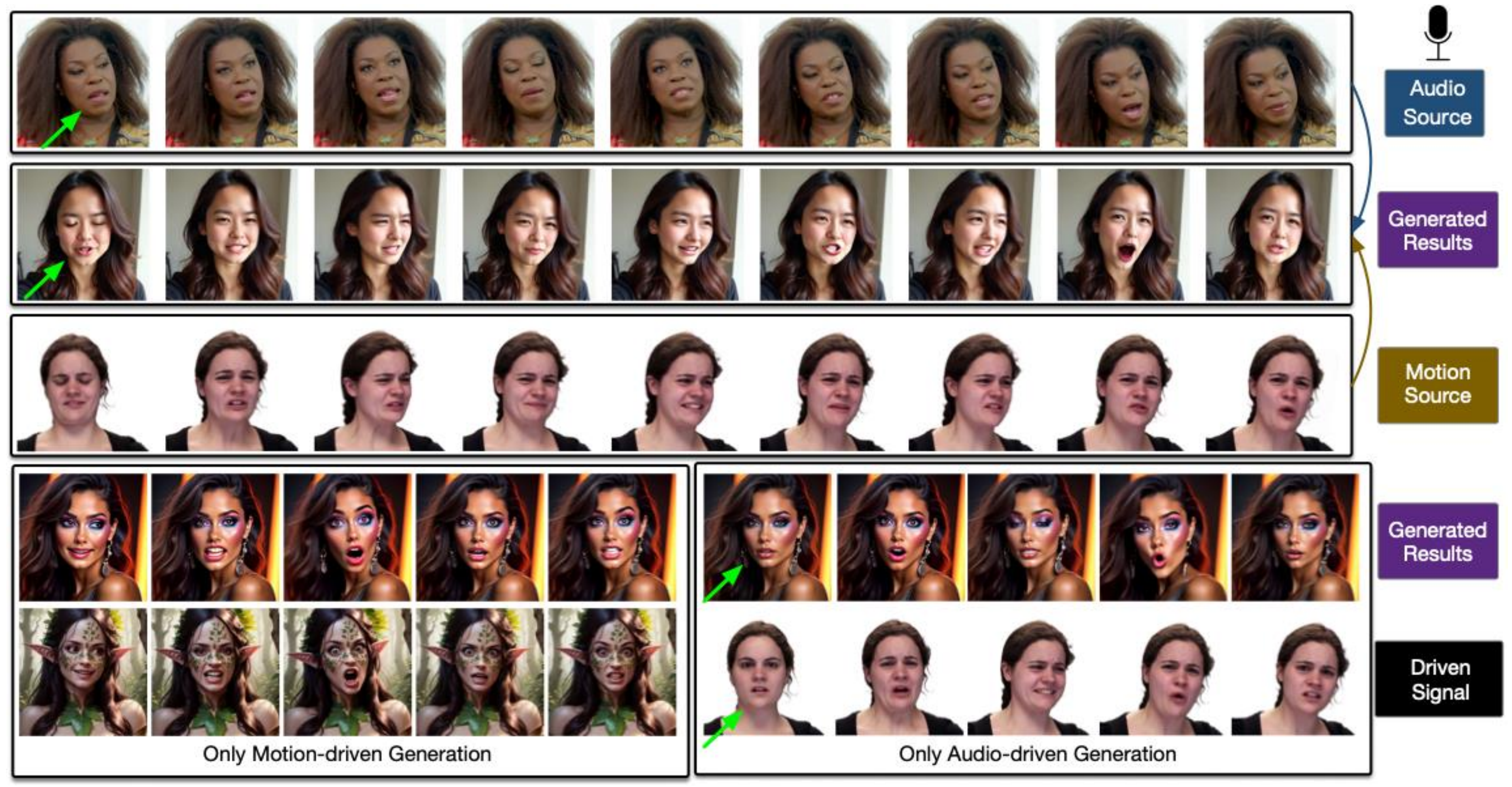
香港科技大學 THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY
腾讯混元
清華大學 Tsinghua University

