ACM TOMM (Transactions on Multimedia Computing, Communications, and Applications)

JambaTalk: Speech-Driven 3D Talking Head Generation based on a Hybrid Transformer-Mamba Model

Farzaneh Jafari[1], Stefano Berretti[2], Anup Basu[1]

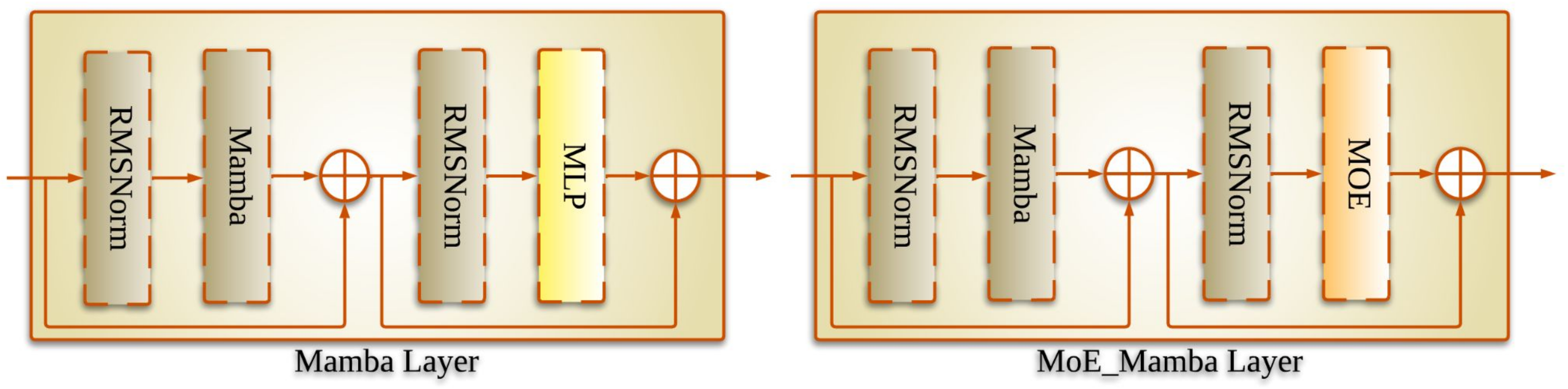
[1]University of Alberta, Canada | [2]University of Florence, Italy

Problem and Motivation

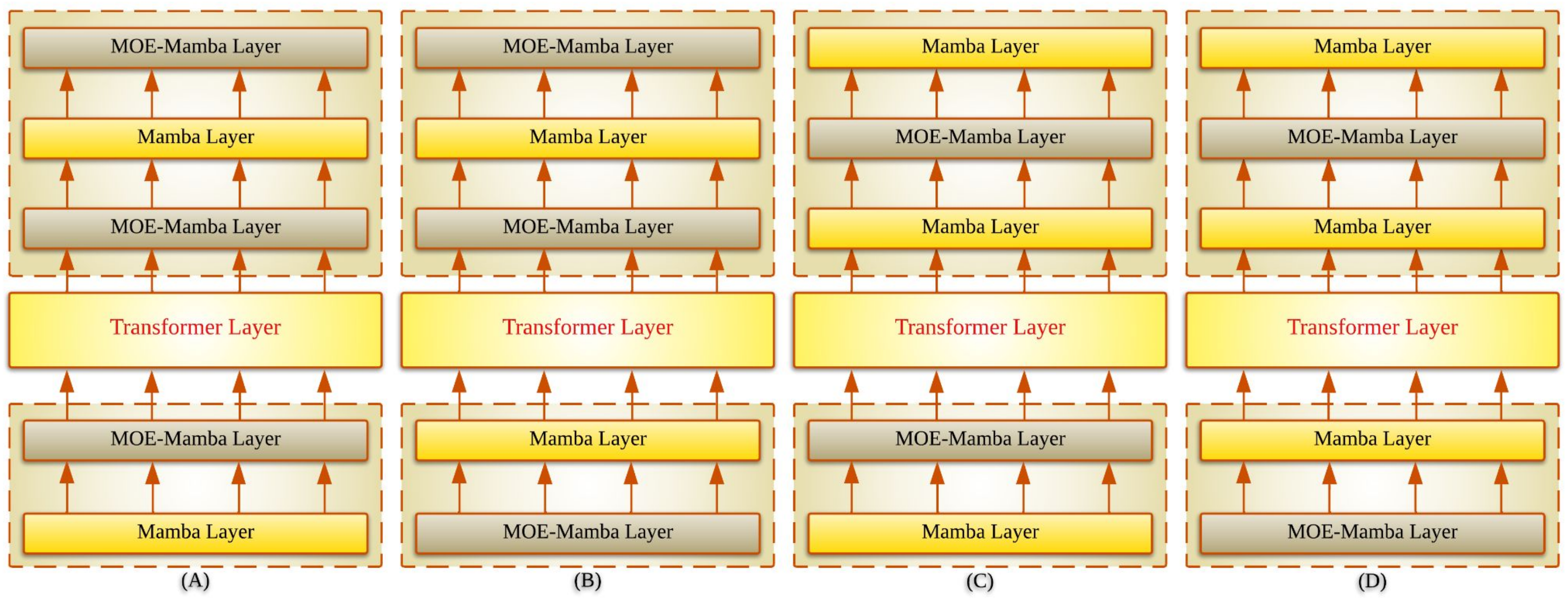
Current talking head generation faces key challenges:

