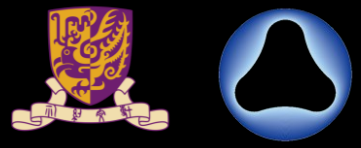


CodeTalker: Speech-Driven 3D Facial Animation with Discrete Motion Prior

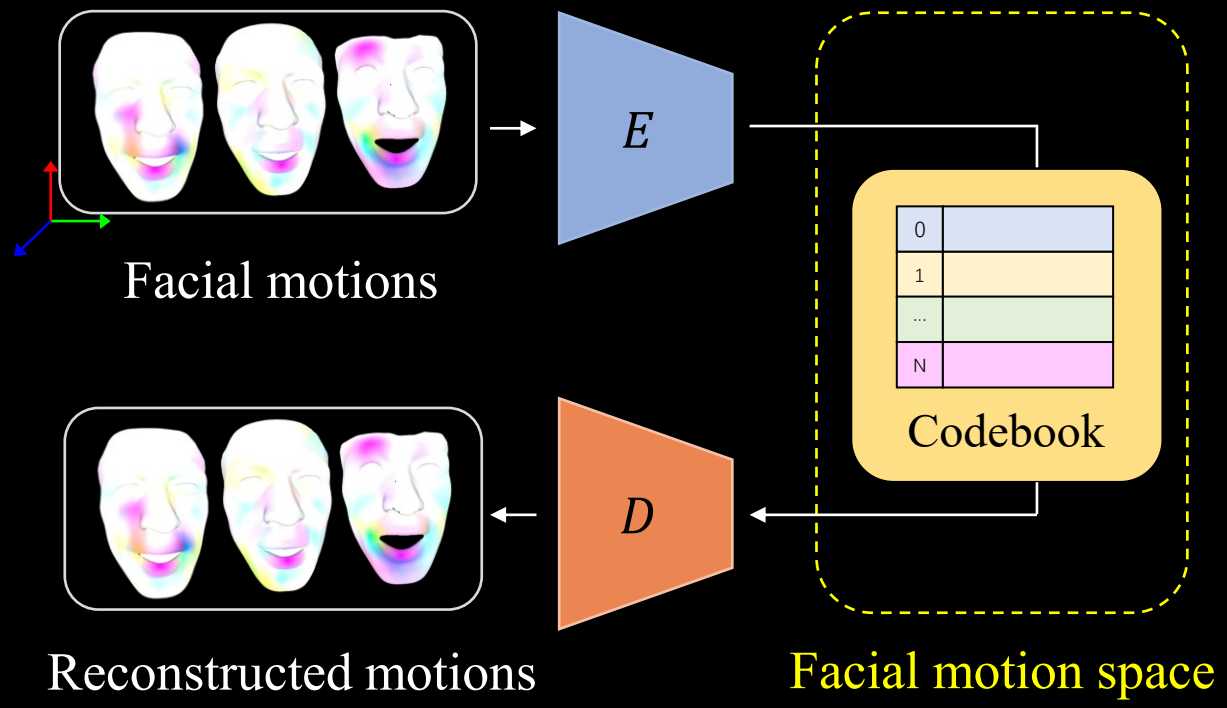
Jinbo Xing[1] Menghan Xia[2] Yuechen Zhang[1] Xiaodong Cun[2] Jue Wang[2] Tien-Tsin Wong[1]

[1]The Chinese University of Hong Kong [2]Tencent AI Lab

WED-PM-041

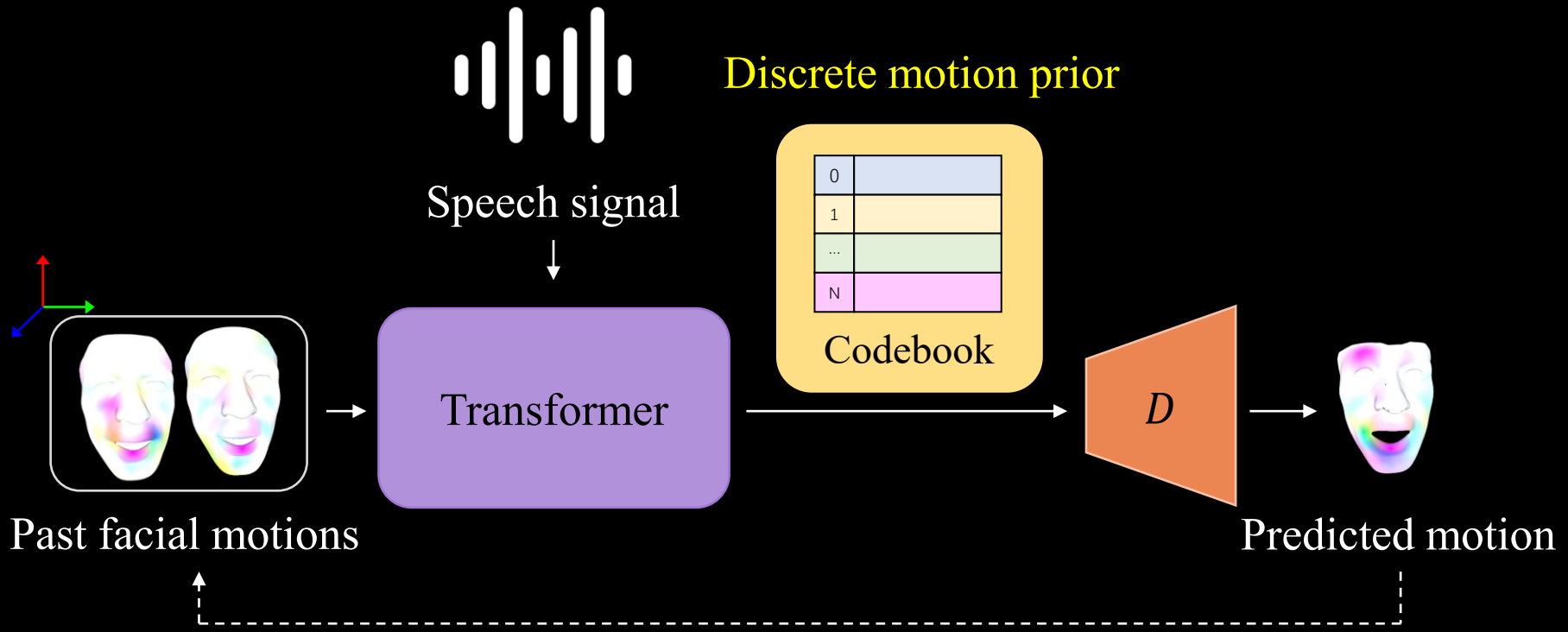
We present CodeTalker, a temporal autoregressive model over
the learned discrete facial motion space for high-quality speech-driven 3D facial animation.