Audio-visual Controlled Video Diffusion with Masked Selective State Spaces Modeling for Natural Talking Head Generation

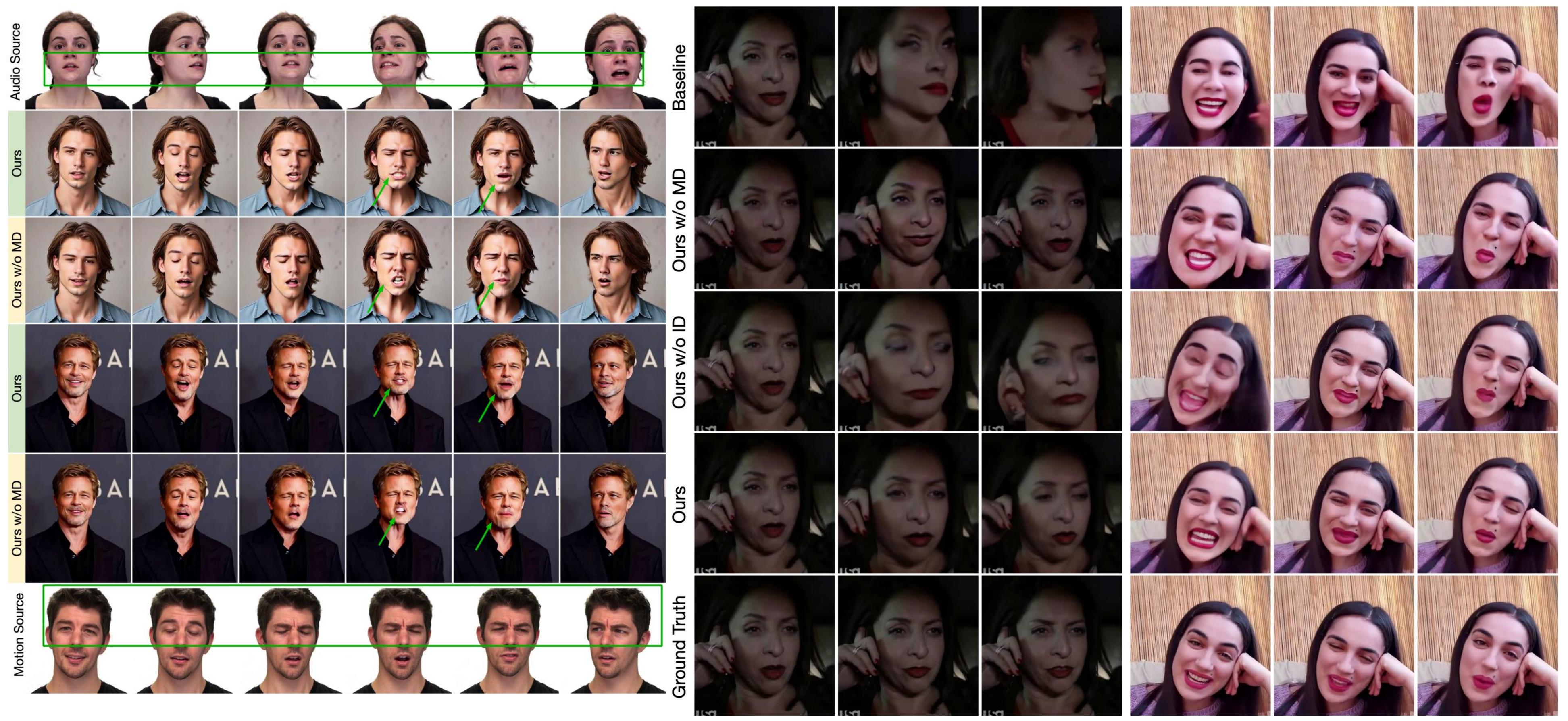
Fa-Ting Hong, Zunnan Xu, Zixiang Zhou, Jun Zhou, Xiu Li, Qin Lin, Qinglin Lu, Dan Xu

