

BodyFormer: Semantics-guided 3D Body Gesture Synthesis with Transformer

Kunkun Pang^{1*}, Dafei Qin^{2*}, Yingruo Fan², Julian Habekost³, Takaaki Shiratori⁴, Junich Yamagishi⁵, Taku Komura^{2,6,†}

1. Institute of Intelligent Manufacturing Guangdong Academy of Sciences,

2. The University of Hong Kong

3. The University of Edinburgh

4. Meta Reality Labs Research ,

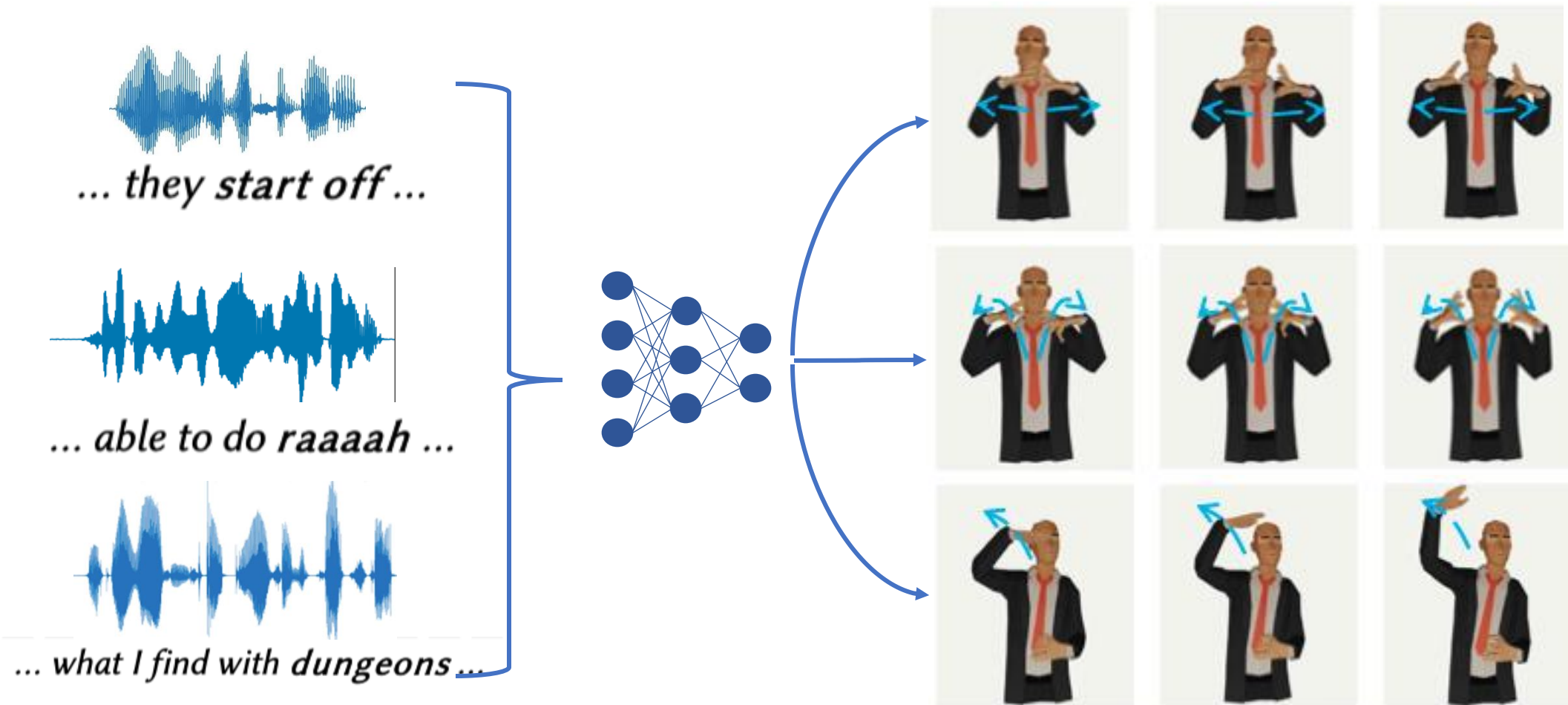
5. National Institute of Informatics,

6. Tohoku University

** Both contributed equally to this research*

† Corresponding author

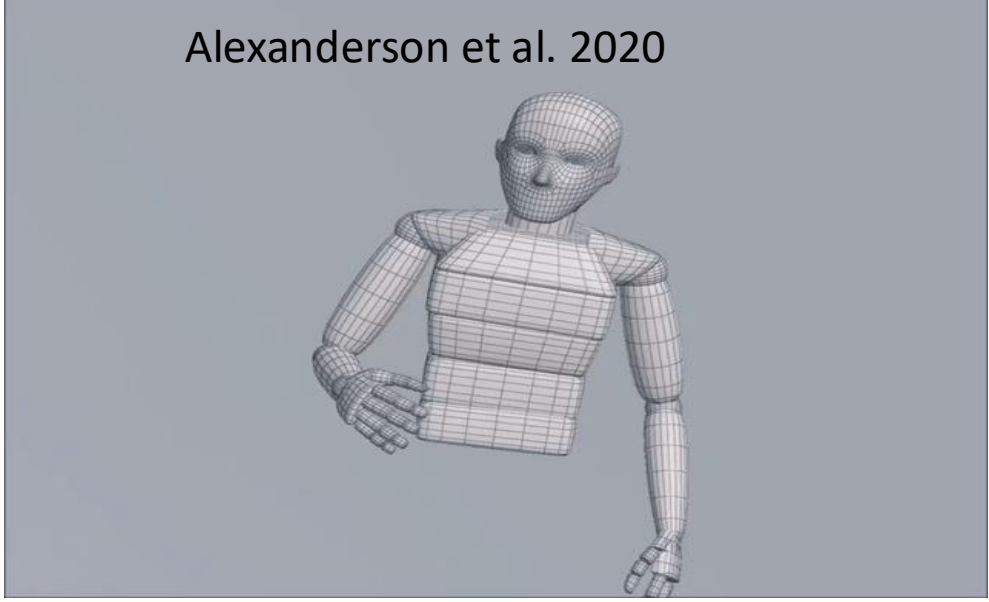
Goal: Automatic Gesture Synthesis from Speech



One-to-One Mapping



Many-to-Many Mapping



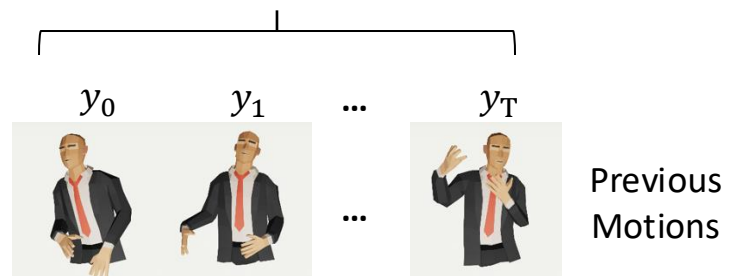
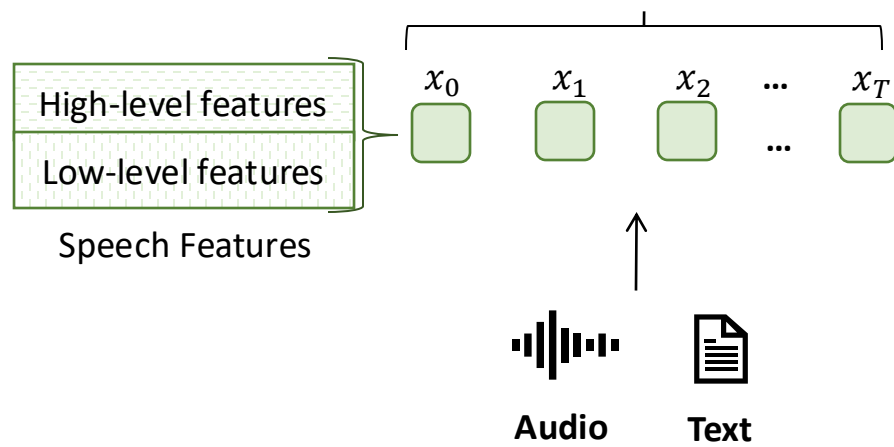
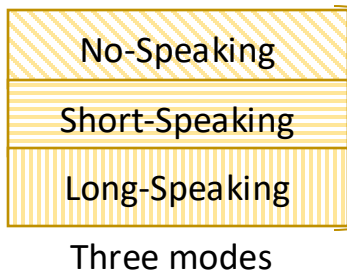
Motivation

