Generative models for human's motion

- Ho Beom Jeon
- UST-ETRI
- Social Robotics Lab

- 23.03.22

Contents

1) What is Generative models?

배경: 생성 모델이란 무엇인가요?

의견 보내기

'세대적 적대 신경망(GAN)'이란 *차세대*는 *차별* 모델과 대조되는 통계 모델 클래스를 설명합니다.

비공식적:

- **생성** 모델은 새 데이터 인스턴스를 생성할 수 있습니다.
- **차별** 모델은 서로 다른 종류의 데이터 인스턴스를 구분합니다.

생성 모델은 진짜 동물처럼 보이는 새로운 동물 사진을 생성할 수 있고, 분류 모델은 고양이에게 개를 말할 수 있습니다. GAN은 생성 모델의 한 가지 종류일 뿐입니다.

더 공식적으로, 데이터 인스턴스 X 집합과 라벨 Y 집합을 지정하면 다음과 같이 됩니다.

- **생성** 모델은 조인 확률 p(X, Y)를 캡처하거나 라벨이 없는 경우 p(X)만 캡처합니다.
- **차별** 모델은 조건부 확률 p(Y | X)를 캡처합니다.

생성 모델은 데이터 자체의 분포를 포함하고 특정 예시의 가능성을 알려줍니다. 예를 들어 시퀀스의 다음 단어를 예측하는 모델은 일반적으로 생성 모델 (일반적으로 GAN보다 훨씬 간단함)입니다. 단어의 시퀀스에 확률을 할당할 수 있기 때문입니다.

구별 모델은 지정된 인스턴스가 발생할 가능성이 있는지에 대한 질문을 무시하고 인스턴스에 라벨이 적용될 가능성을 알려줍니다.

이는 매우 일반적인 정의입니다. 생성 모델에는 여러 종류가 있습니다. GAN은 일종의 생성 모델입니다.

https://developers.google.com/machine-learning/gan/generative?hl=ko

1) What is Generative models?

생성 모델은 어려움

생성 모델은 유사한 차별 모델보다 더 어려운 작업을 처리합니다. 생성 모델은 *더 많이* 모델링해야 합니다.

이미지 생성 모델에서는 배처럼 보이는 물체가 물처럼 보이는 물체 근처에 나타날 가능성이 있고 이마에 눈이 나타나지 않을 가능성이 있습니다. 이것들은 매우 복잡한 분포입니다.

반면에 구별 모델은 몇 가지 이야기 패턴을 찾아서 '범선'과 '보트가 아님'의 차이를 학습할 수 있습니다. 생성 모델이 타당해야 하는 많은 상관관계를 무시할 수 있습니다.

차별 모델은 데이터 공간에 경계를 그리는 반면 생성 모델은 공간 전체에 데이터가 배치되는 방식을 모델링하려고 합니다. 예를 들어 다음 다이어그램은 필기 입력의 구분 및 생성 모델을 보여줍니다.

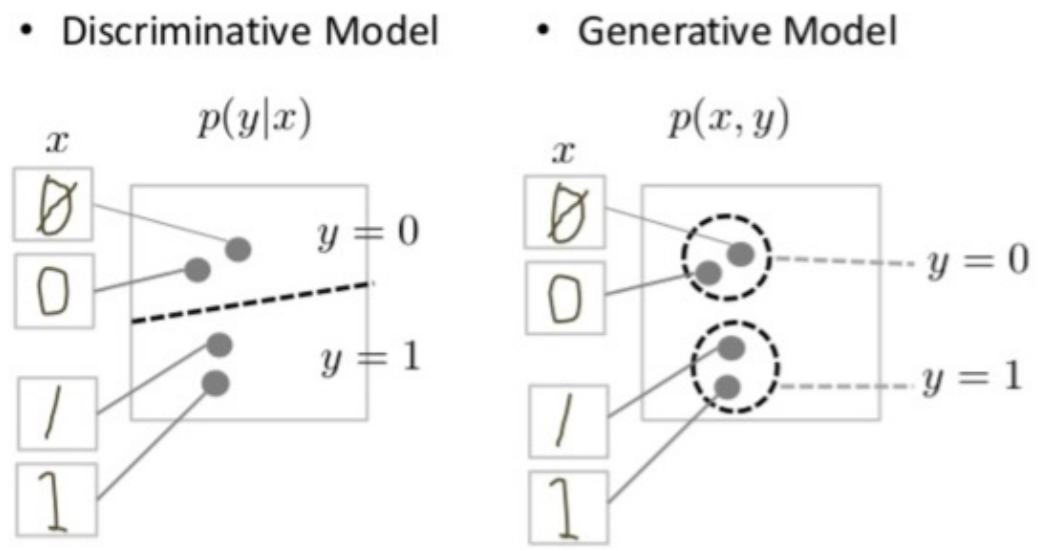


그림 1: 손으로 쓴 숫자의 차별적이고 생성적인 모델

https://developers.google.com/machine-learning/gan/generative?hl=ko

1) What is Generative models?

Stable diffusion model은 이미지를 생성

https://civitai.com/models/4201/realistic-vision-v13

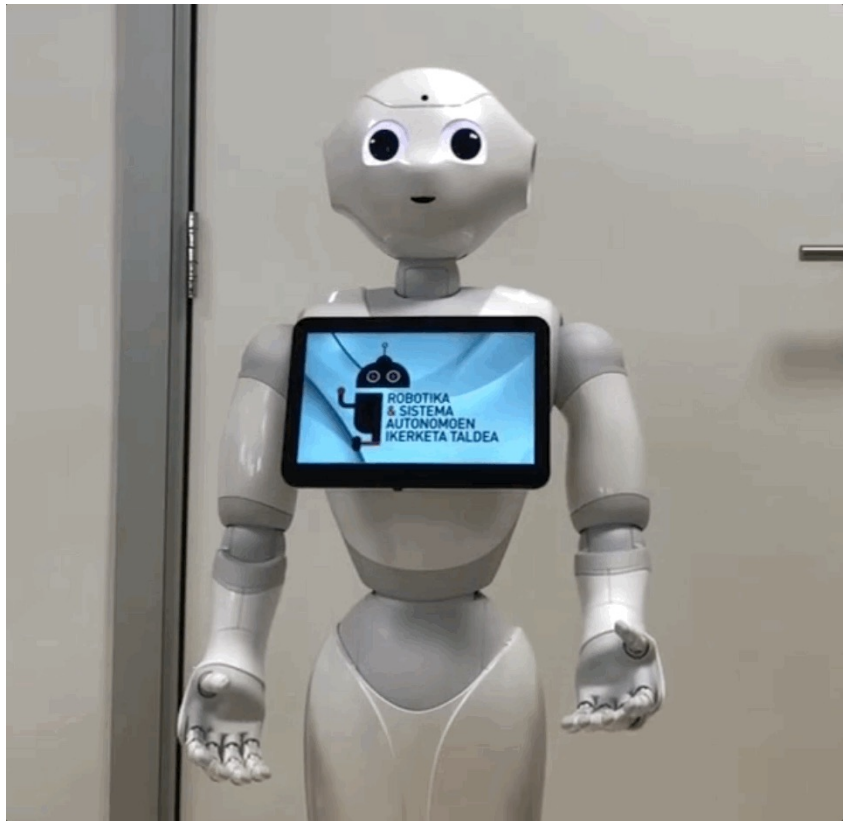
ChatGPT 모델은 텍스트를 생성

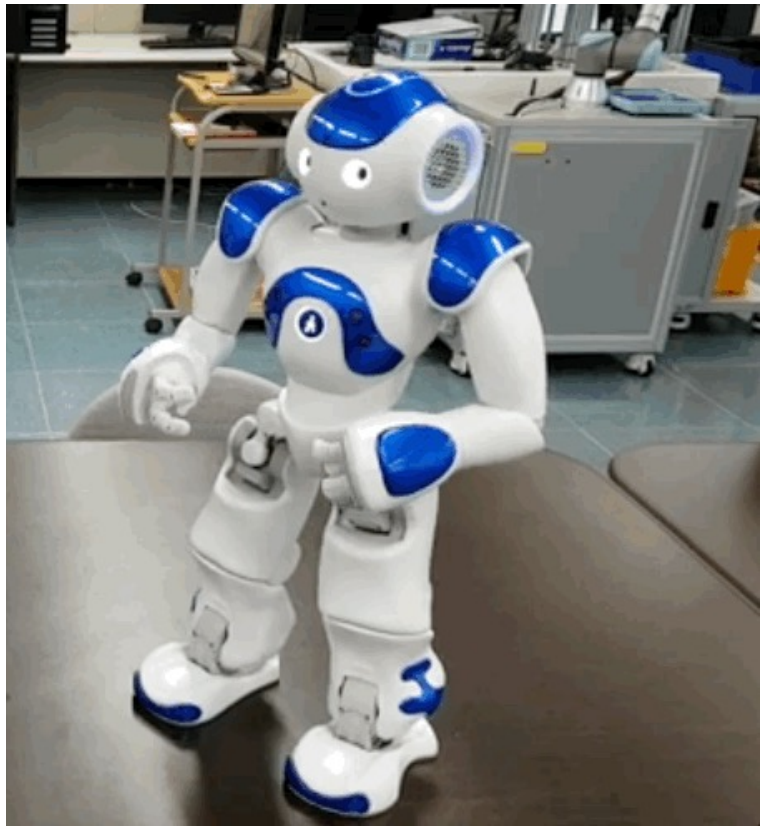
1. Title: Text-based Fire And Arson Detection Using Zero Shot Inference
Summary :
(1): 이 연구는 화재 감지 및 추적 분야에서 새로운 방법을 제안합니다.
(2): 기존 방식은 광범위한 데이터 수집으로 인한 비용문제와 데이터 다양성 부족 문제가 있었습니다. 이에 따라, 이 논문에서는 데이터셋에 대한 의존 없이, 첫 화재 감지와 화재 범인 추적을 할 수 있는 Zero-Shot 종합 임베딩 방법론을 제안합니다.
(3): 제안된 방법은 이미지와 자연어 데이터를 함께 사용하여 사람의 추적 정보, 화재 추적 정보, 텍스트 간 Inference를 통해 Zero-Shot 배제 인식을 수행합니다.
(4): 제안된 방법의 성능은 KISA fire dataset 등으로 검증되어, 다른 기존 학습 방법과 비교하여 높은 분류 성능을 보여주고 있습니다.
Conclusion
- 이 연구는 데이터셋에 대한 의존 없이 화재 및 방화 감지 분야에서 Zero-Shot 방법론을 제안하며, 기존 방식의 한계를 극복하고자 하는 의미가 있다.
- 혁신적인 측면: 제안된 Zero-Shot 방법론은 이전 연구에서 제시된 학습 방식과는 다른 새로운 시각을 제시하고 있으며, 데이터셋 부족 문제를 극복하고자 하는 의지를 보인다.
- 성능 측면: 제안된 방법의 성능은 여러 화재 데이터셋을 사용해 평가되었으며, 다른 기존 방법과 비교하여 높은 분류 성능을 보였다.
- 작업 부하 측면: 논문에서 사용된 데이터셋은 다른 연구들에서 사용되었거나 공개된 것이며, 따라서 작업 부하가 적은 편에 속한다.

https://github.com/kaixindelele/ChatPaper

1) What is Generative models?

사람의 동작(Keypoint Sequence)을 생성하는 모델은?
→ Co-Speech Gesture Generation: 오디오와 발화 문장을 통해 모션 생성





https://maum.ai/
https://sites.google.com/view/youngwoo-yoon/projects/co-speech-gesture-generation?pli=1
https://www.youtube.com/watch?v=AW3BmfS7DIY

1) What is Generative models?

사람의 동작(Keypoint Sequence)을 생성하는 모델은?
→ Conditioned Motion Generation : 행동 라벨 혹은 텍스트 설명을 통해 모션 생성

A tall and skinny female soldier that is arguing.

A skinny ninja that is raising both arms.

A tall and fat Iron Man that is running.

https://github.com/hongfz16/avatarclip

2. Dataset

1) Co-Speech Gesture Generation

TED Gesture Dataset
Robots Learn Social Skills: End-to-End Learning of Co-Speech Gesture Generation for Humanoid Robots. ICRA2019

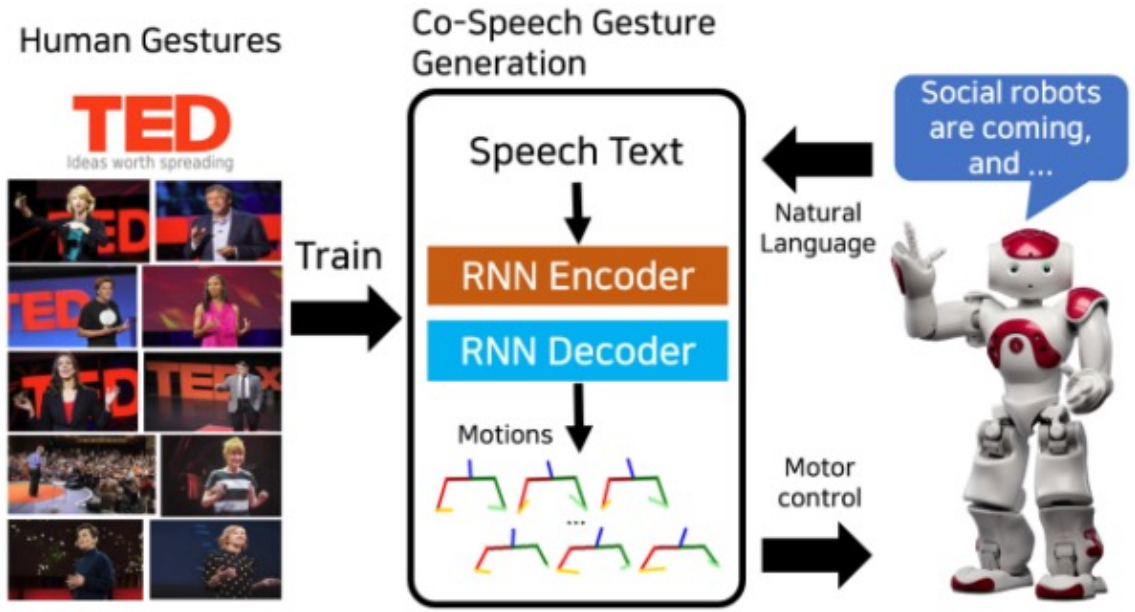


Figure 1. We address a problem of making co-speech gestures for a given speech text. The proposed model generates a sequence of upper-body poses, and it is trained from human gestures in TED talks.

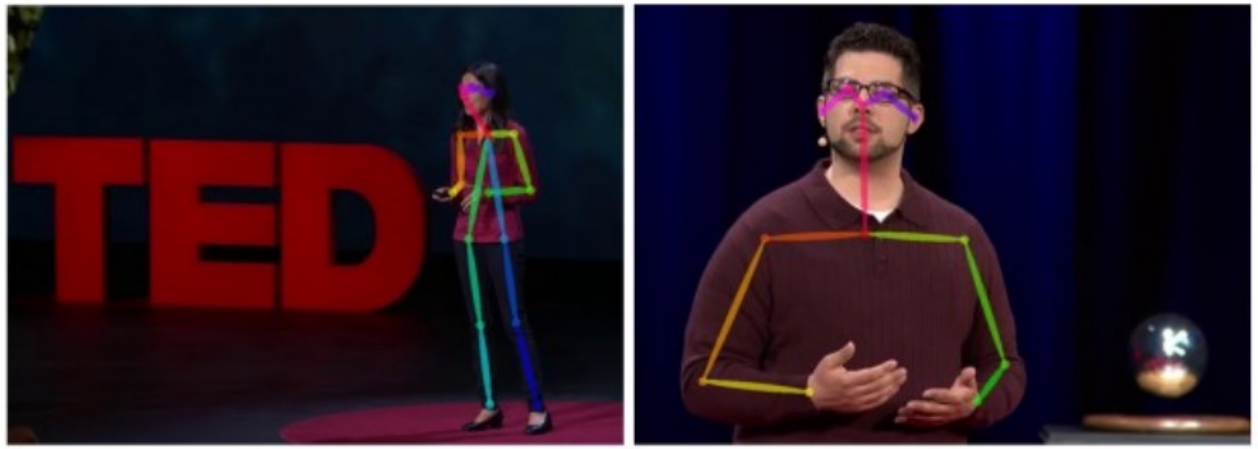
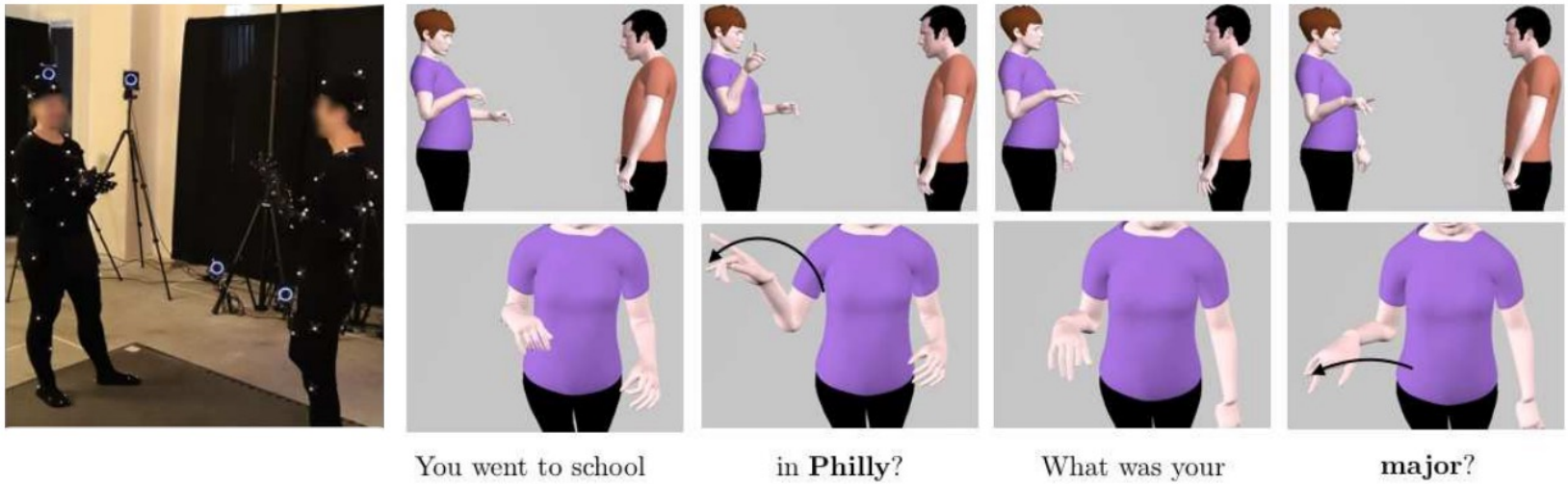


Figure 2. Samples of the TED Gesture Dataset. Extracted human poses are overlaid on the images.

https://github.com/youngwoo-yoon/youtube-gesture-dataset

Talking With Hands 16.2M
A Large-Scale Dataset of Synchronized Body-Finger Motion and Audio for Conversational Motion Analysis and Synthesis, ICCV2019



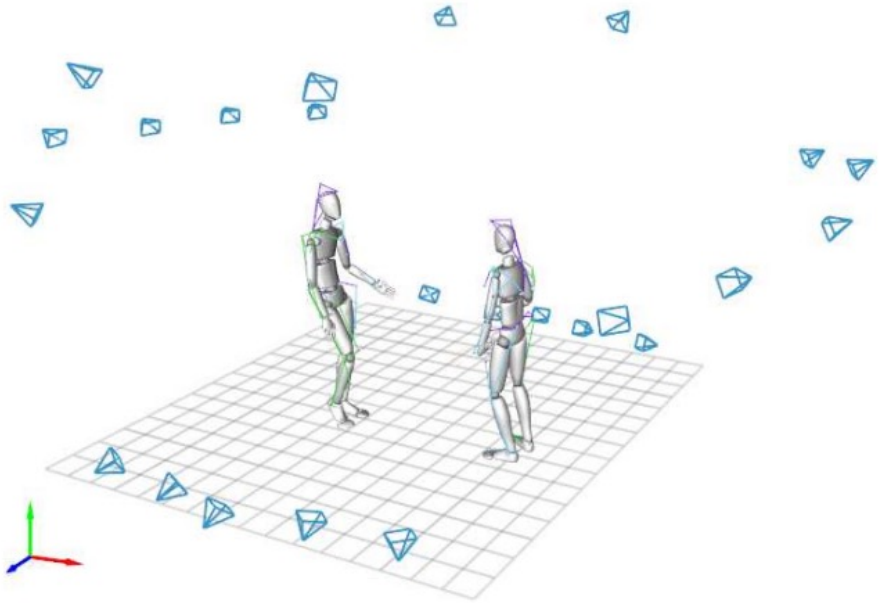


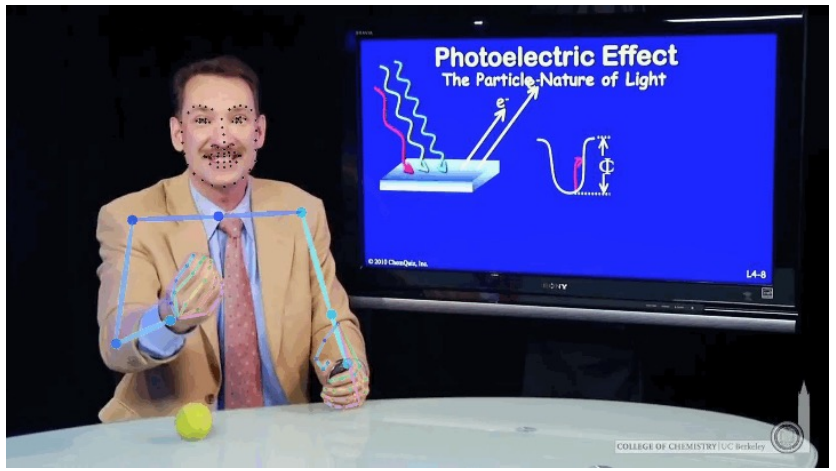
Figure 2: Location of the 24 cameras: 14 cameras were placed on each side of the participants to best capture fingers.