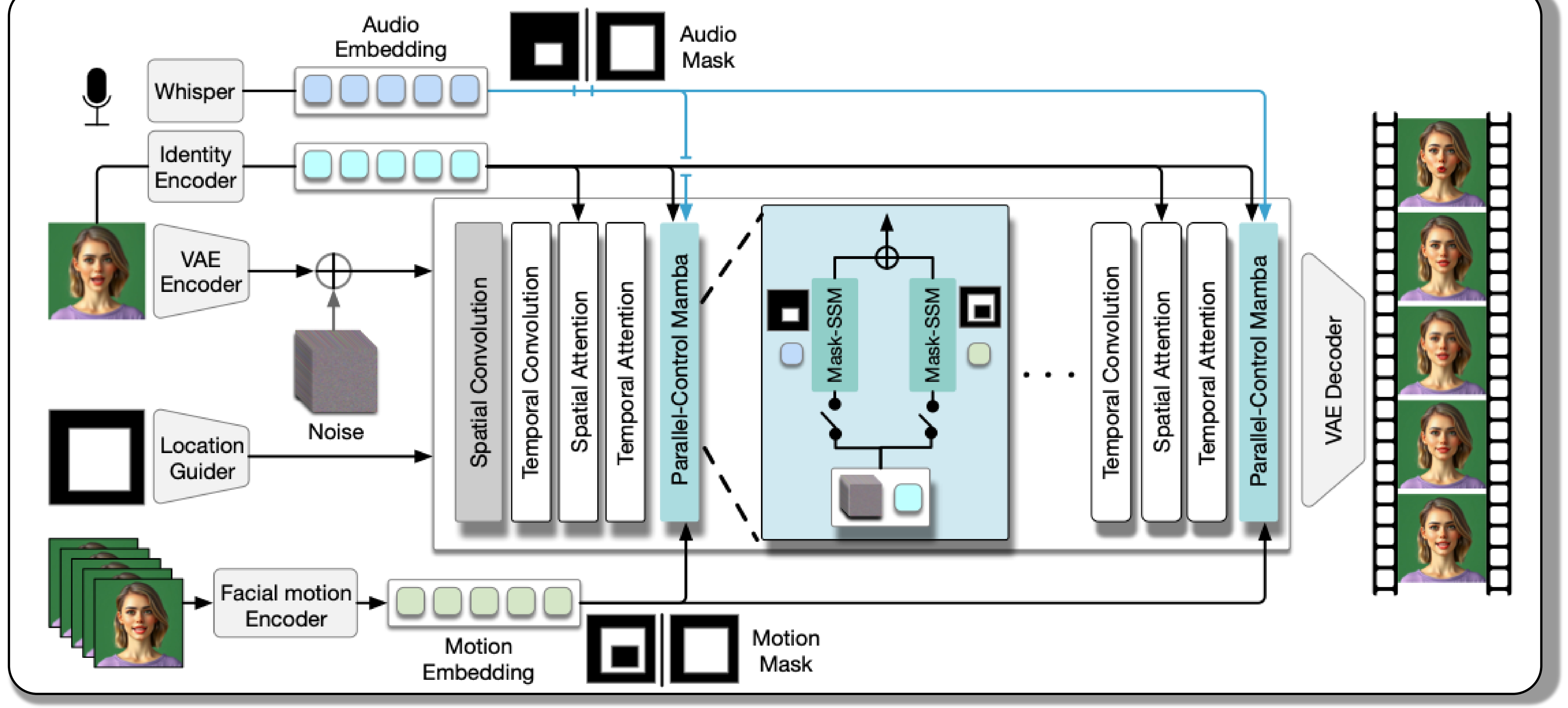
ICCV HONOLULU HAWAII OCT 19-23, 2025



Motivation:

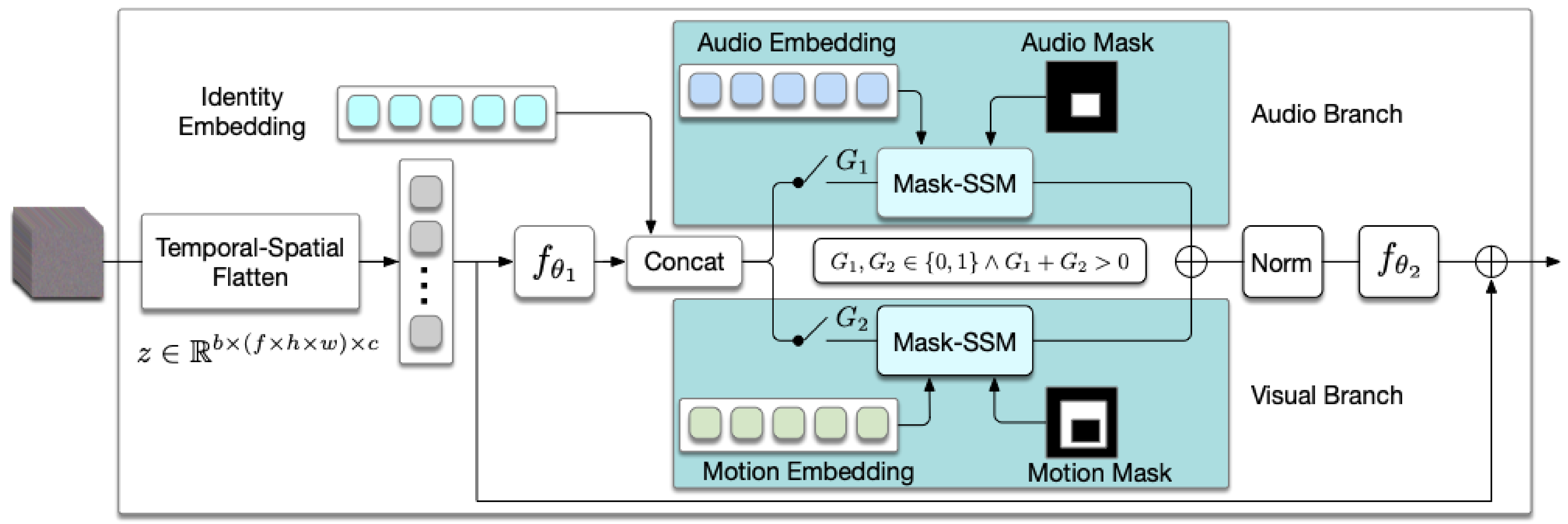
- **Overcoming Single-Modality Limitations:** Existing systems are limited to a single control input which restricts realism.
- **Resolving Conflicting Control Signals:** Combining multiple inputs often leads to conflicting and unnatural facial animations.
- **Enabling Multi-Modal Expressiveness:** We aim to seamlessly integrate audio and motion signals for natural and realistically controlled results.

