The Language of Motion: Unifying Verbal and Non-verbal Language of 3D Human Motion

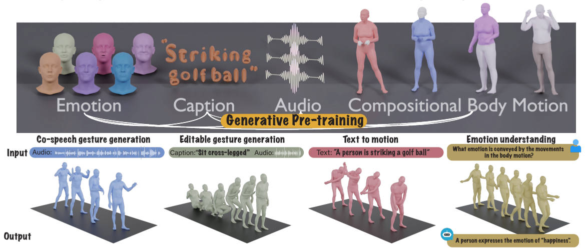
Changan Chen*, Juze Zhang*, Shrinidhi Kowshika Lakshmikanth*, Yusu Fang, Ruizhi Shao, Gordon Wetzstein, Li Fei-Fei, Ehsan Adeli

Stanford University (* denotes equal contribution)

Project page

CVPR Nashville JUNE 11-15, 2025

Unifying Verbal and Non-verbal Languages


"Striking golf ball"
Emotion
Caption
Audio
Compositional Body Motion
Generative Pre-training
Co-speech gesture generation
Editable gesture generation
Text to motion
Emotion understanding
Input
Audio:
Caption:"Sit cross-legged" Audio:
Text: "A person is striking a golf ball"
What emotion is conveyed by the movements in the body motion?
Output
A person expresses the emotion of "happiness".

Motivation: building a foundation model for human behavior

Challenges:

- Existing tasks and models assume specific inputs and outputs
- Misalignment (of the latent space) between different modalities

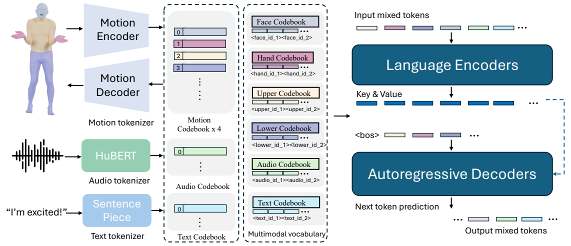
Our ideas: using language models to unify understanding & generation for 3D human motion

- Language models have strong semantic understanding
- Generative pre-training to align latent spaces across modalities

Our Approach

Model pipeline

- Tokenization: each modality is tokenized separately
- Generation: T5 maps the sequence of input tokens to the output token sequence of any modality

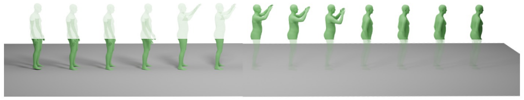


Tokenization

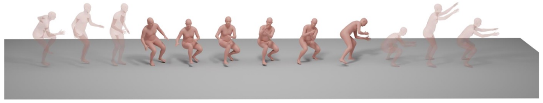
- Compositional body motion tokenization for expressivity
  - Divide the body into four parts: head, upper, lower, and hand
  - One VQ-VAE for each body part
- Speech tokenization: HUBERT
- Text tokenization: WordPiece

Pre-training for Modality Alignment

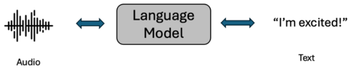
- Compositional body motion alignment: our body motion is inherently compositional
  - Spatial: predict one body part from another



  - Temporal: predict masked off motion segments



- Audio-text alignment
  - Mutually predict audio and text from each other



Post-training for Instruction Following

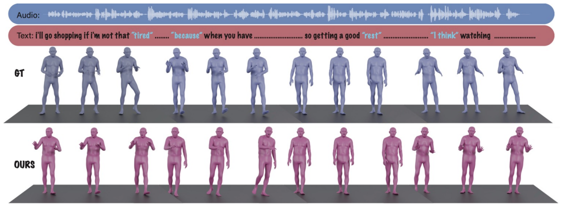
- Compile different tasks into instructions (650 in total)
- Finetune the language model to follow these instructions

Task	Input	Output
Audio-to-Full Motion	Based on [audio], generate a synchronized movement sequence involving both face, hands, upper and lower body. Listen to [audio] and produce movements that involve both the upper and lower body in harmony.	[face][hands] [upper][lower]
Motion-to-Emotion	What emotion is conveyed by the movements in the face, hands, upper body, and lower body within [face][hands][upper][lower]? Examine the face, hand, upper, and lower body movements in [face][hands] [upper][lower] to interpret the emotional tone.	[emotion]

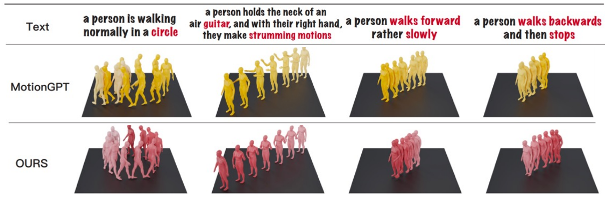
⋮

Experiments

Co-Speech Gesture Generation

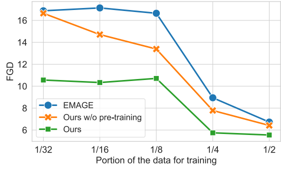

Audio:
Text: i'll go shopping if i'm not that "tired" .... "because" when you have .... so getting a good "rest" .... "I think" watching ....
GT
OURS

Text-to-motion Generation

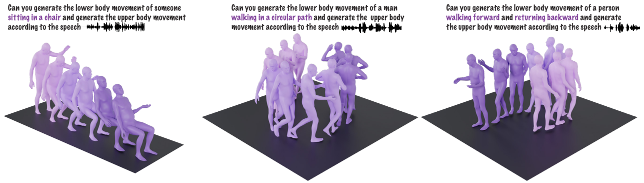

Text
a person is walking normally in a circle
a person holds the neck of an air guitar, and with their right hand, they make strumming motions
a person walks forward rather slowly
a person walks backwards and then stops
MotionGPT
OURS

Effect of Generative Pre-training

	FGD ↓	BC↑	Diversity↑
W/o pre-training	5.501	7.721	14.281
W/o A2T	5.443	7.721	14.499
W/o spatial	6.336	7.381	14.173
W/o temporal	6.800	7.341	13.810
W/o motion	7.776	7.344	14.640
Ours	**5.301**	**7.780**	**15.165**



Editable Gesture Generation


Can you generate the lower body movement of someone sitting in a chair and generate the upper body movement according to the speech
Can you generate the lower body movement of a man walking in a circular path and generate the upper body movement according to the speech
Can you generate the lower body movement of a person walking forward and returning backward and generate the upper body movement according to the speech