Discrete Facial Motion Prior Learning

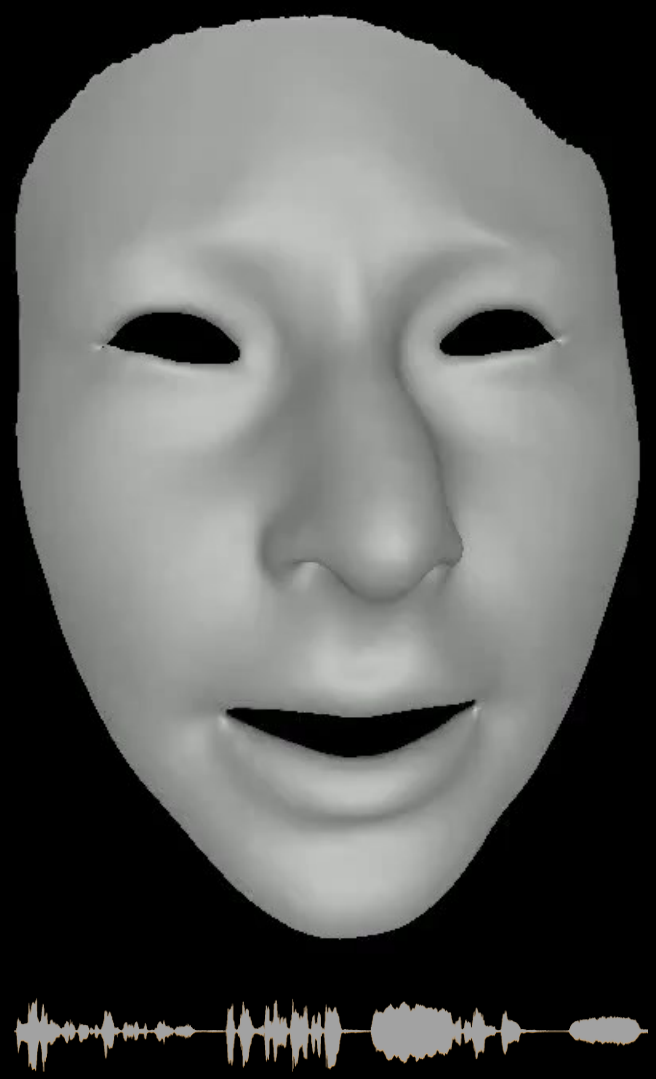


CodeTalker first learns a discrete context-rich facial motion codebook by self-reconstruction learning over real facial motions.

