VibeVoice: Expressive Podcast Generation with Next-Token Diffusion

ICLR 2026 Rio de Janeiro, Brazil

ORAL

Zhiliang Peng, **Jianwei Yu**, **Wenhui Wang**, **Yaoyao Chang**, **Yutao Sun**, **Li Dong**, Yi Zhu, Weijiang Xu, Hangbo Bao, Zehua Wang, Shaohan Huang, Yan Xia, Furu Wei

Microsoft Research | https://aka.ms/GeneralAI

Microsoft

Motivation & Innovations

Long-form, multi-speaker conversational audio (podcasts, audiobooks) demands speaker consistency, natural turn-taking, and expressive cues (breaths, pacing). Prior systems limited to **2 speakers, <10 min**; concatenating single-speaker utterances loses conversational naturalness.

1. **7.5 Hz hybrid tokenizers** — decoupled acoustic (σ-VAE) + semantic (ASR), 6–80× more compact than prior codecs
2. **LLM + Diffusion Head** — Qwen2.5 for dialogue flow, diffusion head for high-fidelity acoustic generation
3. **Podcast data pipeline** — segmentation + diarization + quality filtering; no speech enhancement preserves expressive cues
4. **90 min, 4 speakers** — zero-shot generation, far beyond MoonCast (2 speakers, <10 min)

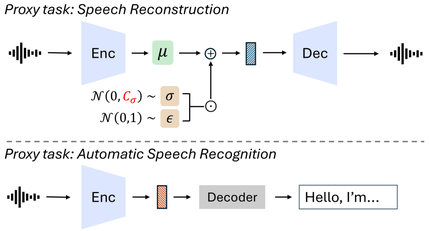
7.5 Hz	3.76	90 min	4
Frame Rate	Best MOS	Max Duration	Speakers

Decoupled Speech Tokenizers

At 7.5 Hz (3200× compression from 24 kHz), a single encoder cannot balance acoustic fidelity and semantic understanding. Coupled tokenizer improves WER (6.22 → 3.55) but speaker similarity collapses (0.68 → 0.45). **Decoupled hybrid achieves the best balance: WER 1.84, SIM-O 0.64.**

$$\mu = \mathrm{Enc}_\phi(\mathbf{x}), \quad z = \mu + \sigma \odot \epsilon, \quad \epsilon \sim \mathcal{N}(0, I), \quad \sigma \sim \mathcal{N}(0, C_\sigma)$$

Unlike standard VAE where σ is learned, σ-VAE samples from a fixed prior — preventing variance collapse in AR generation.

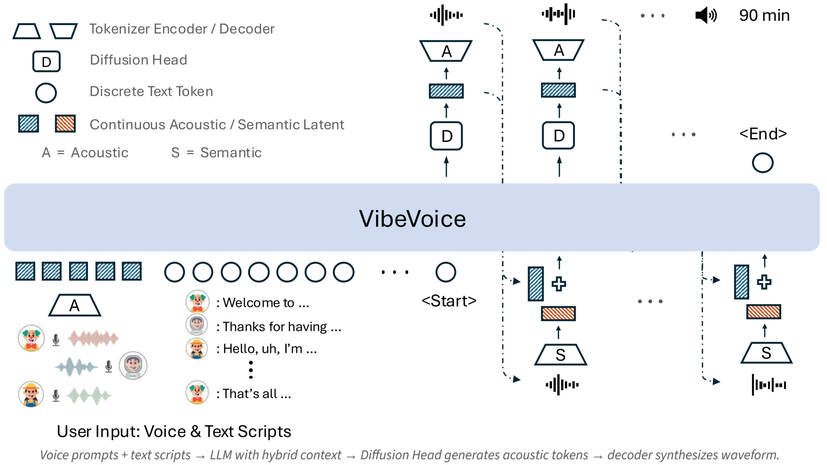


Acoustic Tokenizer (upper): σ-VAE reconstructs waveform. Semantic Tokenizer (lower): ASR proxy task. Both at 7.5 Hz.

Acoustic (σ-VAE): encoder-decoder ~340M params, continuous latent space, trained with DAC discriminator. PESQ 3.068, UTMOS 4.181 at 7.5 Hz.