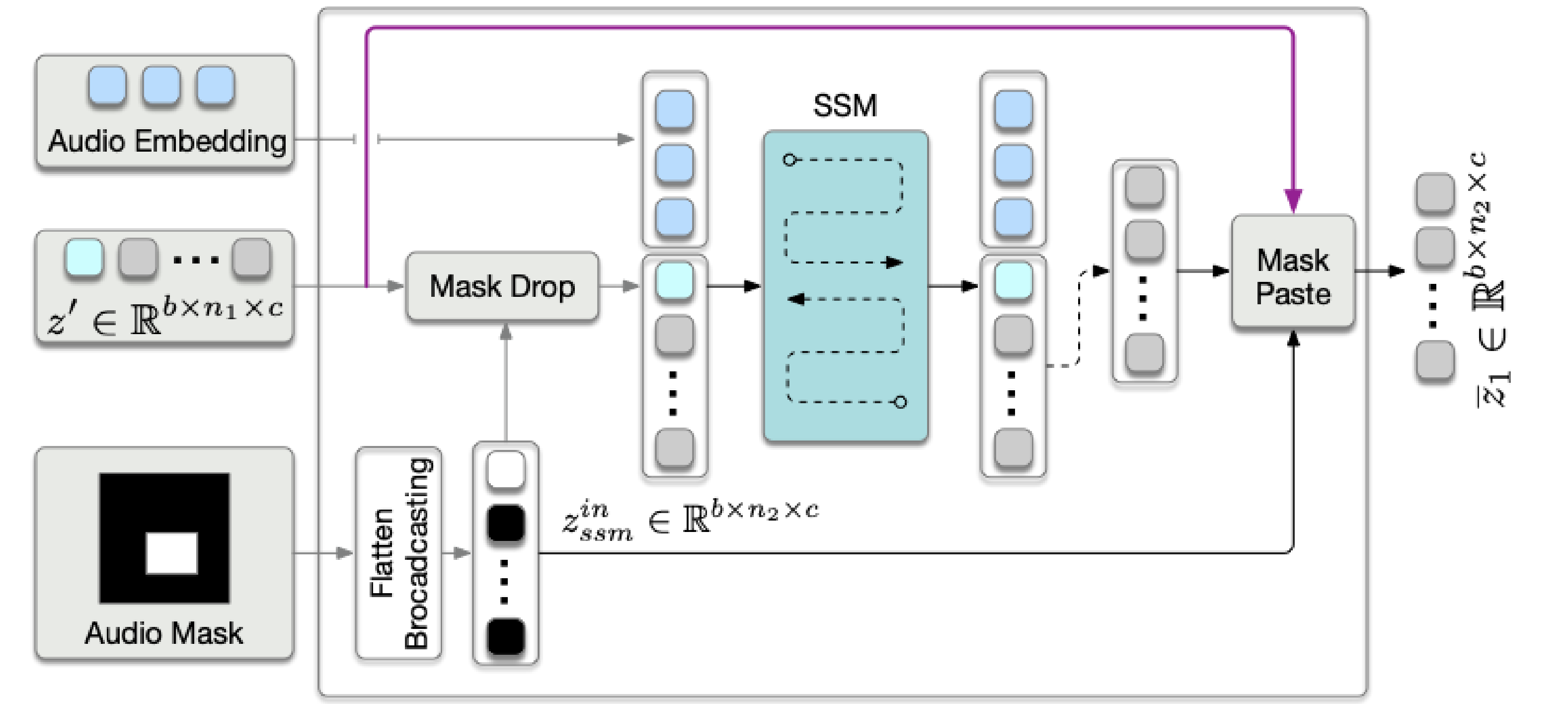




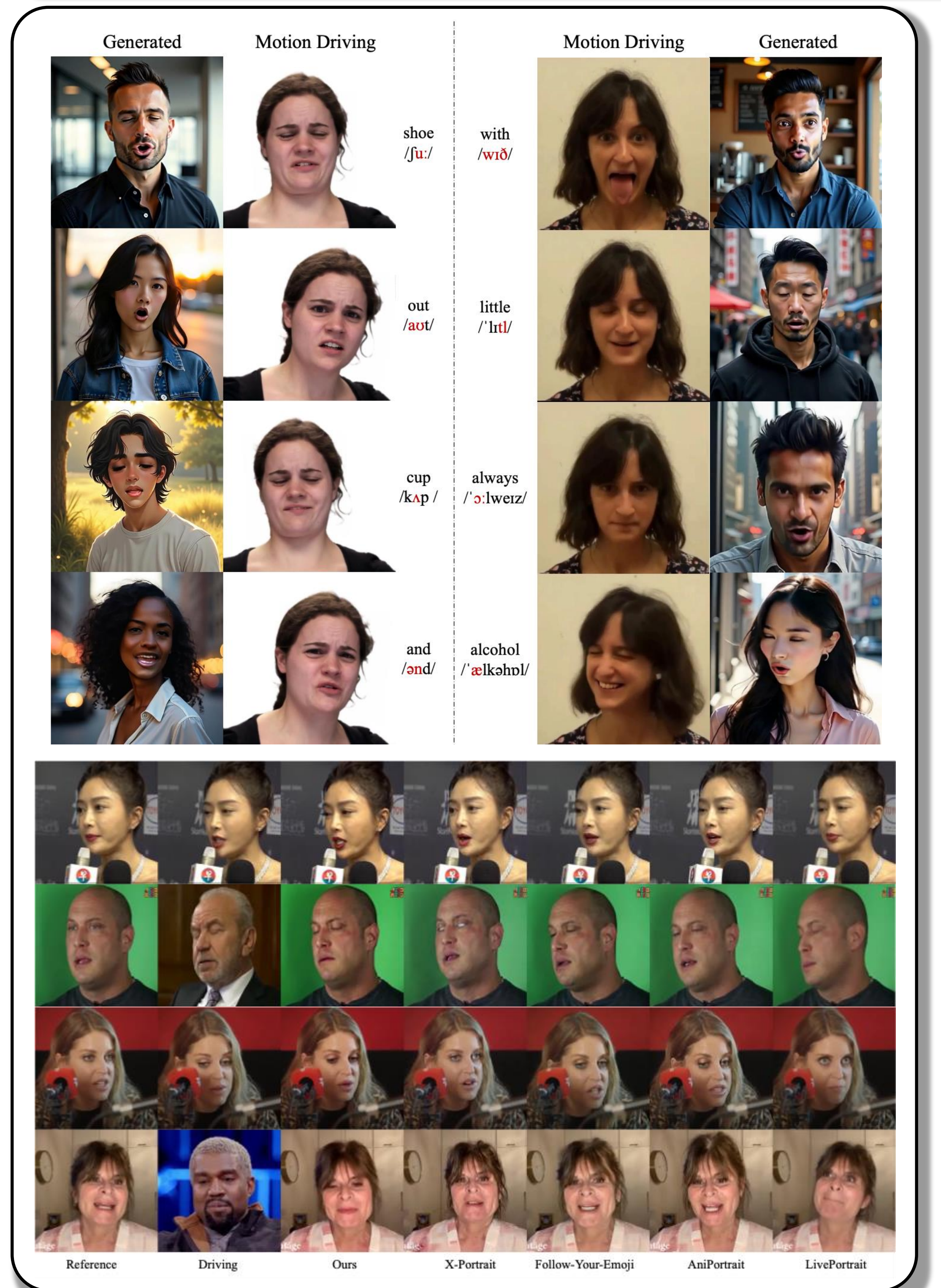
We Propose:

- **A Novel End-to-End Framework:** An end-to-end framework enables seamless and simultaneous control of generated videos using both audio and fine-grained facial motion signals, leading to more realistic and expressive outputs.
- **A Core Mask-SSM :** It coordinates specific driving signals with their relevant facial regions to resolve signal conflicts.
- **A Parallel-control Mamba Layer:** Effectively coordinates multiple driving signals without conflicts, ensuring smooth integration of audio and facial motion signals.




Audio Embedding
$z' \in \mathbb{R}^{b\times n_1\times c}$
Audio Mask
Mask Drop
Flatten Brocadcasting
$z_{ssm}^{in} \in \mathbb{R}^{b\times n_2\times c}$
SSM
Mask Paste
$\bar{z}_1 \in \mathbb{R}^{b\times n_2\times c}$

- **Mask-Drop:** Drops irrelevant tokens to focus on specific control tokens.
- **Mask-Paste:** Connects unmodified tokens to generate a complete image.



Project Page

Personal Page