BiMTokenizer: Preserving Semantic-Acoustic Balance in Low-Bitrate Speech Tokenization via Bidirectional State-Space Modeling

Xin Zhang[1] **Lin Li**[1*] **Chuanbo Liu**[1] **Jianquan Liu**[2,3] **Kong Aik Lee**[4]
[1]Wuhan University of Technology, Wuhan, China
[2]NEC Laboratories Asia Pacific, Singapore [3]NEC Corporation, Japan
[4]The Hong Kong Polytechnic University, Hong Kong
[1]{zx361361,cathylilin,Lchb72}@whut.edu.cn
[2]jianquan_liu@nec.com.sg [3]jqliu@nec.com
[4]kong-aik.lee@polyu.edu.hk

Abstract

Speech codecs serve as bridges between continuous speech signals and large language models, yet face an inherent conflict between acoustic fidelity and semantic preservation. To mitigate this conflict, recent works increasingly adopt dual-tower architectures to decouple semantic and acoustic modeling with separate encoders. However, these dual-tower designs incur substantial architectural overhead. To avoid such complexity, we revisit the single-tower paradigm and propose **BiMTokenizer**, a low-bitrate speech codec (around 1.1 kbps) combining a **bidirectional state-space backbone** with **Residual Spherical Leech Quantization (RSLQ)**. The bidirectional backbone strengthens temporal modeling, while RSLQ offers a fixed, well-separated lattice bottleneck for robust semantic and acoustic tokenization without learned-codebook collapse. Experiments show that BiMTokenizer achieves superior acoustic reconstruction and the lowest WER among low-bitrate codec baselines across both clean and noisy environments, while using less than half the parameters of recent dual-tower baselines. Furthermore, its robust semantic representations yield strong performance on downstream speech understanding tasks, confirming that a well-designed single-tower codec can preserve the semantic–acoustic balance at low bitrates. The code and model weights are available at https://github.com/ZhangXinWhut/BiMTokenizer.

1 Introduction

In recent years, large language models (LLMs) have demonstrated remarkable performance in natural language processing, enabling fluent and natural text-based interactions (Brown et al., 2020). Inspired by this success, researchers have extended LLMs to the speech modality, giving rise to Speech Large Language Models (Speech LLMs) (Lakhotia et al., 2021; Borsos et al., 2023; Chen et al., 2025; Zhang et al., 2023). A key component of Speech LLMs is the speech codec, which converts continuous speech waveforms into discrete tokens compatible with token-based LLMs. Leveraging these discrete tokens, Speech LLMs perform autoregressive modeling over speech sequences, enabling a wide range of downstream speech tasks within a unified framework.

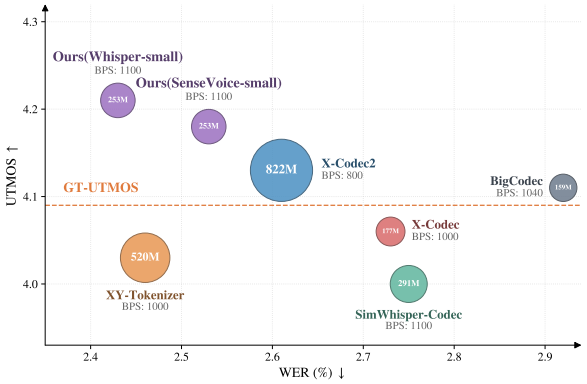

Figure 1: Codec comparison on speech reconstruction: WER (intelligibility, ↓) vs. UTMOS (naturalness, ↑); marker size indicates parameter count. Our variants are annotated by semantic teacher.

Existing speech tokens fall into two categories. Semantic tokens, derived from self-supervised learning (SSL) or supervised speech models (Baevski et al., 2020; Hsu et al., 2021; Chen et al., 2022; Radford et al., 2023; An et al., 2024), capture linguistic content but discard fine acoustic details. Acoustic tokens, in contrast, are produced by reconstruction-oriented neural audio codecs (Zeghidour et al., 2022; Défossez et al., 2023; Kumar et al., 2023), preserving acoustic fidelity but showing weak semantic alignment with text-based LLMs. An ideal codec for Speech LLMs should support both, yet the two objectives tend to compete under low-bitrate budgets, a tension that has shaped much of the recent progress in speech codec design.

*Corresponding author.

Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing, pages 9346–9364
October 24-29, 2026 ©2026 Association for Computational Linguistics

Early efforts stayed within the *single-tower* paradigm: SpeechTokenizer (Zhang et al., 2024) distills SSL features into the first layer of a Residual Vector Quantization (RVQ) bottleneck, and Mimi (Défossez et al., 2024) adopts a split-RVQ structure that dedicates one codebook to semantic content and the rest to acoustic detail. Yet their effectiveness remains constrained at low bitrates: SpeechTokenizer was originally tuned for higher-bitrate regimes and degrades under aggressive compression, whereas Mimi, though designed natively for 1.1 kbps, still attains only a moderate semantic–acoustic balance in this regime. These limitations have shifted recent attention toward *dual-tower* architectures (Li et al., 2025; Chen et al., 2026), which decouple semantic and acoustic modeling into separate encoders; the X-shape family (Ye et al., 2025a,b; Gong et al., 2026) exemplifies this line, pairing a pretrained semantic encoder with an acoustic encoder and fusing their features before quantization. However, such dual-tower designs incur substantial architectural overhead, raising computational complexity relative to a single-tower codec. This raises a natural question: *has the single-tower route truly been exhausted, or have its core components remained under-explored?* We pursue the latter, upgrading the two core components of the single-tower paradigm: the backbone and the quantizer.