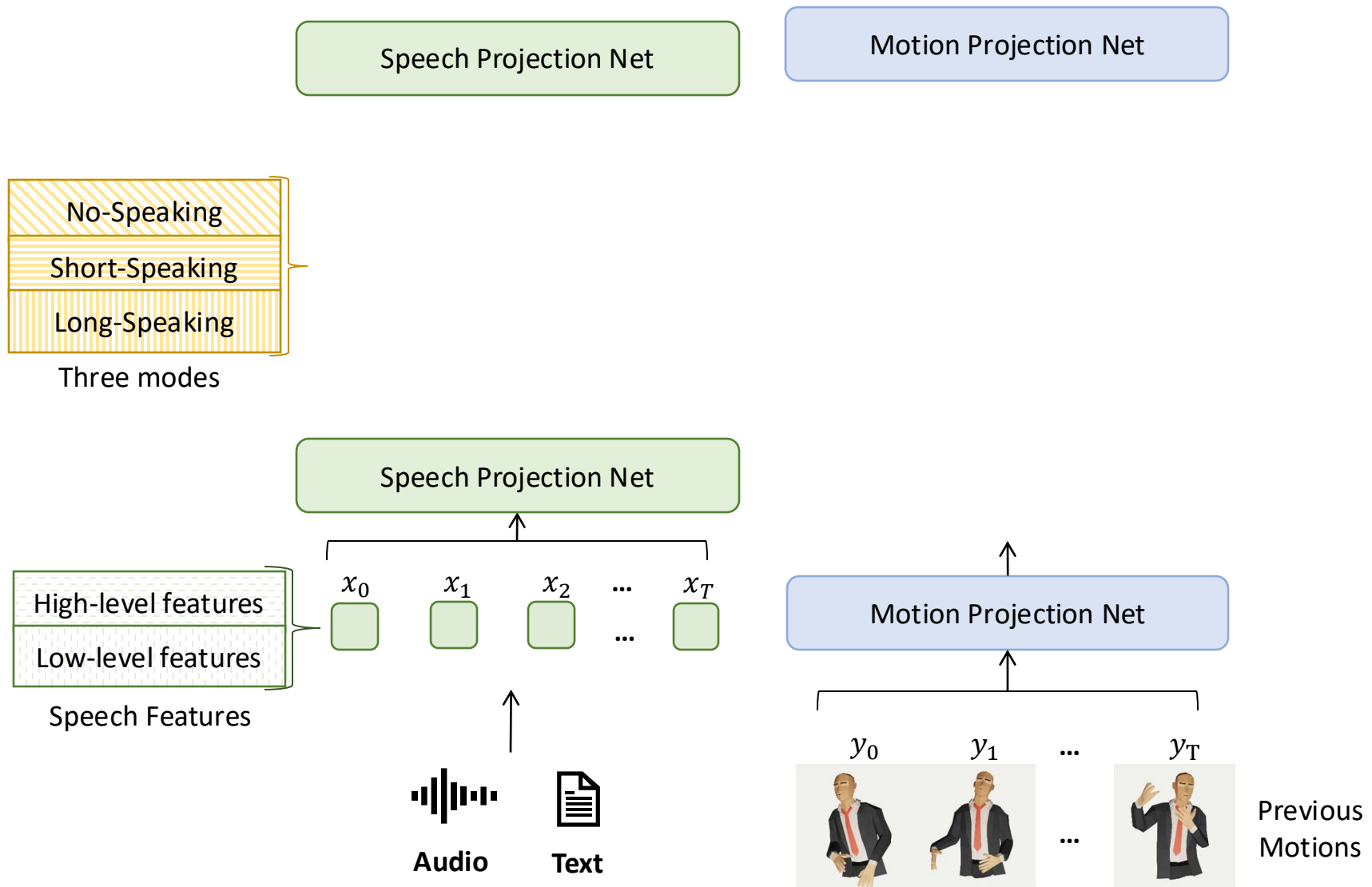
- Problem:

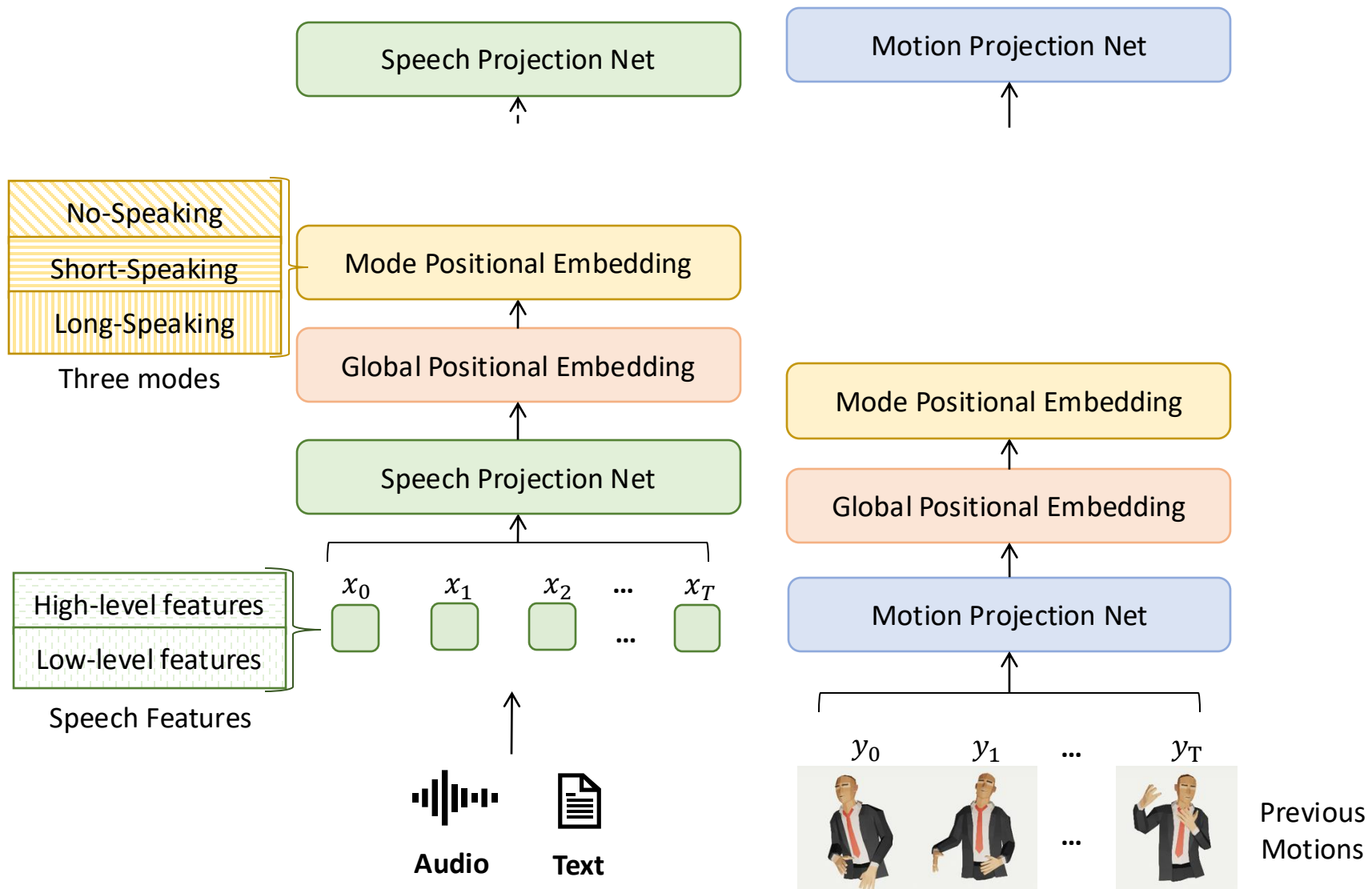
- Few contextualized motion
- Less motion in listening
- Limited training data.

