MemoryTalker: Personalized Speech-Driven 3D Facial Animation via Audio-Guided Stylization

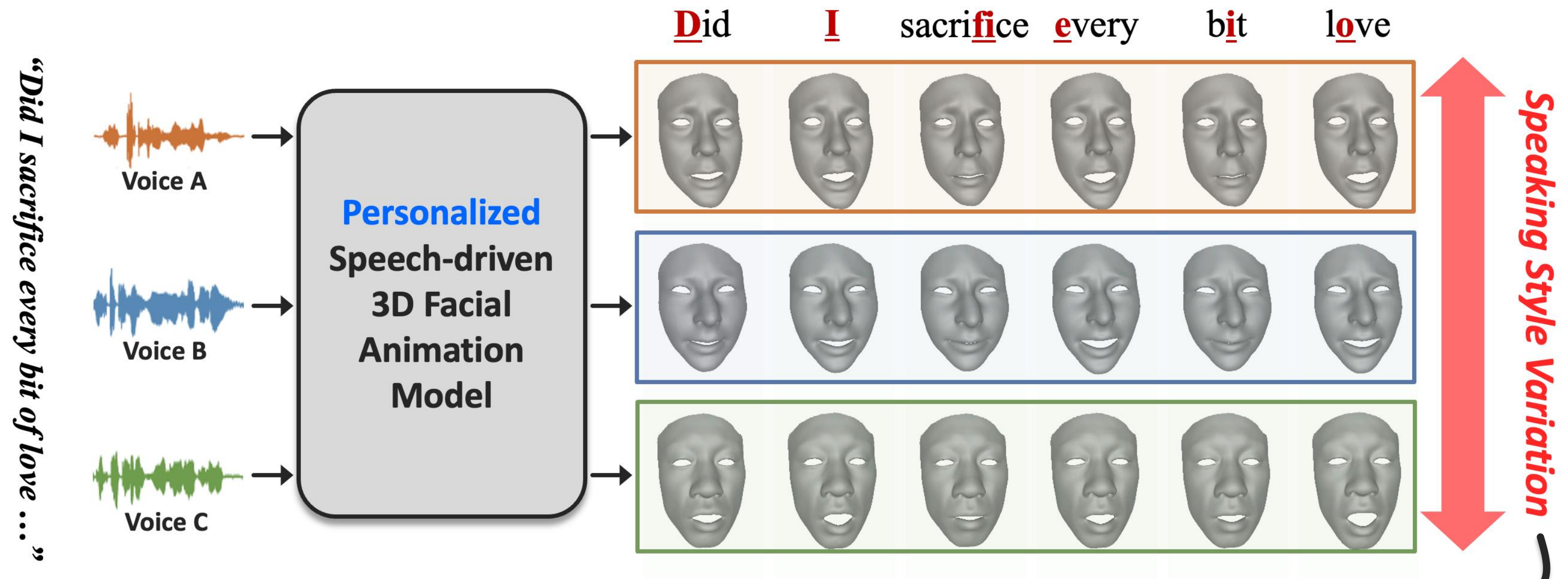
Hyung Kyu Kim[1] Sangmin Lee[2,†] Hak Gu Kim[1,†]

