MambaVoiceCloning (MVC)

Efficient and Expressive Text-to-Speech via State-Space Modeling and Diffusion Control

Sahil Kumar • Namrataben Patel • Honggang Wang • Youshan Zhang*

Yeshiva University, New York, USA | Chuzhou University, Anhui, China

4.18★	21M	1.6×	SSM-only
MOS (VCTK)	Encoder params	Encoder speedup	Inference stack

Published at ICLR 2026

Motivation & Problem

Quadratic Complexity

Attention in text/prosody encoders scales $O(T^2)$ — costly for long sequences.

Recurrent Drift

LSTM/GRU exhibit long-range gradient instability in prosody modeling.

Hybrid Gap

Prior Mamba-TTS systems still retain attention for style and duration modules at inference.

Prior Systems at Inference

System	Attn @ Infer?	SSM-Only?
StyleTTS2	Yes	✗
Zhang'24	Hybrid	✗
MVC (ours)	**None**	✓

Research Question

Can we build a diffusion TTS system with fully SSM-only conditioning — no attention anywhere in the inference stack — while matching or exceeding attention-based quality?

Key Contributions

01

First Fully SSM-Only Diffusion TTS

MVC is the first diffusion TTS system with SSM-only inference conditioning across text, rhythm, and prosody — no attention modules at deployment.

02

Gated Bi-Mamba Fusion + AdaLN

A novel fusion design combining gating and adaptive layer normalization. Ablation shows MOS 4.16 (full) vs 3.64 (concat-only) — a 14% improvement from the fusion design alone.

03

Deployment-Oriented Evaluation

Protocol-matched baselines isolating architectural impact; evaluation covers long-form robustness, finite look-ahead streaming, and linear-time scaling.

04

21M Efficient Encoder

1.6× encoder speedup and 72% peak memory vs StyleTTS2, with linear O(T) conditioning complexity — practical for real-world deployment.