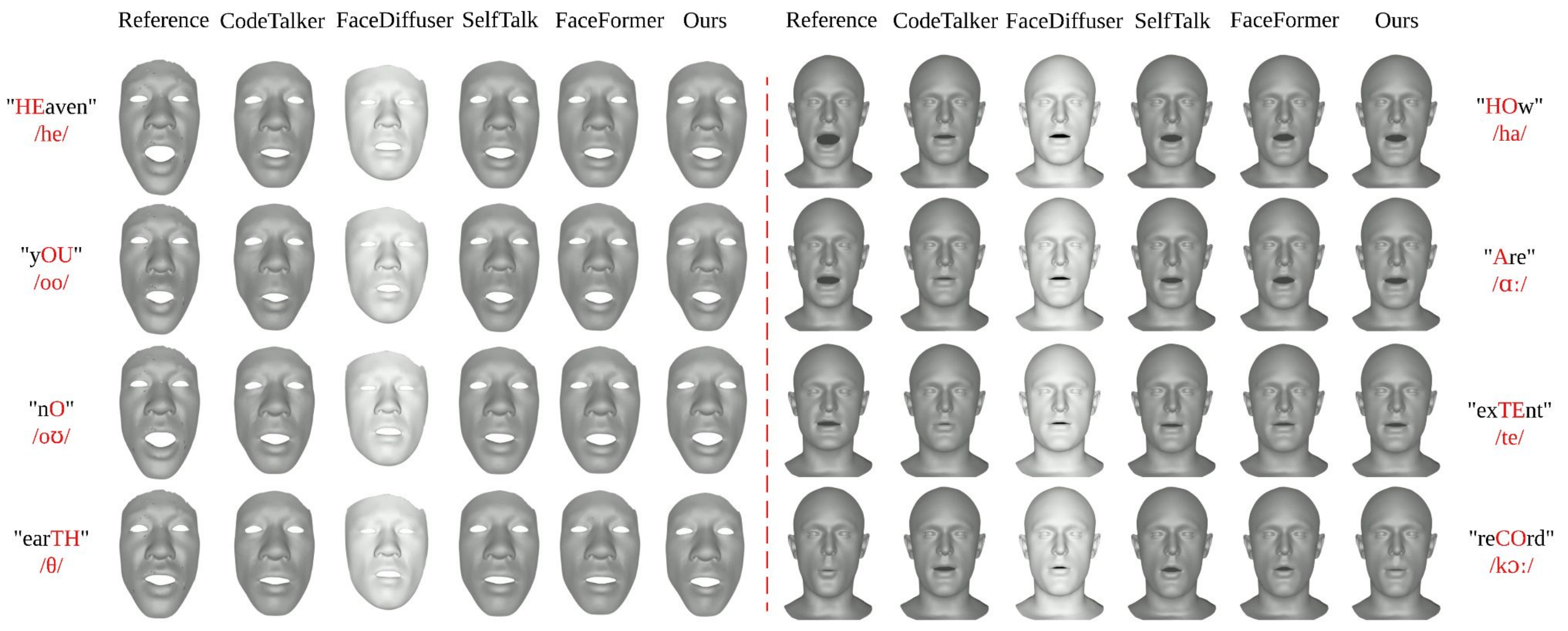
- **Limited sequence length:** Transformers struggle with long sequences due to quadratic complexity
- **Memory constraints:** High computational cost restricts real-time applications
- **Quality trade-offs:** No model excels across all metrics