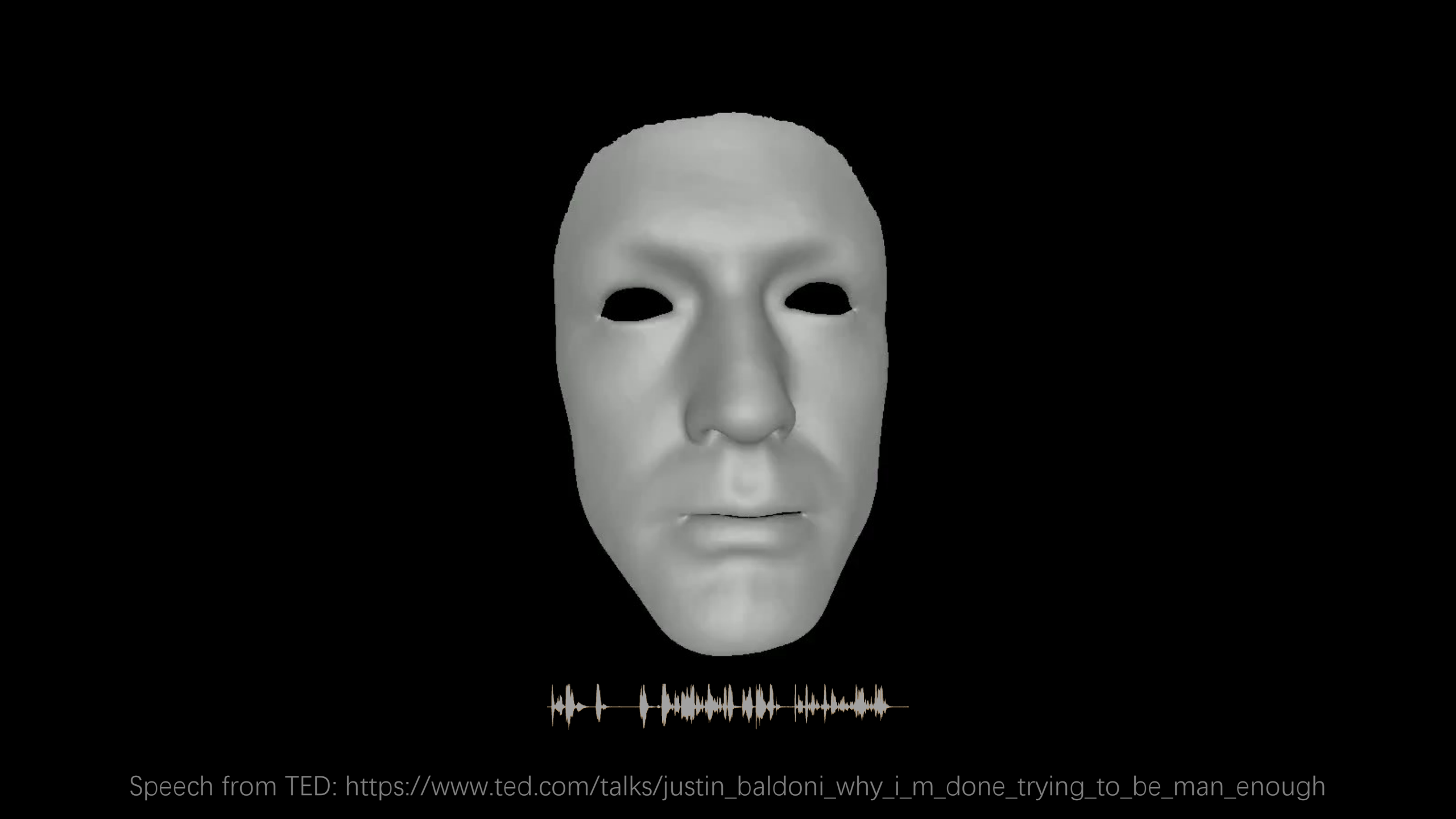
Speech-Driven Motion Synthesis



It then autoregressively synthesize facial motions through *code query* conditioned on both the speech signals and past motions.



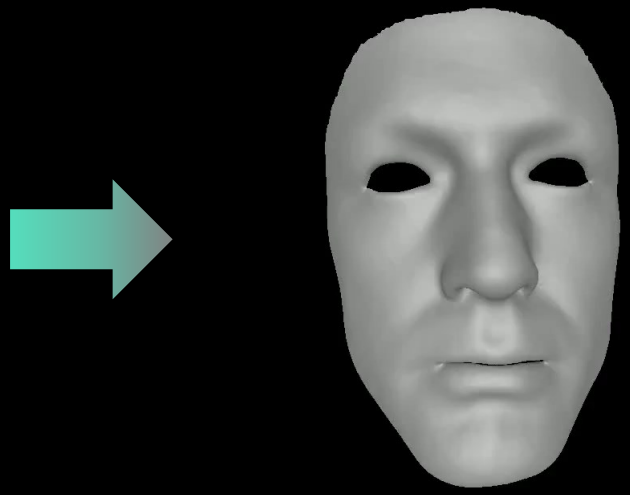
Speech from TEDx: https://www.tedxmelbourne.com/talks/how-singing-together-changes-the-brain



Speech from TED: https://www.ted.com/talks/justin_baldoni_why_i_m_done_trying_to_be_man_enough

Speech-Driven 3D Facial Animation

Speech signal



Facial animation

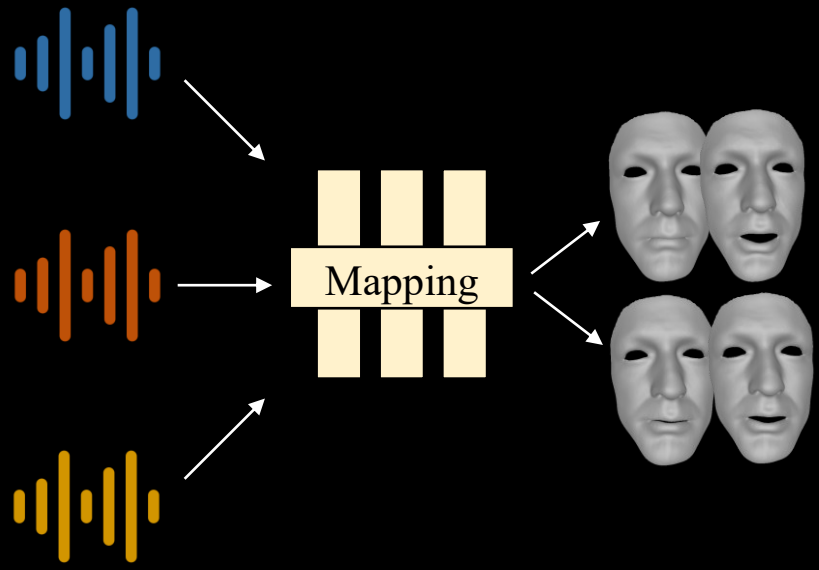
applications

Virtual Reality

Film Production

Game

Background



Linguistics-based

- Xu et al. 2013 & JALI (Edwards et al. 2016)

❌ Manual parameter tuning & limited performance
Complex procedures & Not capable for entire face

Learning-based

- Karras et al. 2017 & Richard et al. 2021

❌ Person-specific & low generality

- VOCA (Cudeiro et al. 2019)