[1]Chung-Ang University [2]Korea University

ICCV HONOLULU HAWAII OCT 19-23, 2025

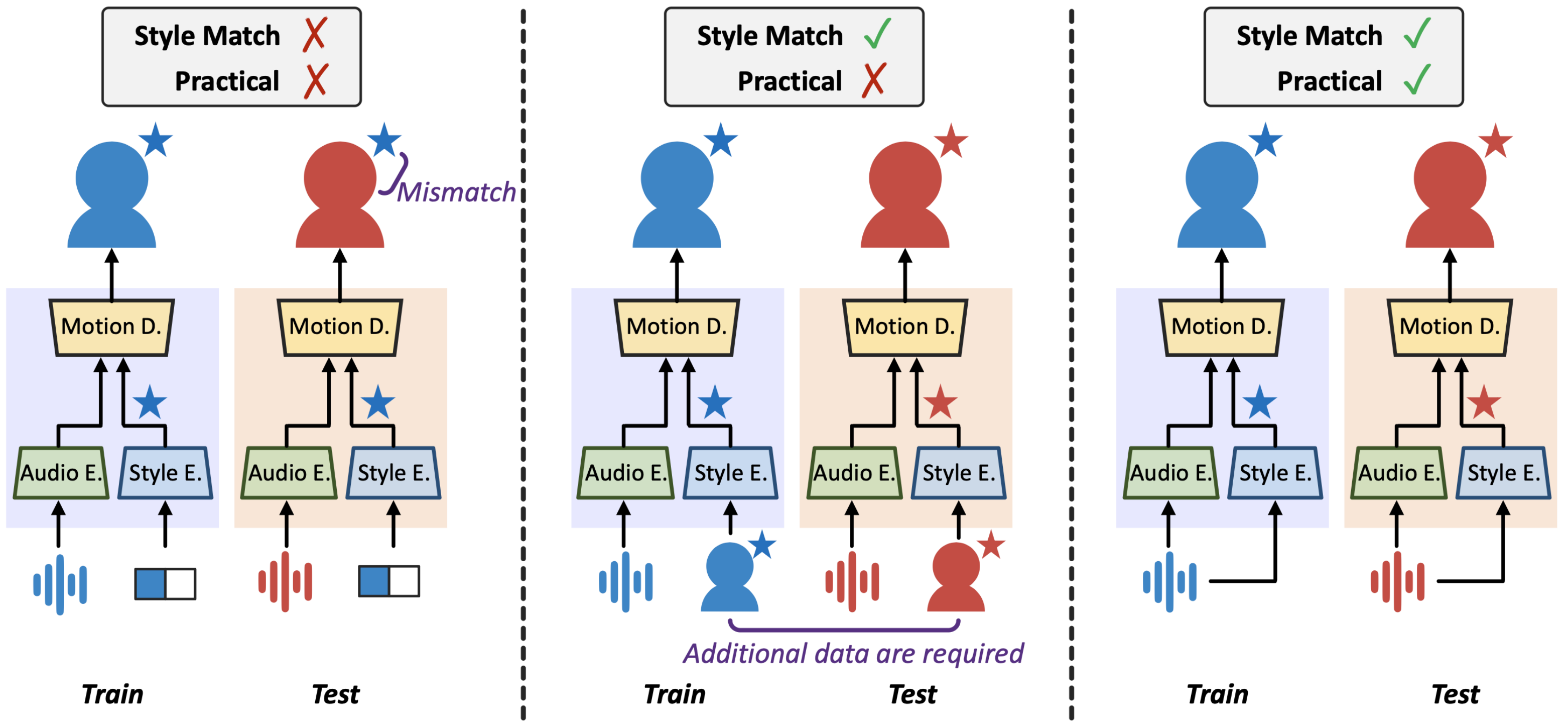
Introduction

- Speech-Driven 3D Facial Animation
  - **Synthesizing realistic 3D facial motion** sequences from given speech signals
- What is Speaking Style for *Personalization*?
  - Even for the same word, speakers differ **in lip shape**, **mouth opening**, and **lip protrusion** (*i.e.*, Speaking style)


"Did I sacrifice every bit of love …"
Voice A
Voice B
Voice C
Personalized Speech-driven 3D Facial Animation Model
Did
I
sacrifice
every
bit
love
Speaking Style Variation