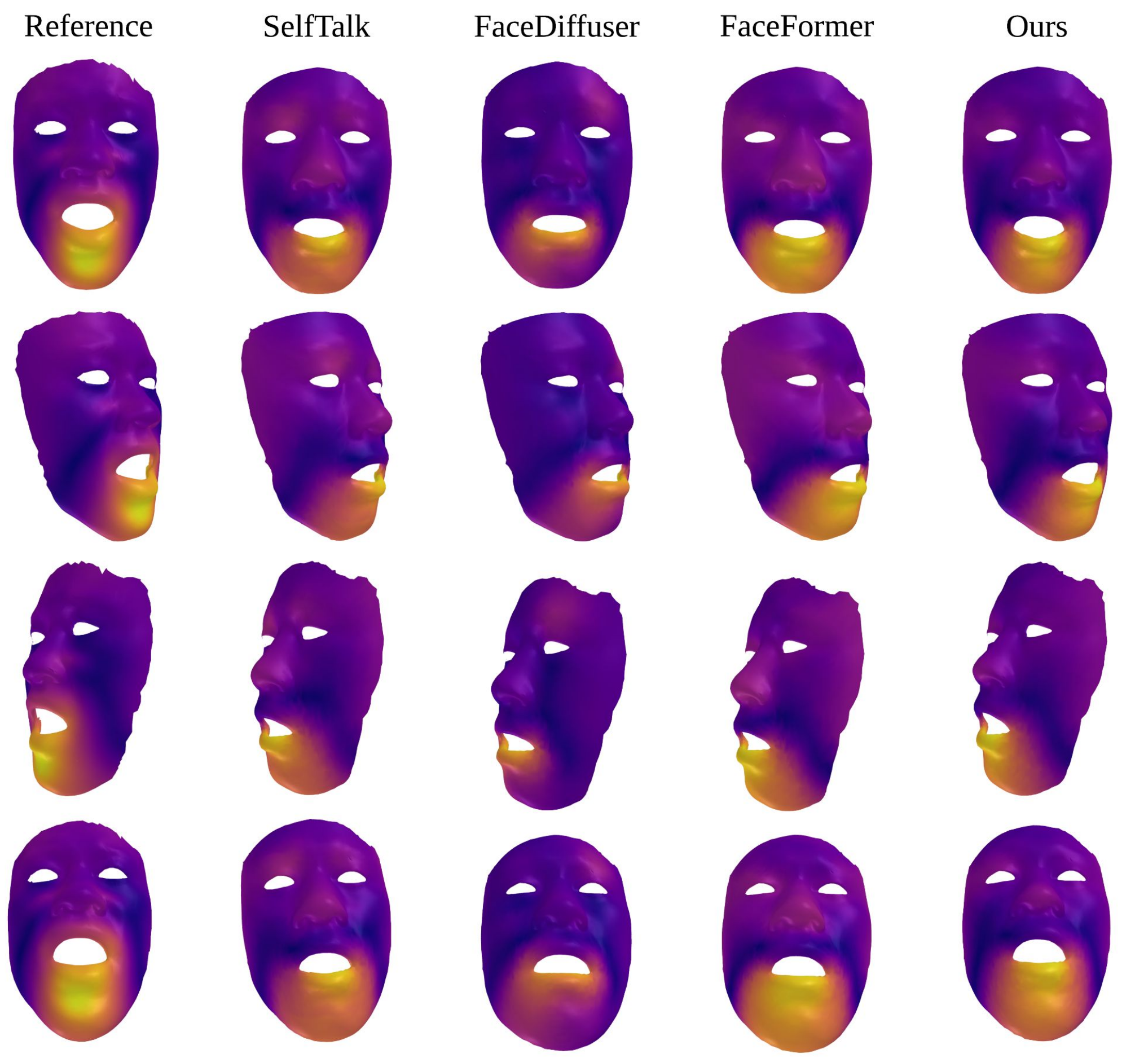
Our Solution: Hybrid Transformer-Mamba architecture combining attention mechanisms with State Space Models for efficient long-sequence processing.



Details of the Mamba and MoE_Mamba layers in the JambaTalk Decoder. Both layers begin with an RMSNorm normalization followed by a Mamba block for sequence modeling and include residual connections to preserve gradient flow. In the standard Mamba Layer (left), the Mamba output is followed by another RMSNorm and a feedforward MLP block. In contrast, the MoE_Mamba Layer (right) replaces the MLP with a Mixture-of-Experts (MoE) module, enabling dynamic expert routing per token and enhancing model capacity while maintaining computational efficiency.