https://github.com/facebookresearch/TalkingWithHands32M

2. Dataset

1) Co-Speech Gesture Generation

PATS Dataset (Pose, Audio, Transcript, Style)
Learning Individual Styles of Conversational Gesture, CVPR2019



We present a large, 144-hour person-specific video dataset of 10 speakers, with frame-by-frame automatically-detected pose annotations. We deliberately pick a set of speakers for which we can find hours of clean single-speaker footage. Our speakers come from a diverse set of backgrounds: television show hosts, university lecturers and televangelists. They span at least three religions and discuss a large range of topics from commentary on current affairs through the philosophy of death, chemistry and the history of rock mu sic, to readings in the Bible and the Qur'an.

http://people.eecs.berkeley.edu/~shiry/speech2gesture/, https://chahuja.com/pats/

2. Dataset

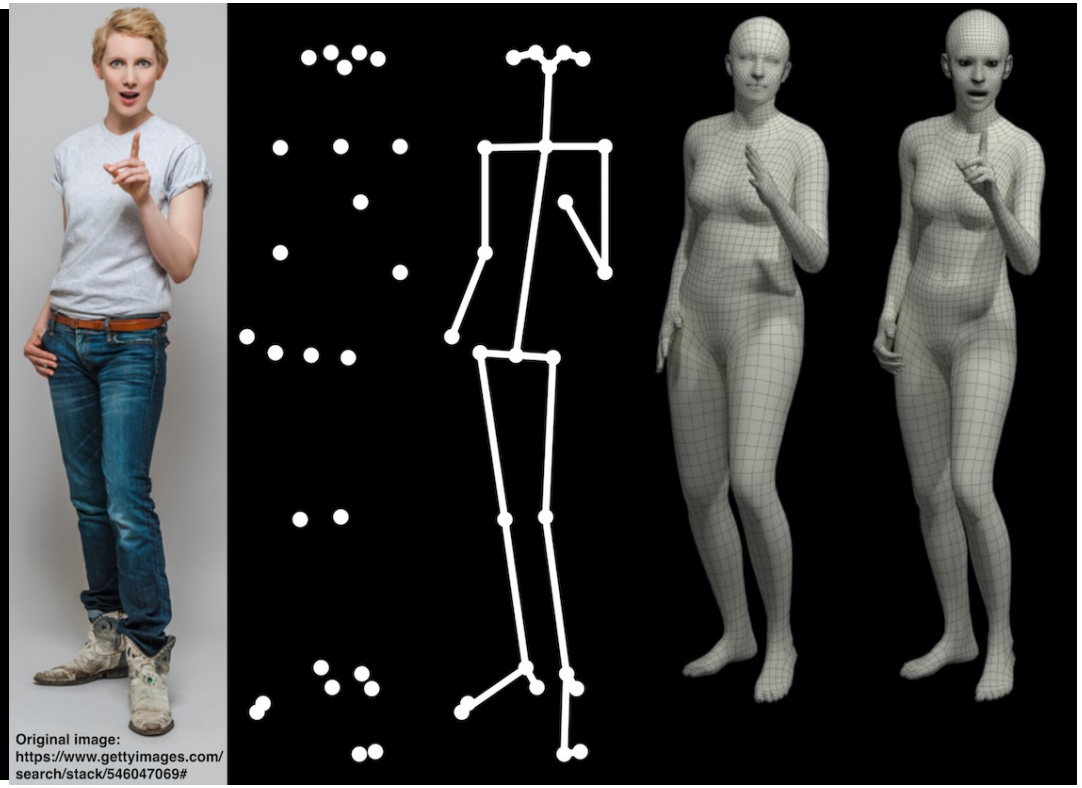
1) Co-Speech Gesture Generation

BEAT: A Large-Scale Semantic and Emotional Multi-Modal Dataset for Conversational Gestures Synthesis
ECCV2022

https://pantomatrix.github.io/BEAT/

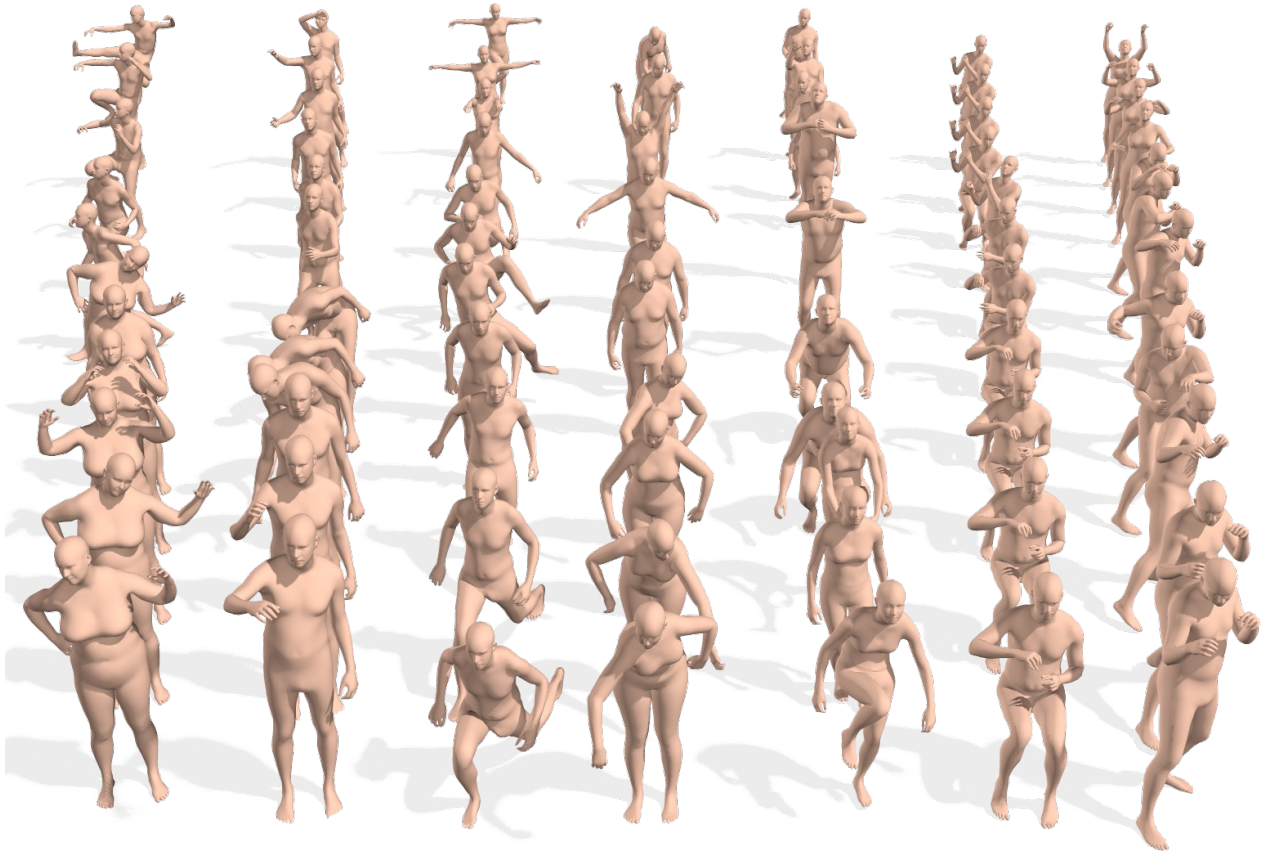
2) Conditioned Motion Generation

SMPL: A Skinned Multi-Person Linear Model
→ SMPL-X, SMPL-H (Expressive Body Capture: 3D Hands, Face, and Body from a Single Image, CVPR 2019)



https://www.youtube.com/watch?v=kuBlUyHeV5U
https://smpl-x.is.tue.mpg.de/

AMASS: Archive of Motion Capture As Surface Shapes
ICCV 2019

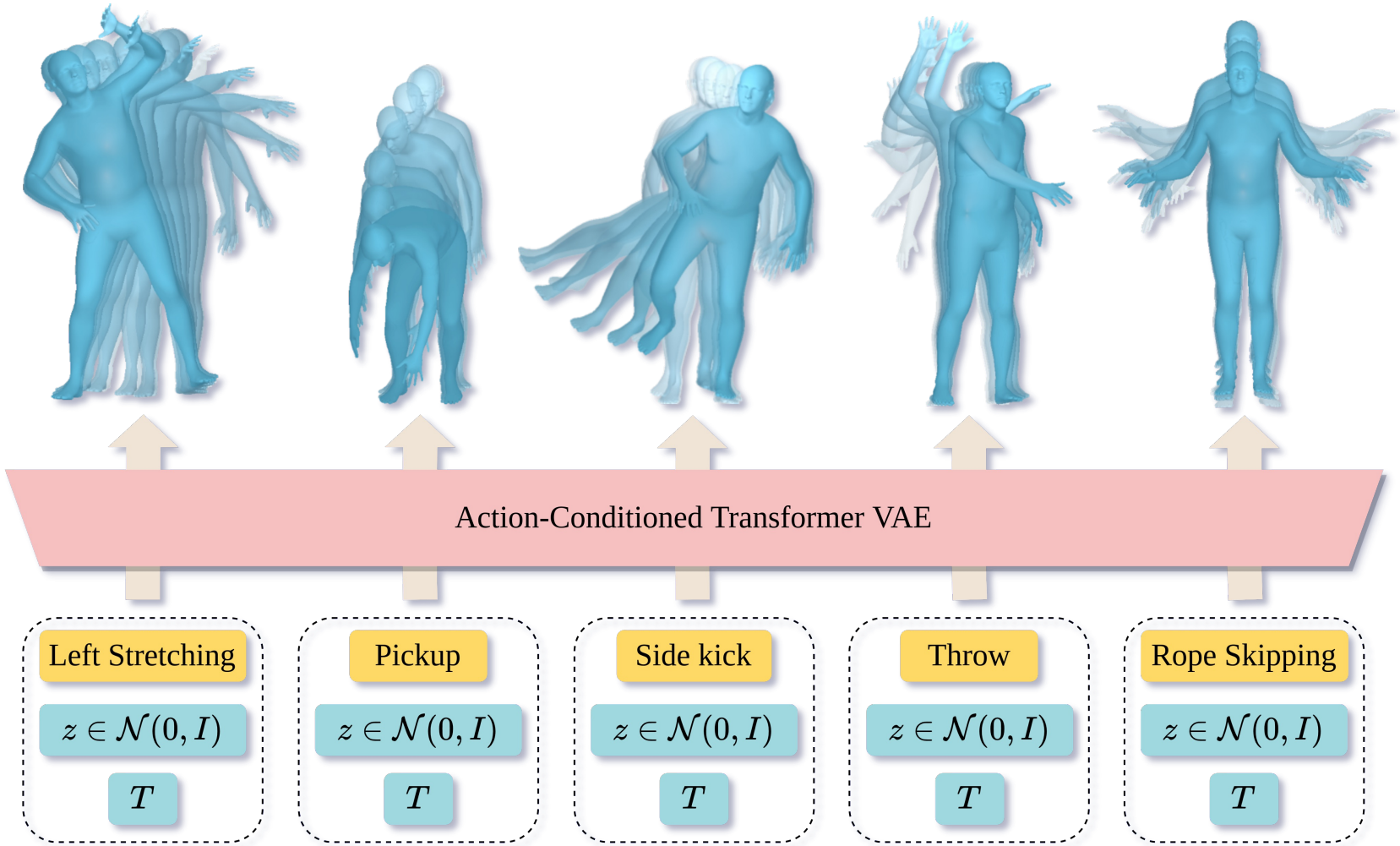


AMASS is a large database of human motion unifying different optical marker-based motion capture datasets by representing them within a common framework and parameterization. AMASS is readily useful for animation, visualization, and generating training data for deep learning.

https://amass.is.tue.mpg.de/

2) Conditioned Motion Generation

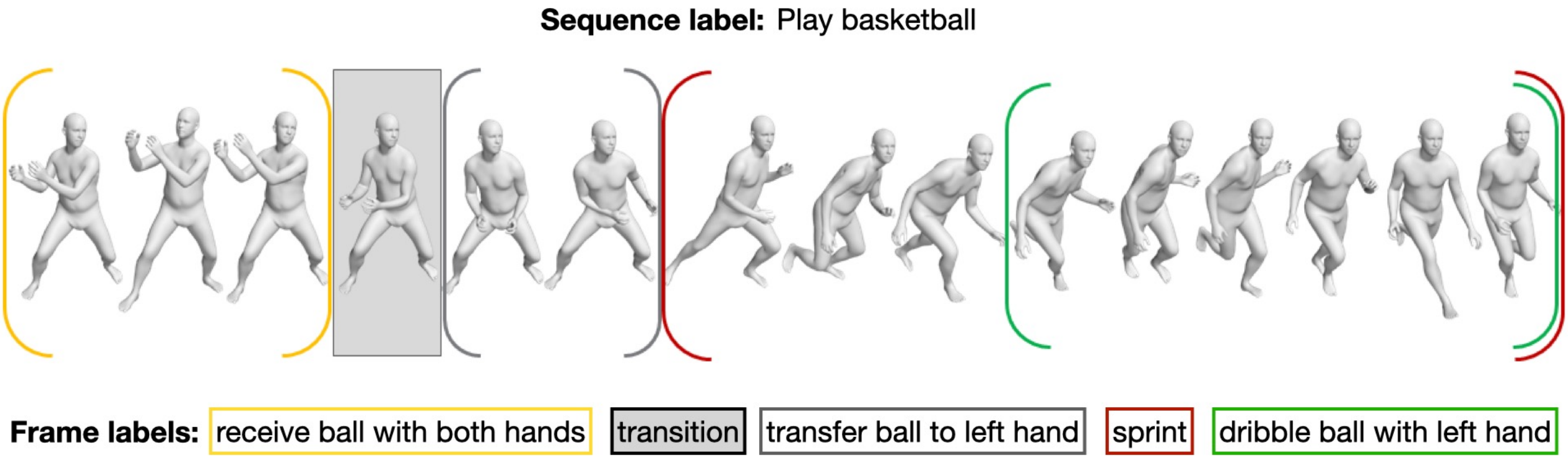
ACTOR: Action-Conditioned 3D Human Motion Synthesis with Transformer VAE
ICCV2021



https://mathis.petrovich.fr/actor/

2) Conditioned Motion Generation

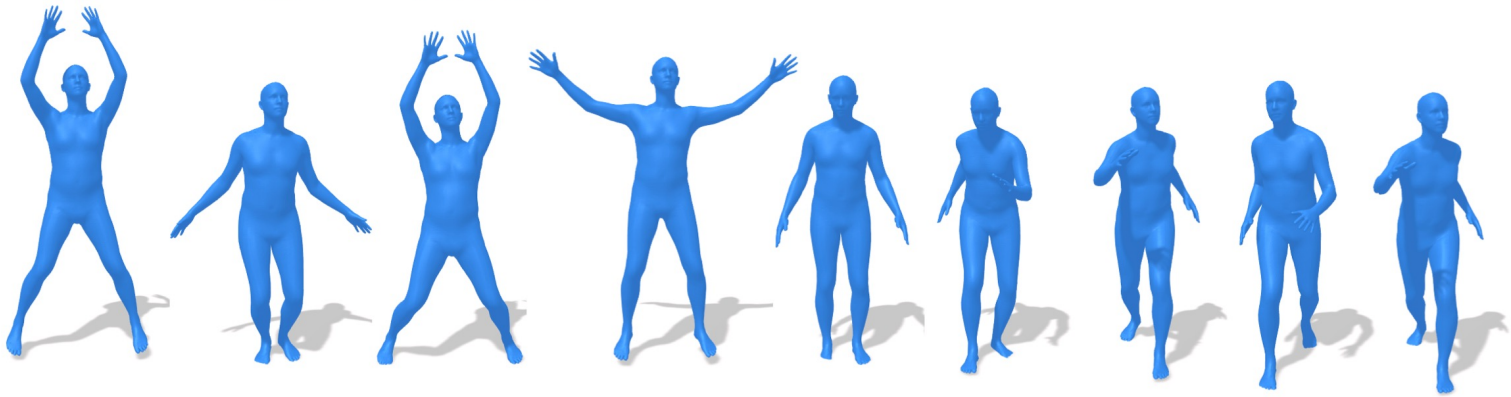
BABEL: Bodies, Action and Behavior with English Labels
CVPR 2021



https://babel.is.tue.mpg.de/

HumanML3D: 3D Human Motion-Language Dataset
3D human motion-language dataset that originates from a combination of HumanAct12 and Amass dataset.
CVPR2022

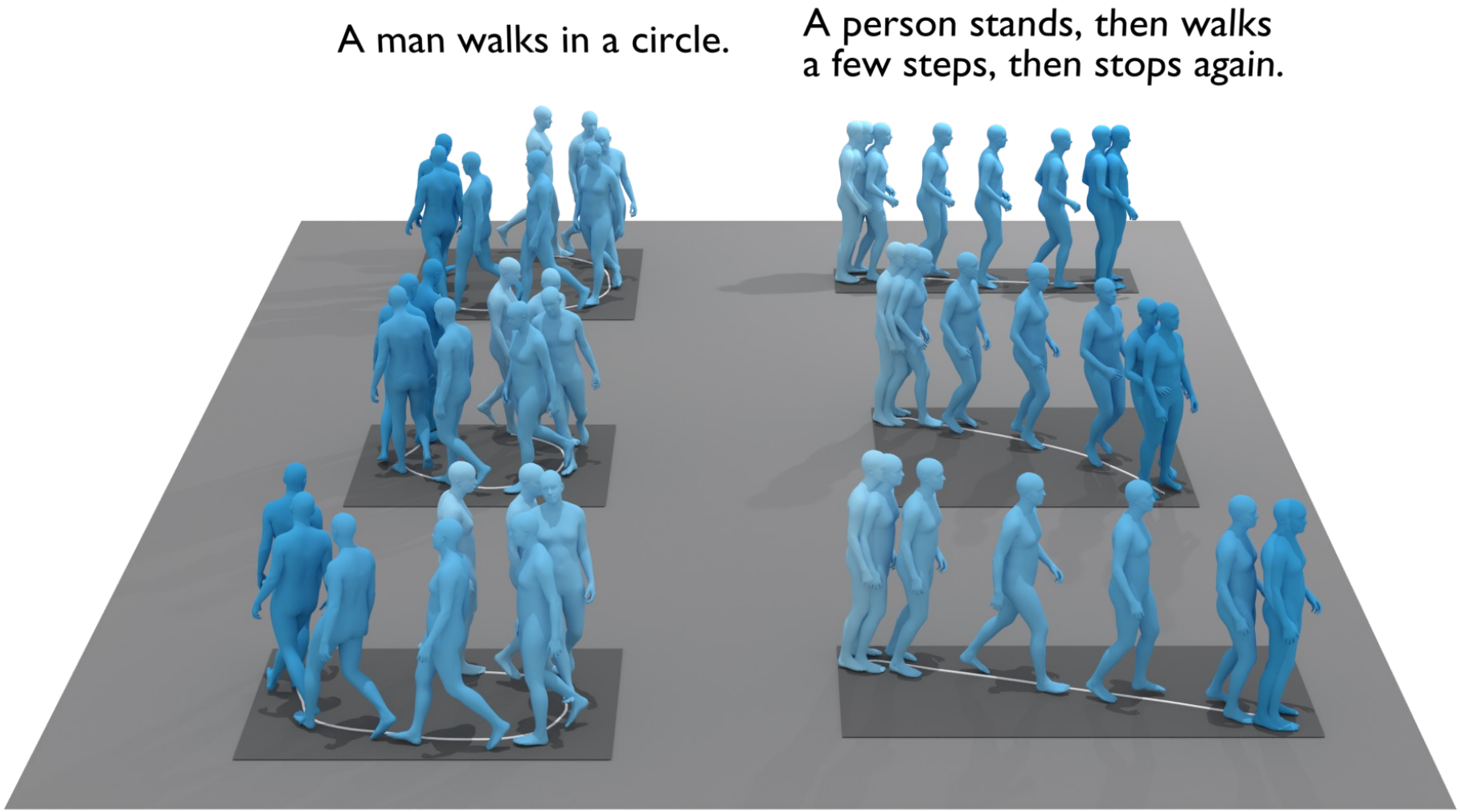
1. The person is leaving at someone with his left hand.
2. A person shakes an item with his left hand.
3. A person waves his left hand repeatedly above his head.