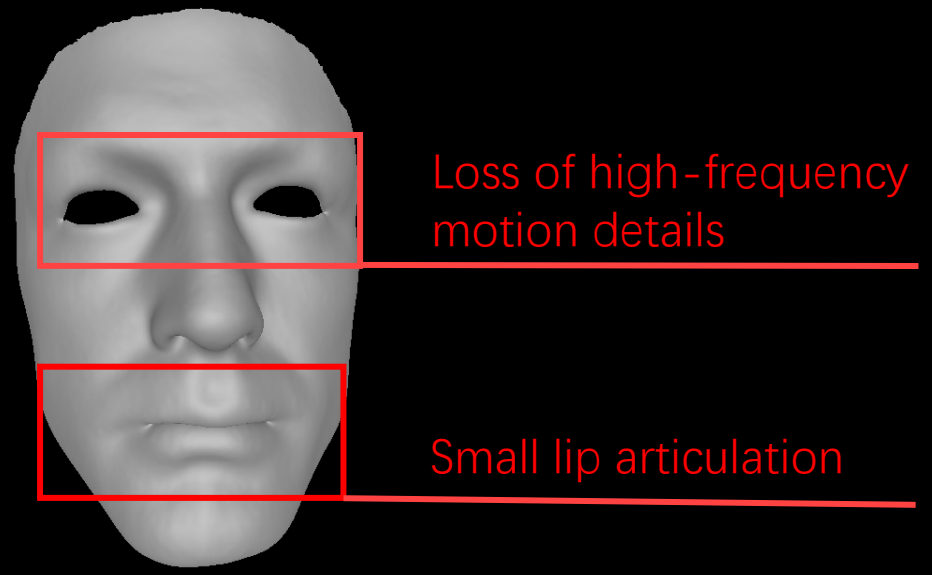
❌ Mild or static facial motions

- FaceFormer (Fan et al. 2022)

❌ Lack of subtle high-frequency motions

- MeshTalk (Richard et al. 2021)

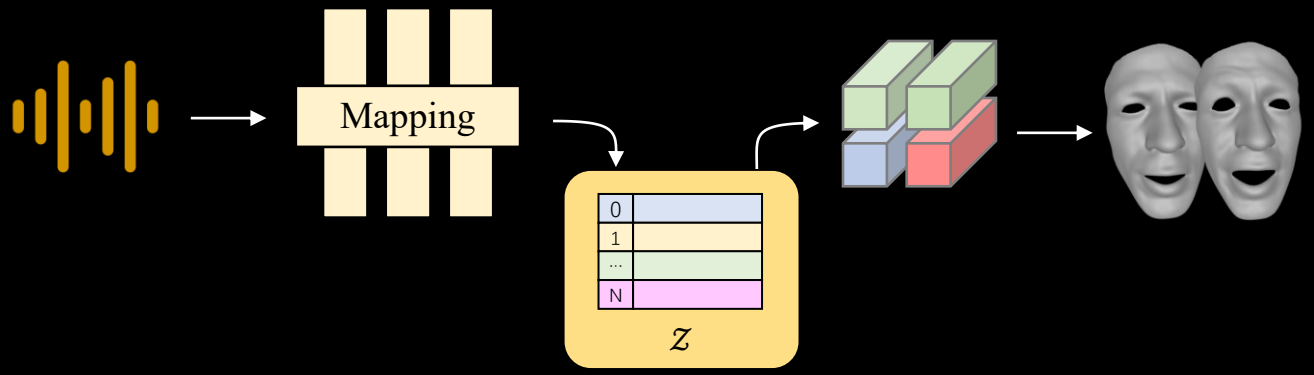
❌ Tricky training & limited expressiveness



Motivation

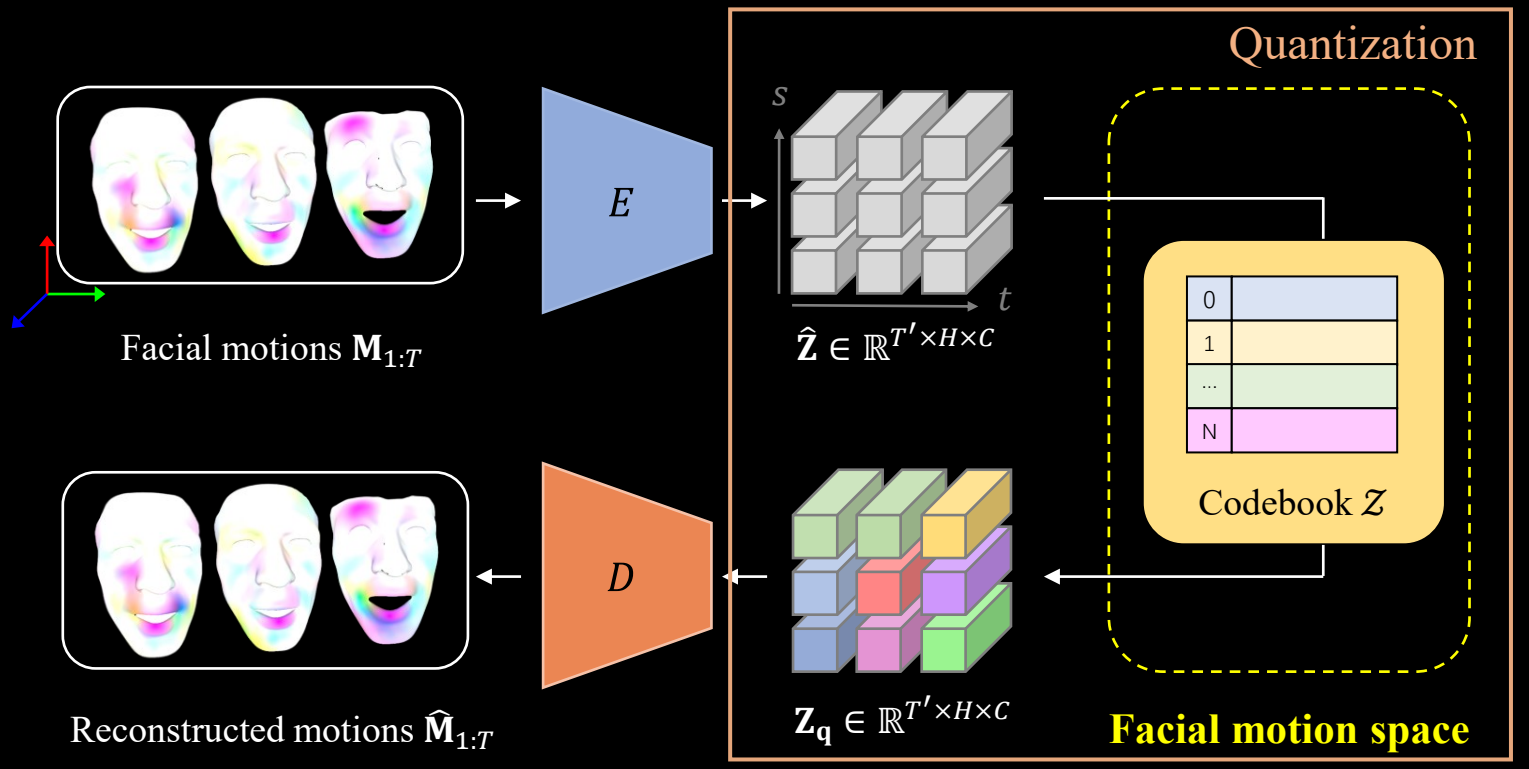
3DMM (Li et al. 2017): Low-dim. representation → Facial expressions