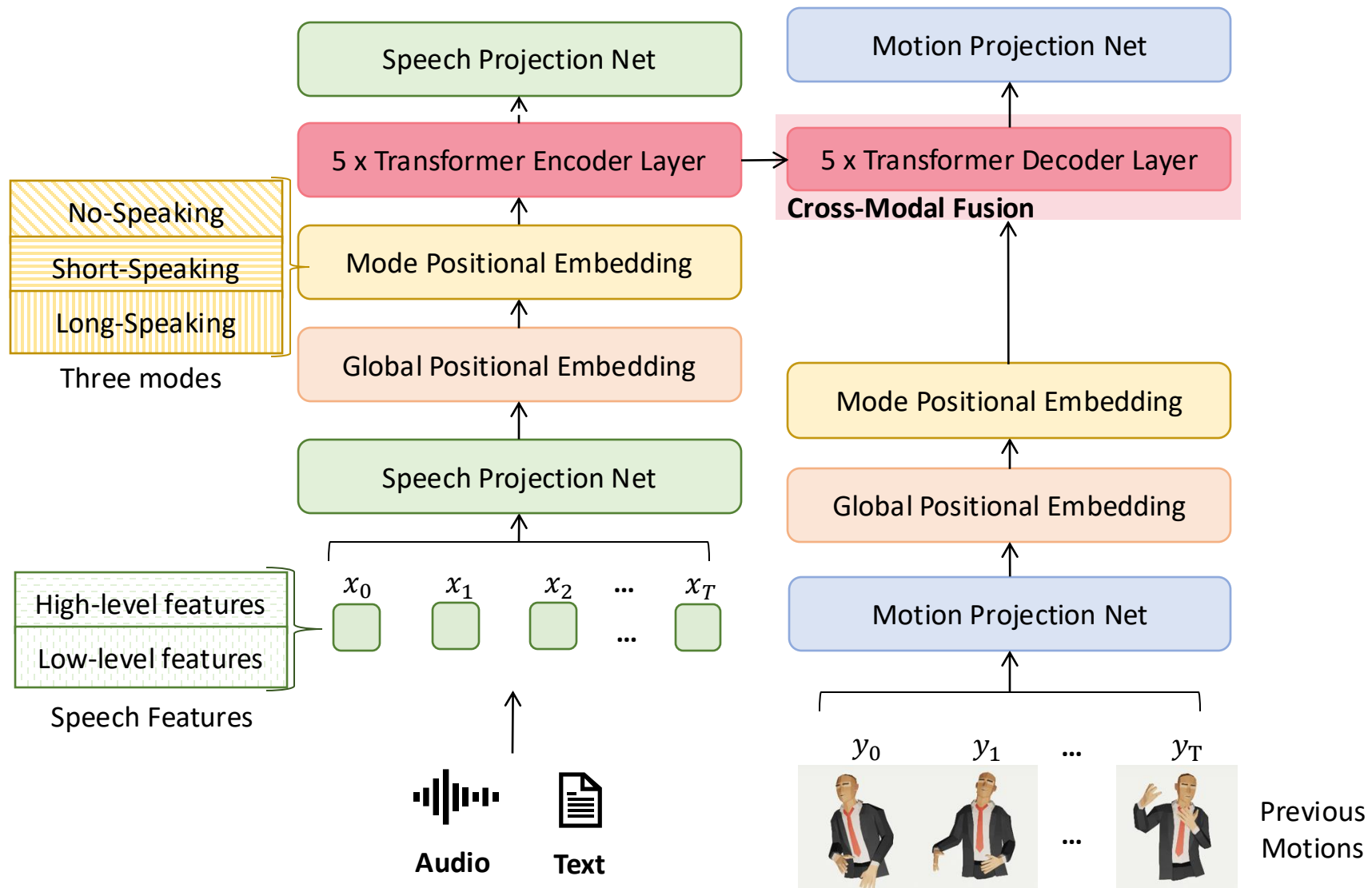
- Solution:

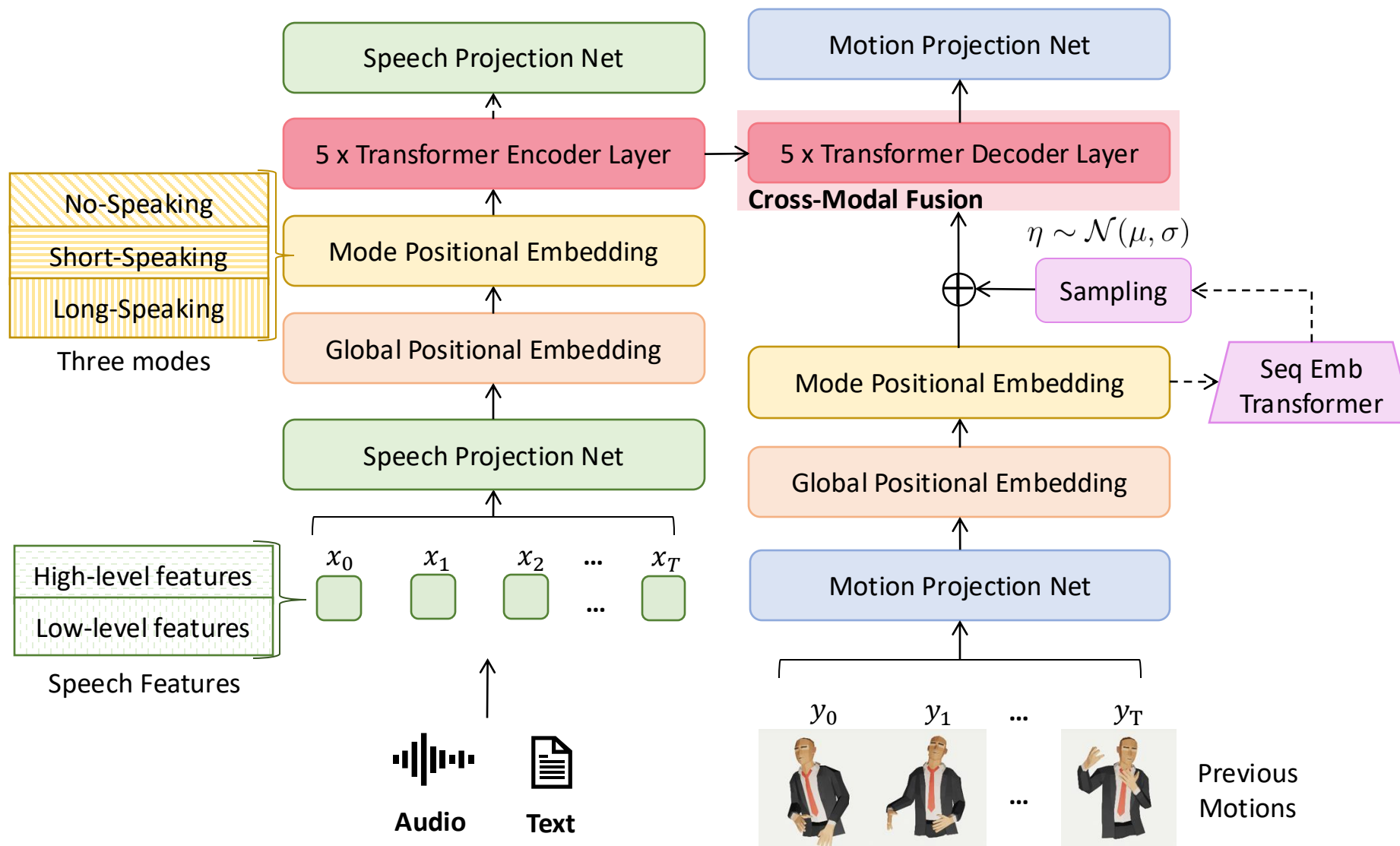
- Transformer
- Mode positional embedding
 - (i.e. Listening mode, speaking mode)
- Intra-modal pretraining