Challenging due to diverse speaking styles

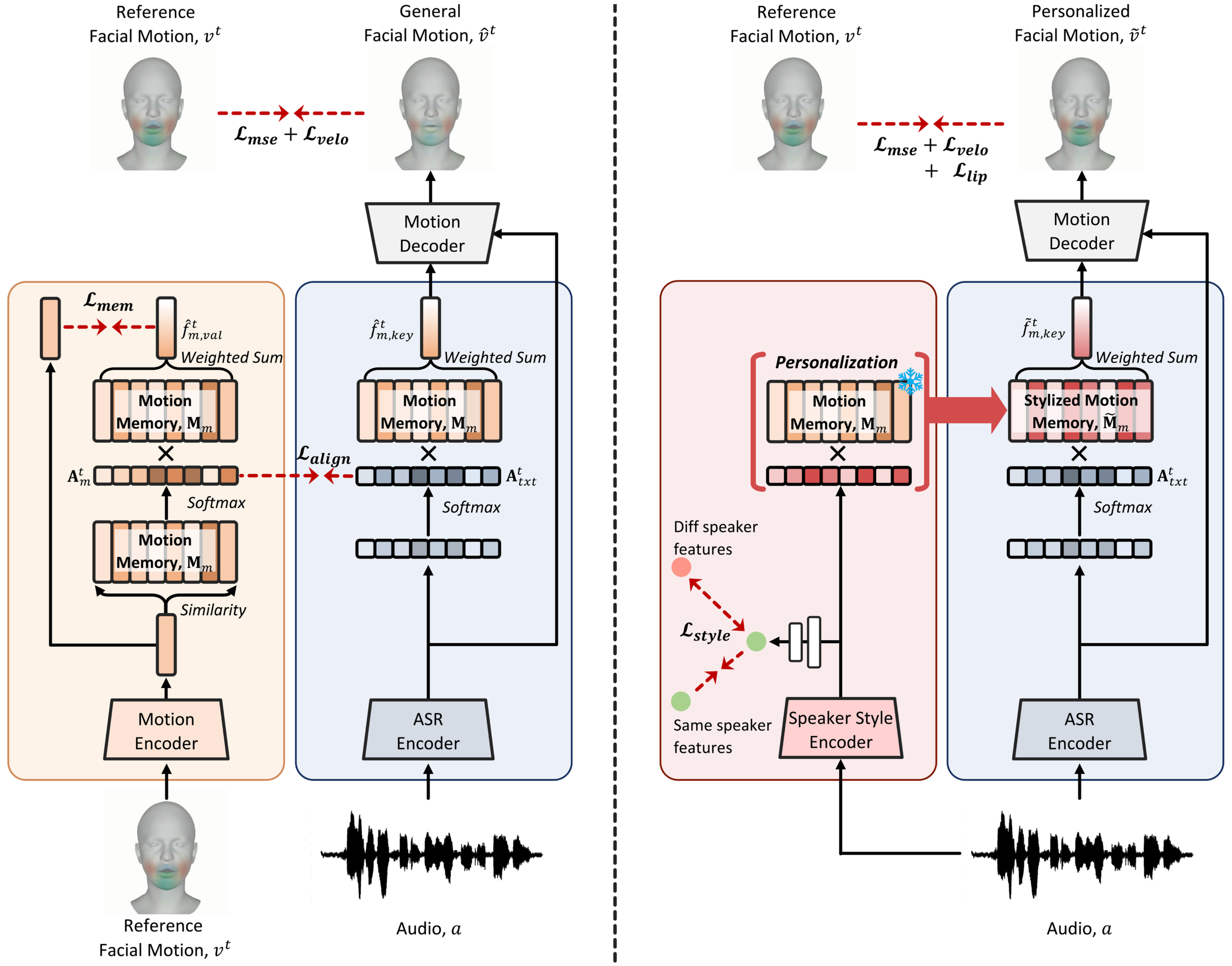
Limitations in Previous Works



(a) Using one-hot encoding (b) Using 3D facial meshes (c) Ours

- (a) **One-hot**: Fails to generalize unseen speakers
- (b) **Mesh**: Requires additional 3D data → Impractical
- (c) **Ours**: Uses only audio, no additional priors

Proposed Method



(a) 1-Stage: Storing and Recalling Facial Motion (b) 2-Stage: Generating Personalized Facial Motion