In this work, we revisit these two components and propose **BiMTokenizer**, a low-bitrate ($\sim$1.1 kbps) single-tower speech codec built upon a bidirectional state-space backbone and Residual Spherical Leech Quantization (RSLQ). Speech evolves continuously along the time axis, and selective state-space models (SSMs) intrinsically capture such dynamics through structured temporal recurrence; equipping them with bidirectional scanning grants each frame access to both past and future context, providing the long-range temporal grounding required for faithful codec reconstruction. On the quantization side, RSLQ replaces the learnable codebooks of RVQ with a fixed, high-capacity codebook grounded in the spherical Leech lattice; its well-separated lattice geometry sidesteps codebook collapse by construction, while offering a representational space large enough to host semantic and acoustic information within a single token stream. Experiments show that BiMTokenizer attains superior acoustic reconstruction and the lowest WER among low-bitrate codec baselines under both clean and noisy conditions, while using fewer than half the parameters of recent dual-tower baselines. These results indicate that the semantic–acoustic balance in low-bitrate speech tokenization can be preserved not only through architectural separation, but also by strengthening the modeling and quantization design within a single-tower codec.

Our contributions can be summarized as follows:

- We propose **BiMTokenizer**, a low-bitrate single-tower speech codec that couples a bidirectional state-space backbone with a fixed-lattice quantizer for joint semantic–acoustic tokenization.
- To the best of our knowledge, this work is the first to introduce Spherical Leech quantization into speech codecs to mitigate the semantic–acoustic conflict
- Experiments on LibriSpeech and the ARCH benchmark demonstrate that BiMTokenizer excels in both acoustic reconstruction and semantic retention. Further analysis on backbone design, quantizer choice, semantic supervision, and efficiency confirm the contribution of each proposed component.

2 Related Work

2.1 From Single-Tower to Dual-Tower Codecs

To bridge continuous speech and LLMs, low-bitrate codecs must balance two competing objectives: preserving semantics for comprehension and maintaining acoustic fidelity for reconstruction. Early semantic-aware codecs mainly explored the single-tower route: SpeechTokenizer (Zhang et al., 2024) distills HuBERT (Hsu et al., 2021) features into a unified encoder–RVQ–decoder pipeline, while Mimi (Défossez et al., 2024) distills WavLM (Chen et al., 2022) into the first level of a split RVQ, leaving the rest for acoustic residuals. SimWhisper-Codec (Zhang et al., 2026) repurposes a frozen, simplified Whisper encoder for low-bitrate speech coding without external semantic supervision. These designs keep inference simple, but their semantic–acoustic balance remains limited at low bitrates. Recent works therefore shift toward dual-tower (dual-encoder) codecs, including X-Codec/XCodec2 (Ye et al., 2025a,b), Dual-Codec (Li et al., 2025), XY-Tokenizer (Gong et al., 2026), and SAC (Chen et al., 2026), which separate semantic and acoustic pathways via dedicated encoders or streams, at the cost of heavier architecture and optimization. BiMTokenizer questions

whether this shift is inevitable. Rather than adding another tower, we strengthen the two shared components of the single-tower paradigm: the temporal backbone and the quantization bottleneck.

2.2 Bidirectional State-Space Modeling in Speech

Speech is inherently temporal, and codec reconstruction must preserve both long-range structure and local transients. State-space models (SSMs), from S4 (Gu et al., 2022) to Mamba (Gu and Dao, 2024), provide a natural inductive bias via structured recurrent dynamics, whereas self-attention (Vaswani et al., 2017) relies on content-based pairwise matching. In speech, bidirectional variants such as BiMamba have been used to replace or complement attention in ASR and enhancement (Zhang et al., 2025b; Lin et al., 2026), and Mamba-SEUNet exploits bidirectional dependencies for monaural enhancement (Wang et al., 2025a). Moreover, Zhang et al. (2025a) suggest that Mamba is particularly effective for reconstruction-oriented tasks (e.g., enhancement and spectrum reconstruction), while classification tasks may require additional modules. To our knowledge, no prior work uses bidirectional SSMs as the backbone of an end-to-end low-bitrate speech codec. BiMTokenizer fills this gap by using full-context bidirectional SSM blocks to improve reconstruction within a single stream.

2.3 Quantization Methods for Discrete Tokenizers

Modern speech codecs still rely largely on learnable Vector Quantization (VQ) and Residual Vector Quantization, inherited from Vector-Quantized Variational Autoencoders (VQ-VAEs) (van den Oord et al., 2017) and popularized by SoundStream (Zeghidour et al., 2022), EnCodec (Défossez et al., 2023), and DAC (Kumar et al., 2023). However, learnable codebooks often require auxiliary objectives to maintain stable code utilization, and this issue becomes more pronounced when semantic and acoustic information are compressed into a single low-bitrate stream. Recent non-parametric visual tokenizers provide an alternative design space by replacing learned codebooks with fixed implicit ones. Lookup-Free Quantization (LFQ) (Yu et al., 2024) and Binary Spherical Quantization (BSQ) (Zhao et al., 2025) remove learnable codebooks but still rely on entropy regularization for utilization; Finite Scalar Quantization (FSQ) (Mentzer et al., 2024) avoids both codebook learning and complex entropy penalties, yet its per-dimension level design remains heuristic and less natural for residual coding. Spherical Leech Quantization (SLQ) (Zhao et al., 2026) casts non-parametric quantization through the lens of lattice coding and uses the highly symmetric first shell of the Leech lattice to construct a large, fixed, and well-separated codebook, reducing the need for learned-codebook optimization and auxiliary codebook losses. Motivated by this geometry, BiMTokenizer adapts it to speech as RSLQ, providing a high-capacity fixed bottleneck for semantic supervision and acoustic residuals.

3 BiMTokenizer

To jointly support high-fidelity reconstruction and semantic preservation within a single-tower codec, we propose **BiMTokenizer**, which strengthens the two components where the semantic–acoustic conflict is most directly expressed: the shared temporal backbone and the shared quantization bottleneck. As illustrated in Fig. 2, BiMTokenizer is an end-to-end, single-stage, GAN-based neural speech codec that comprises a bidirectional Mamba encoder–decoder backbone, a Residual Spherical Leech Quantization (RSLQ) bottleneck, and an auxiliary semantic supervision branch driven by a frozen semantic encoder. The semantic branch is active only during training, while the acoustic codec operates autonomously at inference. We describe each component and the training objectives in the following subsections.