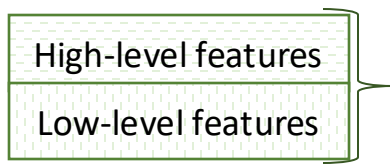
Speech Features

Low-level features

Speech Features



Audio



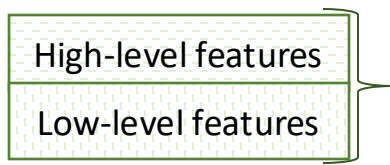
Speech Features



Audio



Text



Speech Features



Audio



Text

y_0

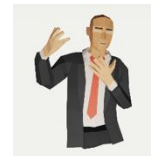


y_1

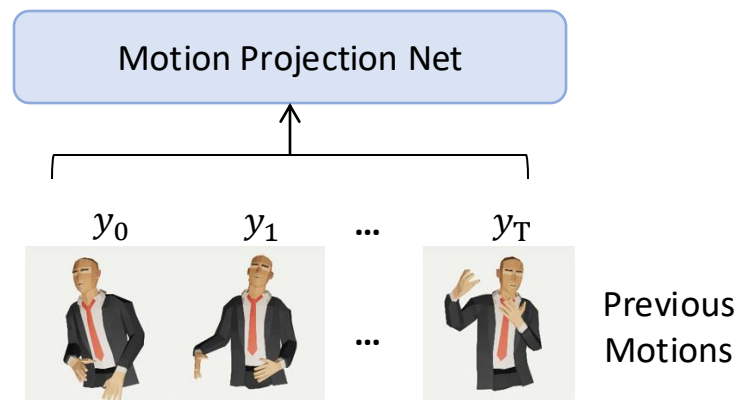
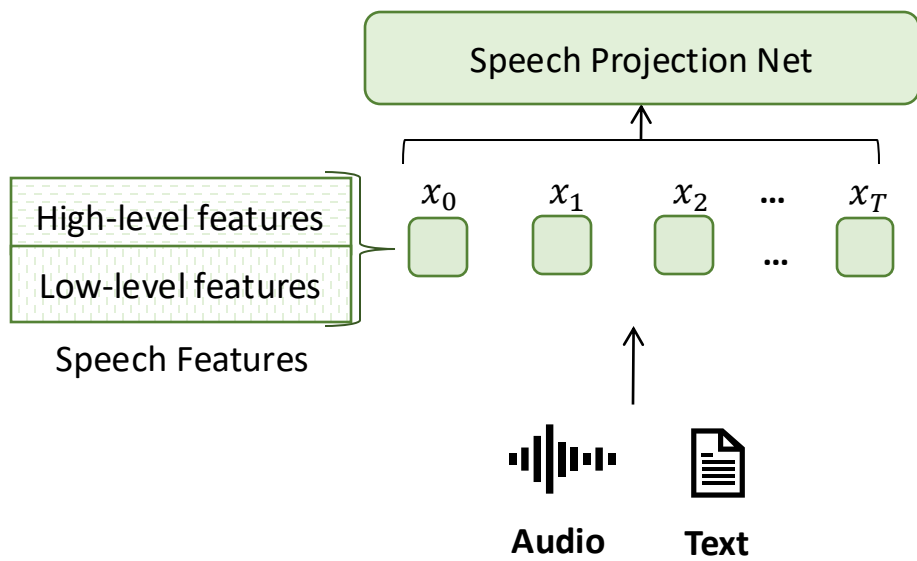


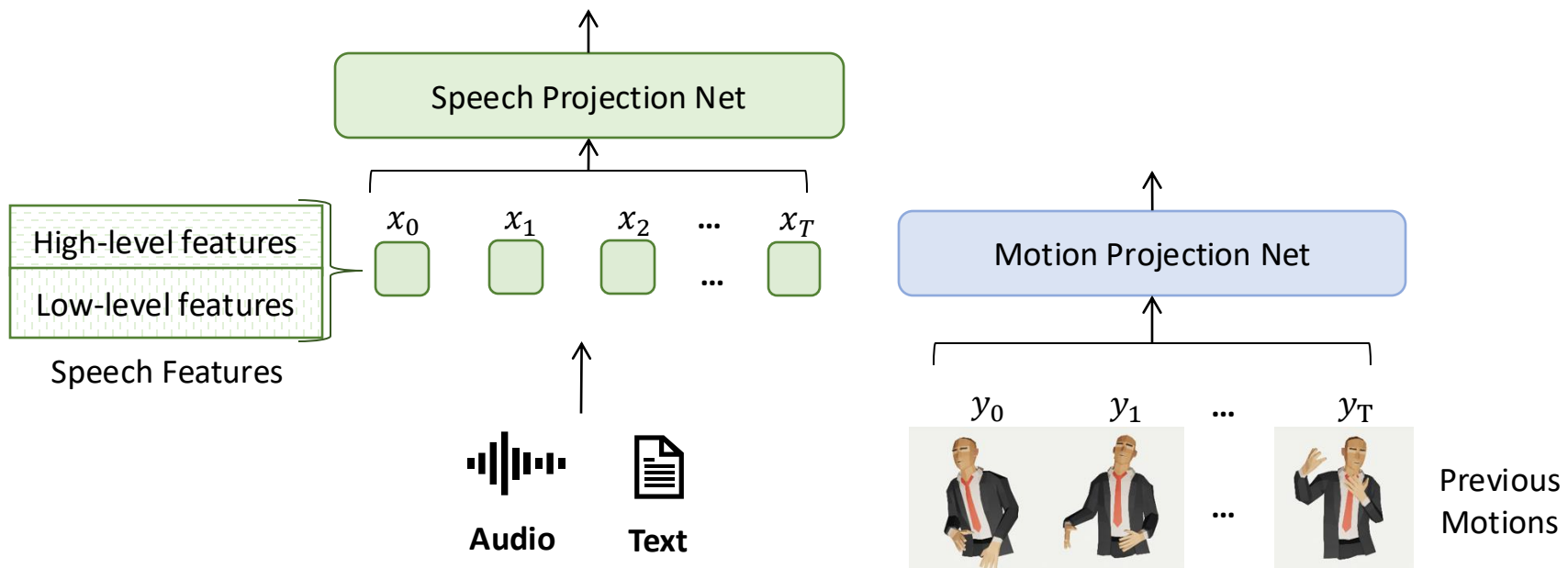
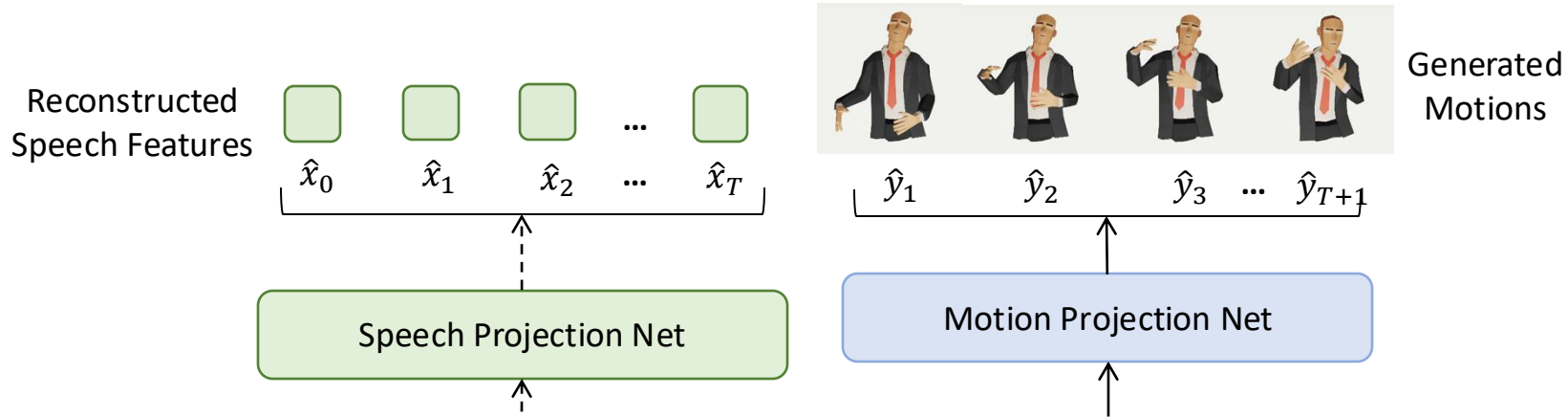
...