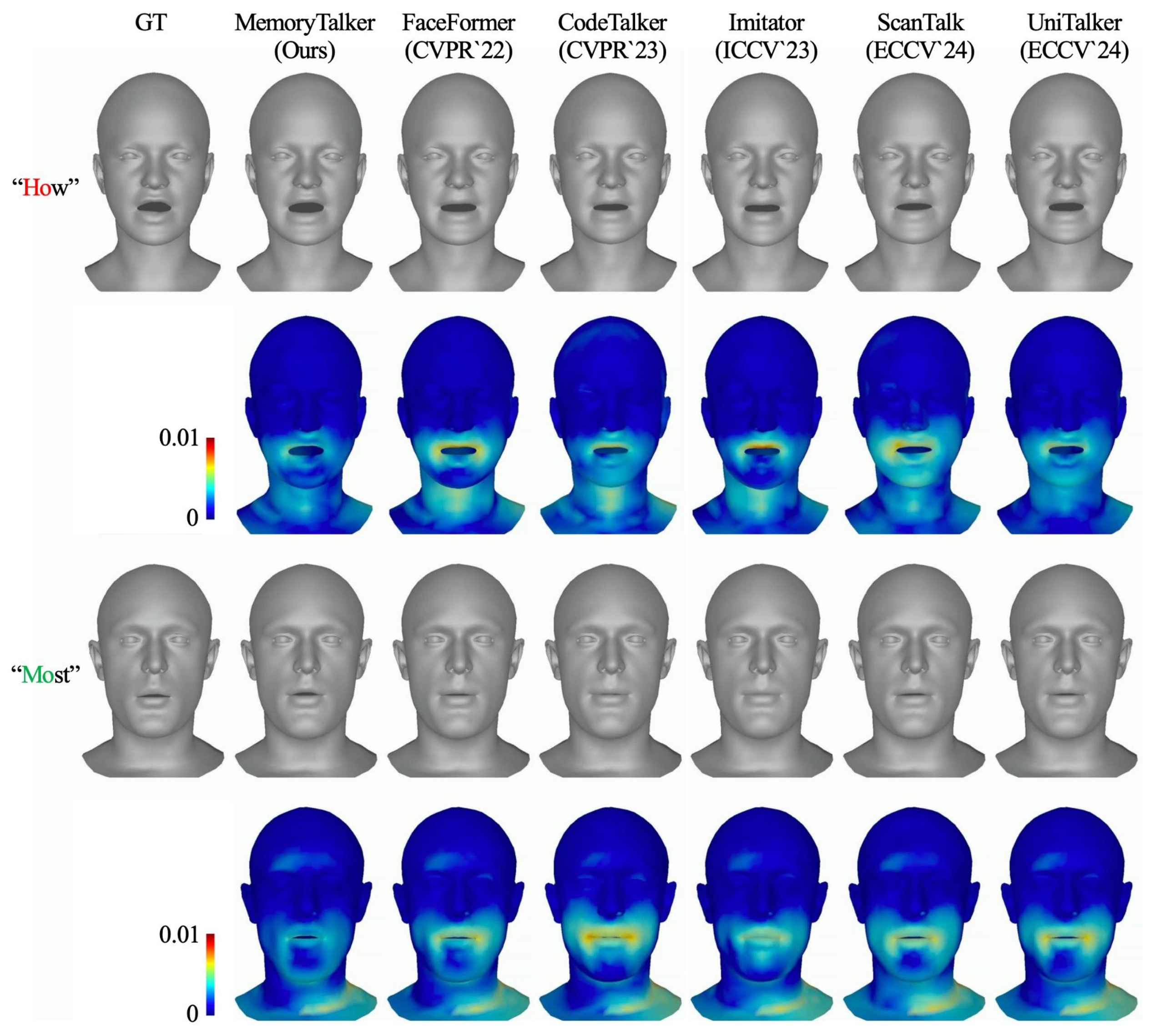
- **1-Stage (*Memorizing*)**: Store general facial motions aligned with phonemes into a motion memory
- **2-Stage (*Animating*)**: Stylize the memory with audio-driven speaking styles for generating personalized motions
- This two-stage design enables motion generation that is phonetically accurate while reflecting speaking styles
- As a result, our model generates realistic and personalized 3D facial animation without one-hot identities and additional 3D data

Experimental Results

Quantitative Evaluation on VOCASET

Method	VOCASET [6] FVE ↓ ($\times10^{-6}$)	LVE ↓ ($\times10^{-5}$)	FID ↓ ($\times10^{-1}$)	LDTW ↓ ($\times10^{-5}$)	Lip-max ↓ ($\times10^{-4}$)
FaceFormer [12]	0.639	0.413	3.583	0.507	0.452
CodeTalker [45]	0.721	0.498	3.713	0.554	0.484
SelfTalk [33]	0.593	0.382	3.279	0.475	0.416
Imitator [41]	0.686	0.456	3.918	0.554	0.472
ScanTalk [31]	0.609	0.375	3.623	0.457	0.420
UniTalker [11]	0.570	0.382	3.256	0.507	0.407
MemoryTalker	**0.506**	**0.293**	**3.045**	**0.418**	**0.331**

