Imperial College London

# Content and Style Aware Audio-Driven Facial Animation

Qingju Liu[1], Hyeongwoo Kim[2], Gaurav Bharaj[1]

[1]Flawless AI, UK          [2]Imperial College London, UK

**Motivation**: An Audio-Driven Facial Animation (ADFA) method which enables content and style reenactment.

**Challenges**: To model the diverse styles disentangled from content with limited data.
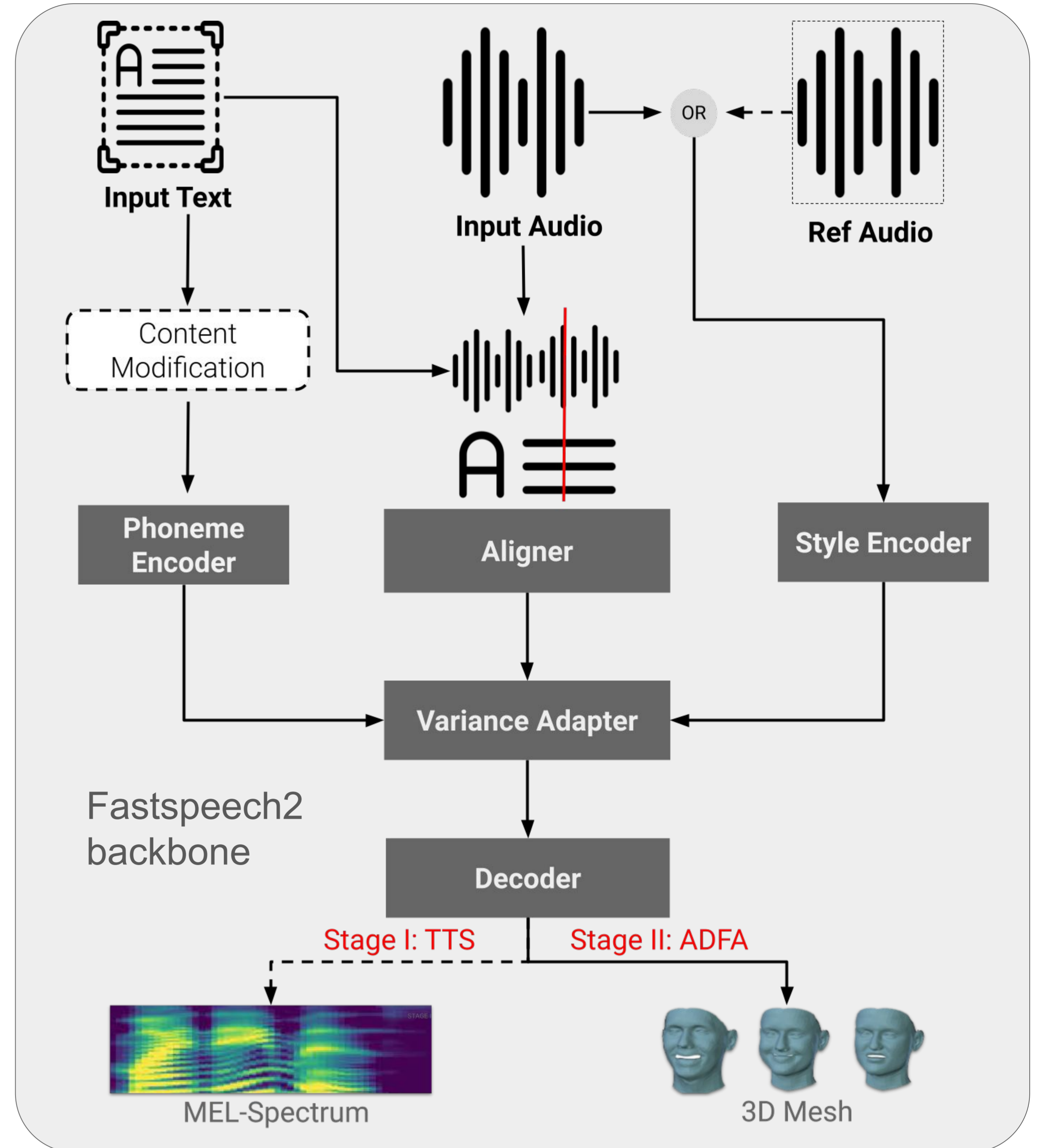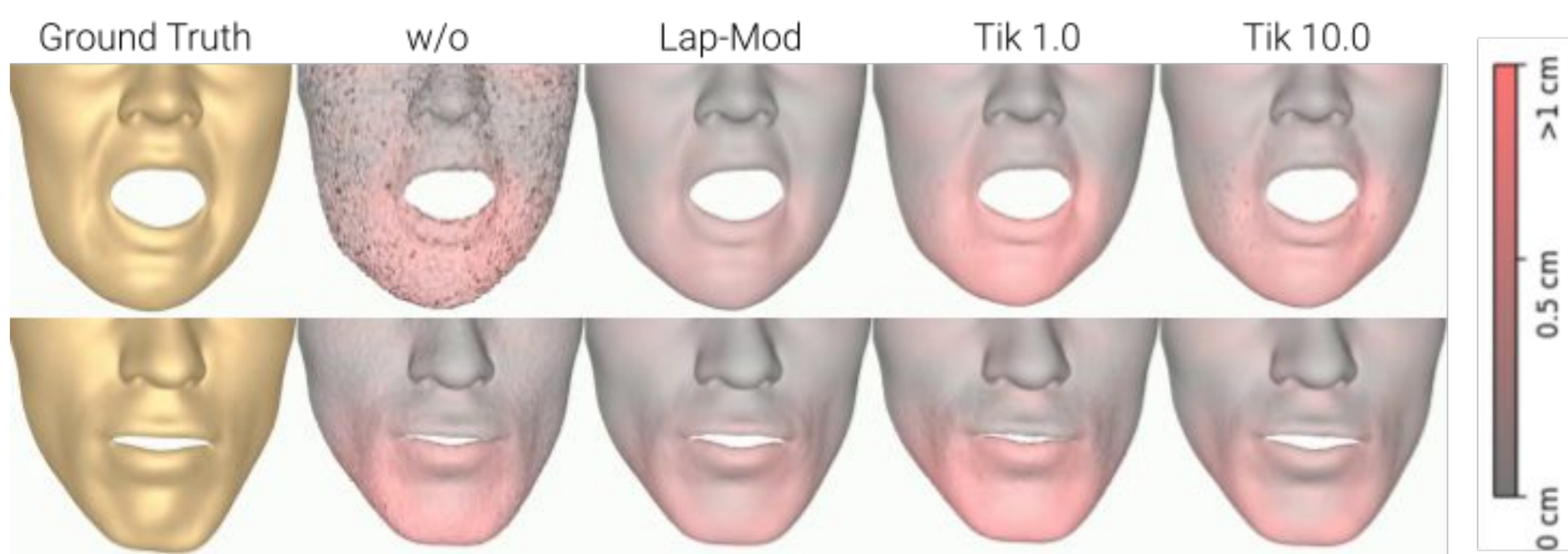
**Novelties**:

  Knowledge transfer from TTS (speech synthesis) to ADFA with two-stage training.

  Styles evolve from audio styles (how does it sound) to visual style (how do articulations look like).
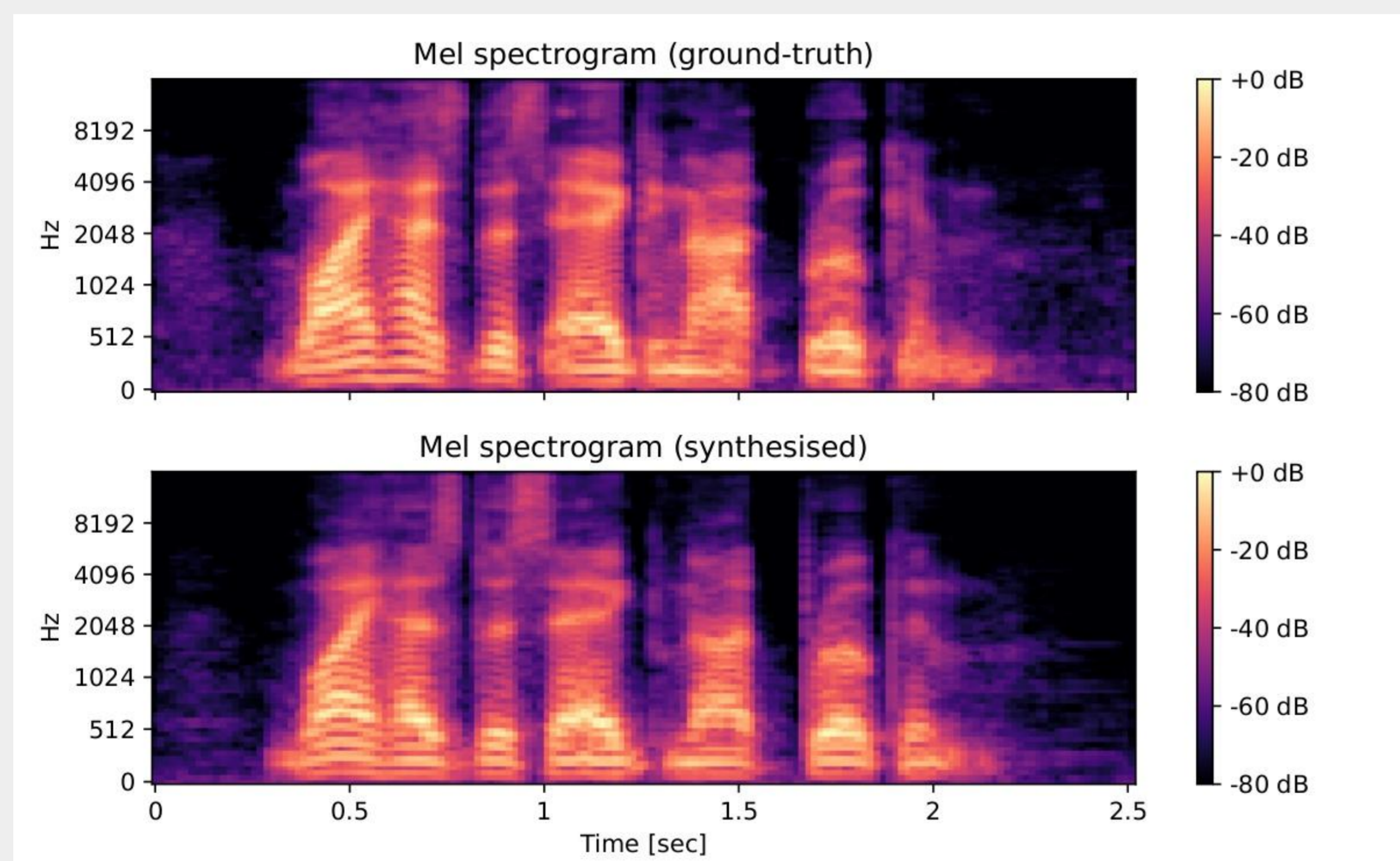
  Modified Laplacian loss to reduce computational complexity.