1. A person doing jumping jacks and then running on the spot.
2. A person is doing jumping jacks, then starts jogging in place.
3. A person does four jumping jacks then three front lunges.

https://github.com/EricGuo5513/HumanML3D

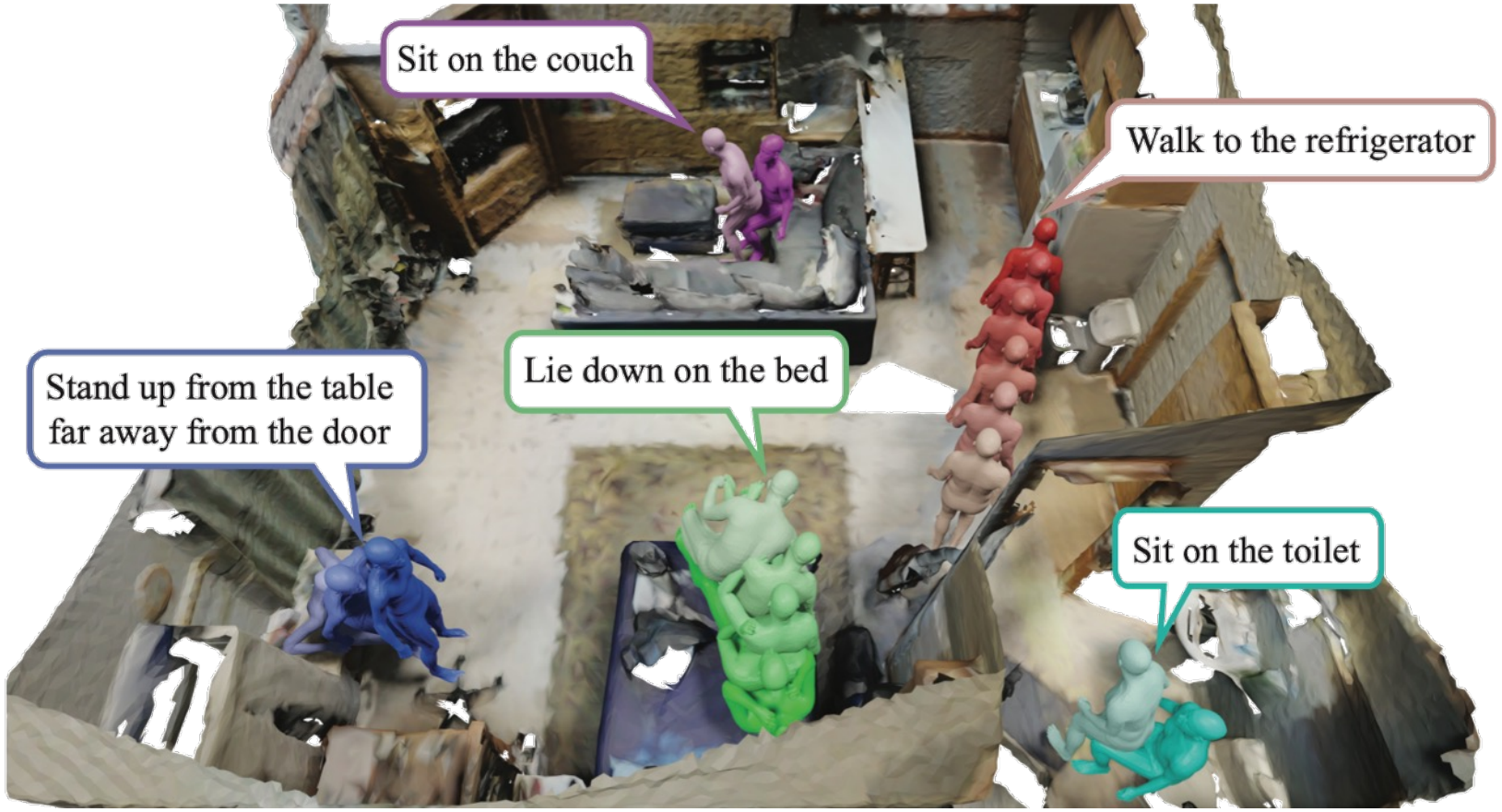
TEMOS: Generating diverse human motions from textual descriptions
ECCV 2022



https://mathis.petrovich.fr/temos/

2) Conditioned Motion Generation

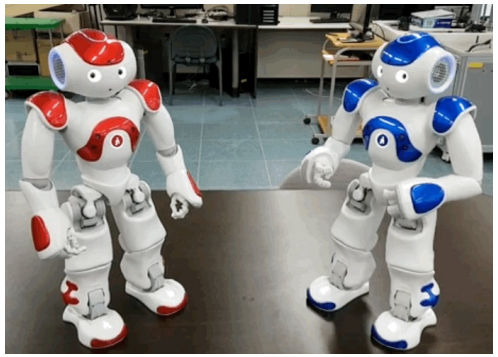
HUMANISE: Language-conditioned Human Motion Generation in 3D Scenes
NeurIPS 2022



https://silvester.wang/HUMANISE/

3. Approach

1-1) Attention Seq2Seq



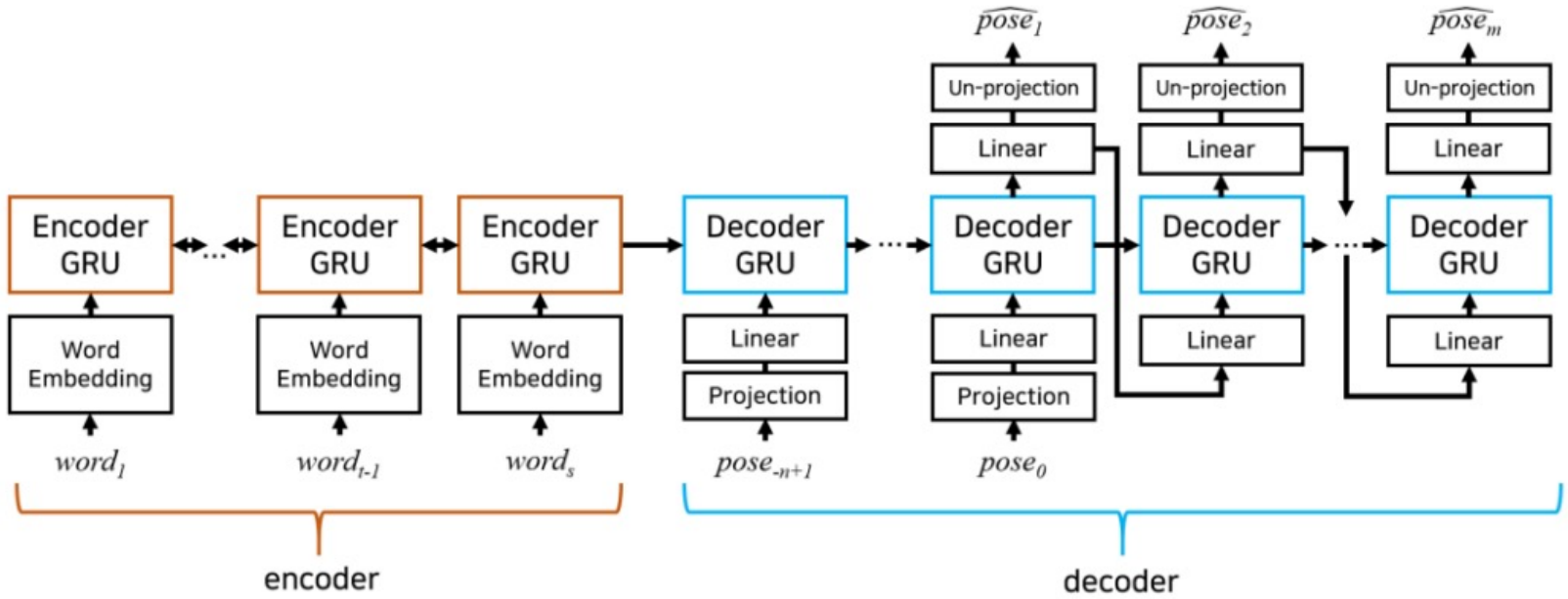
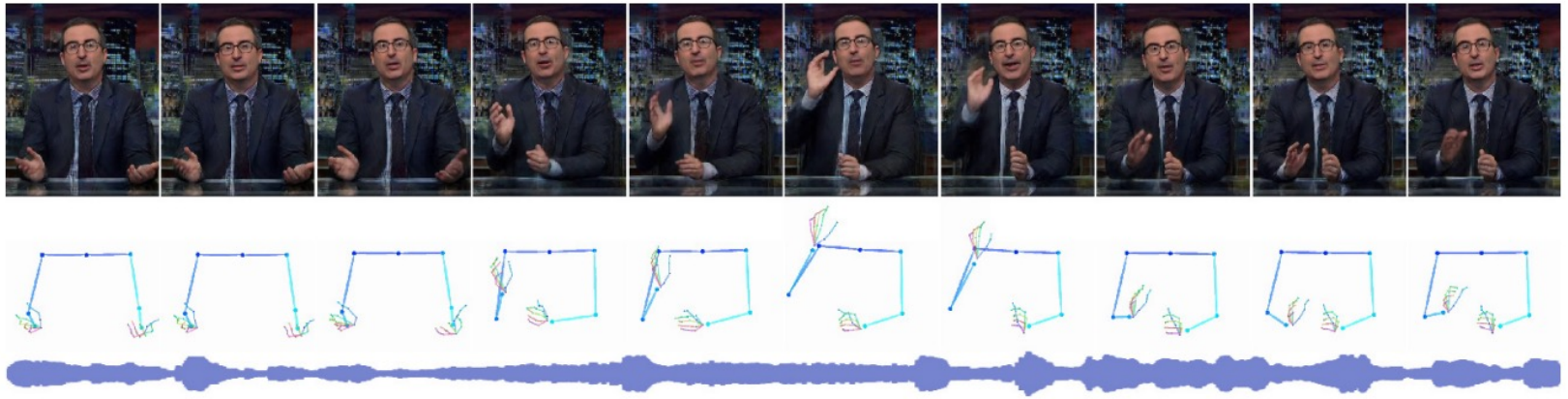


Figure 4. Proposed network architecture. The encoder GRU interprets *s* speech words, and the decoder GRU generates *m* human poses of gestures. The decoder GRU inputs *n* previous poses to make the series of poses continuous. The soft attention mechanism is used but not depicted here.

1-2) Speech2Gestures



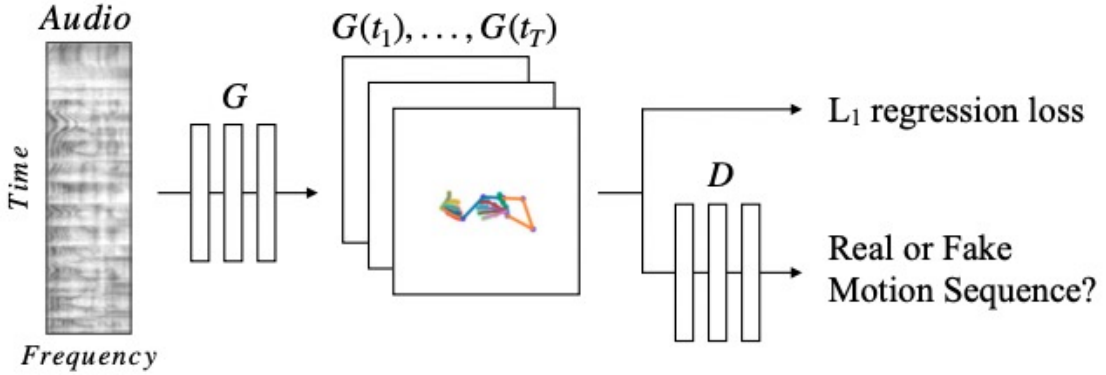


Figure 3: *Speech to gesture translation model.* A convolutional audio encoder downsamples the $2D$ spectrogram and transforms it to a $1D$ signal. The translation model, G, then predicts a corresponding temporal stack of $2D$ poses. L_1 regression to the ground truth poses provides a training signal, while an adversarial discriminator, D, ensures that the predicted motion is both temporally coherent and in the style of the speaker.

$$\mathcal{L}_{L_1}(G) = \mathbb{E}_{\mathbf{s},\mathbf{p}}[||\mathbf{p} - G(\mathbf{s})||_1].$$

$$\mathcal{L}_{GAN}(G, D) = \mathbb{E}_{\mathbf{m}}[\log D(\mathbf{m})] + \mathbb{E}_{\mathbf{s}}[\log(1 - G(\mathbf{s}))], \quad (2)$$

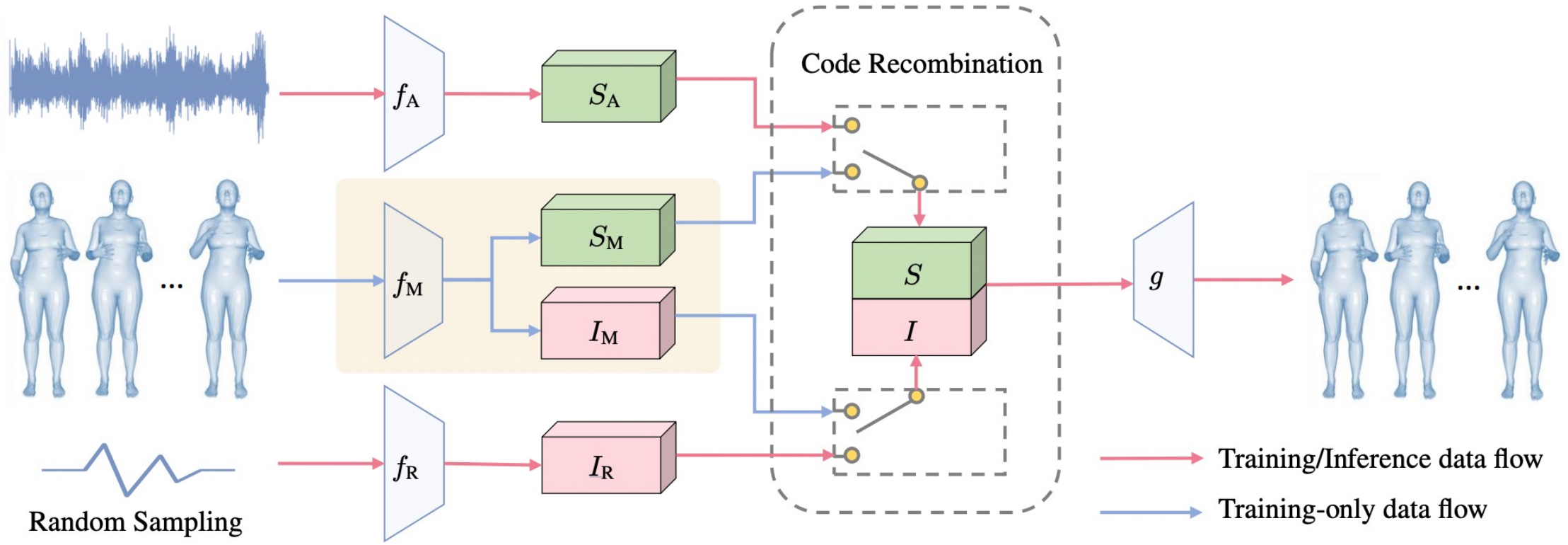
$$\min_G \max_D \mathcal{L}_{GAN}(G, D) + \lambda \mathcal{L}_{L_1}(G). \quad (3)$$

Learning Individual Styles of Conversational Gesture, CVPR2019

1-2) Speech2Gestures

Learning Individual Styles of Conversational Gesture, CVPR2019

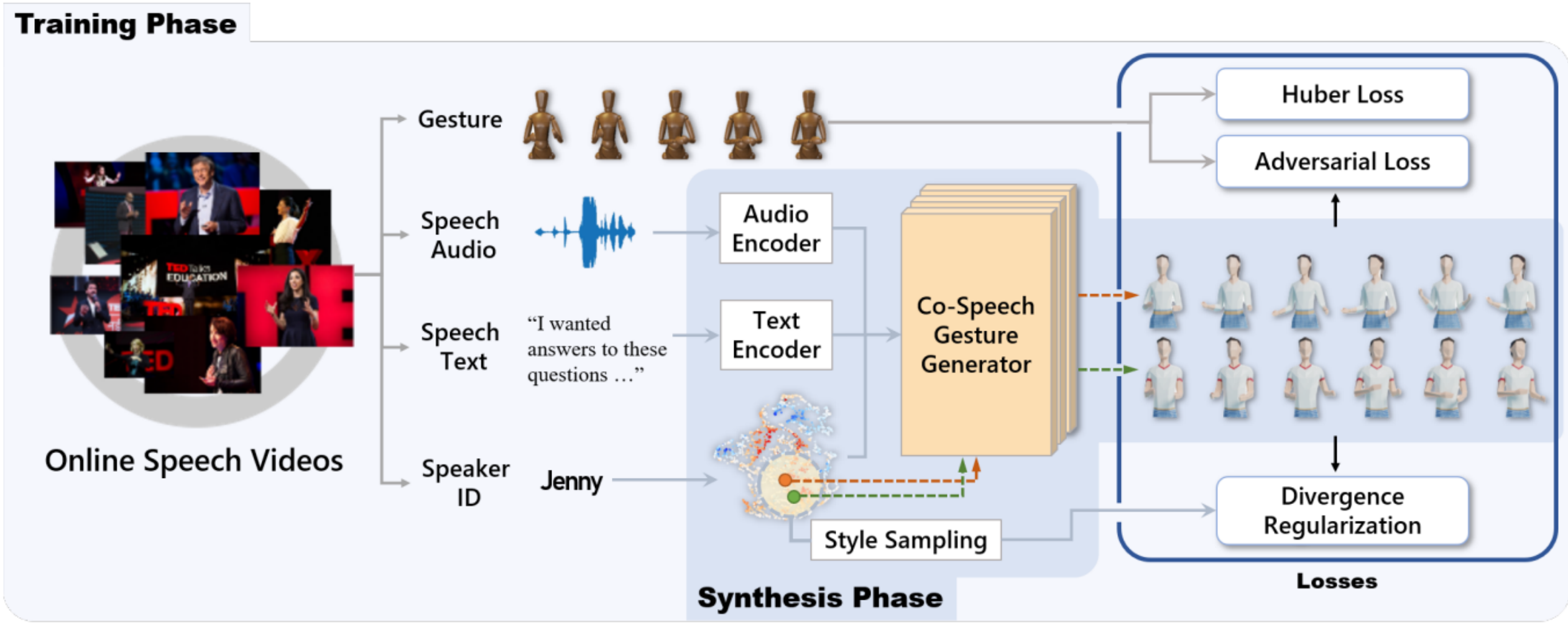
1-3) Audio2Gestures



Audio2Gestures: Generating Diverse Gestures from Speech Audio with Conditional Variational Autoencoders, ICCV2021

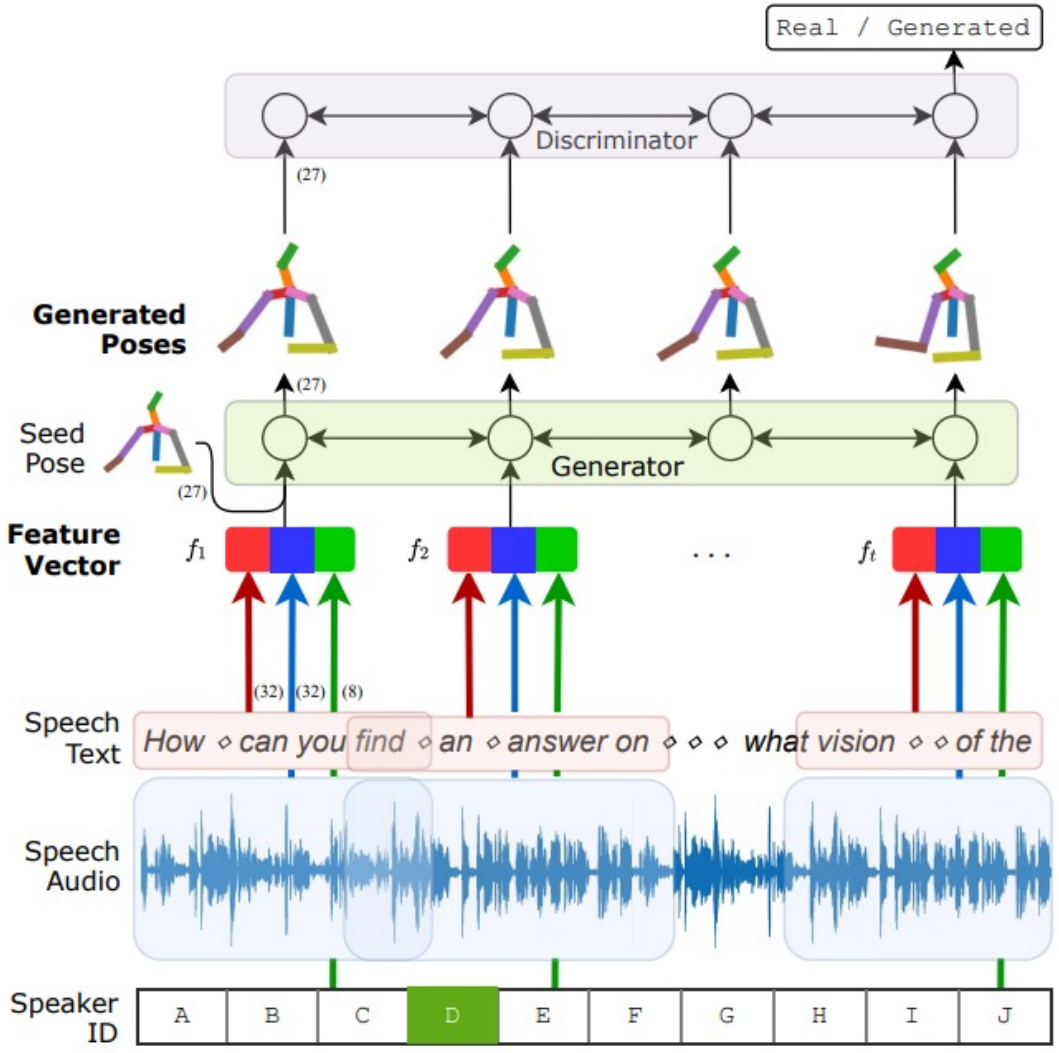
3. Approach

1-4) Trimodal (TriCon, MultiContext)



Speech Gesture Generation from the Trimodal Context of Text, Audio, and Speaker Identity, SIGGRAPH ASIA2020
Youngwoo Yoon, Bok Cha, Joo-Haeng Lee, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee

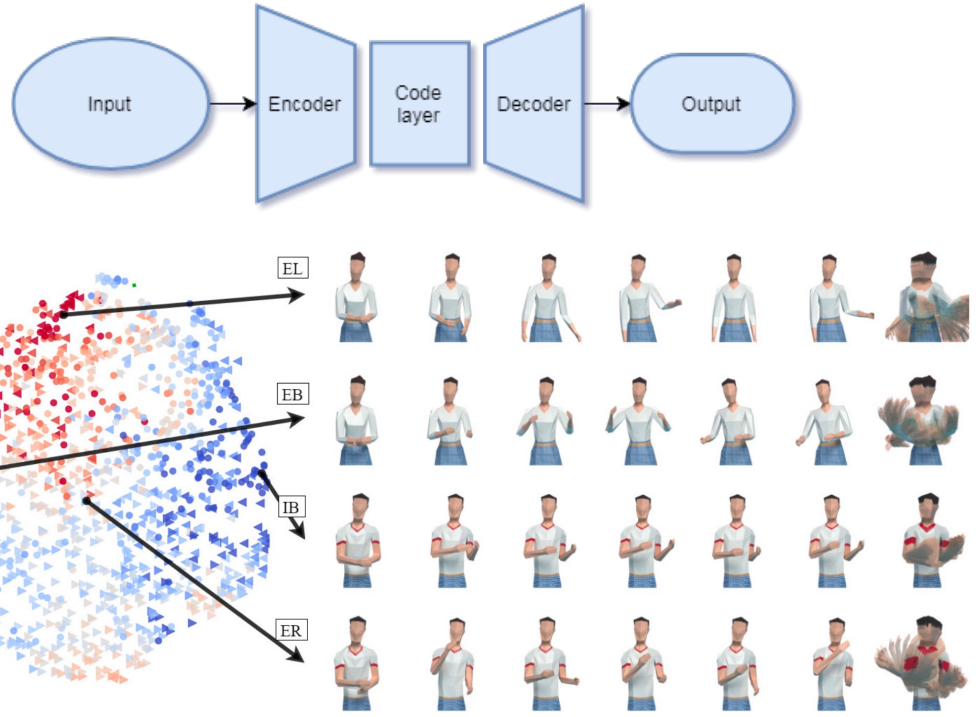
1-4) Trimodal (TriCon, MultiContext)



New Metric: Fréchet Motion Distance

The FMD measures the distance between the distribution of a ground truth and synthetic motion dataset.