Qualitative Evaluation on VOCASET

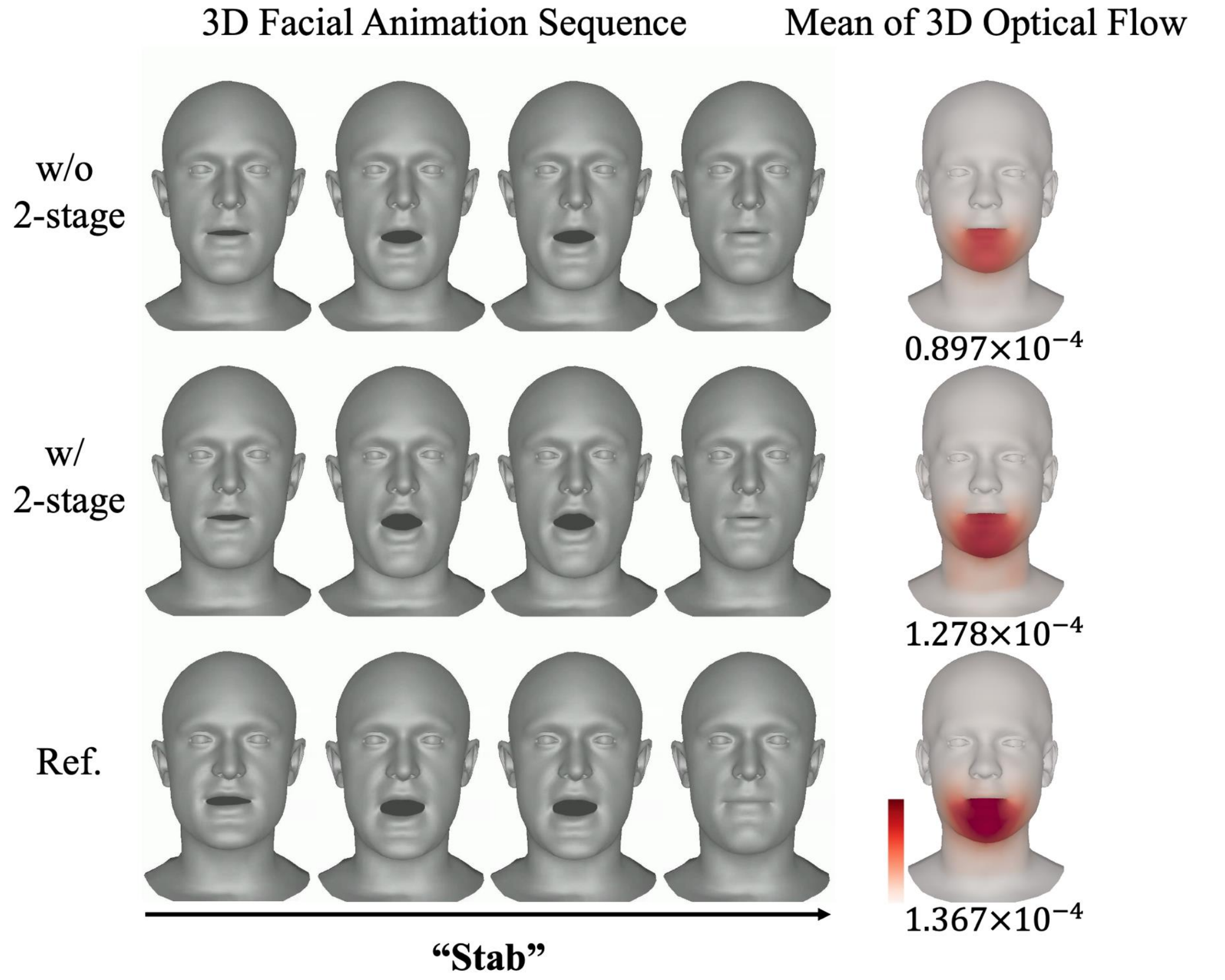


(a) VOCASET

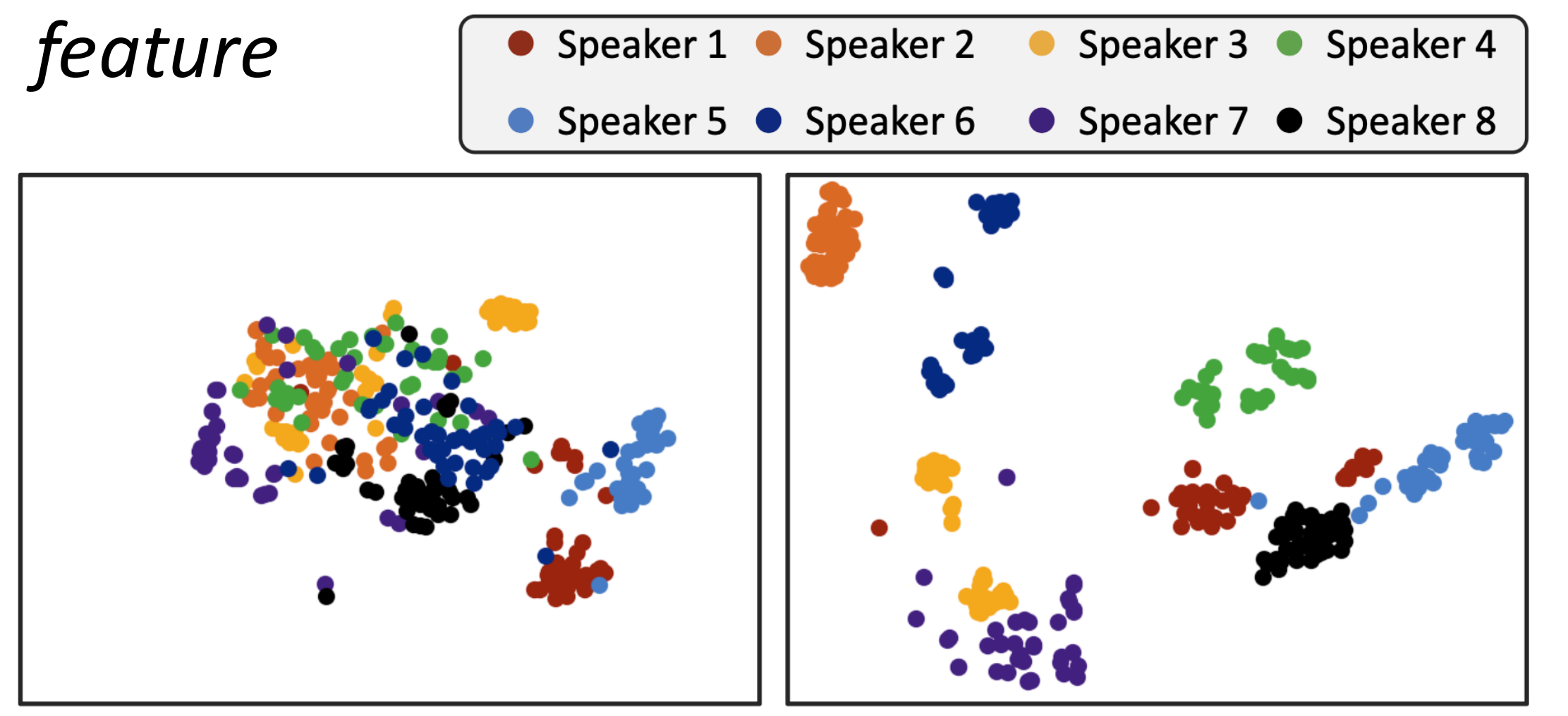
Quantitative Ablation study (1- / 2-stage)

Proposed 1-stage training	Proposed 2-stage training	FVE ↓ ($\times10^{-6}$)	LVE ↓ ($\times10^{-5}$)
✗	✗	0.638	0.460
✓	✗	0.531	0.313
✓	✓	**0.506**	**0.293**

Qualitative Ablation study (1- / 2-stage)



The t-SNE Visualization of recalled motion feature



(a) MemoryTalker (w/o 2-stage) (b) MemoryTalker (w/ 2-stage)

Conclusions

- Our *MemoryTalker* achieves superior personalization **from audio alone**
- This work enables practical 3D facial animation **for VR and the Metaverse**