Speech-Driven Facial animation → **Code query** in a finite proxy space of the learned **discrete motion prior**



- ✔ Significantly reduce the uncertainty
- ✔ Complement realistic high-frequency motions

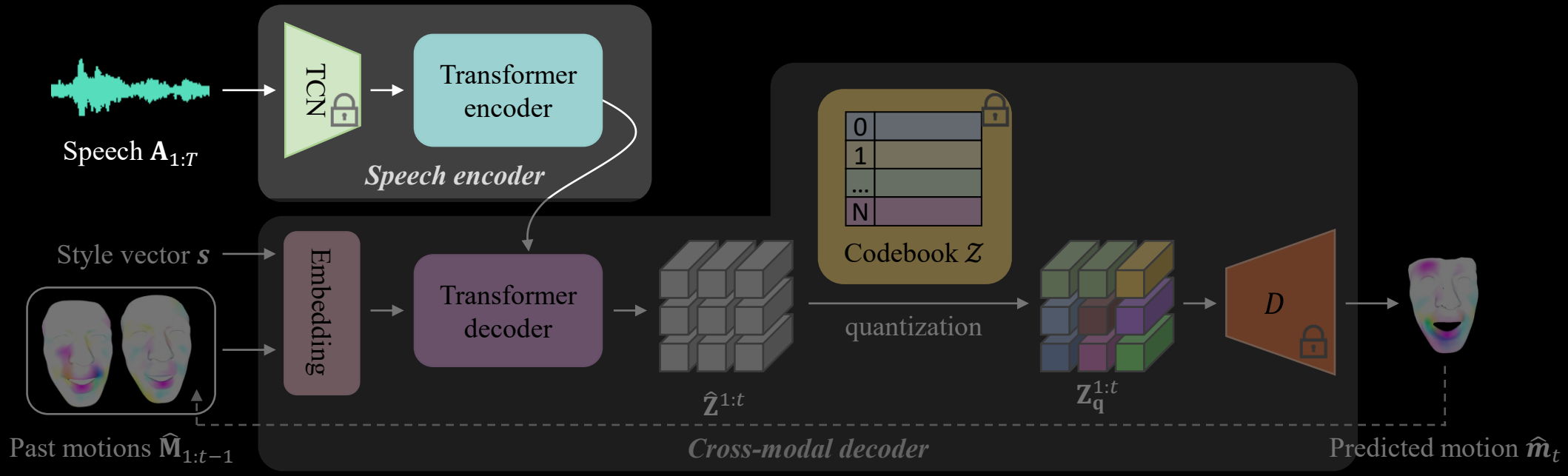
Discrete Facial Motion Space



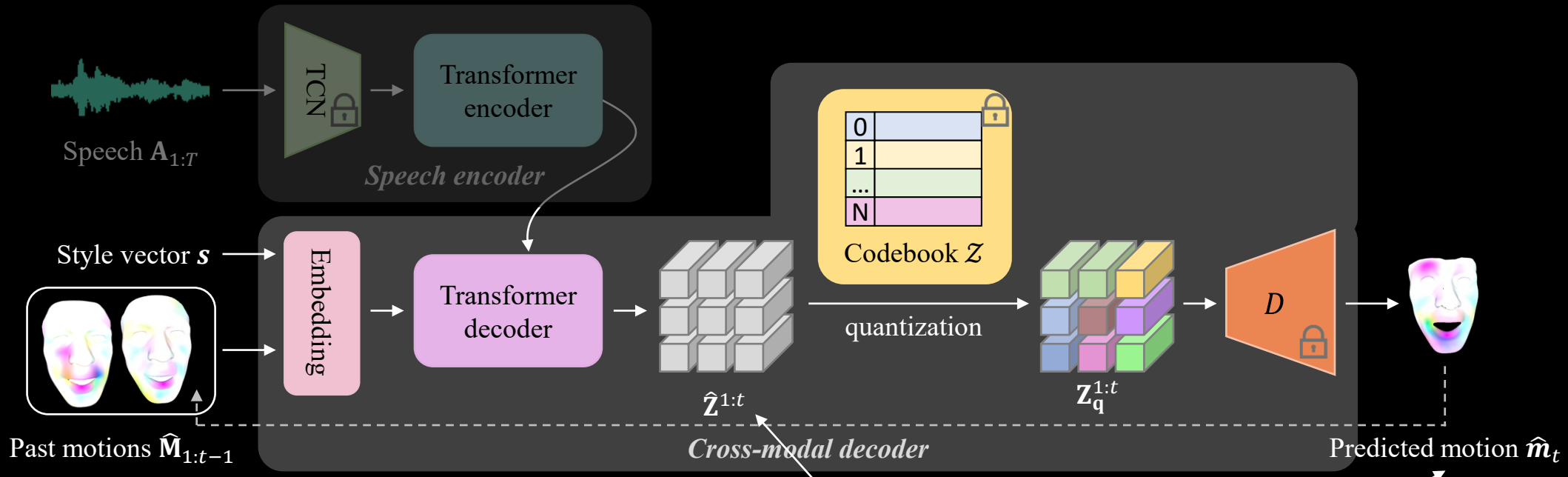
$$\mathcal{L}_{\text{VQ}} = \|\mathbf{M}_{1:T} - \hat{\mathbf{M}}_{1:T}\|_1 + \underbrace{\|\text{sg}(\hat{\mathbf{Z}}) - \mathbf{Z_q}\|_2^2 + \beta\|\hat{\mathbf{Z}} - \text{sg}(\mathbf{Z_q})\|_2^2}$$

Reconstruction — Codebook

Speech-Driven Motion Synthesis



Speech-Driven Motion Synthesis



