance increase if the model sees duration variation on the training.

Generating variable length sequences



Trained with fixed duration
(60 frames)



Fine-tuned with variable duration
(between 60-100 frames)

Petrovich et al. ICCV'21.

Comparison with previous work

Method	NTU-13				HumanAct12			
	$FID_{tr}\downarrow$	Acc.$\uparrow$	Div.$\rightarrow$	Multimod.$\rightarrow$	$FID_{tr}\downarrow$	Acc.$\uparrow$	Div.$\rightarrow$	Multimod.$\rightarrow$
Real [Action2Motion]	0.03	99.9	7.11	2.19	0.09	99.7	6.85	2.45
Real*	0.02	99.8	7.07	2.25	0.02	99.4	6.86	2.60
CondGRU	28.31	7.80	3.66	3.58	40.61	8.0	2.38	2.34
Two-stage GAN	13.86	20.2	5.33	3.49	10.48	42.1	5.96	2.81
Act-MoCoGAN	2.72	**99.7**	6.92	0.91	5.61	79.3	6.75	1.06
Action2Motion	0.33	94.9	7.07	2.05	2.46	92.3	7.03	2.87
ACTOR (ours)	**0.11**	97.1	7.08	2.08	**0.12**	**95.5**	6.84	2.53

Limitations

- The maximum duration it can generate depends on computational resources since it outputs all the sequence at once.
- A set of actions (No open-vocabulary actions)
- No collision check (Mesh interpenetration)
- No motion generation from unconstrained video
- No detailed motion (hands)

Summary

Summary

- Sequence-level motion embedding
- Not autoregressive (gives the output in a single shot)
- Trained on noisy motion estimates
- Denoising as a side effect
- Parametric body model

Questions?