3.1 Model Architecture

BiMTokenizer is a single-tower encoder–quantizer–decoder speech codec, as shown in Fig. 2. The encoder converts an input mel-spectrogram into a 50 Hz feature sequence through a two-layer convolutional stem and a stack of bidirectional Mamba blocks, after which a lightweight downsampler reduces the sequence to 12.5 Hz. At this rate, RSLQ encodes each frame into one semantic-aligned token followed by several residual acoustic tokens. The quantized representation is then upsampled and processed by a mirrored decoder of bidirectional Mamba blocks and transposed convolutions, producing a reconstructed mel-spectrogram that is converted to waveform by a Vocos vocoder (Siuzdak, 2024). During training, a frozen semantic teacher (Whisper-small or SenseVoice-small) (Radford et al., 2023; An et al., 2024) supervises the

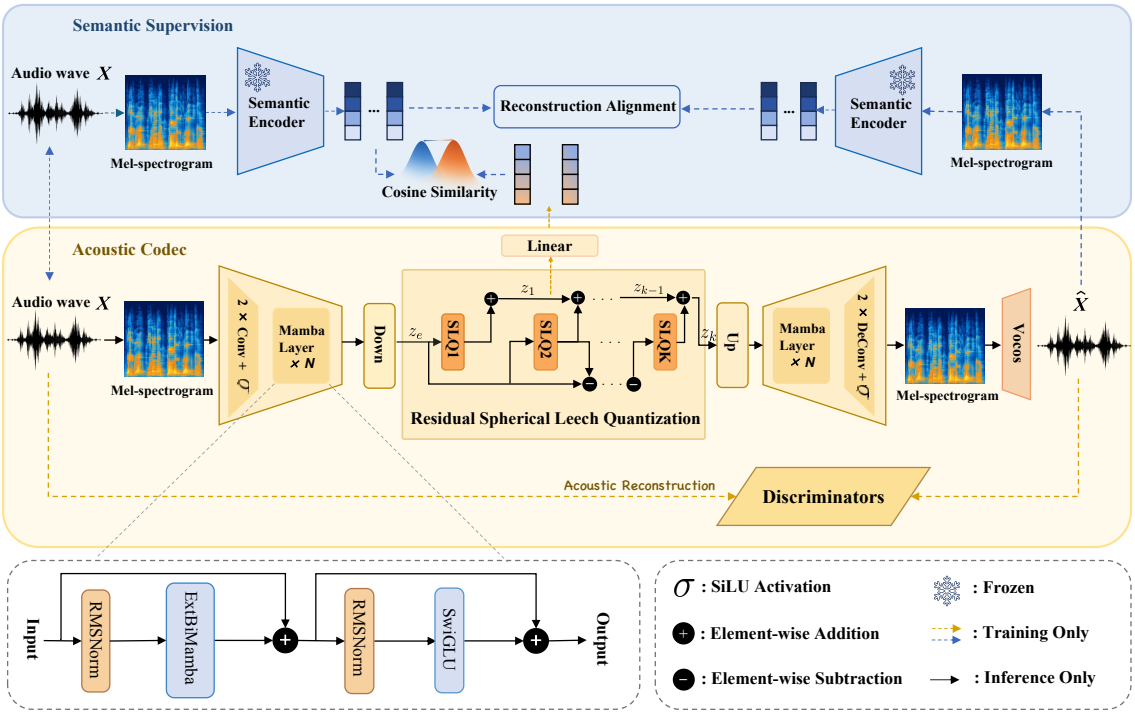

Figure 2: Overview of BiMTokenizer. The encoder and decoder use stacked bidirectional Mamba blocks, while the bottleneck is replaced by RSLQ. A frozen semantic teacher, instantiated as Whisper-small or SenseVoice-small, provides training-only semantic supervision and reconstruction alignment.

codec's representation via cosine alignment and a reconstruction-alignment loss; this branch is removed at inference, leaving a single 12.5 Hz discrete token stream. Further architectural details are provided in Appendix A.

3.2 Bidirectional Mamba Backbone

A speech codec backbone must capture long-range temporal dependencies while preserving fine-grained local detail. Selective state-space models (Gu and Dao, 2024) provide a strong inductive bias for such sequential, locally-correlated signals at linear time complexity, and recent work shows that Mamba-based backbones can match or surpass attention-based counterparts on reconstruction-style speech tasks (Zhang et al., 2025a). We therefore adopt Mamba as the temporal backbone of BiMTokenizer. Because we target the offline (non-streaming) codec setting, in which the entire utterance is available at encoding time, we adopt a *bidirectional* variant so that each latent frame integrates context from both preceding and subsequent frames.

Let $h \in \mathbb{R}^{T \times d}$ denote the input hidden sequence to a backbone block. Each block is defined as

$$h' = h + \mathrm{ExtBiMamba}(\mathrm{RMSNorm}(h)), \quad (1)$$

$$h_{\mathrm{out}} = h' + \mathrm{SwiGLU}(\mathrm{RMSNorm}(h')), \quad (2)$$

following the pre-norm, residual layout popularized by LLaMA (Touvron et al., 2023). Here $\mathrm{ExtBiMamba}(\cdot)$ denotes the external bidirectional Mamba layer of Zhang et al. (2025b): the input and its time-reversed copy are processed by two Mamba branches with independent input and output projections, and the two outputs are fused by element-wise addition. Compared with vanilla unidirectional Mamba and the inner-fused InnBiMamba variant, ExtBiMamba has been reported to deliver stronger empirical performance on speech tasks while maintaining a simpler and faster realization (Zhang et al., 2025b). The SwiGLU branch contributes per-frame non-linearity that complements the linear-in-time state-space operator, providing the additional expressive capacity required by the semantic-alignment objective in our single-tower design. We stack N such blocks in both the encoder and the decoder.