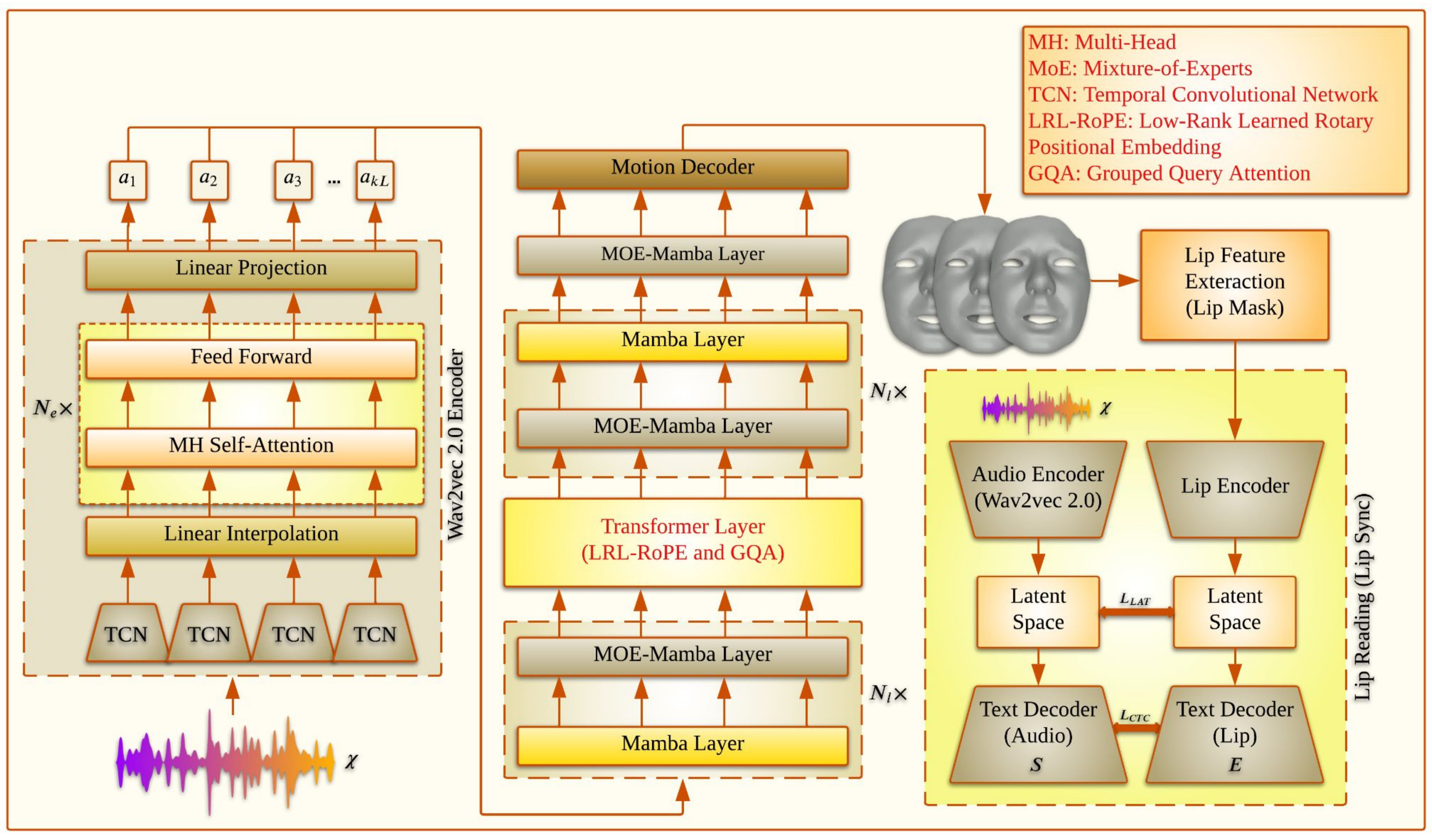






A visual comparison of frames from synthesized facial animation sequences produced by various methods, alongside reference frames from the ground-truth sequence. The red utterances are depicted in the visual frames. Our approach generates lip shapes that closely resemble the reference frames. Left: *BIWI*_6 Test-B. Right: Vocaset Test.