→ AutoEncoder의 Feature space에서의 Distance



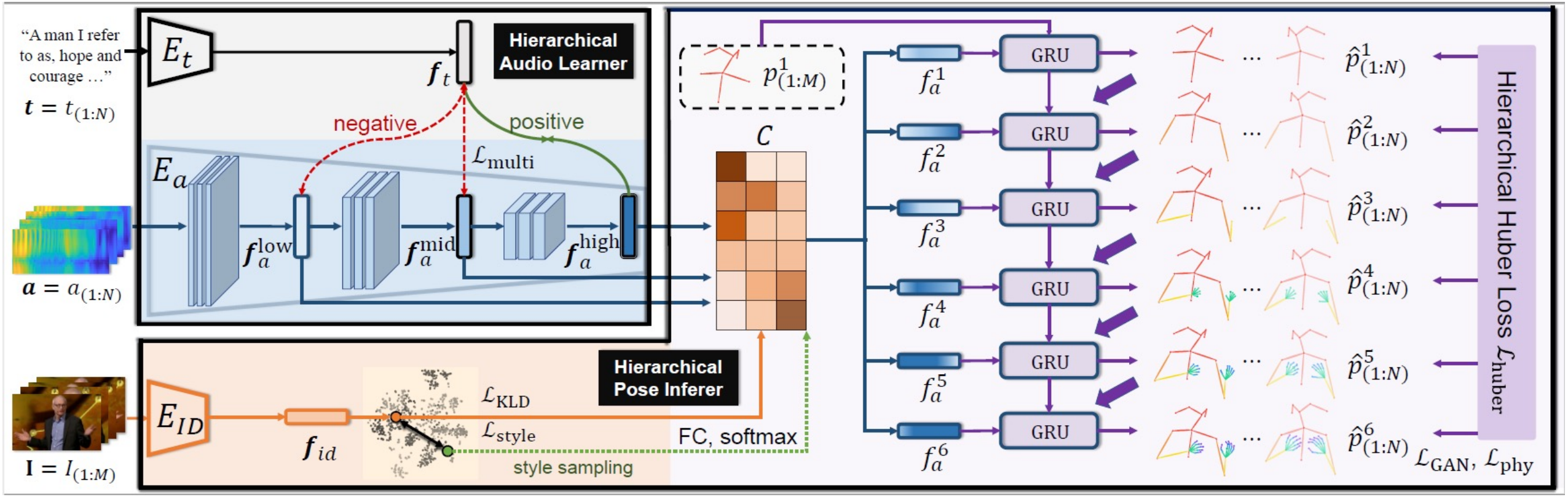
Speech Gesture Generation from the Trimodal Context of Text, Audio, and Speaker Identity, SIGGRAPH ASIA2020

Youngwoo Yoon, Bok Cha, Joo-Haeng Lee, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee

SIGGRAPH ASIA 2020

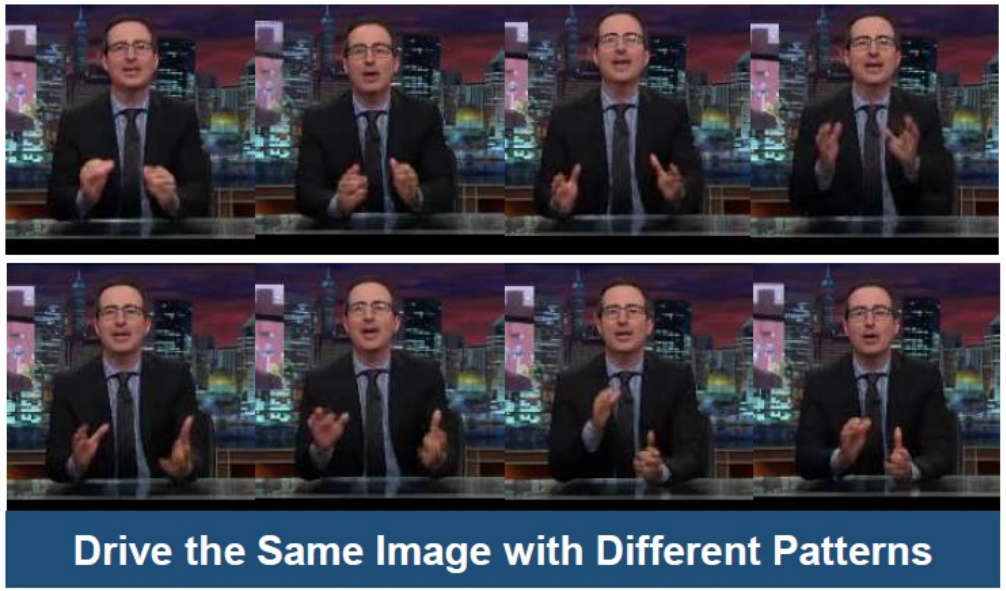
Speech Gesture Generation from the Trimodal Context of Text, Audio, and Speaker Identity

Youngwoo Yoon, Bok Cha, Joo-Haeng Lee, Minsu Jang, Jaeyeon Lee, Jaehong Kim, Geehyuk Lee



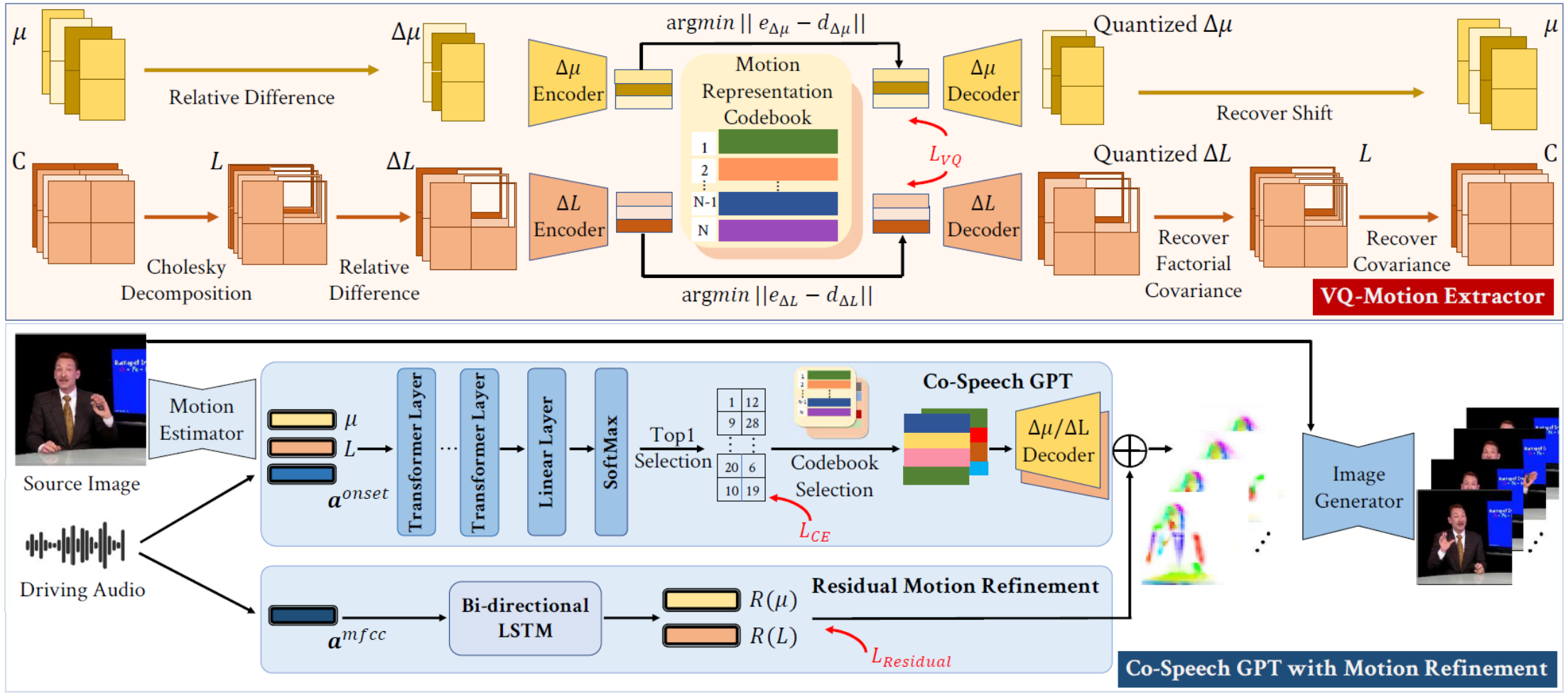
Learning Hierarchical Cross-Modal Association for Co-Speech Gesture Generation, CVPR2022

Co-Speech Gesture Image Generator



Drive the Same Image with Different Patterns

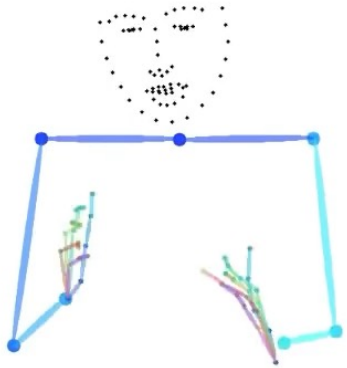
Drive Different Images with the Same Pattern



Audio-Driven Co-Speech Gesture Video Generation, NeurIPS2022

1-6) ANGIE

Predicted



Audio-Driven Co-Speech Gesture Video Generation, NeurIPS2022

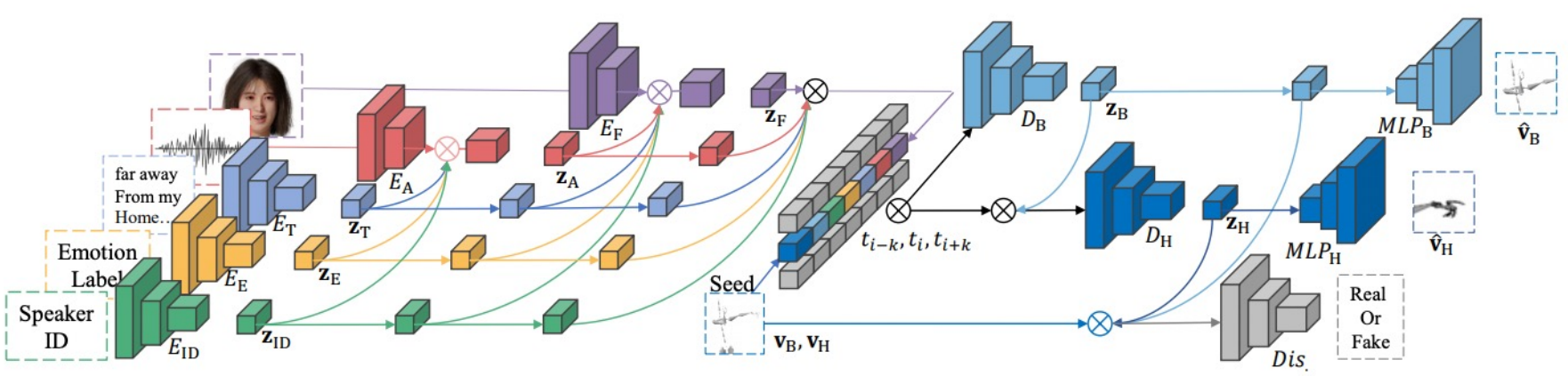


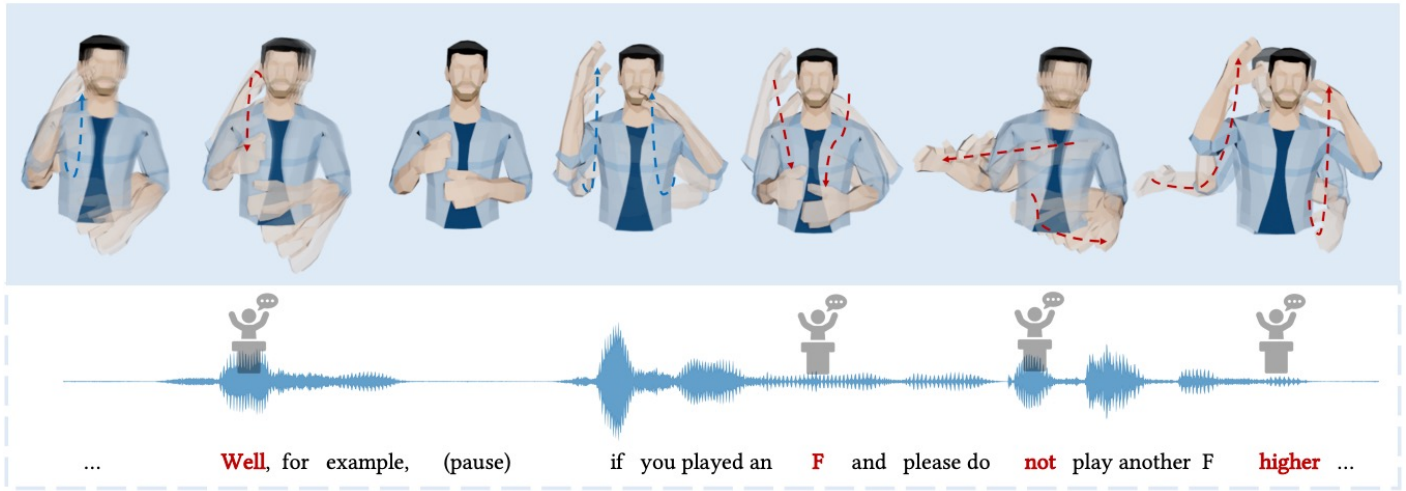
Fig. 5. Cascaded Motion Network (CaMN). As a multi-modal gesture synthesis baseline, CaMN inputs *text, emotion label, speaker ID, audio* and *facial blendweight* in a cascaded architecture, the audio and facial feature will be extracted by concatenating the features of previous modalities. The fused feature will be reconstructed to body and hands gestures by two cascaded LSTM+MLP decoders.

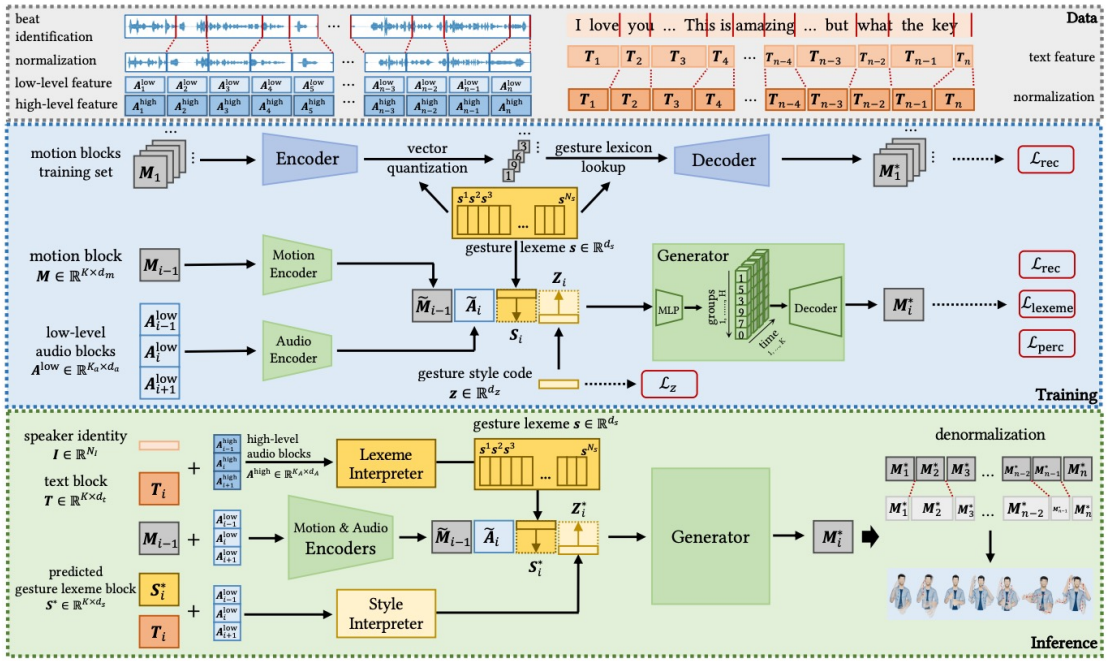
New Metric: Semantic-Relevant Gesture Recall
→ PCK (Probability of Correct Keypoint) 기반 평가

$$D_{SRGR} = \lambda \sum \frac{1}{T \times J} \sum_{t=1}^{T} \sum_{j=1}^{J} \mathbf{1} \left[\left\| p_t^j - \hat{p}_t^j \right\|_2 < \delta \right],$$

3. Approach

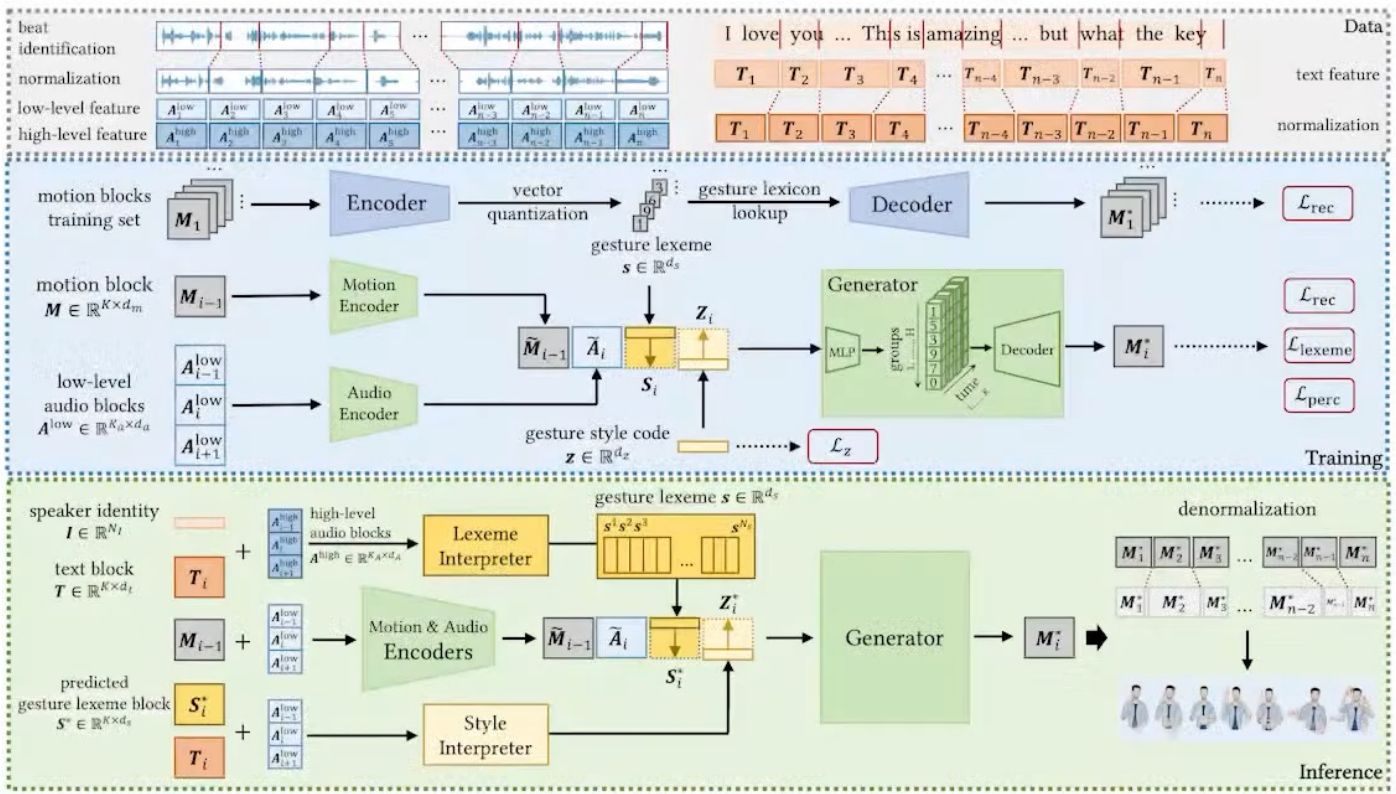
1-8) Rhythmic Gesticulator





Rhythmic Gesticulator: Rhythm-Aware Co-Speech Gesture Synthesis with Hierarchical Neural Embeddings,
SIGGRAPH ASIA2022 https://www.youtube.com/watch?v=qy2MrNhsoIs