Architecture: SSM-Only Conditioning Stack

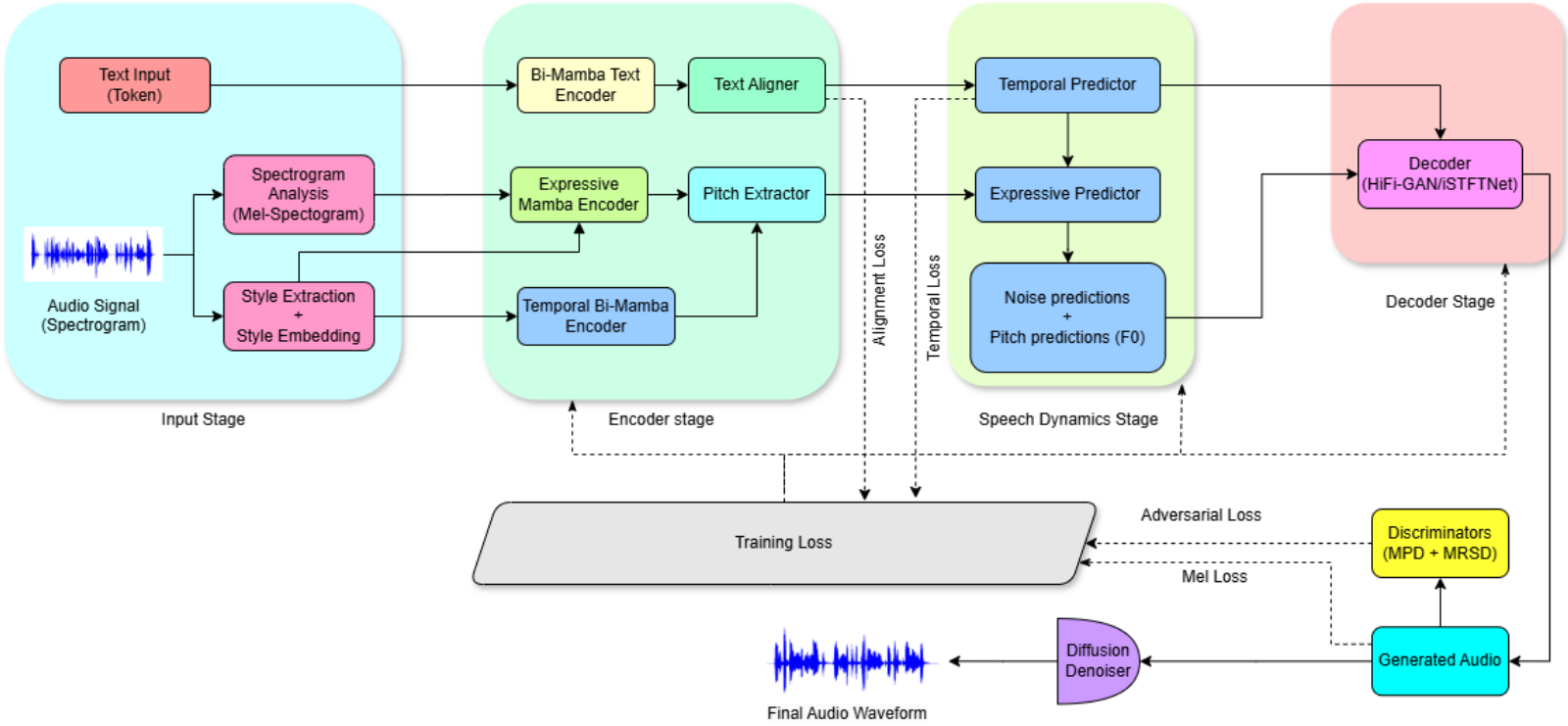


Figure 1: Overview of MambaVoiceCloning (MVC). The framework uses Bi-Mamba Text Encoders for phoneme modeling, a Temporal Bi-Mamba for rhythmic alignment, and an Expressive Mamba for prosodic control. A lightweight aligner (dotted box) provides phoneme–frame supervision only during training, ensuring an SSM-only encoder at inference. Conditioning features drive a diffusion decoder and vocoder for waveform synthesis.

Demo

Single-Speaker (LJSpeech, Out-Of-Distribution Texts)

Text: *Then leaving the corpse within the house they go themselves to and fro about the city and beat themselves, with their garments bound up by a girdle.*

Ground Truth

MVC

StyleTTS 2

JETS

Main Results: Quality & Generalization

VCTK Zero-Shot (Subjective MOS)

Model	MOS-N ↑	MOS-S ↑
VITS	3.66	3.53
StyleTTS2	4.12	4.01
MVC (ours)	**4.18 ★**	**4.09 ★**

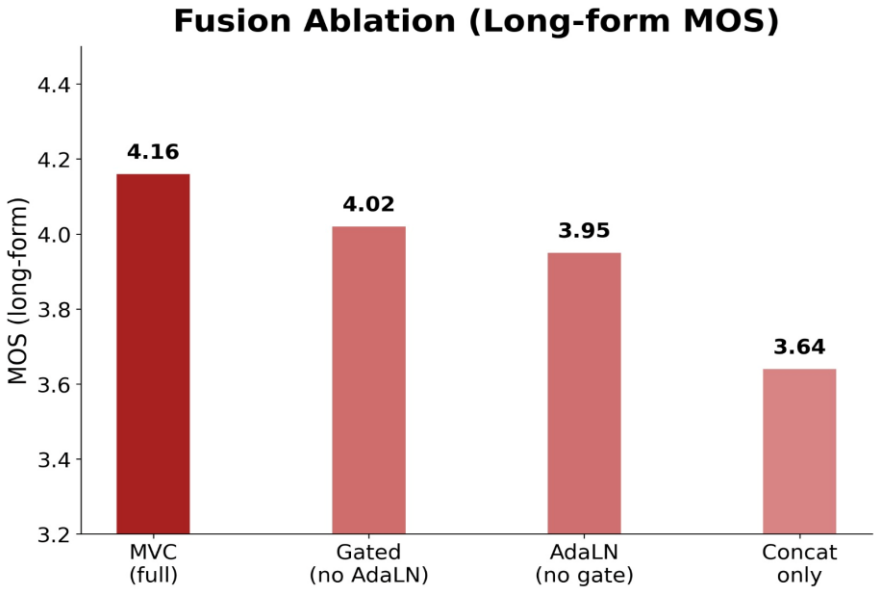
+0.07 MOS over StyleTTS2 on VCTK
Consistent gains across all objective metrics
Lowest RTF = fastest synthesis