$$\mathrm{F}_{\mathrm{emb}}^{1:t-1} = \mathcal{P}_\theta\left(\hat{\mathbf{M}}_{1:t-1}\right) + \mathbf{B} \cdot \frac{\mathbf{s}}{\|\mathbf{s}\|_1}$$

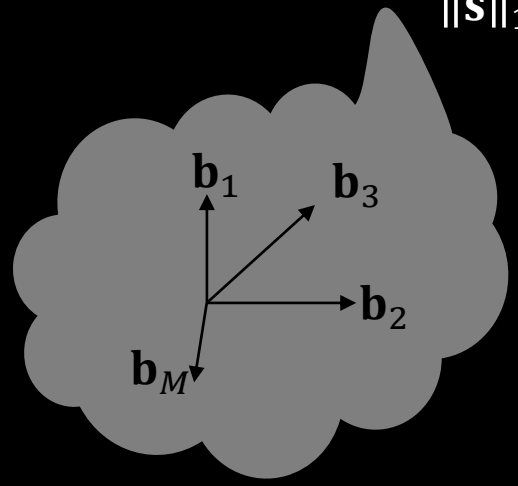
$$\mathcal{L}_{\mathrm{syn}} = \left\|\hat{\mathbf{Z}}^{1:T} - \mathrm{sg}\left(\mathbf{Z}_{\mathbf{q}}^{1:T}\right)\right\|_2^2 + \left\|\mathbf{M}_{1:T} - \hat{\mathbf{M}}_{1:T}\right\|_2^2$$

Feature regularity

Motion

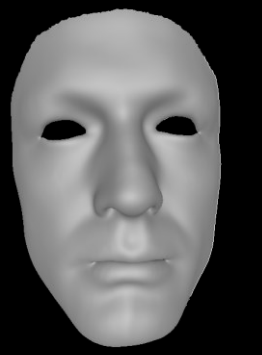
Style Embedding Space

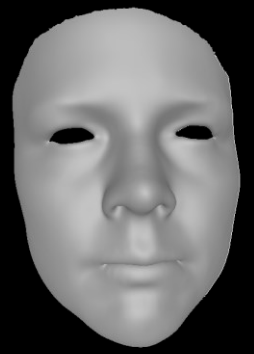
$$\mathbf{B} \cdot \frac{\mathbf{s}}{\|\mathbf{s}\|_1} := \mathbf{e}_i$$



Style embedding space

Linearly spanned by M learned basis vectors



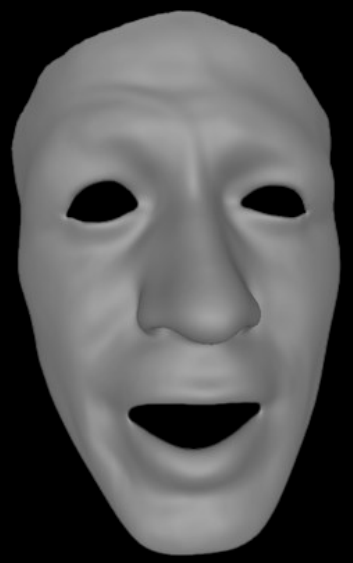


...