3.3 Residual Spherical Leech Quantization

To balance semantic preservation and acoustic reconstruction within a single low-bitrate stream, we propose **Residual Spherical Leech Quantization (RSLQ)**. RSLQ adapts the fixed spherical Leech codebook recently introduced for visual tokenization (Zhao et al., 2026) into a split residual bottleneck for speech codecs. Unlike standard RVQ,

which learns its codebook embeddings, RSLQ quantizes against a fixed codebook $\mathcal{C} \subset \mathbb{S}^{23}$ on the 24-dimensional unit hypersphere, taken to be the $|\mathcal{C}| = 196{,}560$ minimal vectors of the Leech lattice; this corresponds to $\log_2 |\mathcal{C}| \approx 17.58$ bits per token. Since the Leech lattice realizes the densest known sphere packing in 24 dimensions (Conway and Sloane, 1988), $\mathcal{C}$ provides a high-capacity yet uniformly distributed discrete space, which sidesteps the codebook-collapse issues commonly observed in learnable VQ at low bitrates (van den Oord et al., 2017).

For each RSLQ layer, an input feature $f \in \mathbb{R}^C$ is projected to the 24-dimensional Leech space, normalized onto the unit hypersphere, quantized by maximum inner product, and mapped back to the codec feature space:

$$\tilde{f} = \frac{W_{\text{down}} f}{\|W_{\text{down}} f\|_2}, \tag{3}$$

$$k = \arg\max_j \langle \tilde{f}, c_j \rangle, \qquad c_j \in \mathcal{C}, \tag{4}$$

$$e = \gamma W_{\text{up}} c_k, \tag{5}$$

where W_{down} and W_{up} are learnable projections and γ is a learnable scale. Gradients are propagated through the nearest-codeword assignment using a straight-through estimator (Bengio et al., 2013). Since $\mathcal{C}$ is fixed, only the projection layers and scale parameters are learned.

RSLQ is designed around the semantic–acoustic structure of speech rather than used as a direct visual-tokenizer replacement. Given an encoder frame u, we allocate the first quantization layer to semantic modeling and the remaining layers to acoustic residual coding. The semantic layer quantizes u in parallel and is supervised by the semantic distillation objective in Section 3.4. The acoustic layers form a residual chain:

$$e_0 = \text{RSLQ}_0(u), \qquad r_0 = u, \tag{6}$$

$$e_i = \text{RSLQ}_i(r_{i-1}), \tag{7}$$

$$r_i = r_{i-1} - e_i, \quad i = 1, \ldots, K-1, \tag{8}$$

$$\hat{u} = e_0 + \sum_{i=1}^{K-1} e_i. \tag{9}$$

This split residual design lets a single encoder assign one discrete layer to linguistic content while using the remaining layers for speaker traits, prosody, and fine acoustic detail, avoiding the architectural cost of a separate semantic tower. With a token rate of 12.5 Hz and $K = 5$, BiMTokenizer operates at approximately $5 \times 12.5 \times 17.58 \approx 1.10$ kbps. Additional validation-time codebook statistics are provided in Appendix B.

3.4 Auxiliary Semantic Supervision

To inject linguistic information through training-only supervision, BiMTokenizer uses a frozen semantic teacher $S_\tau(\cdot)$ only during training, where τ denotes either Whisper-small or SenseVoice-small. Given the input and reconstructed waveforms, we extract semantic targets

$$s = S_\tau(x), \qquad \hat{s} = S_\tau(\hat{x}). \tag{10}$$

The semantic RSLQ output is mapped to the teacher feature space by a lightweight projector P_ψ (Han et al., 2020; Li et al., 2021):

$$\tilde{s} = P_\psi(q_{\text{sem}}). \tag{11}$$

Teacher-specific projector details are provided in Appendix C.1. We use a frame-level cosine semantic distillation loss

$$\mathcal{L}_{\text{sem}} = 1 - \frac{1}{T} \sum_{t=1}^{T} \frac{\tilde{s}_t^\top s_t}{\|\tilde{s}_t\|_2 \|s_t\|_2}, \tag{12}$$

Inspired by self-supervised representation reconstruction (SSRR) (Lee et al., 2026), we further enforce a reconstruction alignment loss on the synthesized audio $\hat{x}$ using a frozen ASR model as the feature extractor. This objective encourages the decoder output, rather than only the quantized bottleneck, to preserve the semantic information of the original speech:

$$\mathcal{L}_{\text{align}} = \frac{1}{T} \sum_{t=1}^{T} \|\hat{s}_t - s_t\|_1. \tag{13}$$

Both semantic objectives are training-only. The frozen teacher, the projector, and the alignment path are discarded at inference.

3.5 Training Objectives

BiMTokenizer is optimized with reconstruction, adversarial, feature-matching, and semantic objectives. We use a multi-scale mel-spectrogram reconstruction loss, together with a multi-period discriminator (MPD) (Kong et al., 2020) and a multi-scale short-time Fourier transform discriminator (MS-STFTD) (Défossez et al., 2023). The adversarial objective follows the least-squares generative adversarial network (GAN) formulation (Mao et al.,