LJSpeech Objective Results

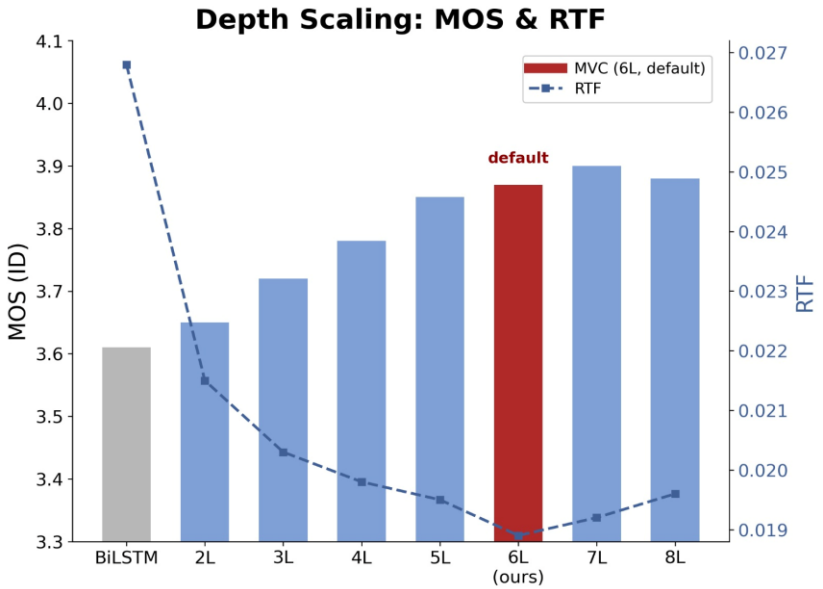
Model	F0 RMSE ↓	MCD ↓	WER ↓	PESQ ↑	RTF ↓
VITS	0.667	4.97	7.23%	3.64	0.0211
StyleTTS2	0.651 ★	4.93	6.50% ★	3.79	0.0174
Hybrid-Mamba	0.659	4.95	6.68%	3.79	0.0189
Bi-Mamba (concat)	0.656	4.93	6.58%	3.82	0.0181
MVC (ours)	0.653	**4.91 ★**	6.52%	**3.85 ★**	**0.0169 ★**

Ablation Studies: Fusion Design & Encoder Depth

Fusion Ablation (Long-form MOS)



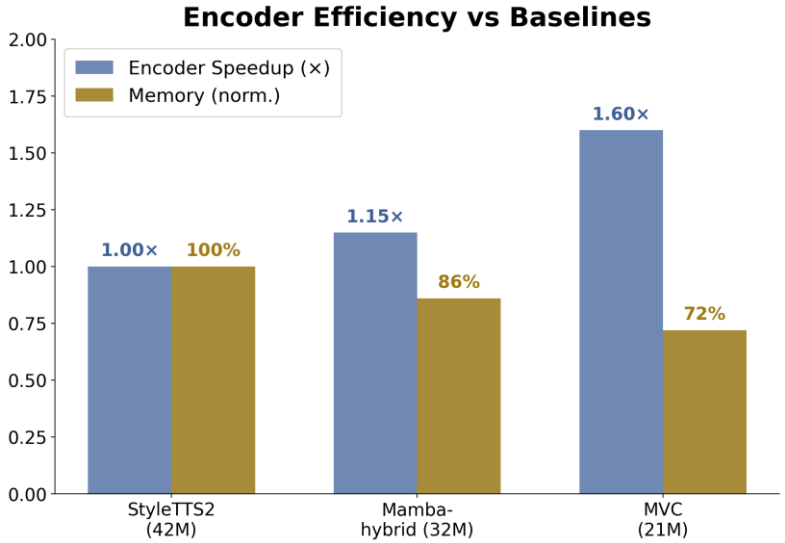
Depth Scaling: MOS & RTF



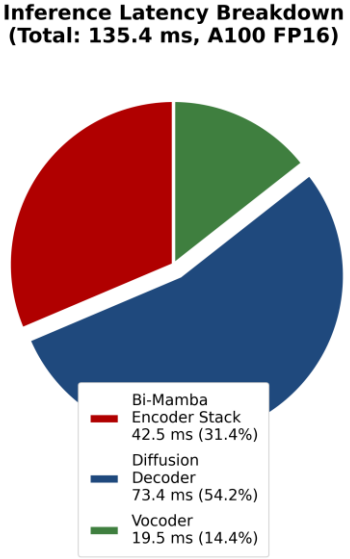
Key Takeaway: Full fusion (Gated + AdaLN) gives MOS 4.16 vs 3.64 for concat-only (+14%). 6-layer encoder chosen as optimal depth — best RTF/quality tradeoff.