  Allowing both audio style transfer and content modification.
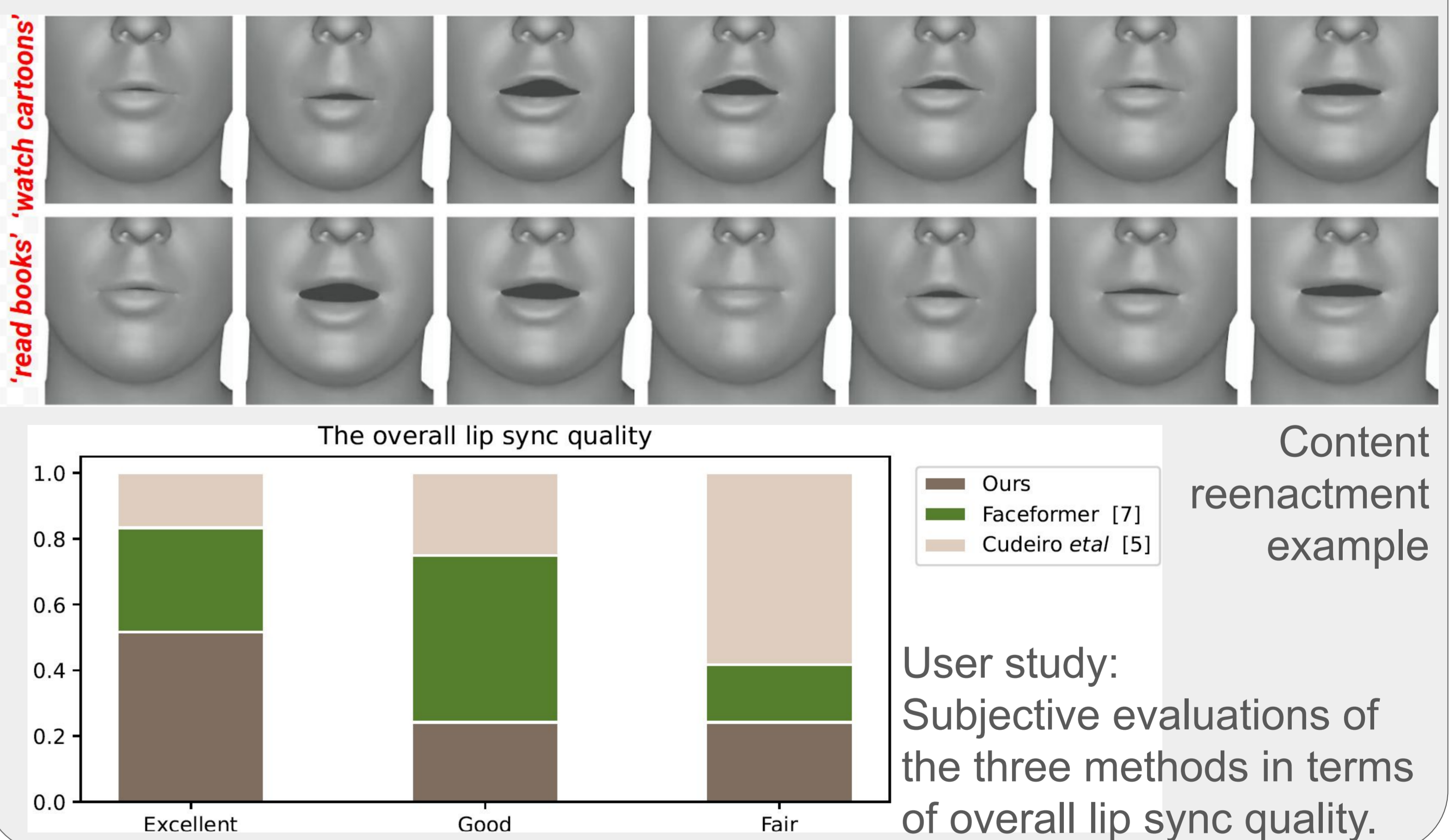
$$\mathcal{L}_{stageII} = \mathcal{L}_{geometry} + \mathcal{L}_{temporal} + \lambda \mathcal{L}_{lap-mod}$$



Ground Truth | w/o | Lap-Mod | Tik 1.0 | Tik 10.0

>1 cm / 0.5 cm / 0 cm



Fastspeech2 backbone

Input Text — Content Modification — Phoneme Encoder

Input Audio — OR — Ref Audio

Aligner — Style Encoder

Variance Adapter

Decoder

Stage I: TTS — Stage II: ADFA

MEL-Spectrum — 3D Mesh

Stage I training: Allows to synthesize natural speech—TTS.



Mel spectrogram (ground-truth)

Mel spectrogram (synthesised)

Stage II training: We can generate facial animation, modify style and content.



'read books' 'watch cartoons'

Content reenactment example

The overall lip sync quality

Ours / Faceformer [7] / Cudeiro etal [5]

Excellent / Good / Fair

User study: Subjective evaluations of the three methods in terms of overall lip sync quality.