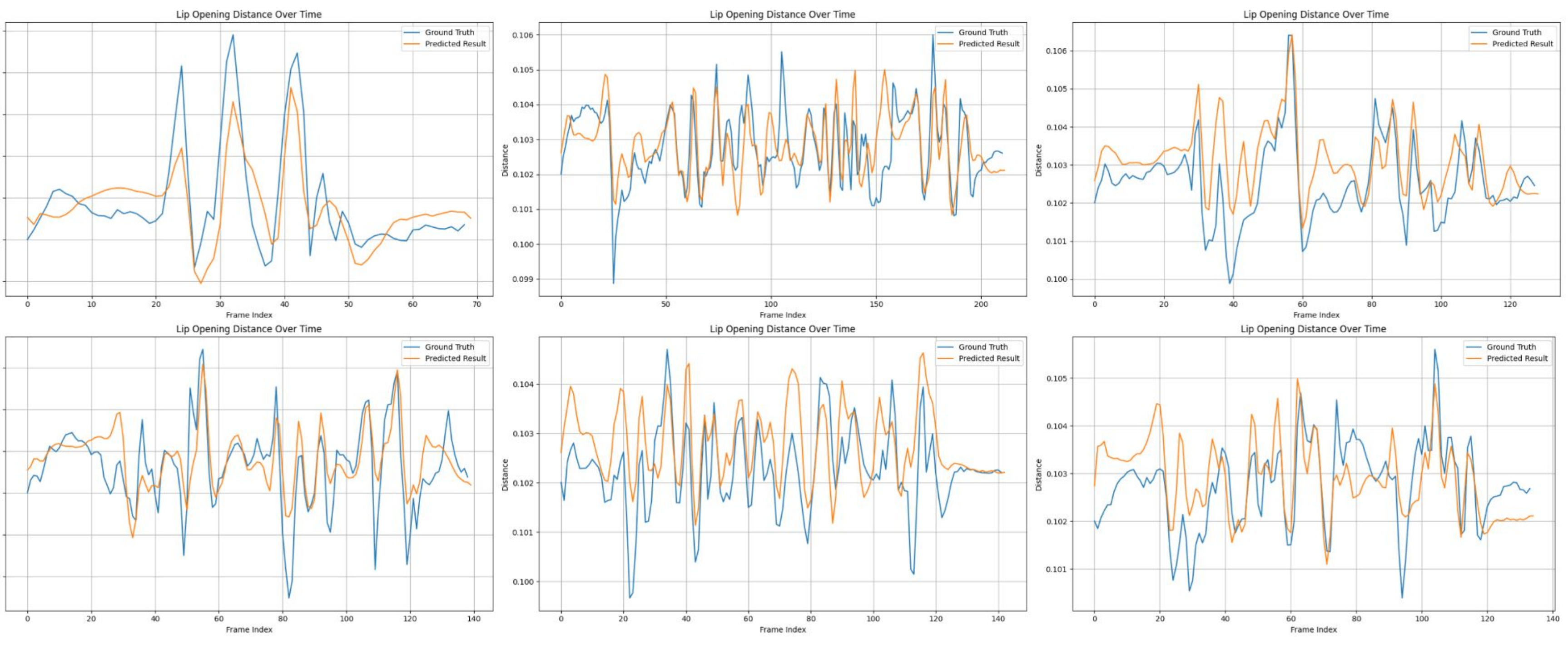
Qualitative comparison of mouth dynamics between the reference and different models (SelfTalk, FaceDiffuser, FaceFormer, and our proposed JambaTalk). The mean heatmaps visualize the lip and jaw motion intensity during speech.