Efficiency: Encoder Speedup & Inference Latency

Encoder Efficiency vs Baselines

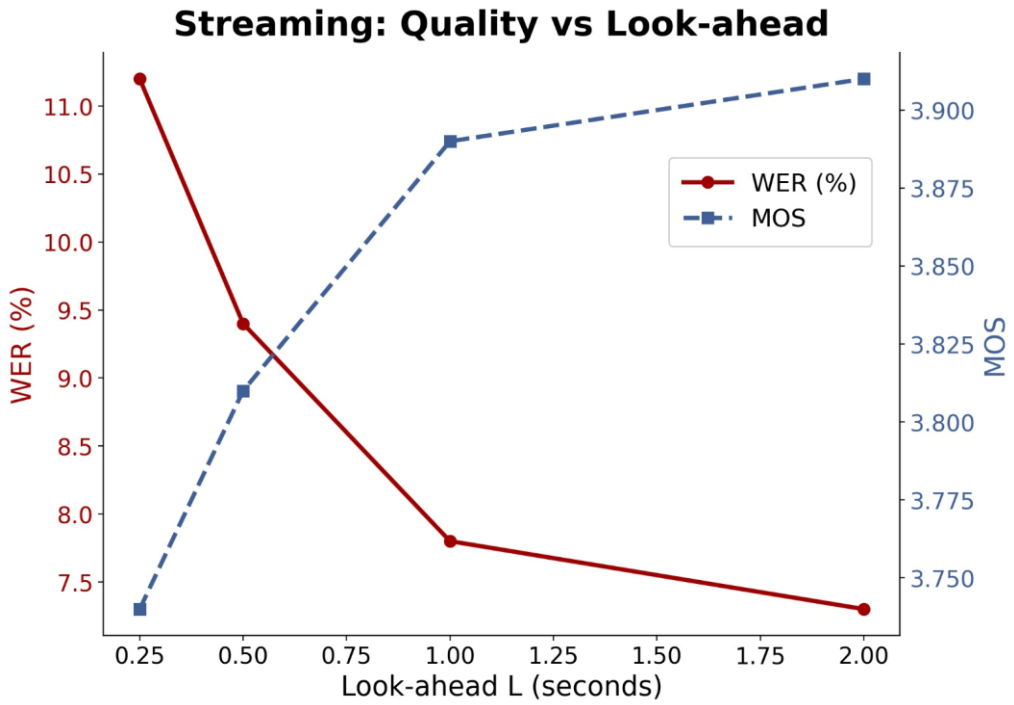


Inference Latency Breakdown (135.4 ms, A100 FP16)



1.6×	72%	21M	O(T)
Encoder speedup vs StyleTTS2	Peak memory vs StyleTTS2	Encoder parameters	Conditioning complexity

Streaming TTS: Quality vs Look-ahead



How Streaming Works

Look-ahead L

Controls how many future seconds the model observes before synthesizing each chunk.

WER vs MOS tradeoff

Shorter look-ahead → lower latency, higher WER. Longer → better quality.

L = 0.5s sweet spot

WER 9.4%, MOS 3.81 — practical for real-time voice applications.

Enabled by SSM

Linear-time processing makes streaming tractable without attention caching overhead.

MVC is the first SSM-based TTS system evaluated under a finite look-ahead streaming protocol.

References

[1] Gu, A. and Dao, T. Mamba: Linear-Time Sequence Modeling with Selective State Spaces. ICLR, 2024.

[2] Li, Y. A., Han, C., Raghavan, V. S., Mischler, G., and Mesgarani, N. StyleTTS 2: Towards Human-Level Text-to-Speech through Style Diffusion and Adversarial Training with Large Speech Language Models. NeurIPS, 2023.

[3] Jiang, X., Li, Y. A., Florea, A. N., Han, C., and Mesgarani, N. Speech Slytherin: Examining the Performance and Efficiency of Mamba for Speech Separation, Recognition, and Synthesis. Interspeech, 2024.

[4] Zhang, X. et al. Mamba in Speech: Towards an Alternative to Self-Attention. arXiv, 2024.

[5] Popov, V., Vovk, I., Gogoryan, V., Sadekova, T., and Kudinov, M. Grad-TTS: A Diffusion Probabilistic Model for Text-to-Speech. ICML, 2021.

[6] Kong, J., Kim, J., and Bae, J. HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis. NeurIPS, 2020.

Conclusions

1. MVC is the first diffusion TTS system with fully SSM-only conditioning — no attention at inference.
2. Gated Bi-Mamba + AdaLN fusion is critical: MOS 4.16 (full) vs 3.64 (concat-only), a +14% gain.
3. 21M encoder delivers 1.6× speedup and 72% memory vs StyleTTS2 with better perceptual quality.
4. Streaming TTS enabled at L ≥ 0.5s with strong WER (9.4%) and MOS (3.81) tradeoff.
5. Robust generalization: consistent gains on VCTK zero-shot and multilingual CSS10 (ES/DE/FR).

Source Code

Thank you for your attention!

Sahil Kumar • Namrataben Patel • Honggang Wang • Youshan Zhang*
Yeshiva University, New York | ICLR 2026