Model	Codebook Size	bps	Frame Rate	N_q	Single Tower	Param	SIM↑	STOI↑	PESQ NB↑	PESQ WB↑	UTMOS↑	ViSQOL↑	WER(%)↓
Ground Truth	–	–	–	–	–	–	1.00	1.00	4.55	4.64	4.09	5.00	2.16
EnCodec (24 kHz)	1024	1500	75	2	✓	15M	0.60	0.85	1.95	1.56	1.58	3.59	5.63
DAC (16 kHz)	1024	1500	50	3	✓	74M	0.47	0.80	1.61	1.25	1.48	3.38	7.80
SpeechTokenizer (16 kHz)	1024	1000	50	2	✓	103.7M	0.36	0.77	1.59	1.25	2.28	3.15	4.49
BigCodec (16 kHz)	8192	1040	80	1	✓	159M	0.84	0.94	3.27	2.68	4.11	4.21	2.92
Mimi (24 kHz)	2048	1100	12.5	8	✓	79M	0.74	0.91	2.80	2.25	3.63	3.85	3.28
WavTokenizer (24 kHz)	4096	900	75	1	✓	80M	0.65	0.90	2.63	2.13	3.79	3.74	4.18
SimWhisper-Codec (16 kHz)	2016	1100	12.5	8	✓	291M	0.83	0.93	3.29	2.72	4.00	4.31	2.75
XCodec (16 kHz)	1024	1000	50	2	×	177M	0.68	0.86	2.68	2.11	4.06	3.72	2.73
XCodec2.0 (16 kHz)	65536	800	50	1	×	820M	0.82	0.92	3.04	2.43	4.13	3.83	2.61
DualCodec (24 kHz)	16384/4096	1075	12.5	1/6	×	664M	0.84	0.93	3.25	2.71	4.12	4.19	2.46
XY-Tokenizer (16 kHz)	1024	1000	12.5	8	×	520M	0.85	0.92	3.10	2.50	4.03	4.22	2.46
BiMTokenizer-Whisper (Ours, 16 kHz)	196560	1100	12.5	5	✓	253M	**0.87**	**0.95**	**3.56**	**3.03**	**4.21**	4.33	**2.44**
BiMTokenizer-SenseVoice (Ours, 16 kHz)	196560	1100	12.5	5	✓	253M	0.85	0.94	3.45	2.85	4.18	**4.34**	2.53

Table 1: Speech reconstruction comparison on LibriSpeech *test-clean*. BiMTokenizer-Whisper and BiMTokenizer-SenseVoice use the same single-tower codec architecture and differ only in the frozen semantic teacher used during training. Best codec results are in bold, and second-best results are underlined.

2017), and feature matching is computed on discriminator intermediate features. The generator objective is

$$\mathcal{L}_G = \lambda_{\text{rec}}\mathcal{L}_{\text{rec}} + \lambda_{\text{adv}}\mathcal{L}_{\text{adv}} + \lambda_{\text{feat}}\mathcal{L}_{\text{feat}} + \lambda_{\text{sem}}\mathcal{L}_{\text{sem}} + \lambda_{\text{align}}\mathcal{L}_{\text{align}}. \tag{14}$$

Detailed reconstruction, adversarial, and feature-matching losses are provided in Appendix C. Because RSLQ uses a fixed geometric codebook, the objective contains no VQ codebook loss, commitment loss, or entropy regularization.

4 Experimental Setup

4.1 Dataset and Training Details

Following common practice in low-bitrate neural speech codecs (Xin et al., 2024; Défossez et al., 2023), we train BiMTokenizer on the full LibriSpeech (Panayotov et al., 2015) training set (960 hours, sampled at 16 kHz), using randomly cropped 2-second segments as input. The model contains 253M parameters in total. BiMTokenizer is instantiated with two choices of semantic teacher, yielding the variants BiMTokenizer-Whisper and BiMTokenizer-SenseVoice, which use Whisper-small (Radford et al., 2023) and SenseVoice-small (An et al., 2024), respectively, as the frozen teacher throughout training.

Training runs for 1,000,000 steps in a single stage on two NVIDIA H100 GPUs, with a per-GPU batch size of 64 and gradient accumulation of 1, giving an effective batch size of 128. Both the generator and the discriminator are optimized with AdamW (Loshchilov and Hutter, 2019) (β_1=0.8, β_2=0.9, weight decay 0.01), under a cosine-annealing schedule (Loshchilov and Hutter, 2017) that decays the learning rate from 1×10^{-4} to 0 after 30k warmup steps. Additional training details are provided in Appendix C.

4.2 Evaluation Details

An effective speech codec for Speech LLMs should reconstruct speech with high fidelity while preserving discrete representations that remain useful for speech understanding. We therefore evaluate BiMTokenizer from two complementary perspectives: speech reconstruction and semantic representation.

Speech Reconstruction. We evaluate reconstruction quality on LibriSpeech *test-clean* (Panayotov et al., 2015), reporting codec configuration, intelligibility, perceptual quality, and speaker similarity metrics: STOI, WER, PESQ-NB/PESQ-WB, UTMOS, ViSQOL, and SIM. Metric definitions, evaluation models, and implementation links are provided in Appendix D. The comparison includes representative low-bitrate speech codecs with similar operating ranges, covering reconstruction-oriented codecs (Défossez et al., 2023; Kumar et al., 2023; Xin et al., 2024; Défossez et al., 2024), semantic-aware single-tower tokenizers (Zhang et al., 2024; Ji et al., 2025; Zhang et al., 2026), and dual- or X-shaped tokenizers (Ye et al., 2025a,b; Li et al., 2025; Gong et al., 2026).