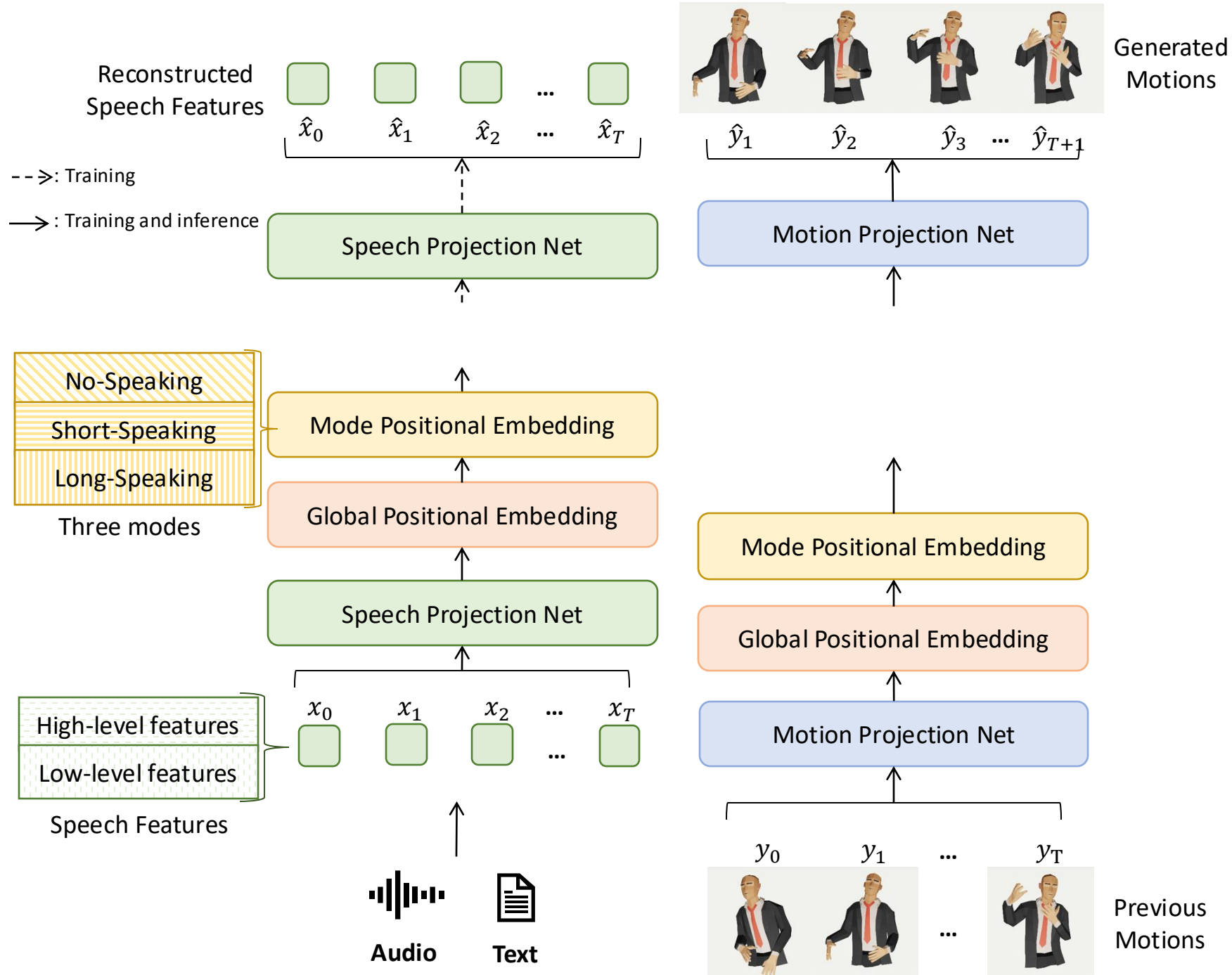
y_T

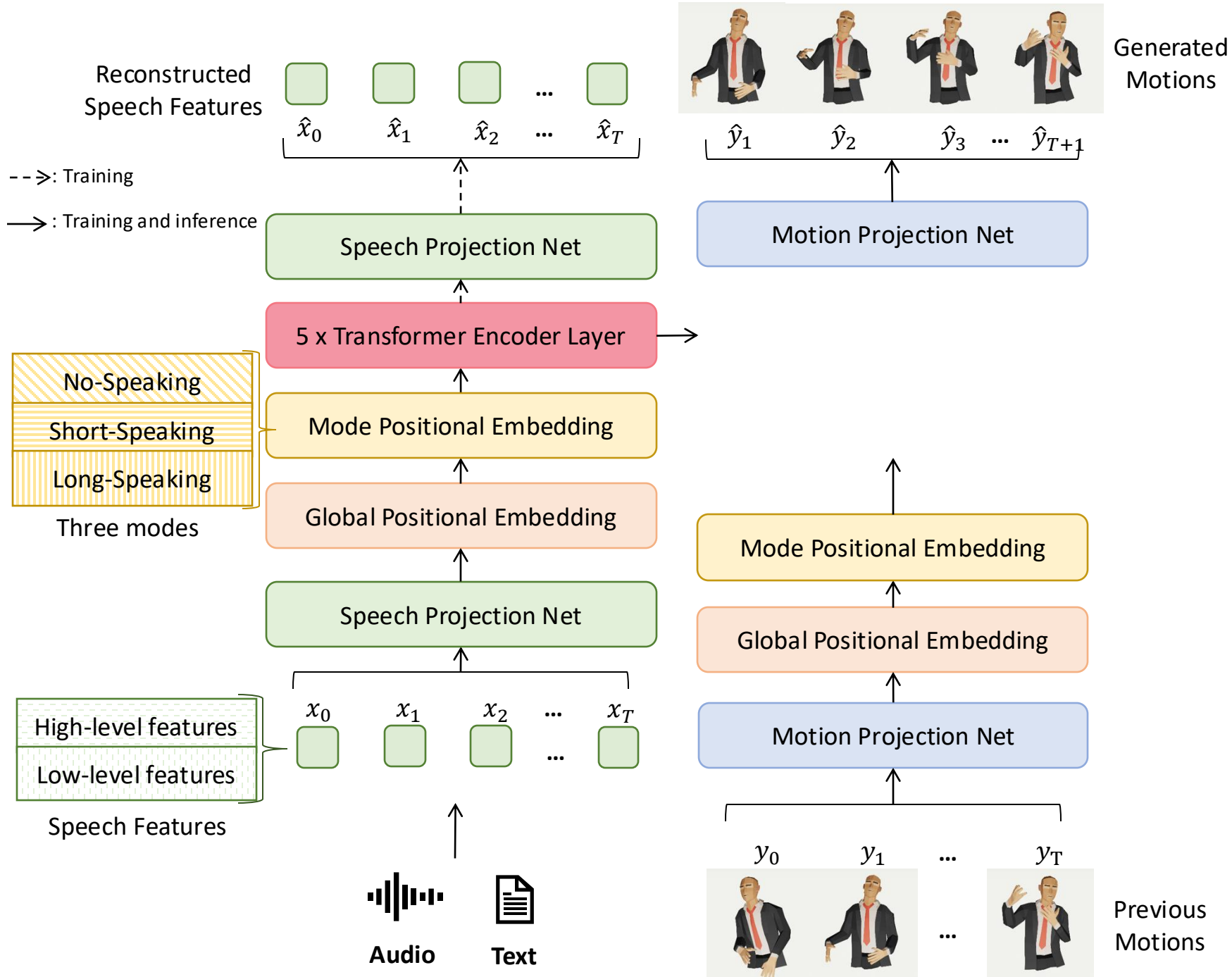