1-8) Rhythmic Gesticulator



In this paper, we propose a new rhythm and semantics-aware

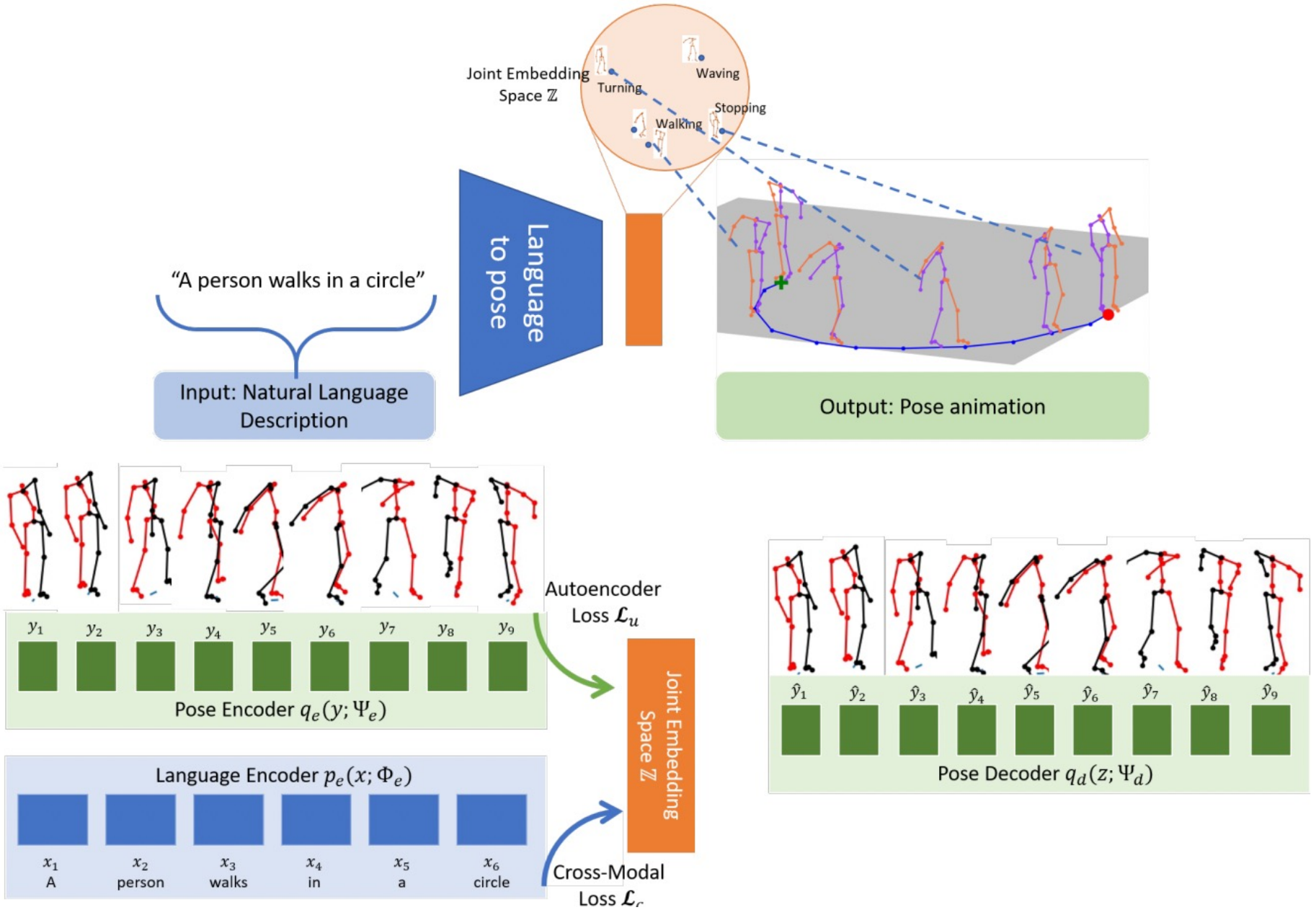
1-8) Rhythmic Gesticulator

Long Sample Results (1/2) – Obama Speech

* Pre-trained model trained on the Trinity and Chinese dataset.

Rhythmic Gesticulator: Rhythm-Aware Co-Speech Gesture Synthesis with Hierarchical Neural Embeddings, SIGGRAPH ASIA2022

2-0) Language2Pose (JointEmbedding)



Language2Pose: Natural Language Grounded Pose Forecasting, 3DV2019

2-1) Action2Motion

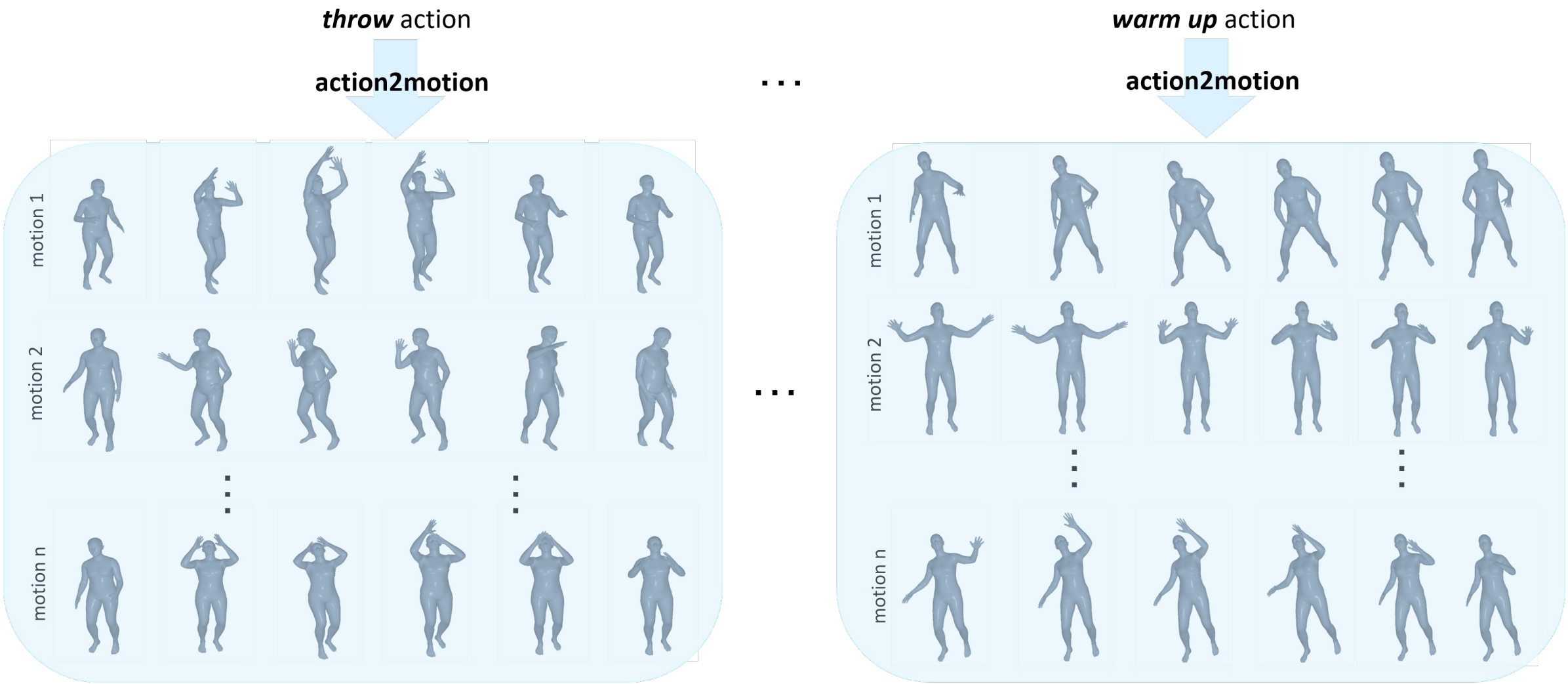
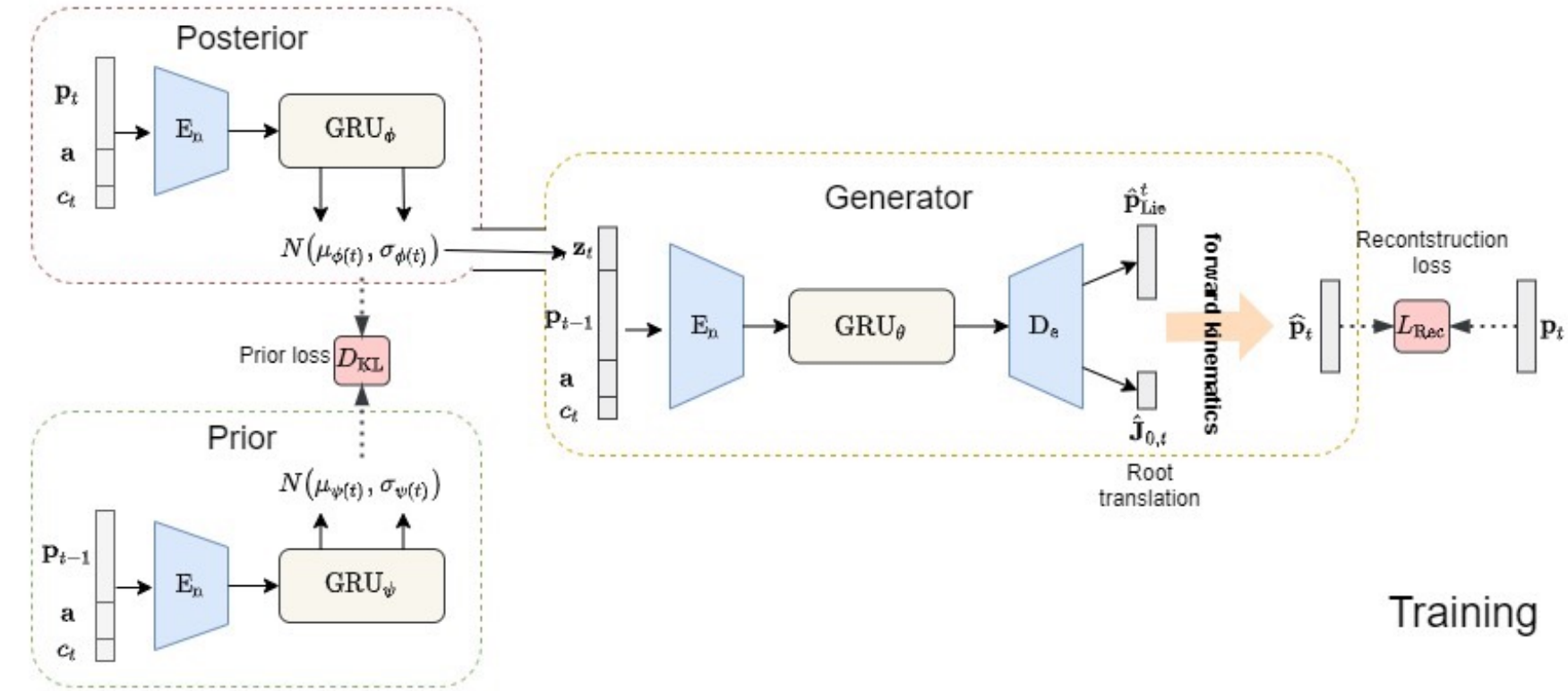


Figure 1: Conditioned on an action category (such as throw, warm up), our approach can generate a diverse set of natural 3D human motions.

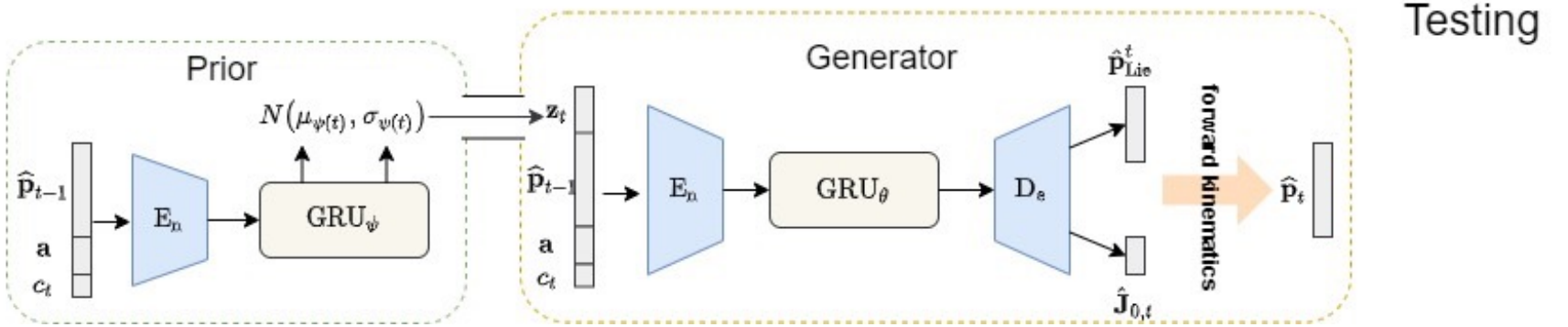
Action2Motion: Conditioned Generation of 3D Human Motions (ACM MM 2020)

2-1) Action2Motion



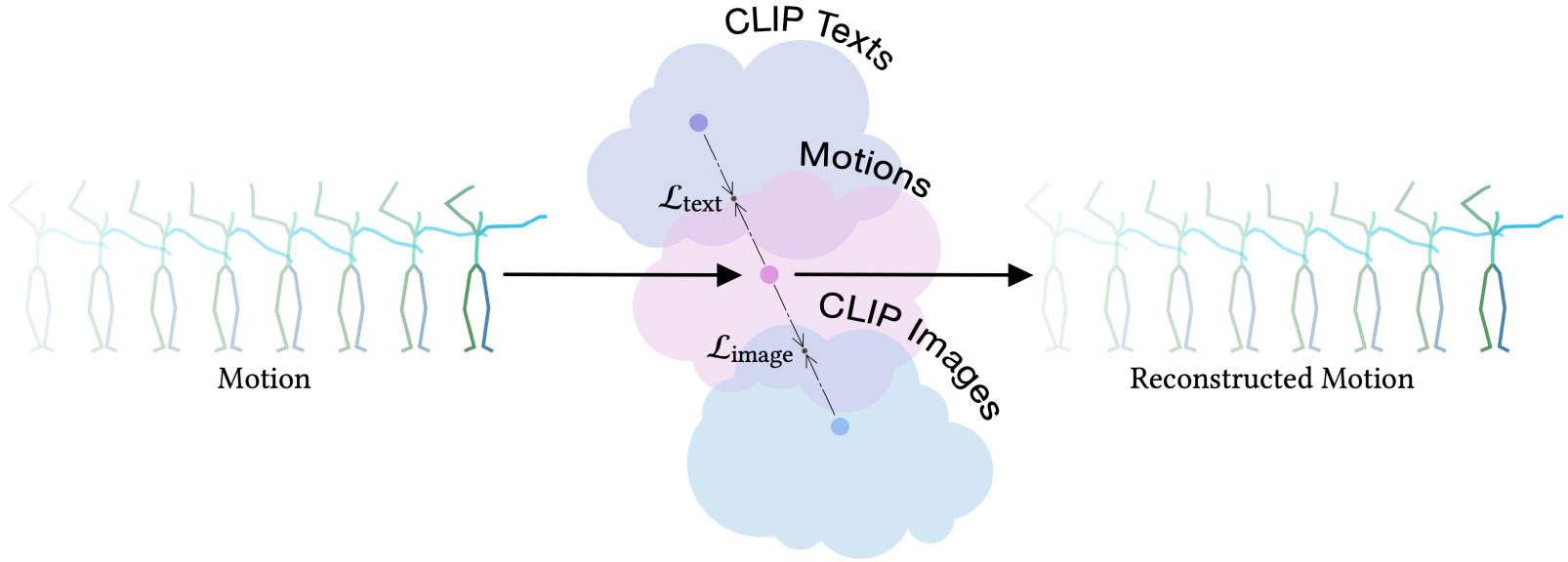
Training

Testing



Action2Motion: Conditioned Generation of 3D Human Motions (ACM MM 2020)

2-2) MotionClip



"Usain Bolt"

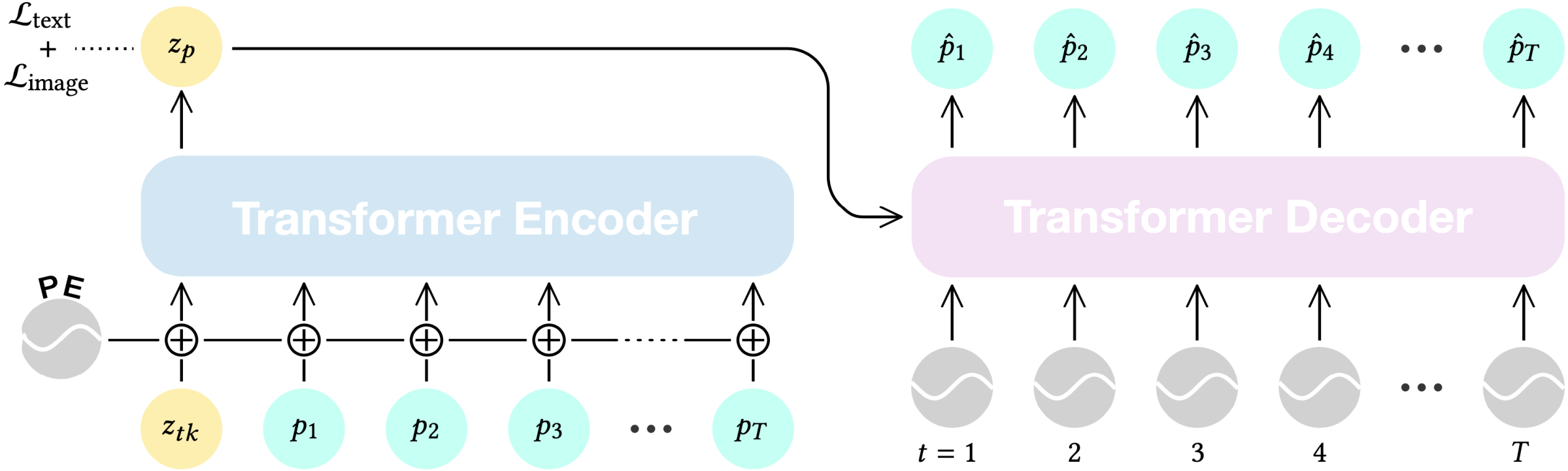
"Swan lake"

"Gollum"

"Spiderman in action!"