Speech Representation. We evaluate the semantic quality of codec representations using the speech-domain protocol of the ARCH benchmark (La Quatra et al., 2024), following recent codec evaluations (Ji et al., 2025; Chen et al., 2026). ARCH covers emotion recognition with RAVDESS (Livingstone and Russo, 2018) and EMOVO (Costantini et al., 2014), intent classifi-

Category	Model	Token Rate	BPS	RAVDESS↑	EMOVO↑	SLURP↑	AM↑	Avg.↑
SSL Models†	wav2vec 2.0	–	–	55.32	31.80	14.37	86.38	46.97
	data2vec	–	–	48.03	27.27	**43.57**	99.06	54.48
	HuBERT	–	–	65.28	40.48	33.75	**99.58**	59.77
	WavLM	–	–	**67.94**	**43.08**	30.98	99.50	**60.38**
ASR Models	Whisper-small	–	–	**83.33**	53.06	63.30	99.83	74.88
	SenseVoice-small	–	–	82.64	**56.29**	**71.47**	**99.98**	**77.60**
Codec Models	EnCodec	150	1500	30.90	27.72	8.37	72.57	34.89
	DAC	150	1500	31.25	21.60	7.98	69.46	32.57
	BigCodec	80	1040	31.94	15.99	7.72	65.83	30.37
	XCodec2.0	50	800	32.99	27.21	7.66	66.14	33.50
	XY-Tokenizer	100	1000	39.58	24.49	18.39	96.34	44.70
	WavTokenizer	75	900	32.55	**31.63**	8.02	69.57	35.44
	SimWhisper-Codec	100	1100	42.71	24.15	8.13	79.76	38.69
	BiMTokenizer-Whisper	62.5	1100	**43.19**	27.08	12.51	86.20	42.25
	BiMTokenizer-SenseVoice	62.5	1100	40.28	26.87	**18.48**	**98.02**	**45.91**

Table 2: Semantic representation evaluation on the speech domain of ARCH. Best results in each category are in bold, and second-best results are underlined. ASR models provide teacher upper-bound references; codec rows use pooled quantized representations. † SSL results are taken from the original ARCH benchmark.

cation with SLURP (Bastianelli et al., 2020), and digit recognition with AudioMNIST (Becker et al., 2024). For each codec, quantized representations are average-pooled over time and used as input to a linear classifier. To contextualize the gap between discrete codec tokens and continuous speech representations, we also report self-supervised learning (SSL) references, including wav2vec 2.0 (Baevski et al., 2020), data2vec (Baevski et al., 2022), HuBERT (Hsu et al., 2021), and WavLM (Chen et al., 2022). Since BiMTokenizer uses frozen ASR encoders as semantic teachers during training, we additionally include Whisper-small and SenseVoice-small as teacher upper-bound references rather than codec baselines.

5 Experimental Results and Discussion

5.1 Speech Reconstruction Results

Quantitative Results. Table 1 compares BiMTokenizer with representative low-bitrate neural codecs on LibriSpeech *test-clean*. At 1.1 kbps and 12.5 Hz, BiMTokenizer-Whisper achieves state-of-the-art reconstruction among compared low-bitrate codecs, with the best SIM, STOI, PESQ-NB, PESQ-WB, UTMOS, and WER. In particular, it reaches 3.56 PESQ-NB and 3.03 PESQ-WB, substantially exceeding all codecs below 1.5 kbps. BiMTokenizer-SenseVoice also obtains the best ViSQOL and remains close to the Whisper variant on most reconstruction metrics. Figure 3 further shows that Mimi and SpeechTokenizer recover the dominant energy regions but lose some narrow harmonic bands, while XY-Tokenizer preserves more structure yet still exhibits visible smoothing of local transitions in the enlarged region. BiMTokenizer retains sharper bands and clearer transitions, making its reconstruction visually closer to the ground truth. This qualitative evidence complements the PESQ and UTMOS gains. Appendix F reports consistent gains on the noisier LibriSpeech *test-other* set and the out-of-distribution Seed-TTS-Eval benchmark in English and Mandarin (Anastassiou et al., 2024).

Variant Comparison. The two variants differ only in the frozen teacher. Whisper-small's 50 Hz features match the encoder's pre-downsampling rate; this denser supervision yields the strongest PESQ and WER. SenseVoice-small's lower-rate features align more closely with the 12.5 Hz bottleneck, delivering stronger semantic transfer, as shown in Table 2. This contrast suggests that the teacher's frame rate, rather than its training objective alone, shapes how supervision interacts with the codec bottleneck.

5.2 Semantic Representation Results

Table 2 evaluates the semantic representation quality of codec tokens on the speech-domain tasks of ARCH. BiMTokenizer-SenseVoice obtains the highest average accuracy, surpassing the second-best XY-Tokenizer by 1.21 absolute points while remaining a single-tower codec. Its advantage is most clear on text-related tasks: it achieves the best codec results on SLURP and AudioMNIST, indicating that the quantized representations retain useful linguistic information even under a 1.1 kbps bitrate. BiMTokenizer-Whisper, in turn, achieves the best codec result on RAVDESS, consistent with its stronger reconstruction in Table 1 and suggesting

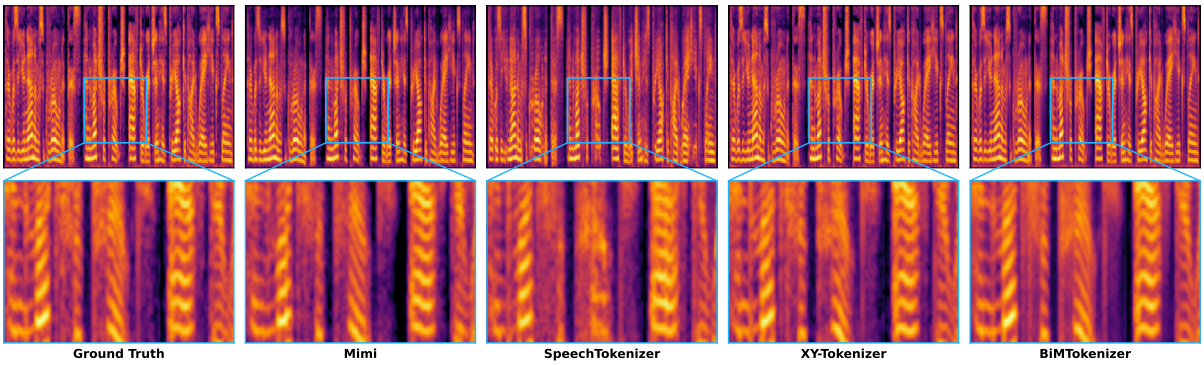

Figure 3: Mel-spectrogram reconstructions. The zoomed regions show that BiMTokenizer preserves local harmonic and spectral structure more faithfully than prior low-bitrate codecs.

that denser Whisper supervision benefits paralinguistic cues. More broadly, both variants improve substantially over reconstruction-oriented codecs without explicit semantic supervision (EnCodec, DAC, BigCodec), confirming that the single-tower design yields competitive codec-level semantic representations on this benchmark.