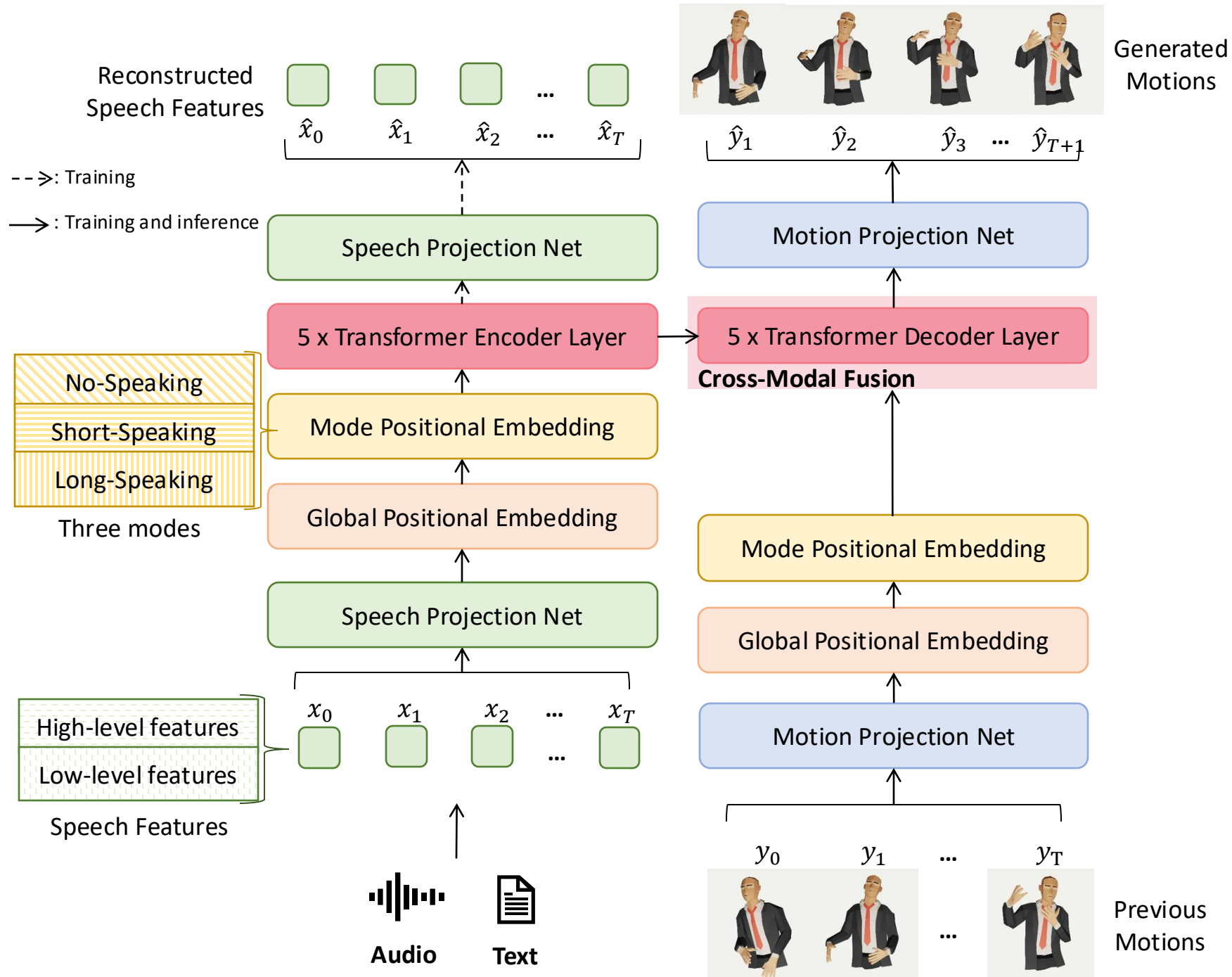
Previous
Motions

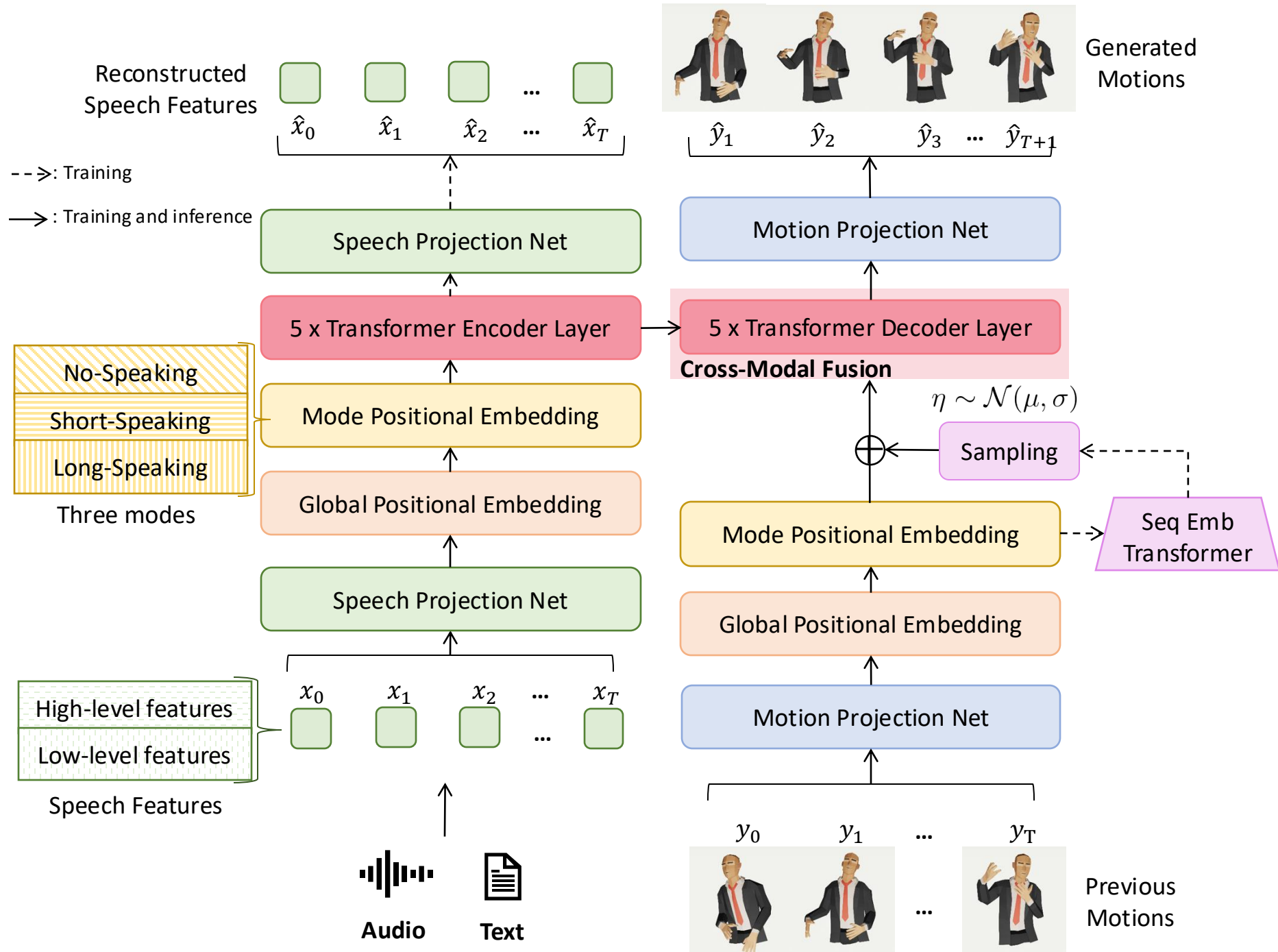


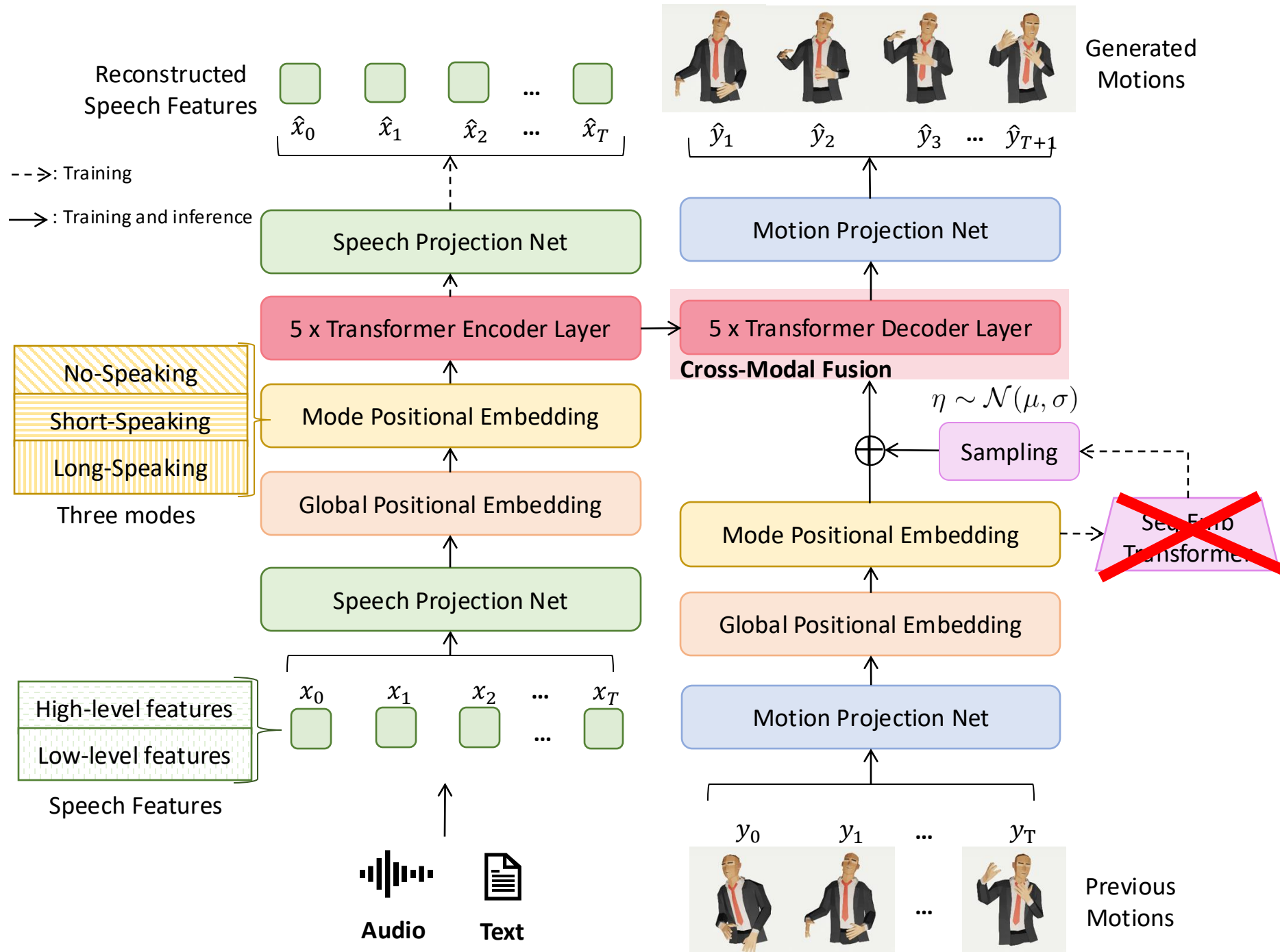