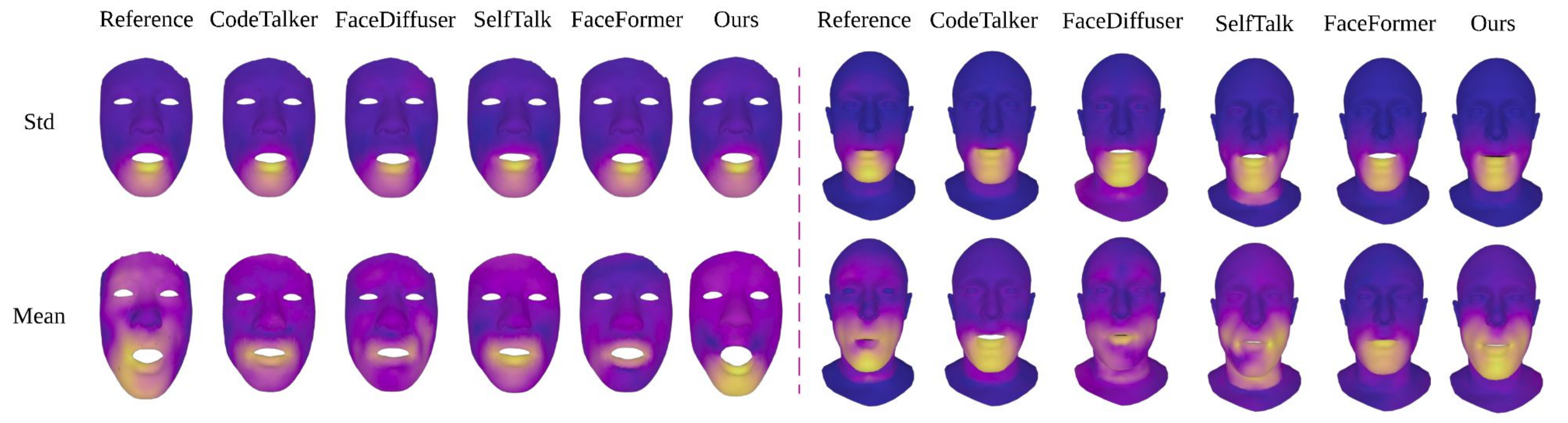
Overview of JambaTalk: The Wav2Vec 2.0 model is used to extract features from the input speech, with the encoder initialized using pre-trained weights from the original model. These encoded features are passed to the JambaTalk decoder, which generates a sequence of animated 3D face meshes. The Transformer layer incorporates Low-Rank Learned Rotary Positional Embedding (LRL-RoPE) and Grouped-Query Attention (GQA), providing a computation-efficient alternative to traditional Transformers. The lip feature extraction block then converts motion decoder outputs into lip deformation features by selecting lip vertices with a lip mask, which are processed by a Transformer-based lip encoder in the lip reader module to synchronize lip shapes.

Algorithm 1 Low-Rank Learned RoPE (LRL-RoPE)

```
Require: $x \in \mathbb{R}^{B\times T\times d}$, pos_idx $\in \mathbb{R}^{T}$, $W_1 \in \mathbb{R}^{d\times r}$, $W_2 \in \mathbb{R}^{r\times d}$
function LowRankLearnedRoPE($x$, pos_idx, $W_1$, $W_2$)
    $\theta \leftarrow 1/(10000^{\mathrm{arange}(0,d,2)/d})$        ▷ predefined frequencies, size $[d/2]$
    angles ← pos_idx[:, None] · $\theta$[None, :]        ▷ $[T, d/2]$
    sin_pos ← sin(angles)
    cos_pos ← cos(angles)
    base_emb ← concat(sin_pos, cos_pos, dim=-1)        ▷ $[T, d]$
    learned_emb ← base_emb · $W_1^\top \cdot W_2^\top$        ▷ low-rank learnable correction with transposes
    $\Delta$sin, $\Delta$cos ← split(learned_emb, 2)
    sin_final ← sin_pos + $\Delta$sin
    cos_final ← cos_pos + $\Delta$cos
    $x_{\mathrm{even}} \leftarrow x[\ldots, 0 :: 2]$
    $x_{\mathrm{odd}} \leftarrow x[\ldots, 1 :: 2]$
    $x_{\mathrm{rot\_even}} \leftarrow x_{\mathrm{even}} * \mathrm{cos\_final}[None, :, :] - x_{\mathrm{odd}} * \mathrm{sin\_final}[None, :, :]$
    $x_{\mathrm{rot\_odd}} \leftarrow x_{\mathrm{even}} * \mathrm{sin\_final}[None, :, :] + x_{\mathrm{odd}} * \mathrm{cos\_final}[None, :, :]$
    return interleave($x_{\mathrm{rot\_even}}$, $x_{\mathrm{rot\_odd}}$)
end function
```



Lip opening distance over time, showing the variation in 3D Euclidean distance between the upper and lower lip landmarks for each video frame. Peaks indicate moments when the mouth is open wider, while valleys correspond to smaller openings or closed lips on *BIWI*_6 Test-B dataset.



The temporal statistics (mean and standard deviation) of motion variations between adjacent frames in the sequence on Vocaset Test and *BIWI*_6 Test-B datasets.

Project Page

arXiv

GitHub

Contact: farzane1@ualberta.ca | Funded by NSERC and Alberta Innovates | © 2026 JambaTalk Research Project. All rights reserved.