5.3 LLM-Based Speech Generation

We evaluate BiMTokenizer on Seed-TTS-Eval (Anastassiou et al., 2024) using an autoregressive decoder-only model. Because the original tokenizer targets codec evaluation, we retrain a TTS-oriented variant for speech generation, as detailed in Appendix E. We report word error rate (WER), speaker similarity (SIM), and UTMOS on the English and Chinese test sets.

As shown in Table 3, BiMTokenizer achieves 1.87% WER, 0.59 SIM, and 4.11 UTMOS on *test-en*, together with 1.72% WER, 0.65 SIM, and 3.30 UTMOS on *test-zh*. These results confirm that its discrete representations support autoregressive generation in both languages.

5.4 Further Analysis

Computational Efficiency. We assess computational efficiency at both the codec and backbone levels. At the codec level, Table 4 shows that BiMTokenizer contains 253M parameters and requires 14.82G MACs, compared with 664M parameters and 40.98G MACs for DualCodec and 520M parameters and 82.02G MACs for XY-Tokenizer. Relative to these two dual-tower codecs, BiMTokenizer reduces total MACs by 63.8% and 81.9%, respectively, with particularly low encoding cost for large-scale corpus tokenization. The total RTF of BiMTokenizer is 0.007, matching that of XY-Tokenizer and remaining substantially below the 0.019 RTF of DualCodec. At the backbone level,

Model	WER(%)↓	SIM↑	UTMOS↑
Seed-TTS *test-en*			
Llasa-1B-80k	3.71	0.54	4.06
Llasa-1B-160k	3.60	0.56	4.05
Llasa-1B-250k	2.99	0.57	4.07
Spark-TTS	1.98	0.58	3.94
Ours	**1.87**	**0.59**	**4.11**
Seed-TTS *test-zh*			
Llasa-1B-80k	2.69	0.65	3.27
Llasa-1B-160k	2.22	0.66	3.28
Llasa-1B-250k	1.89	**0.67**	3.28
Spark-TTS	**1.20**	**0.67**	3.27
Ours	1.72	0.65	**3.30**

Table 3: LLM-based speech generation performance on the Seed-TTS-Eval benchmark.

Figure 4 shows that BiMamba gains efficiency as the input duration grows, whereas bidirectional RoPE-based self-attention (Su et al., 2024) incurs steeper growth in computation and runtime. Together, these results show that the single-tower design reduces end-to-end computational overhead and scales favorably to long-form audio. Detailed measurement settings and additional analysis are provided in Appendix G.

Component Analysis. Table 7 in Appendix H isolates the backbone and quantizer effects without semantic supervision. Replacing BiMamba with bidirectional self-attention raises WER from 2.55% to 4.15% and lowers PESQ from 3.56/2.99 to 3.28/2.77. Replacing RSLQ with GroupFSQ or RVQ likewise degrades all reconstruction metrics. Thus, BiMamba benefits both efficiency and reconstruction, while RSLQ outperforms the alternative quantizers at the same bitrate.

Teacher-supervised ablations reveal a semantic–acoustic trade-off. For SenseVoice at λ_{sem}=15,

Model	Params	Encode		Decode		Total	
		MACs↓	RTF↓	MACs↓	RTF↓	MACs↓	RTF↓
EnCodec	15M	2.80G	0.005	**2.80G**	0.004	**5.60G**	0.009
DAC	74M	12.27G	0.005	43.22G	0.003	55.49G	0.008
BigCodec	159M	25.52G	0.008	35.56G	0.006	61.08G	0.014
Mimi	79M	**2.72G**	**0.003**	8.60G	**0.001**	11.32G	**0.004**
DualCodec	664M	31.83G	0.015	9.15G	0.004	40.98G	0.019
XY-Tokenizer	520M	63.59G	0.005	18.43G	0.002	82.02G	0.007
BiMTokenizer (Ours)	253M	4.79G	0.004	10.03G	0.003	14.82G	0.007

Table 4: Codec-level computational efficiency on LibriSpeech *test-clean*. MACs are computed for a 1-second audio input, and RTF denotes real-time factor. Total costs sum encoding and decoding. Lower values indicate higher efficiency.

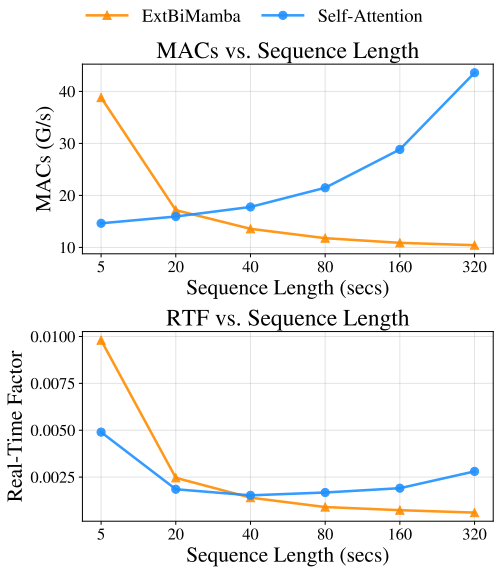

Figure 4: Length-scaling comparison between the Bi-Mamba backbone and the bidirectional Transformer counterpart with rotary position embeddings (RoPE).

reconstruction alignment lowers WER from 2.58% to 2.53% and raises AudioMNIST from 97.79 to 98.02, whereas increasing the semantic weight to 20 improves SLURP at the cost of SIM and PESQ. Whisper shows the same trend, with λ_{sem}=15 giving the best overall acoustic performance. We set λ_{sem}=15 and λ_{align}=1 in both final variants.