Semantic (ASR proxy): mirrors acoustic encoder architecture, trained on ASR task for content-centric features. Decoder discarded after training.

Architecture Overview



Voice prompts + text scripts → LLM with hybrid context → Diffusion Head generates acoustic tokens → decoder synthesizes waveform.

VibeVoice uses an LLM backbone with **Next-Token Diffusion** for audio generation. At each autoregressive step, the LLM reads text + previous audio context, then a lightweight diffusion head generates the next continuous audio token — decoded into waveform by the acoustic decoder.

Next-Token Diffusion

Standard LLMs predict the next *word* via softmax over a vocabulary. But audio lives in **continuous space** — there is no dictionary to choose from. VibeVoice replaces softmax with a **diffusion head**: a small network (4 layers, ~123M params) that iteratively shapes random noise into an audio token, guided by the LLM's hidden state.

LLM Context h_i = Transformer(text, $z_{<i}$) → **Diffusion Head** noise → … → clean → **Audio Token** $z_{a,i}$ → Decoder → Audio

Standard LLM: h_i → softmax → discrete word | **VibeVoice:** h_i → diffusion → continuous audio vector

Train: $\mathcal{L}_{\text{Diff}} = \mathbb{E}\left\|\epsilon - \epsilon_\theta(z_a^{(t)}, t, h_i)\right\|^2$ ← predict the noise added to clean token

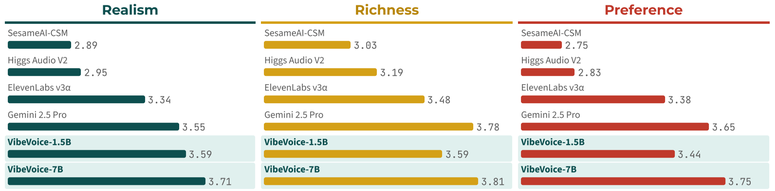
Infer (CFG): $\hat{\epsilon} = \epsilon_{\text{uncond}} + w \cdot (\epsilon_{\text{cond}} - \epsilon_{\text{uncond}})$ ← amplify LLM conditioning (w=1.3)

Why σ-VAE Enables Next-Token Diffusion

The Transformer runs **once** per token; only the tiny diffusion head iterates. Critically, **σ-VAE enables Next-Token Diffusion**: standard VAE's learned variance collapses to near-zero, leaving no room for generation error. σ-VAE enforces a fixed variance $\sigma \sim \mathcal{N}(0, C_\sigma)$, creating a **tolerance zone** in latent space that absorbs the inevitable error accumulation from autoregressive generation — enabling stable synthesis over 90 minutes.

Main Results

VibeVoice-Eval benchmark: 108 podcast samples (1–30 min). Subjective: 24 annotators, ~6h audio.



Objective — Best WER-W **1.11** (VibeVoice-1.5B) | Best SIM-O **0.692** (VibeVoice-7B)

Tokenizer Reconstruction

7.5 Hz with 3200× compression. Continuous σ-VAE achieves best PESQ and UTMOS on LibriTTS test-clean.

Model	N_q	Rate	PESQ ↑	UTMOS ↑
Encodec	8	600	2.72	3.04
DAC	4	400	2.738	3.433
WavTokenizer	1	75	2.373	4.049
Ours (Acoustic)	N/A	7.5	**3.068**	**4.181**

Scalability

Model	Short (0~12 min)		Long (12~30 min)	
	WER ↓	SIM-O ↑	WER ↓	SIM-O ↑
CosyVoice2 (concat)	4.27	0.73	4.95	0.74
MoonCast	10.4	0.55	crashes †	–
VibeVoice-1.5B	**2.11**	**0.59**	**1.55**	**0.59**
VibeVoice-7B	**0.66**	**0.75**	**1.24**	**0.75**

VibeVoice supports up to **90 min** duration with **4** distinct speakers — MoonCast is limited to 2 speakers and crashes on long-form (>12 min) generation.

Conclusion

VibeVoice: zero-shot, long-form, multi-speaker podcast generation. 7.5 Hz tokenizers + LLM next-token diffusion → SOTA on subjective and objective metrics. **90 min, 4 speakers.**

Code & Checkpoints