Dataset:
Stage I: ESD + LJSpeech
Stage II: VOCAset, BIWI



VOCAset

BIWI
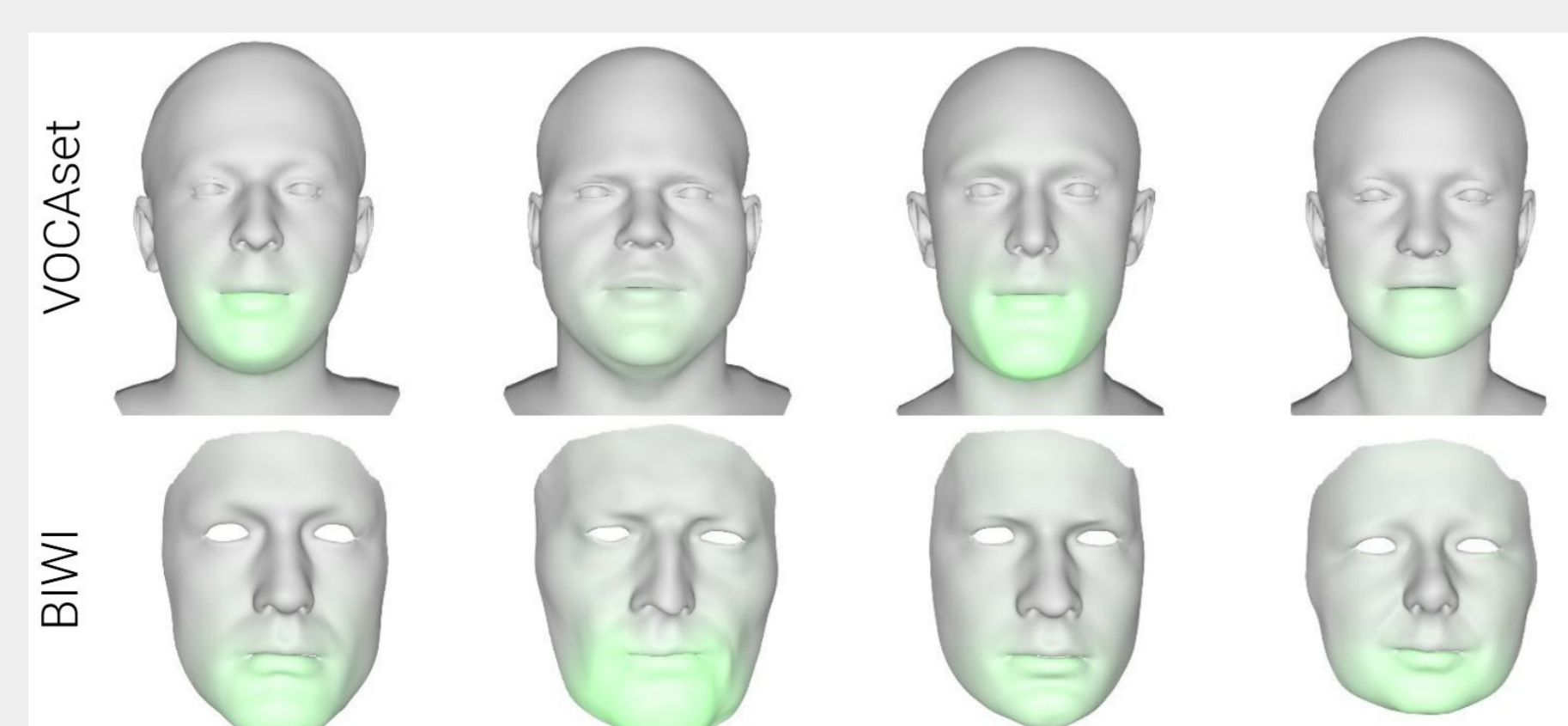
Comparisons of bilabial mouth closure for "our ex*p*eri*m*ent's *p*ositive outco*m*e"



Ground Truth

Ours

CodeTalker [50]

Cudeiro [5]

Faceformer [7]

EmoTalk [33]