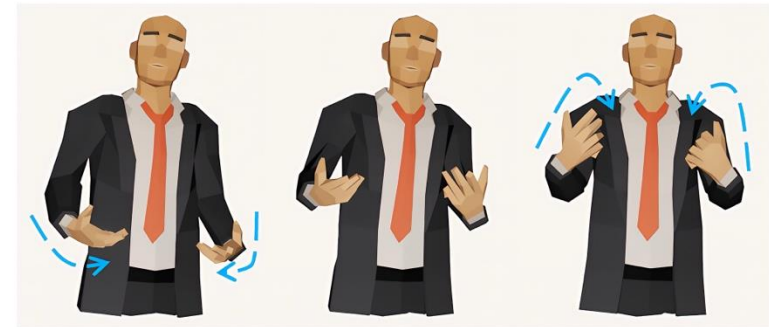
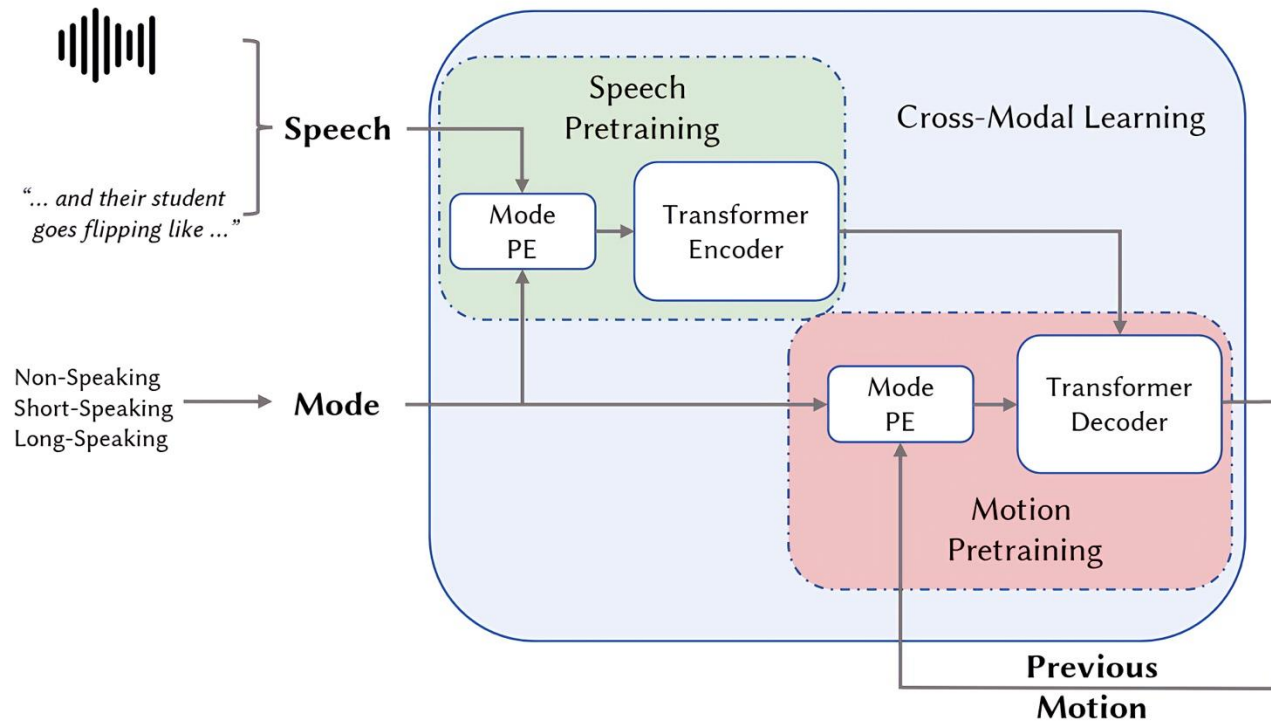