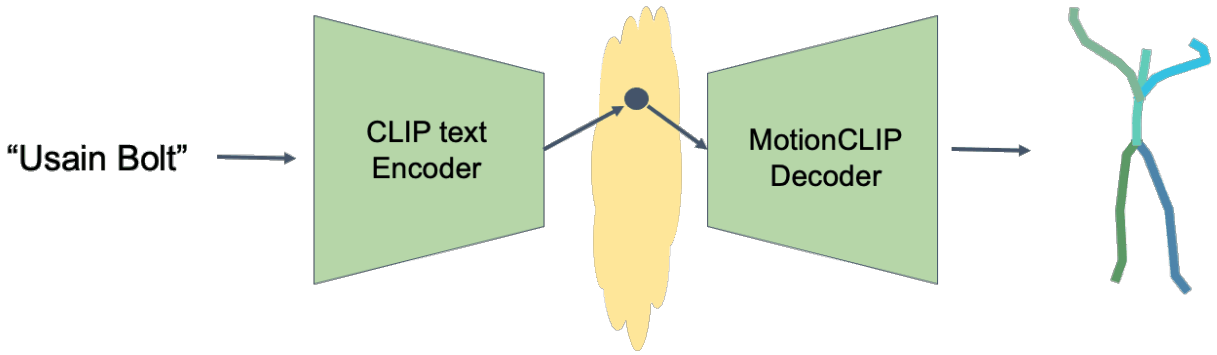
MotionCLIP: Exposing Human Motion Generation to CLIP Space, ECCV2022

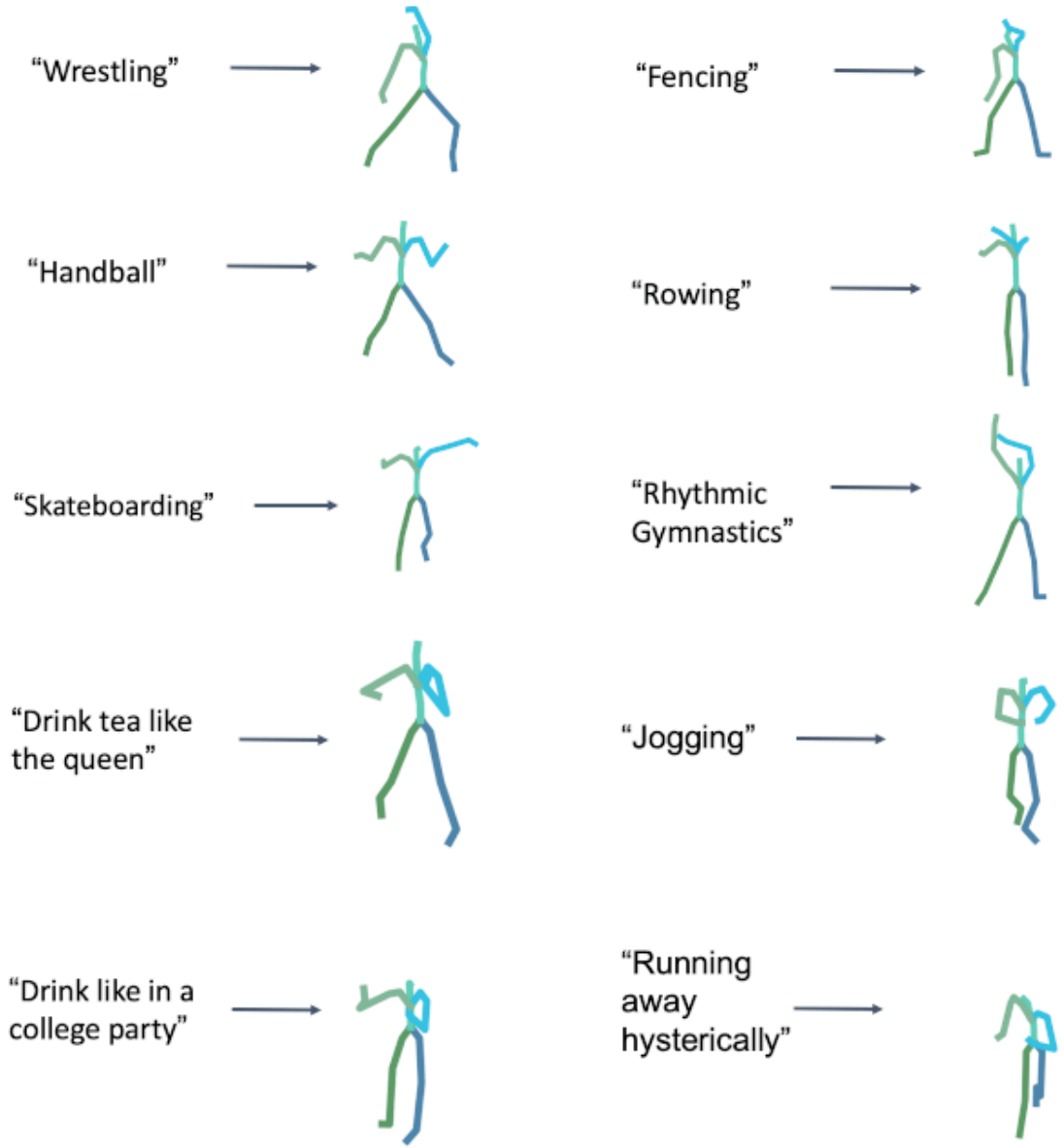
2-2) MotionClip



MotionCLIP: Exposing Human Motion Generation to CLIP Space, ECCV2022

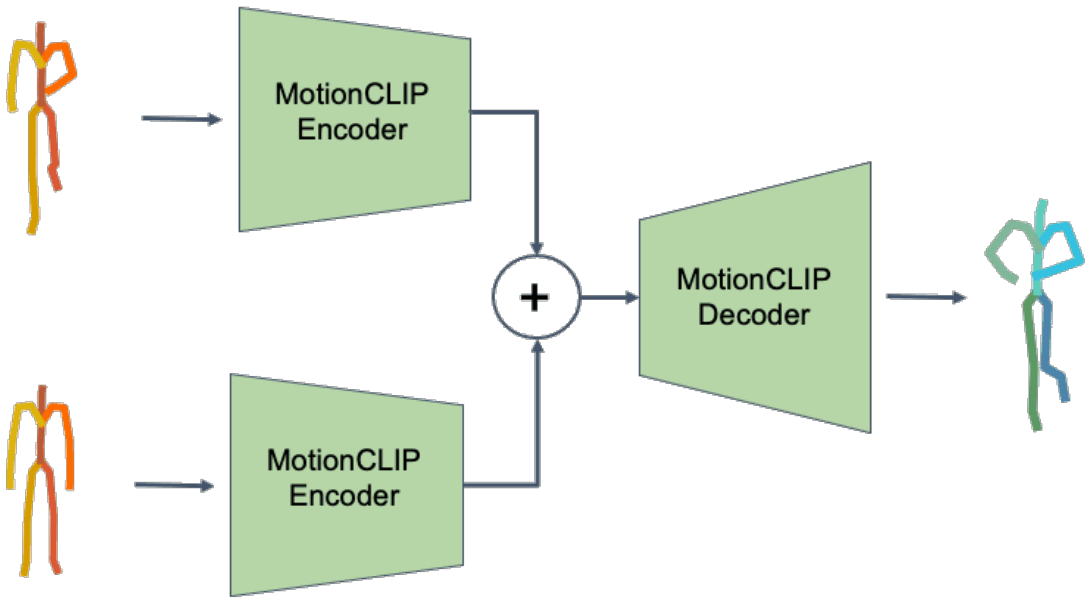
2-2) MotionClip

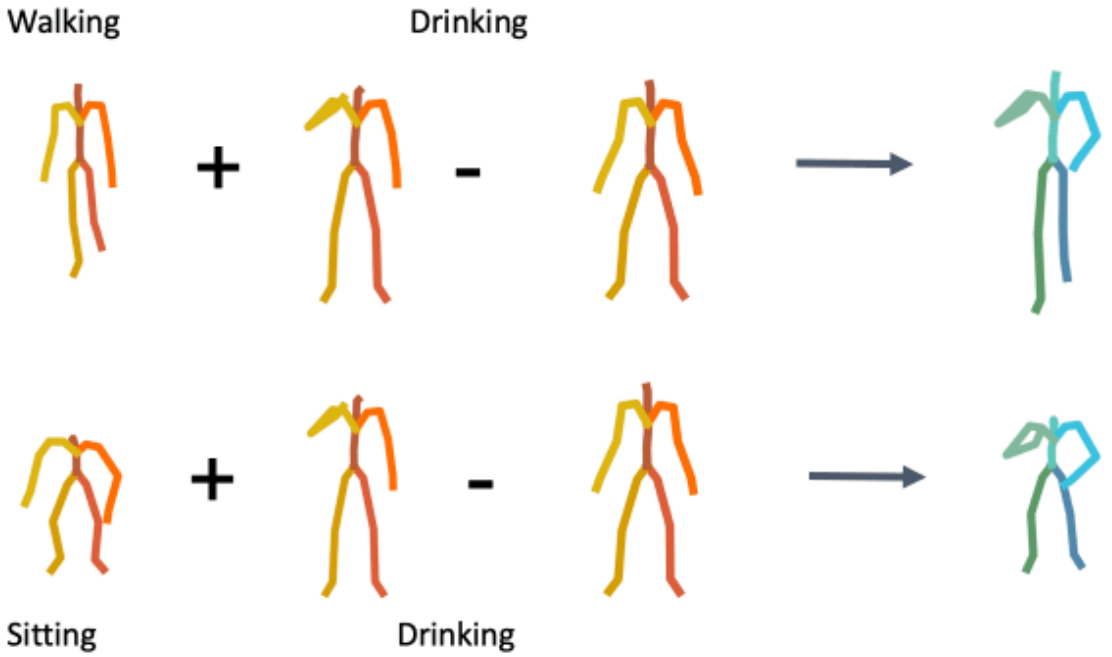


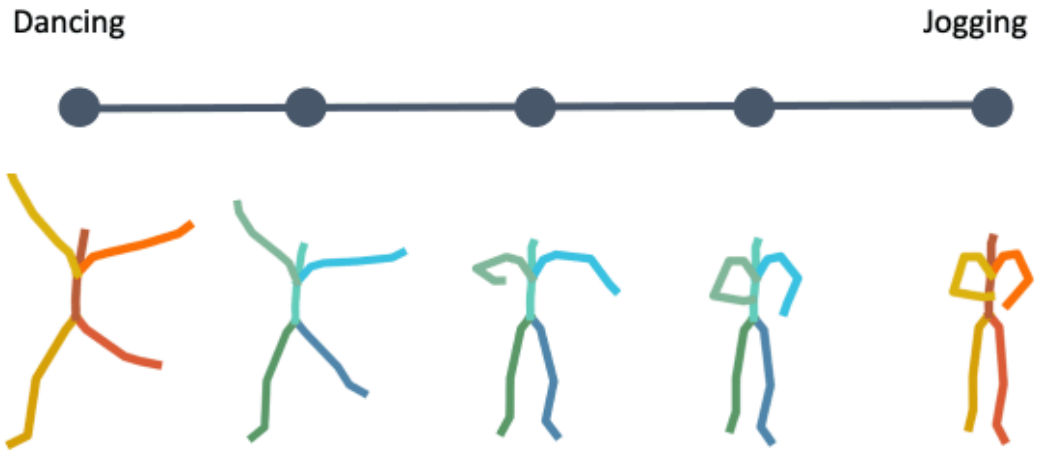


MotionCLIP: Exposing Human Motion Generation to CLIP Space, ECCV2022

2-2) MotionClip

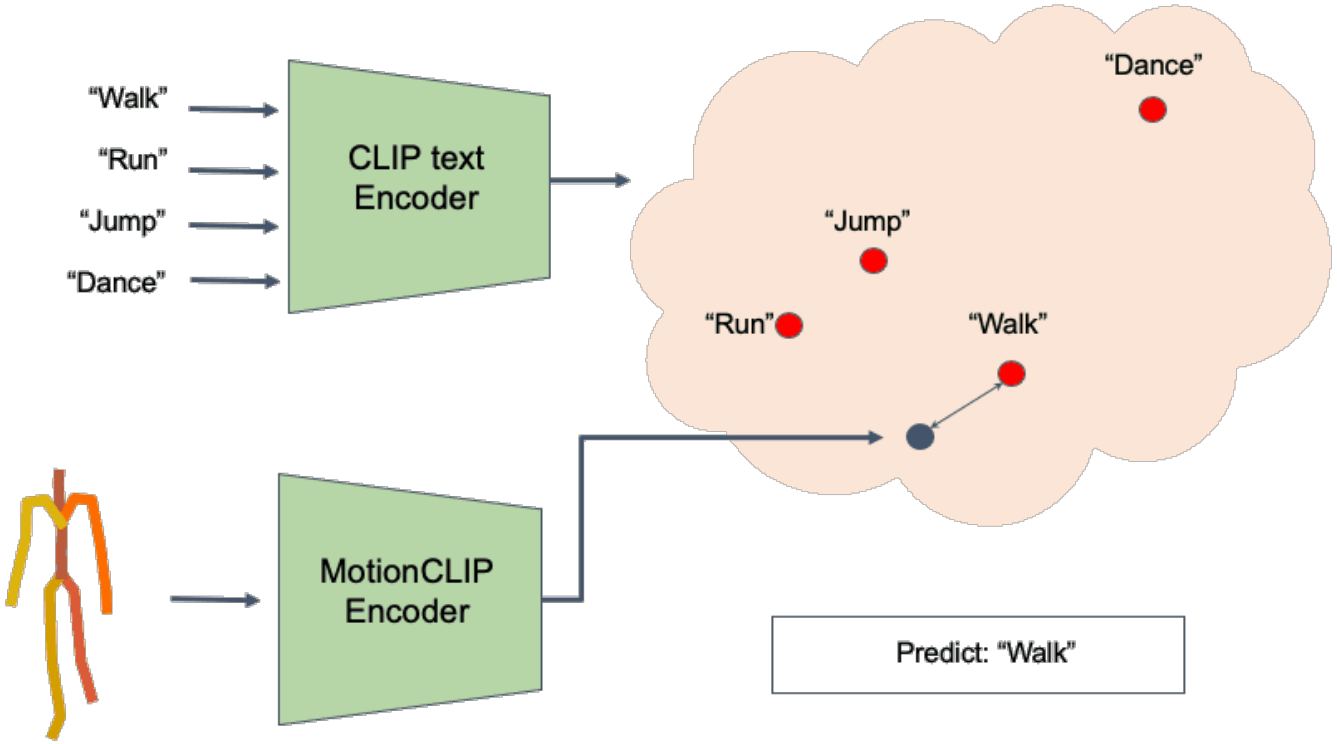






[MotionCLIP: Exposing Human Motion Generation to CLIP Space](), ECCV2022

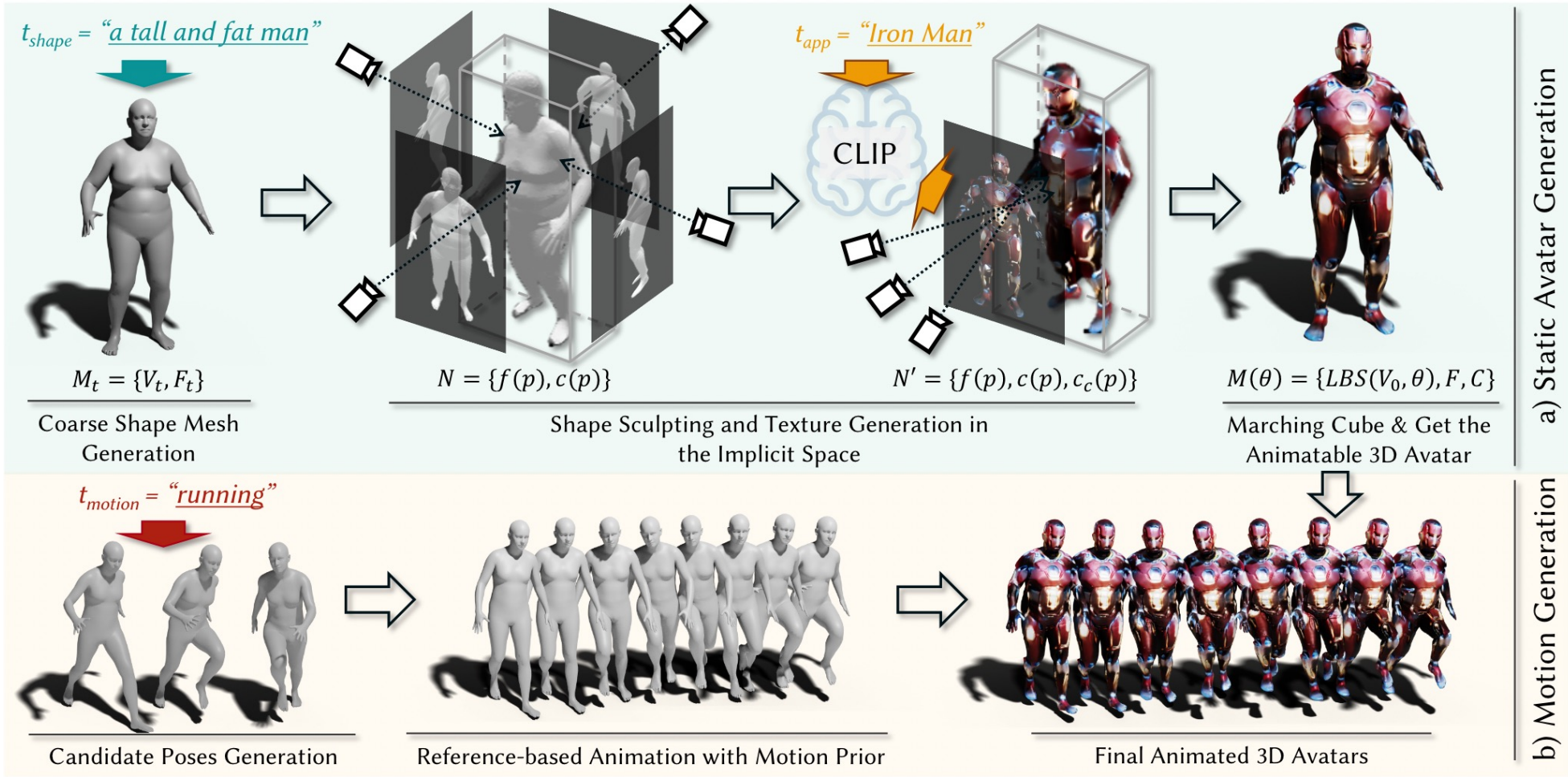
2-2) MotionClip



	Top-1 acc.	Top-5 acc.
MotionCLIP	40.9 %	57.71%
W.O. image loss	35.05%	50.26%
W.O. text loss	4.54%	18.37%
2s-AGCN [2019]	41.14%	73.18%

MotionCLIP: Exposing Human Motion Generation to CLIP Space, ECCV2022

2-3) AvatarCLIP



AvatarCLIP: zero-shot text-driven generation and animation of 3D avatars, SIGGRAPH2022

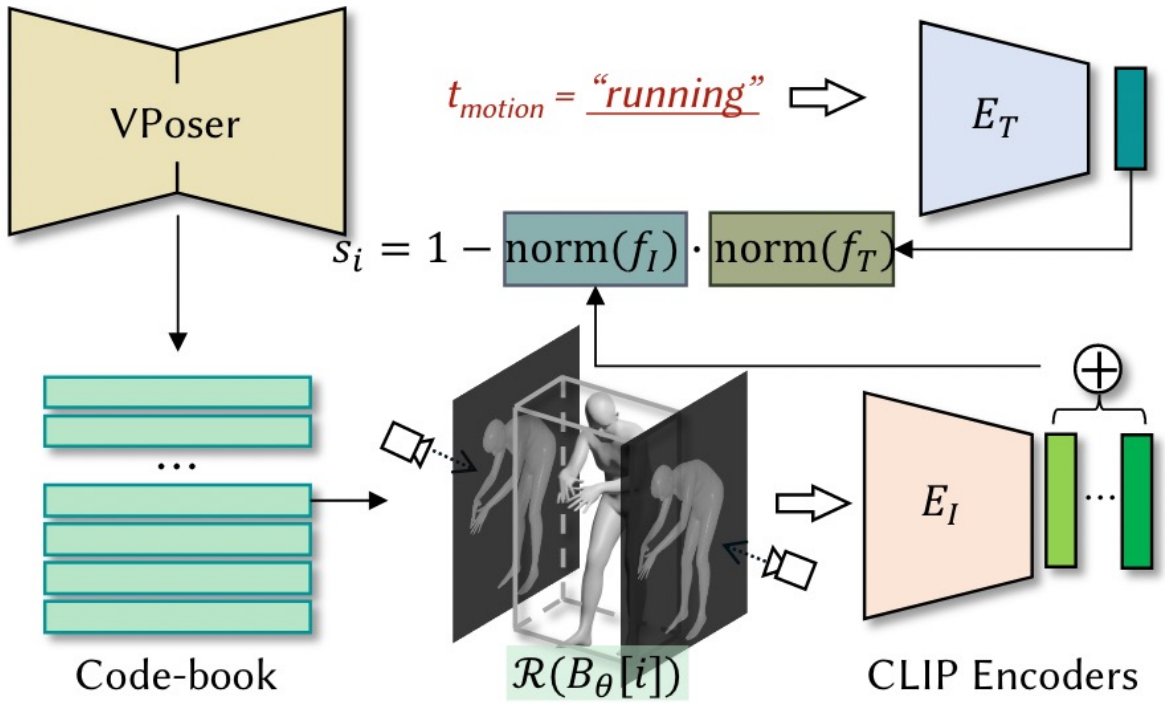


Fig. 8. **Detailed Pipeline of Candidate Poses Generation.** The pre-trained VPoser is first used to build a code-book. Given text description t_{motion}, each pose feature f_I from the code-book is used to calculate the similarity with the text feature f_T, which is used to select Top-K entries as candidate poses.

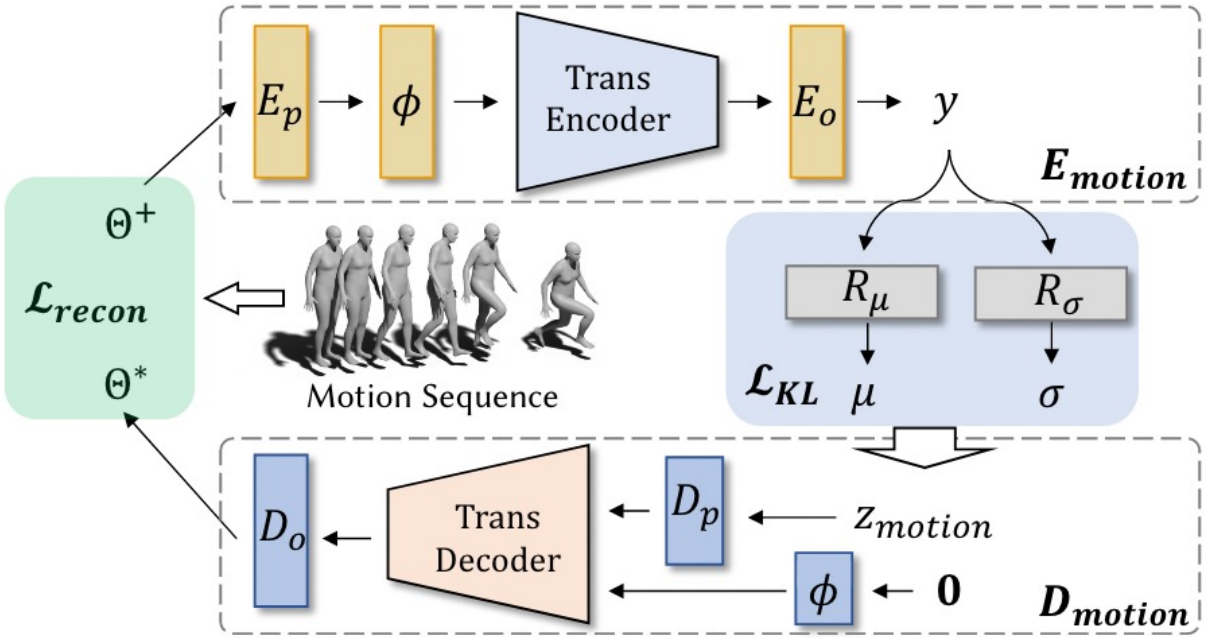


Fig. 9. **Structure of the Motion VAE.** The motion VAE contains three parts: the encoder E_{motion}, the decoder D_{motion}, and a reparameterization module. The reconstruction loss $\mathcal{L}_{\text{recon}}$ and the KL-divergence term $\mathcal{L}_{\text{KL}}$ are used for the motion VAE training.