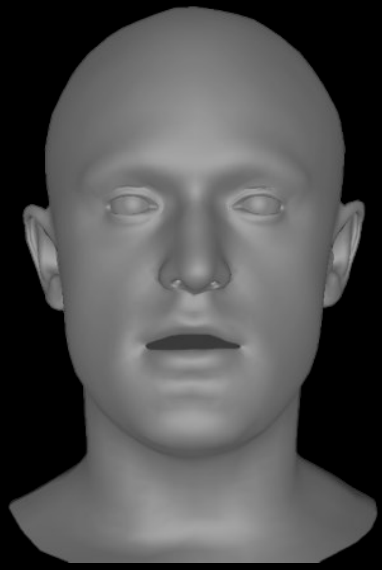
Dataset

Specification	BIWI	VOCASET
Subjects	14	12
Sequences	560	480
Unique sentences	40	255
FPS	25	60
Avg. duration	4.67s	3~4s
Vertices	23370	5023
Language	English	

Sample frame from BIWI:

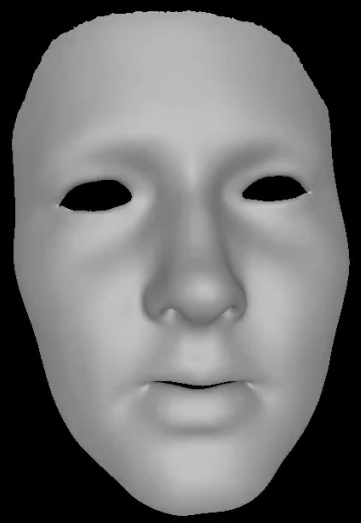


Sample frame from VOCASET:

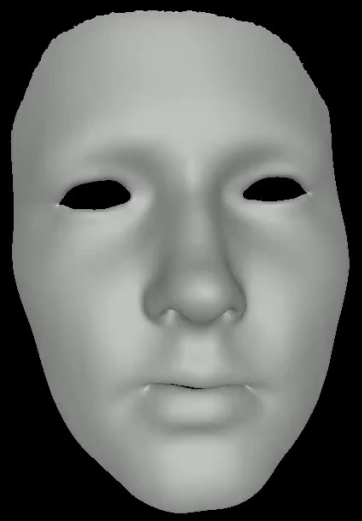


Comparison to SOTA Methods on BIWI

VOCA

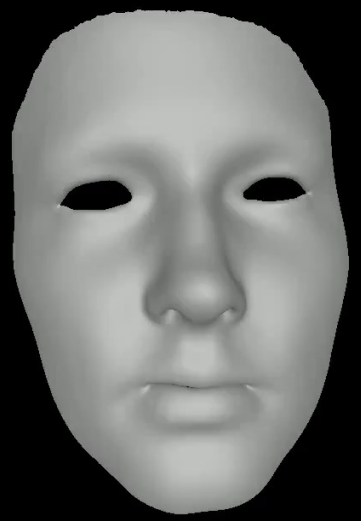


MeshTalk

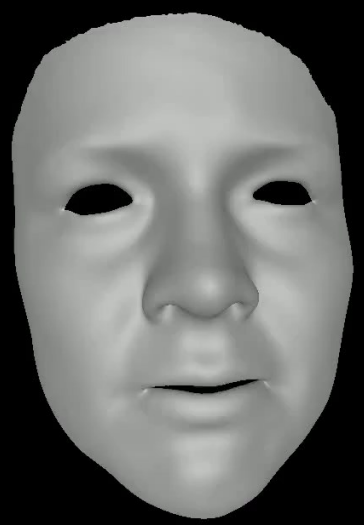


FaceFormer

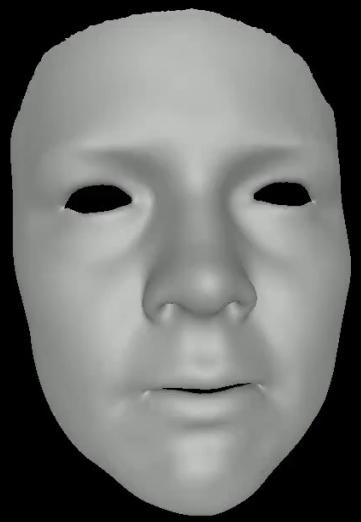
CodeTalker (Ours)



Reference



VOCA



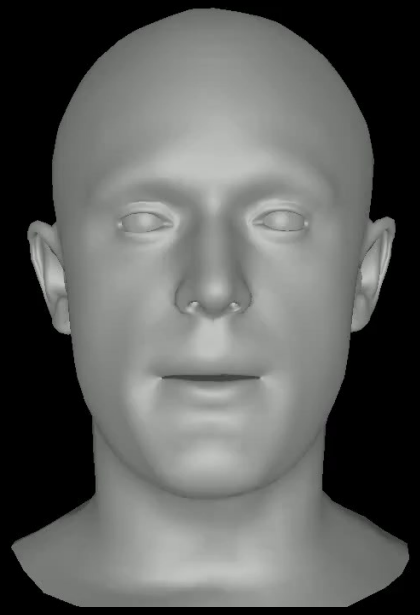
MeshTalk

FaceFormer

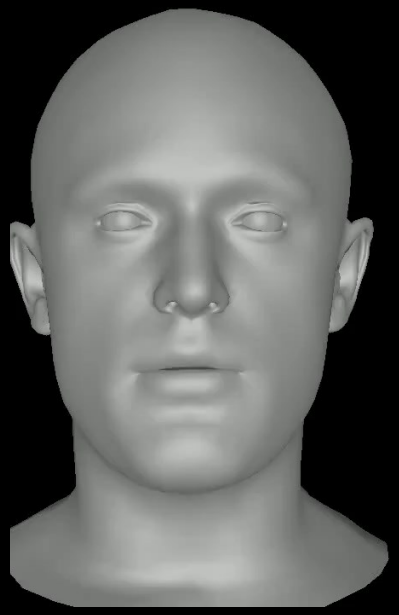
CodeTalker (Ours)