Intra-modal Pre-training

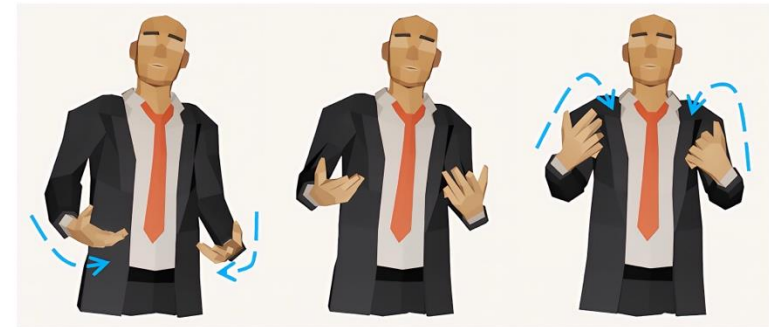
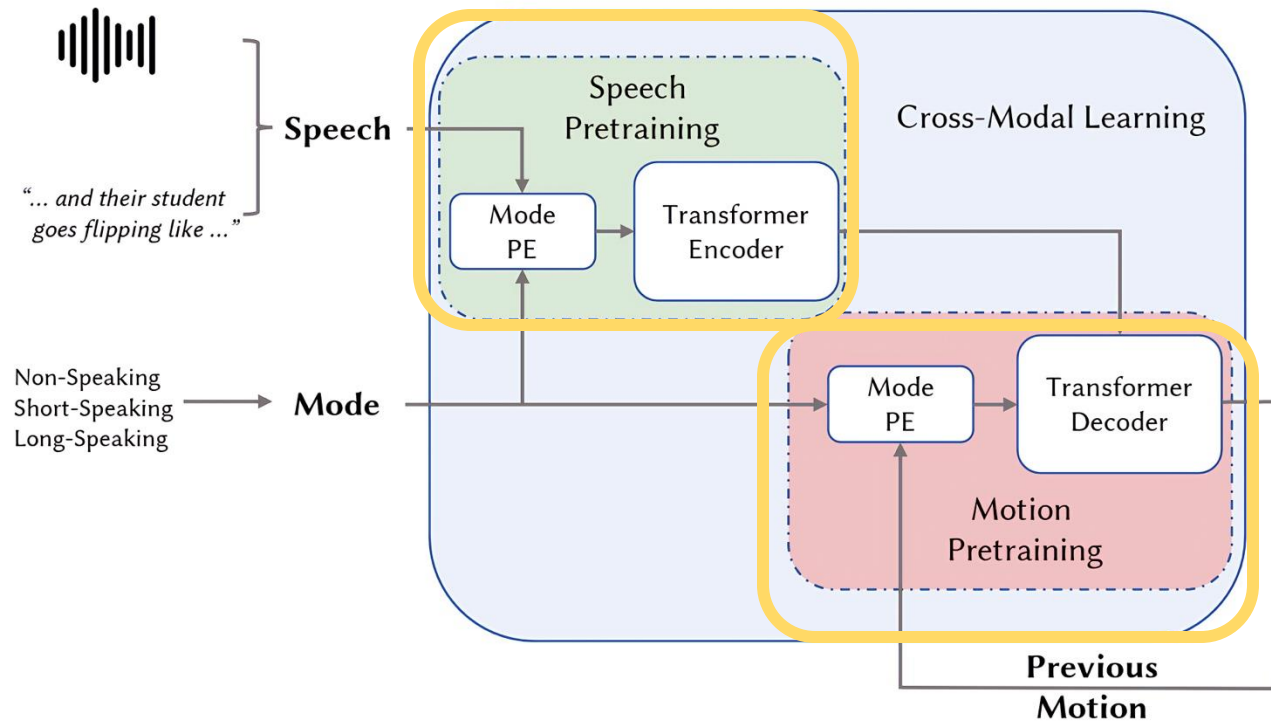


"... and *their* ---- students go ---- *flipping*



like off ---- *into the* ---- *wall ...*

Intra-modal Pre-training

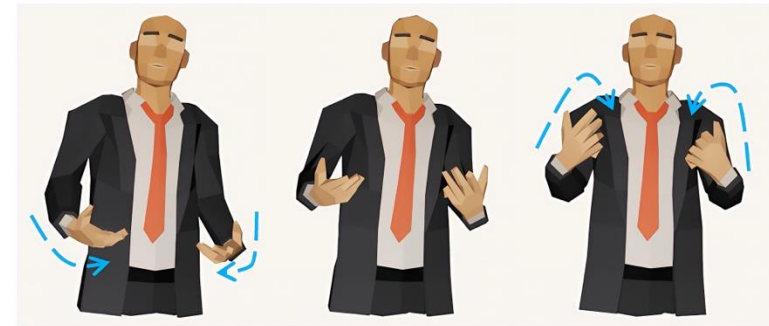
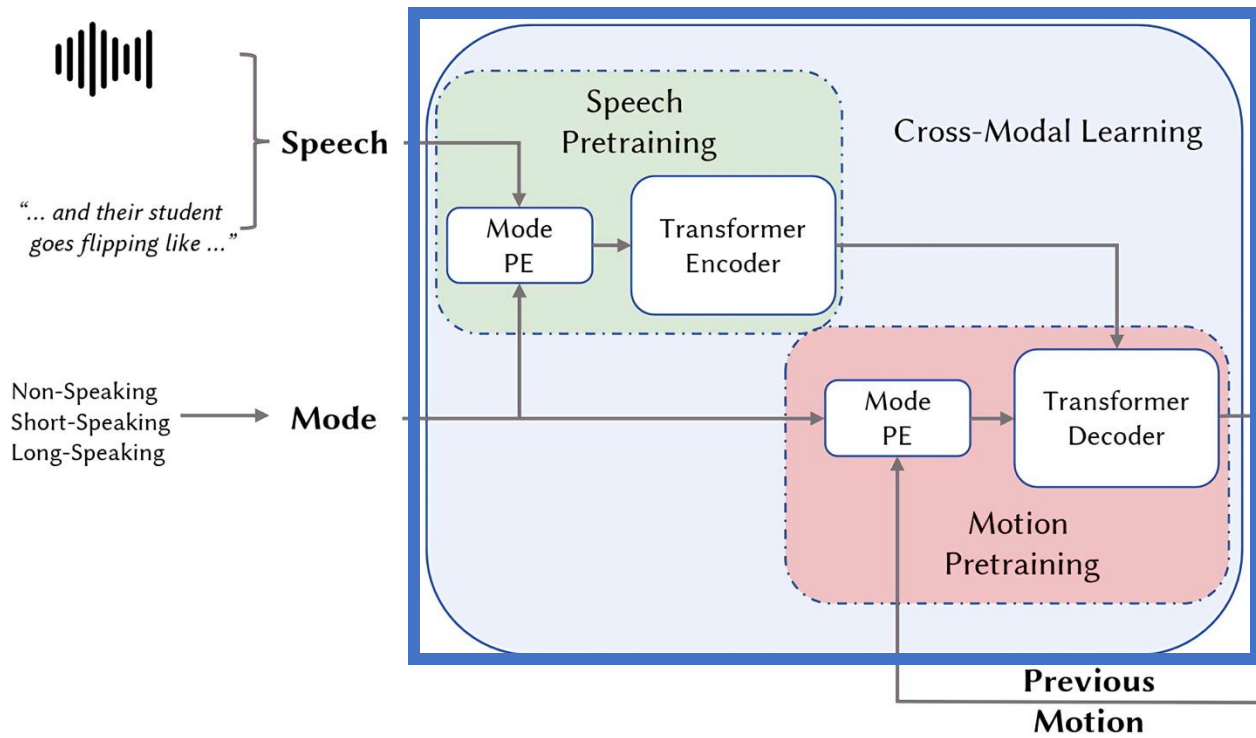


"... and *their* ---- students go ---- *flipping*"



like off ---- *into the* ---- *wall ...*"

Intra-modal Pre-training



"... and *their* ---- students go ---- *flipping*"



like off ---- into the ---- *wall ...*"

Comparison on Trinity

Ablation Study I: Intra-modal pre-training

Ablation Study II: Mode Embedding

Comparison to Rhythem Gesticulator

Comparison to Rhythmic Gesticulator

Comparison on Talking With Hands 16.2M

Summary

- We propose a transformer-based method to capture the long-term context information to generate human pose.
- We introduce the mode positional embedding to deal with the long-tailed distribution of the gesture patterns.
- Then, we introduce the intra-modal pre-training in order to train the transformer on small-scaled dataset.
- Our approach can produce high-quality movement in response to the rhythm and context of the speech.

Thanks you!