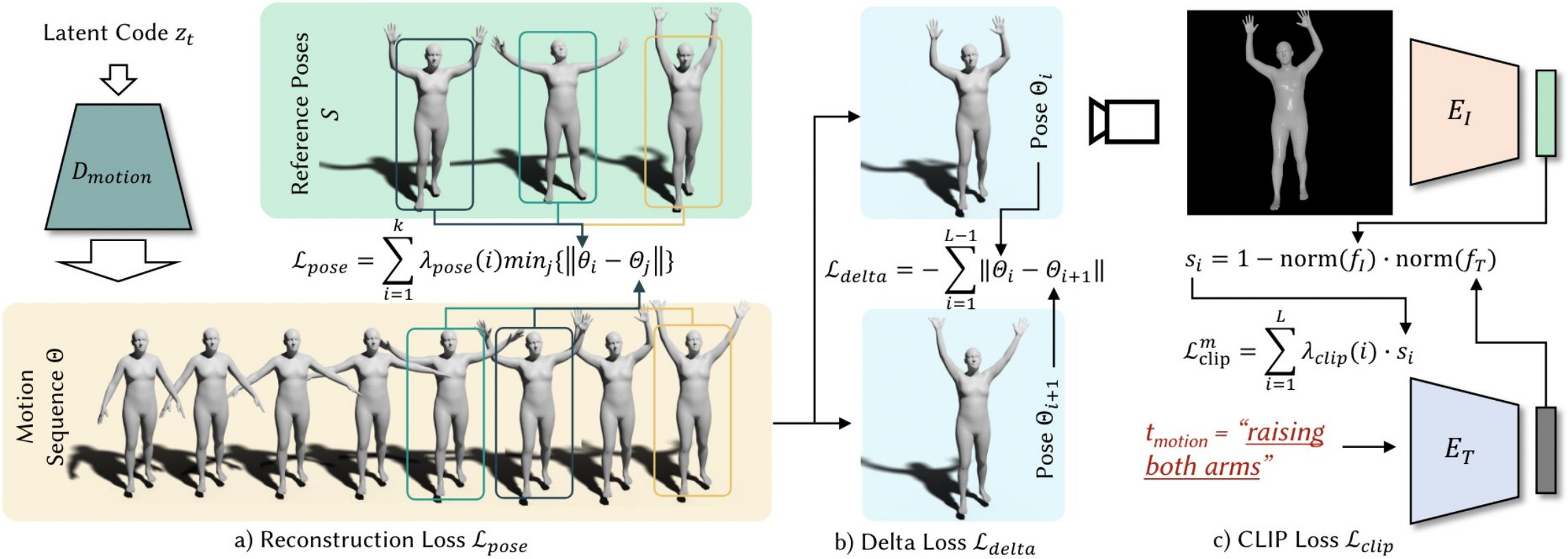
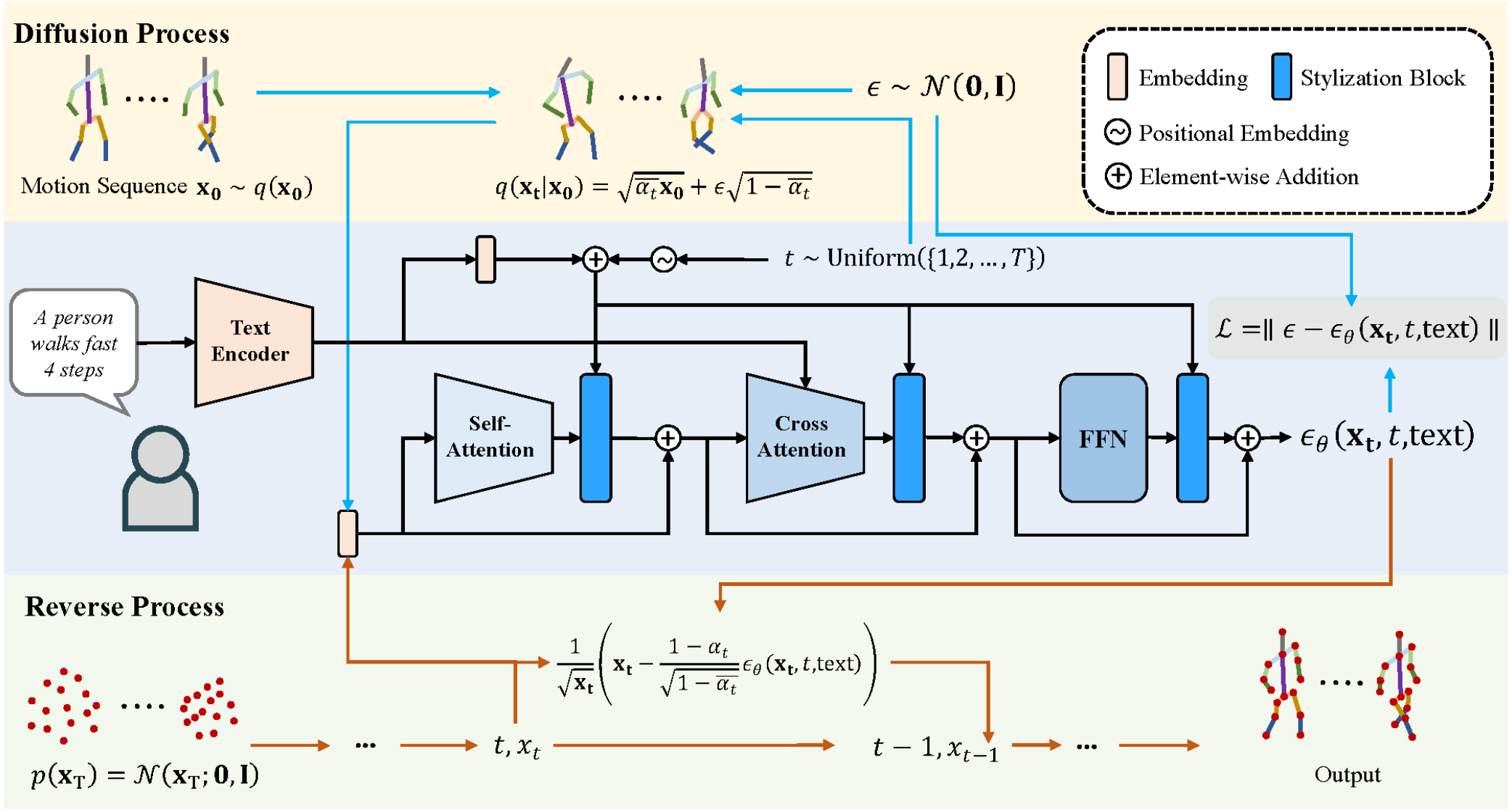


Fig. 10. **Detailed Pipeline of Reference-Based Animation with Motion Prior.** Three constraint terms are designed to optimize the latent code z_t. For the motion sequence Θ decoded by D_{motion}, $\mathcal{L}_{\text{pose}}$ is used to minimize the distance between each candidate pose and the nearest pose in Θ. $\mathcal{L}_{\text{delta}}$ is an adjustable loss item that measures the differences between adjacent poses and is capable of controlling the intensity of motion. $\mathcal{L}^m_{\text{clip}}$ measures the similarity between description t_{motion} and each pose in Θ.

2-4-1) MotionDiffuse



MotionDiffuse: Text-Driven Human Motion Generation with Diffusion Model, arXiv2022, 528 star

2-4-2) MDM: Motion Diffusion Model

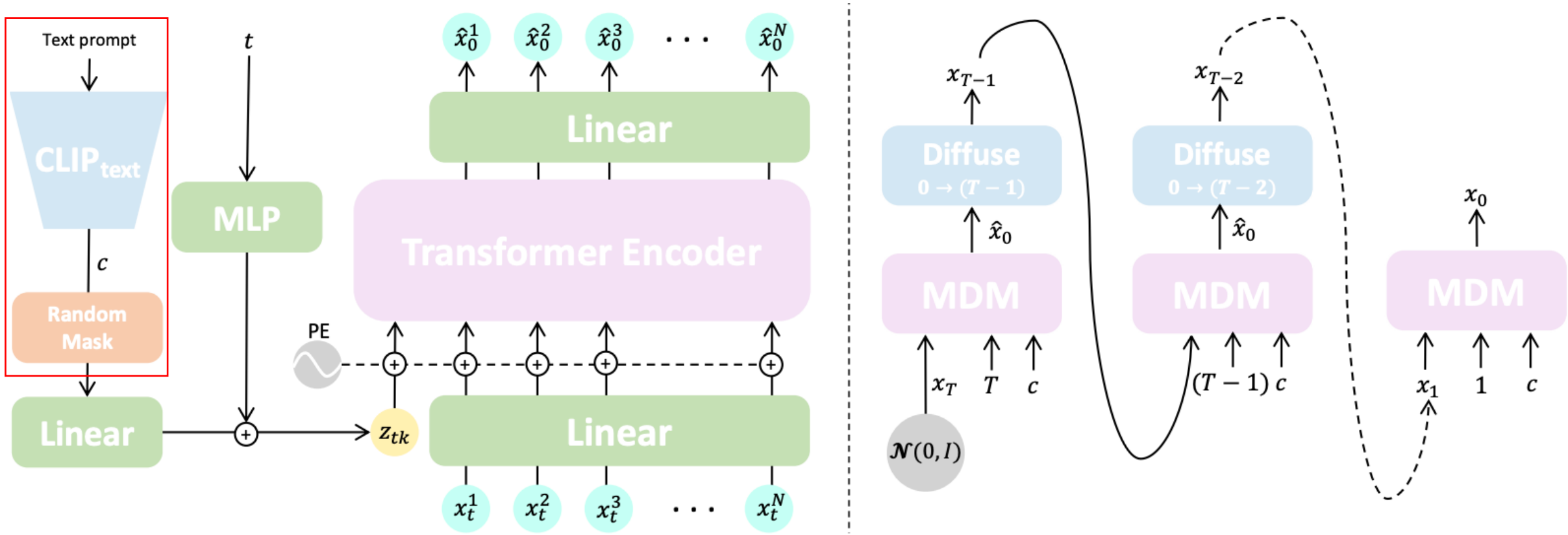
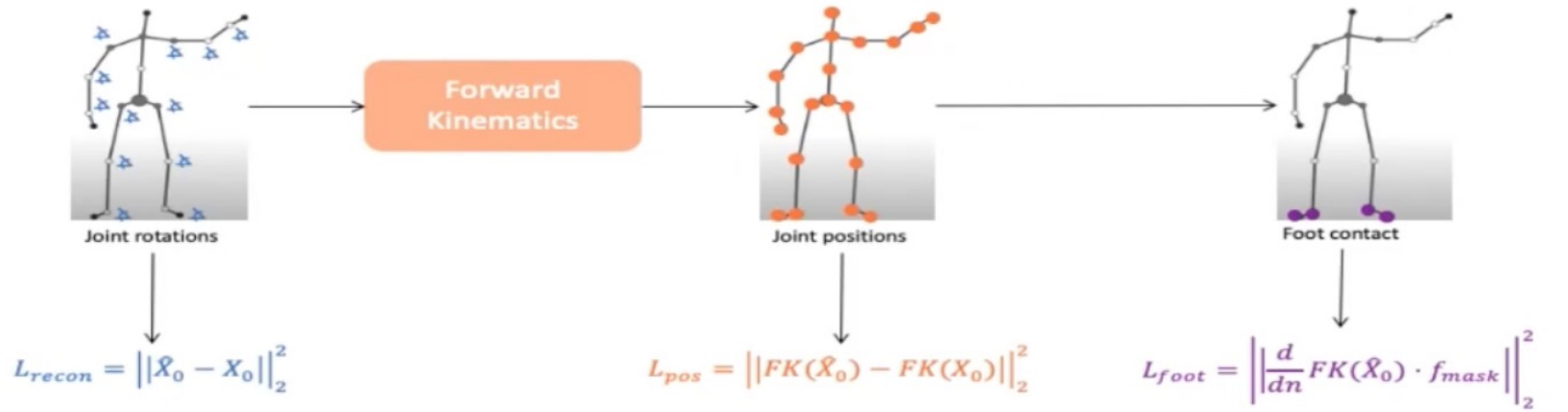
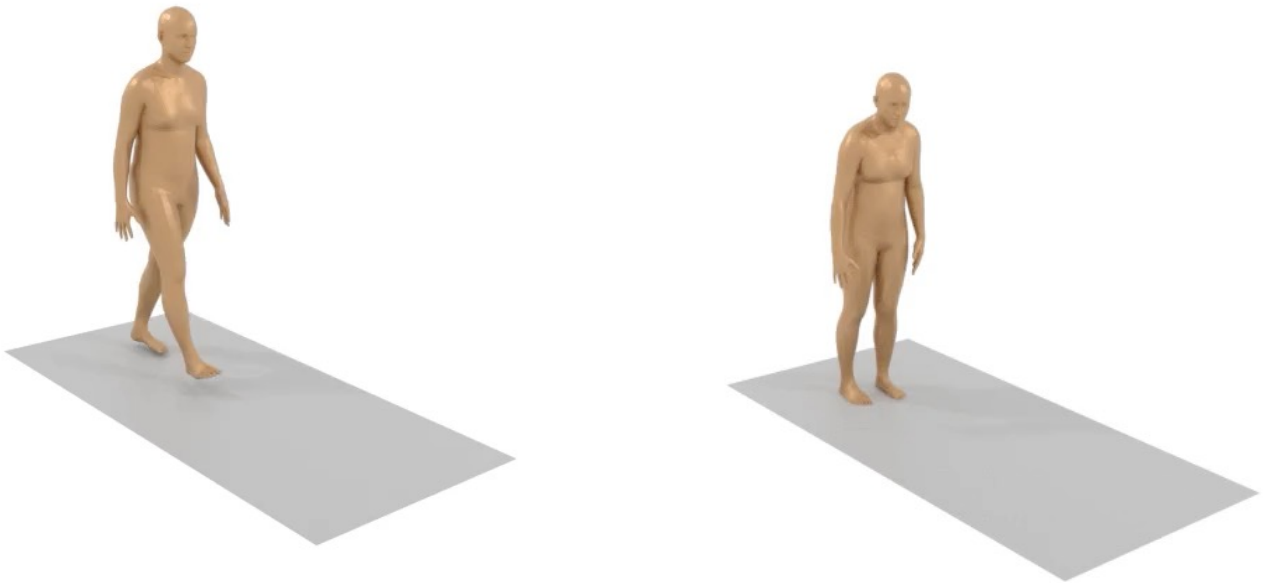


Figure 2: **(Left) Motion Diffusion Model (MDM) overview.** The model is fed a motion sequence $x_t^{1:N}$ of length N in a noising step t, as well as t itself and a conditioning code c. c, a CLIP (Radford et al., 2021) based textual embedding in this case, is first randomly masked for classifier-free learning and then projected together with t into the input token z_{tk}. In each sampling step, the transformer-encoder predicts the final clean motion $\hat{x}_0^{1:N}$. **(Right) Sampling MDM.** Given a condition c, we sample random noise x_T at the dimensions of the desired motion, then iterate from T to 1. At each step t, MDM predicts the clean sample $\hat{x}_0$, and diffuses it back to x_{t-1}.

MDM: Human Motion Diffusion Model, ICLR2023, 1.9K star

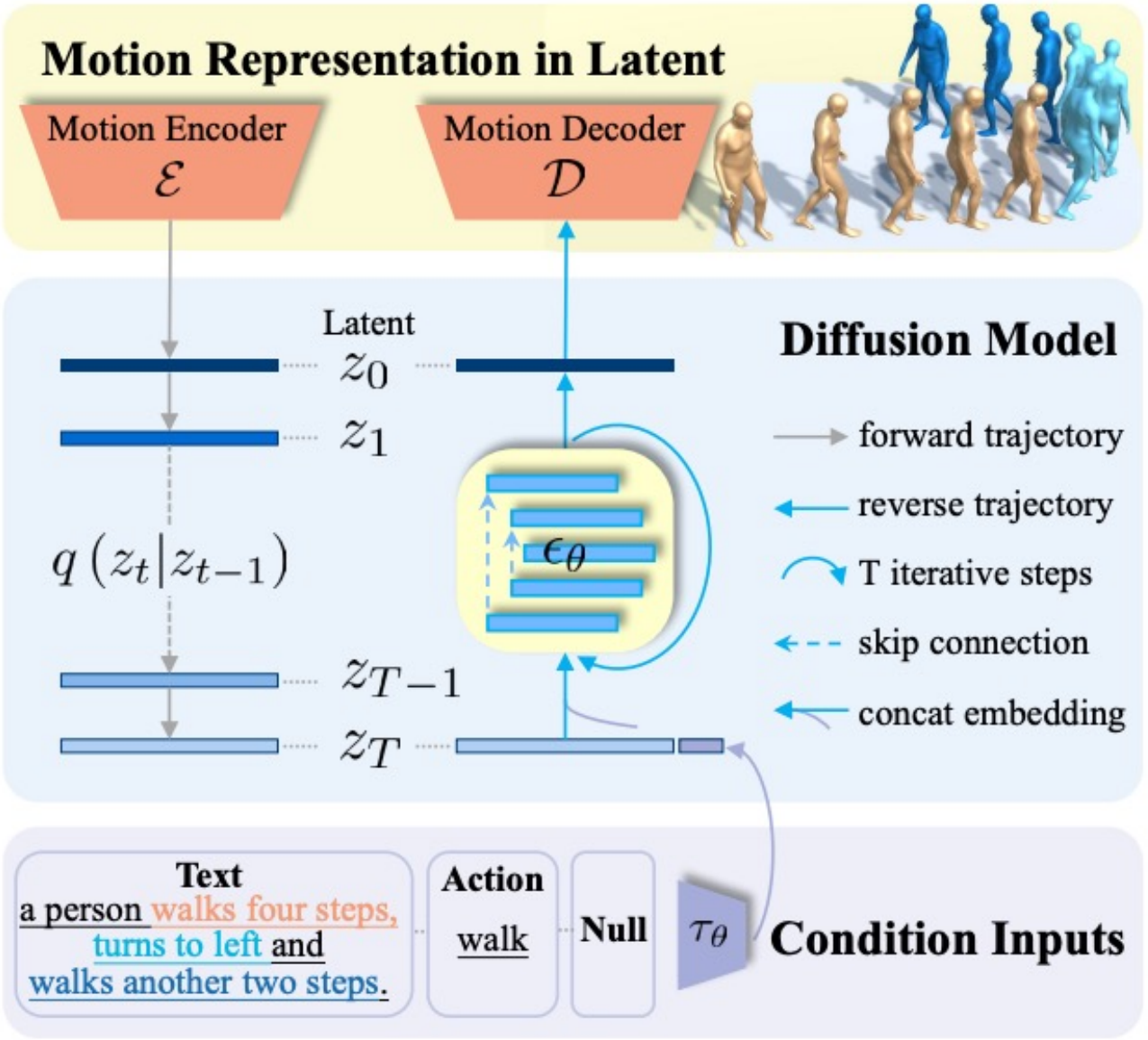
2-4-2) MDM: Motion Diffusion Model





[MDM: Human Motion Diffusion Model](), ICLR2023, 1.9K star

2-4-3) MLD: Motion latent diffusion



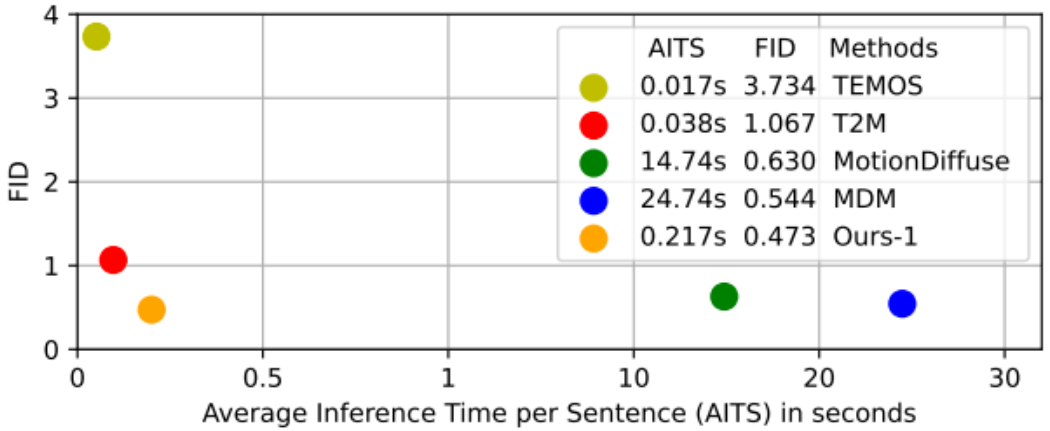
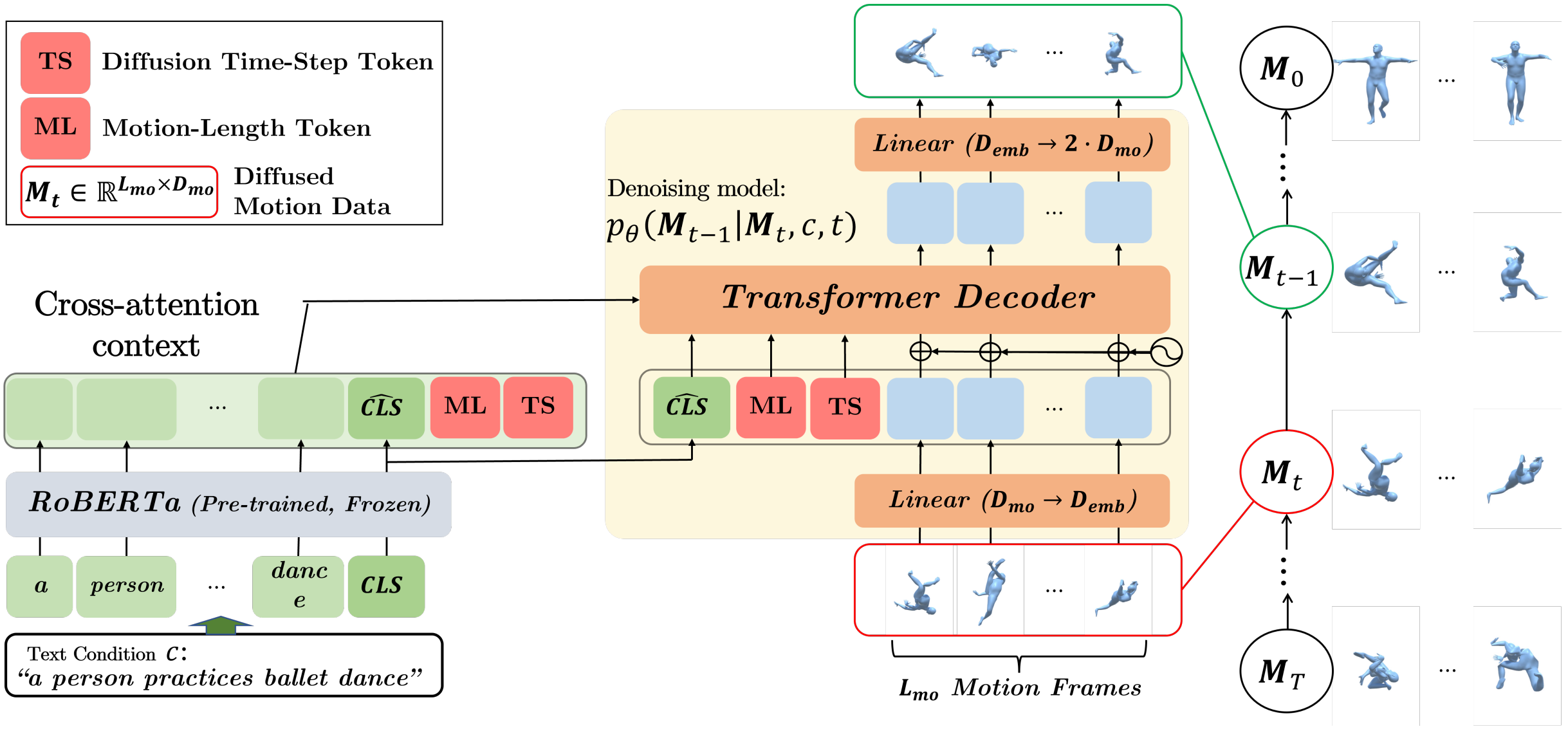


Figure 6. Comparison of the inference time costs on text-to-motion. We calculate AITS on the test set of HumanML3D [17] without model or data loading parts. All tests are performed on the same Tesla V100. The closer the model is to the origin the better.