Reference

Comparison to SOTA Methods on VOCASET

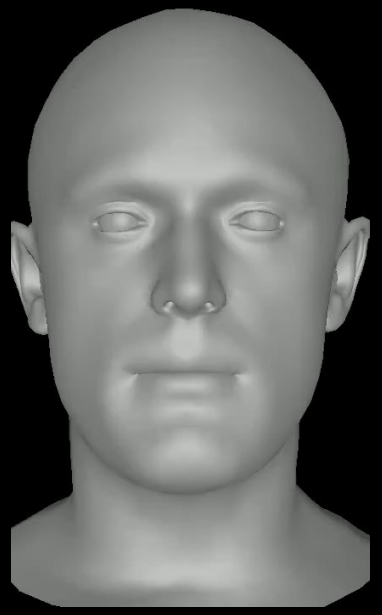


VOCA



MeshTalk

FaceFormer



CodeTalker (Ours)

Reference

VOCA

MeshTalk

FaceFormer

CodeTalker (Ours)

Reference

Comparison to Previous Methods

Karras et al.
SIGGRAPH'17

Taylor et al.
SIGGRAPH'17

VOCA
CVPR'19

MeshTalk
ICCV'21

FaceFormer
CVPR'22

CodeTalker
(Ours)

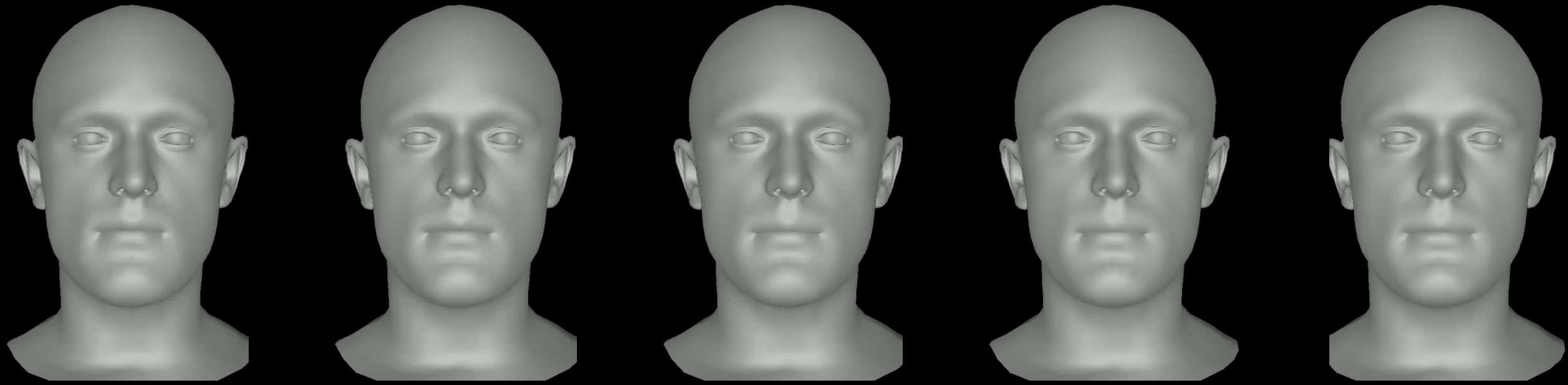
Materials of previous methods from the supplementary video of FaceFormer, CVPR'22

Talking Style Interpolation

Style 1
Large mouth amplitude

Style 1
Large mouth amplitude

Style 2
Small mouth amplitude



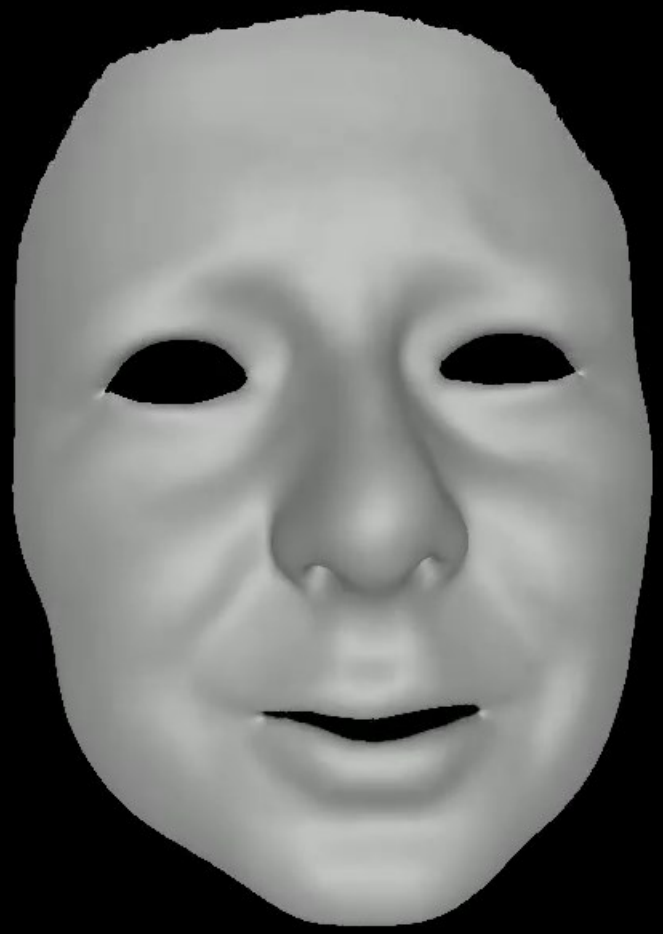
Style 1
Large mouth amplitude

Interpolation

Style 2
Small mouth amplitude

Different Languages

Spanish



Japanese


JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

CodeTalker: Speech-Driven 3D Facial Animation with Discrete Motion Prior

Jinbo Xing Menghan Xia Yuechen Zhang Xiaodong Cun Jue Wang Tien-Tsin Wong

Thanks for watching!

Project & Video

Code