Executing your Commands via Motion Diffusion in Latent Space, CVPR 2023, 174 star

2-4-4) FLAME

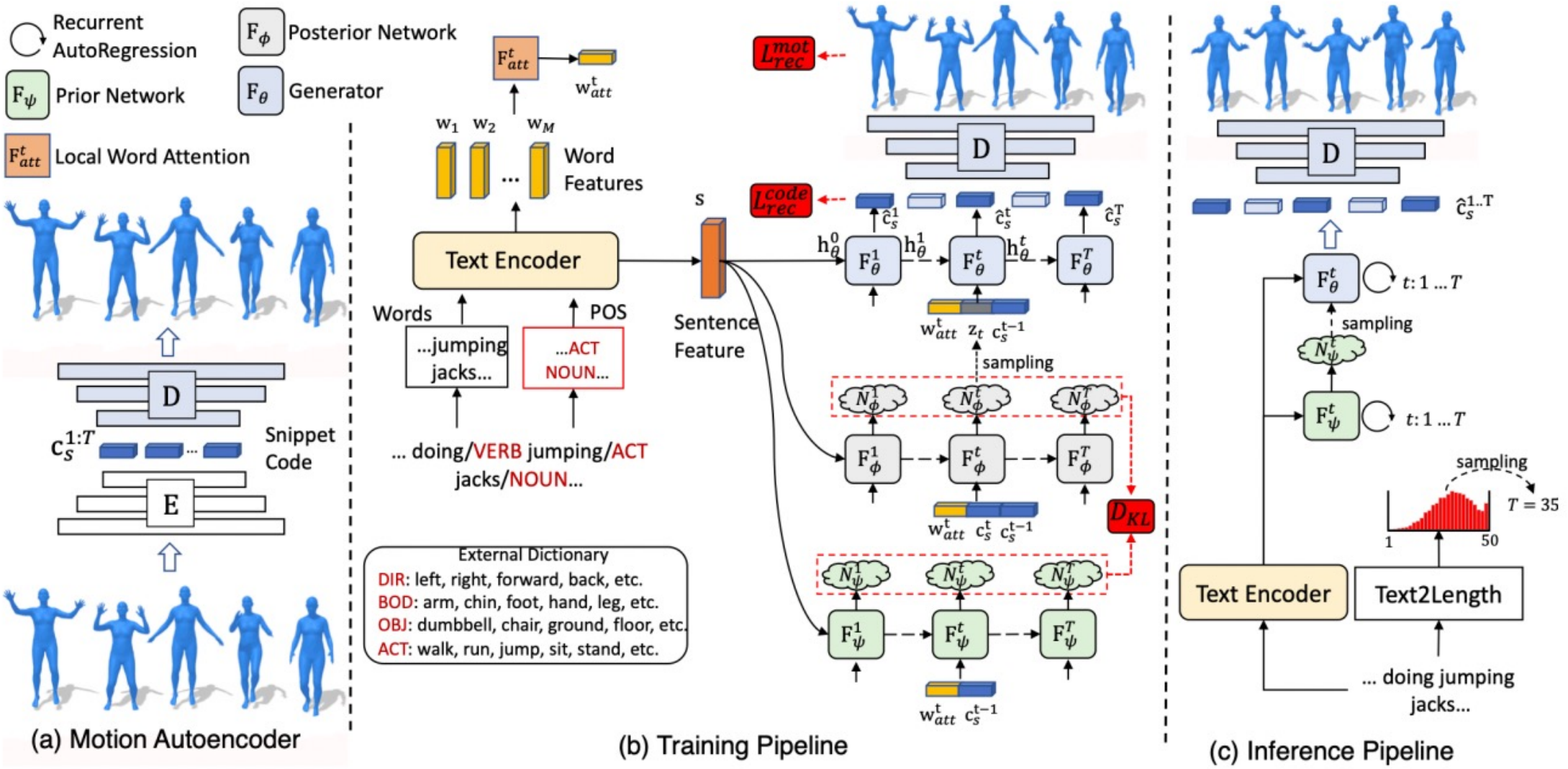


FLAME: Free-form Language-based Motion Synthesis & Editing, AAAI 2023, star 61

Diffusion Steps at $t = 0$

FLAME: Free-form Language-based Motion Synthesis & Editing, AAAI 2023, star 61

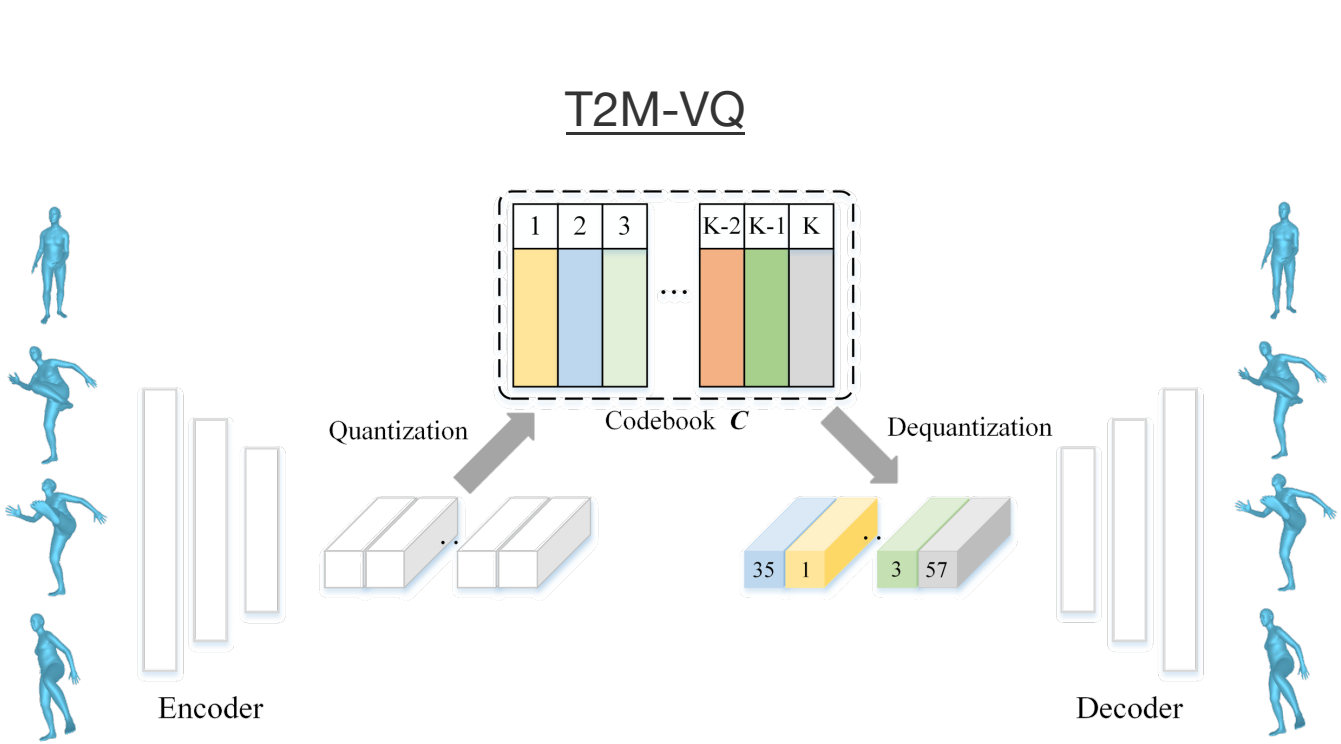
2-5) VQ-VAE

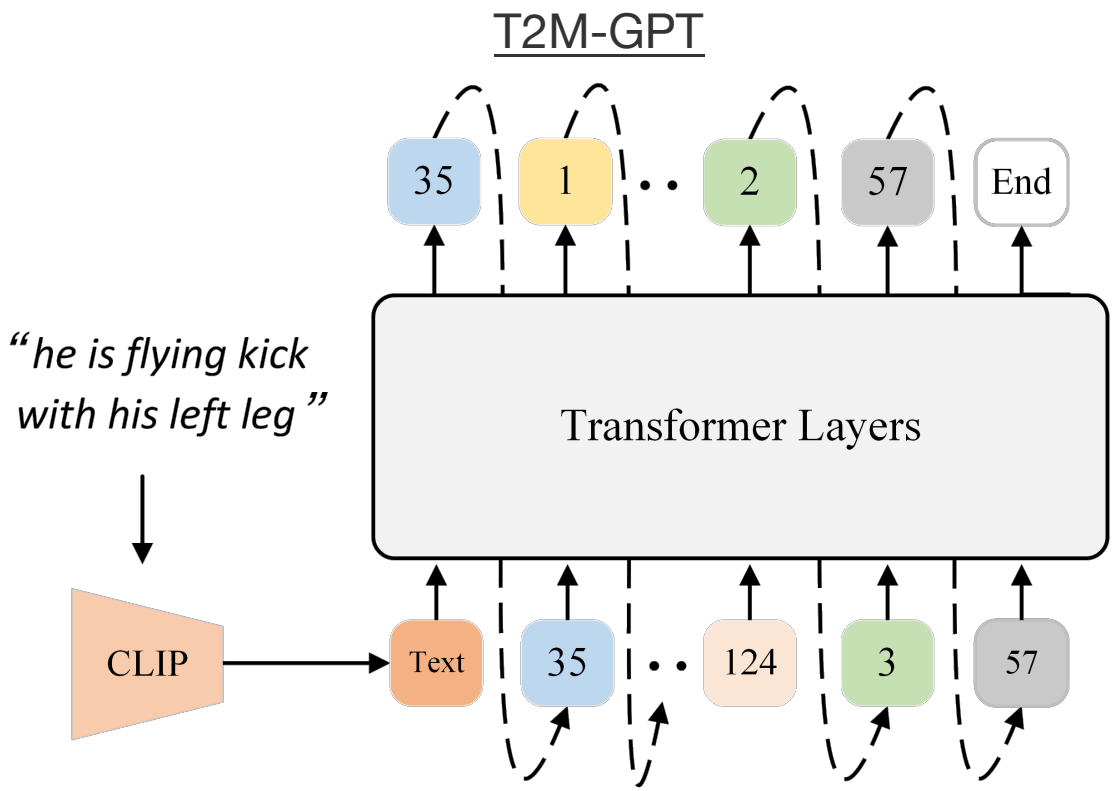


Generating Diverse and Natural 3D Human Motions from Text (CVPR 2022), CVPR2022

T2M-GPT: Generating Human Motion from Textual Descriptions with Discrete Representations, CVPR2023

2-5) VQ-VAE





Generating Diverse and Natural 3D Human Motions from Text (CVPR 2022), CVPR2022

T2M-GPT: Generating Human Motion from Textual Descriptions with Discrete Representations, CVPR2023

3. Approach

1+2) MoFusion

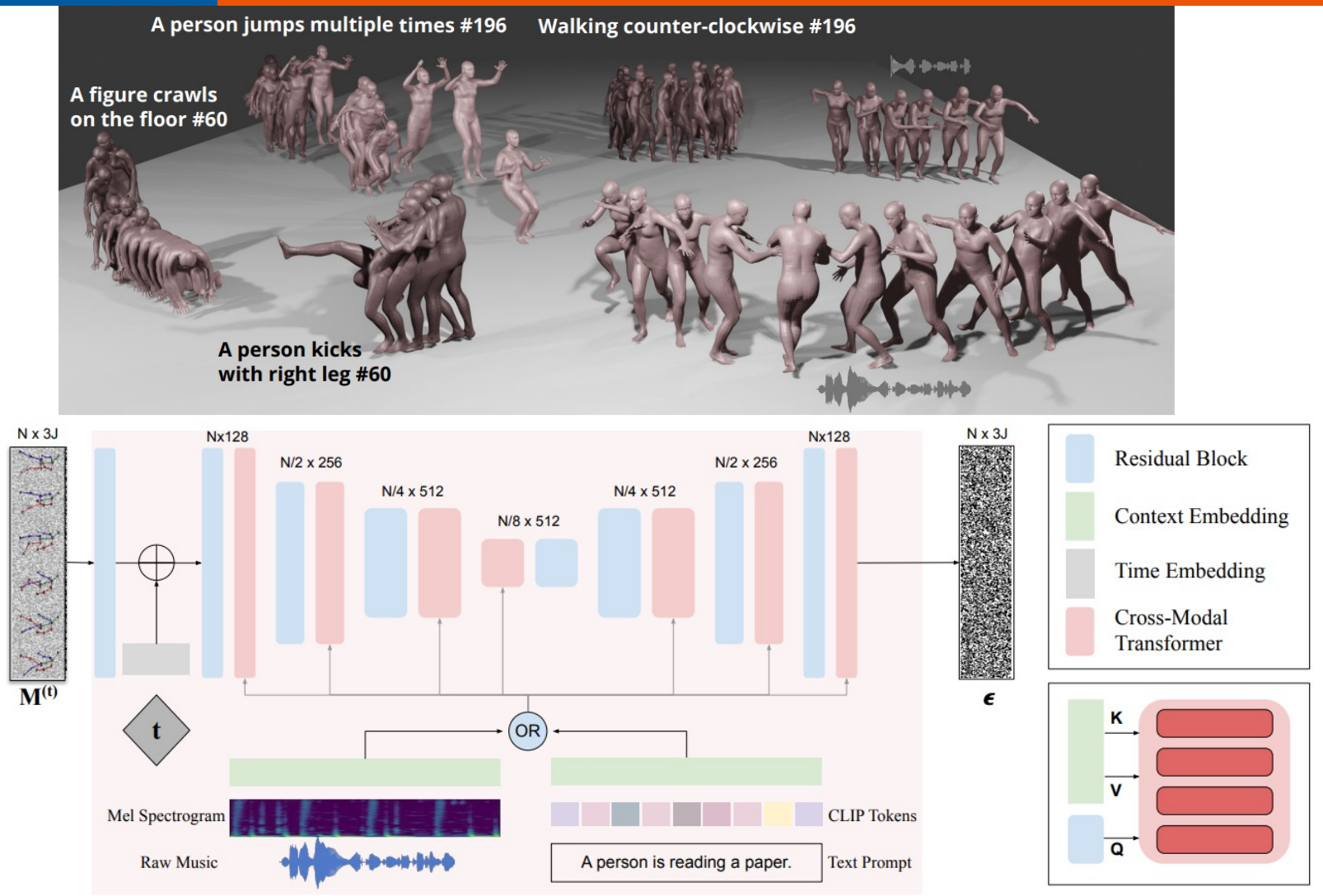
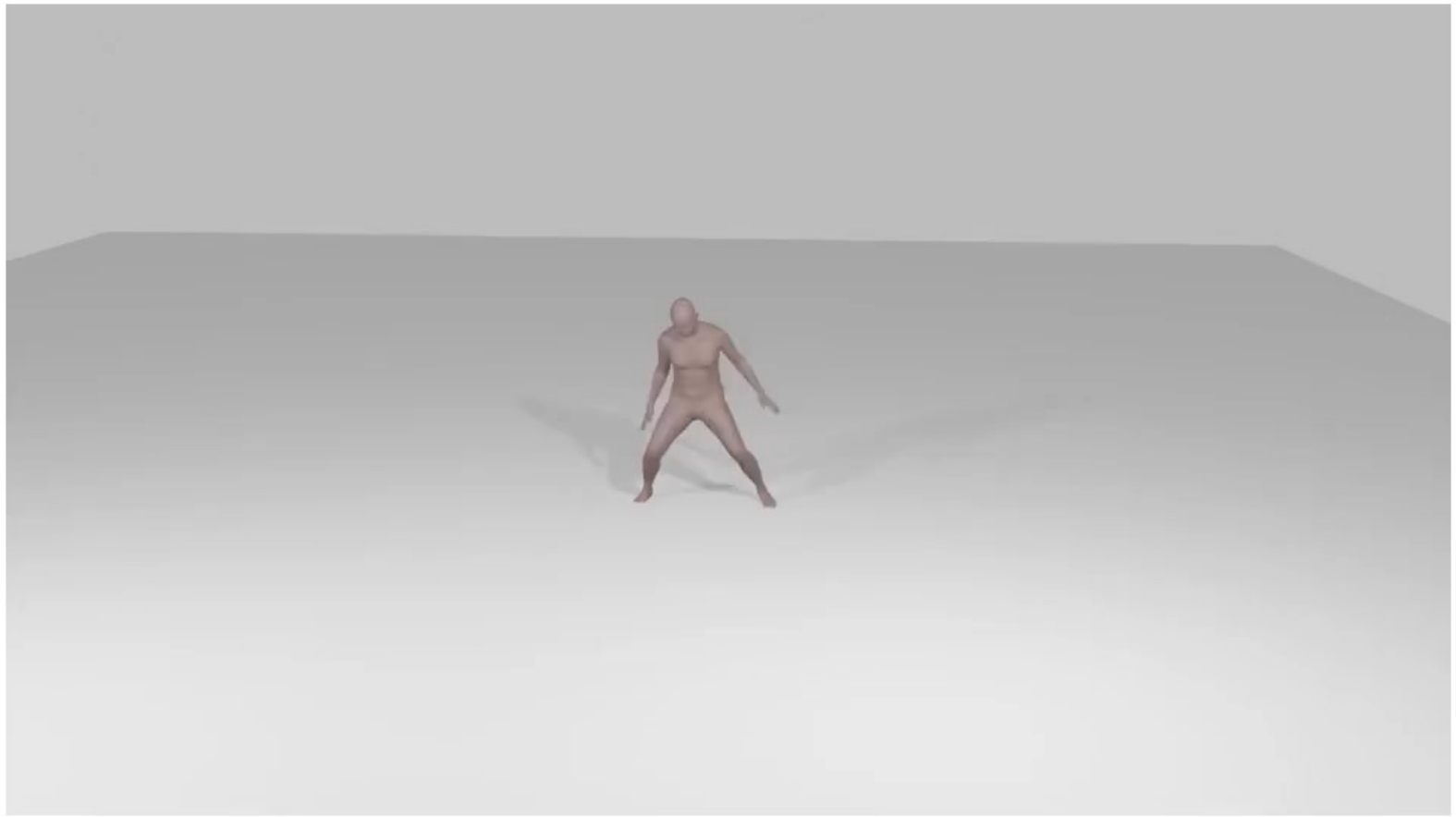


Figure 3: Illustration of the 1D U-Net architecture with cross-modal transformer blocks with multi-head attention (bottom right). The network's input is a noisy motion sample at timestep t, and the output is an estimate of the noise ϵ. Additionally, it can be conditioned on either music or text prompts. In both cases, we learn a projection function to map the conditioning features to 1D U-Net features.

MoFusion: A Framework for Denoising-Diffusion-based Motion Synthesis, CVPR2023

1+2) MoFusion

Music-to-Dance Synthesis



MoFusion results
after IK
3

[MoFusion: A Framework for Denoising-Diffusion-based Motion Synthesis](), CVPR2023

4. Conclusion

What we do next??

→ 우리는 생성 모델을 어떤 Task에 응용할 수 있을까?

→생성모델의 Contrastive, Condition Feature가 우리 Task에서 활용가능한가?