6 Conclusion

We presented BiMTokenizer, a 1.1 kbps single-tower speech codec built on a bidirectional state-space backbone and Residual Spherical Leech Quantization. Experiments on reconstruction and semantic benchmarks show that BiMTokenizer matches or surpasses recent dual-tower codecs while using far fewer parameters. These results offer a counterpoint to the recent shift toward dual-tower designs, suggesting that careful redesign of a single-tower codec's shared backbone and bottleneck remains an equally viable path for low-bitrate speech tokenization.

Acknowledgments

This work was partially supported by the Guangxi Science and Technology Major Program under Grant No. AA24206067. A large language model was used solely to assist with language polishing and improving the clarity of writing.

Limitations

Several limitations remain. First, RSLQ uses a fixed 196,560-entry codebook. The resulting vocabulary provides high capacity and separation but enlarges the prediction space for autoregressive TTS, potentially requiring more training data and model capacity to learn infrequent codec symbols. This capacity–predictability trade-off warrants further study. Second, the current codec combines convolutional encoder/decoder modules with Mamba blocks, resulting in a heterogeneous CNN–Mamba architecture. An architecturally homogeneous, fully Mamba-based codec is a promising direction for future work. Finally, the bidirectional backbone requires future frames, limiting the codec to offline tokenization. Causal or chunk-wise variants are needed for low-latency streaming.

Ethical Considerations

This work studies a speech codec at the representation level and is intended for research purposes. Ethical considerations such as privacy, bias, and misuse are primarily determined by downstream applications rather than the codec itself. We encourage responsible use according to ethical guidelines.

References

Keyu An, Qian Chen, Chong Deng, Zhihao Du, Changfeng Gao, Zhifu Gao, Yue Gu, Ting He, Hangrui Hu, Kai Hu, Shengpeng Ji, Yabin Li, Zerui Li, Heng Lu, Haoneng Luo, Xiang Lv, Bin Ma, Ziyang Ma, Chongjia Ni, and 14 others. 2024. FunAudioLLM: Voice understanding and generation foundation models for natural interaction between humans and LLMs. *Preprint*, arXiv:2407.04051.

Philip Anastassiou, Jiawei Chen, Jitong Chen, Yuanzhe Chen, Zhuo Chen, Ziyi Chen, Jian Cong, Lelai Deng, Chuang Ding, Lu Gao, Mingqing Gong, Peisong Huang, Qingqing Huang, Zhiying Huang, Yuanyuan Huo, Dongya Jia, Chumin Li, Feiya Li, Hui Li, and 27 others. 2024. Seed-tts: A family of high-quality versatile speech generation models. *Preprint*, arXiv:2406.02430.

Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. 2022. data2vec: A general framework for self-supervised learning in speech, vision and language. In *International Conference on Machine Learning*, pages 1298–1312. PMLR.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems*.

Emanuele Bastianelli, Andrea Vanzo, Pawel Swietojanski, and Verena Rieser. 2020. SLURP: A spoken language understanding resource package. In *Empirical Methods in Natural Language Processing*, pages 7252–7262. Association for Computational Linguistics.

Sören Becker, Johanna Vielhaben, Marcel Ackermann, Klaus-Robert Müller, Sebastian Lapuschkin, and Wojciech Samek. 2024. AudioMNIST: Exploring explainable artificial intelligence for audio analysis on a simple benchmark. *J. Frankl. Inst.*, 361(1):418–428.

Yoshua Bengio, Nicholas Léonard, and Aaron Courville. 2013. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*.

Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matthew Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi, and Neil Zeghidour. 2023. AudioLM: A language modeling approach to audio generation. *IEEE ACM Trans. Audio Speech Lang. Process.*, 31:2523–2533.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*.

Edresson Casanova, Ryan Langman, Paarth Neekhara, Shehzeen Hussain, Jason Li, Subhankar Ghosh, Ante Jukic, and Sang-gil Lee. 2025. Low frame-rate speech codec: a codec designed for fast high-quality speech LLM training and inference. In *International Conference on Acoustics, Speech and Signal Processing, ICASSP 2025*, pages 1–5. IEEE.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. 2022. WavLM: Large-scale self-supervised pre-training for full stack speech processing. *IEEE J. Sel. Top. Signal Process.*, 16(6):1505–1518.

Sanyuan Chen, Chengyi Wang, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, Lei He, Sheng Zhao, and Furu Wei. 2025. Neural codec language models are zero-shot text to speech synthesizers. *IEEE ACM Trans. Audio Speech Lang. Process.*, 33:705–718.

Wenxi Chen, Ruiqi Yan, Yushen Chen, Zhikang Niu, Ziyang Ma, Xiquan Li, Yuzhe Liang, Wenhanlin, Shunshun Yin, Ming Tao, Xinsheng Wang, and Xie Chen. 2026. SAC: Neural speech codec with semantic-acoustic dual-stream quantization. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3030–3048, San Diego, California, United States. Association for Computational Linguistics.

Michael Chinen, Felicia S. C. Lim, Jan Skoglund, Nikita Gureev, Feargus O'Gorman, and Andrew Hines. 2020. ViSQOL v3: An open source production ready objective speech and audio metric. In *Quality of Multimedia Experience*, pages 1–6. IEEE.

John H. Conway and Neil J. A. Sloane. 1988. *Sphere Packings, Lattices and Groups*, volume 290 of *Grundlehren der mathematischen Wissenschaften*. Springer.

Giovanni Costantini, Iacopo Iaderola, Andrea Paoloni, and Massimiliano Todisco. 2014. EMOVO corpus: an italian emotional speech database. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, pages 3501–3504. European Language Resources Association (ELRA).

Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. 2023. High fidelity neural audio compression. *Trans. Mach. Learn. Res.*, 2023.

Alexandre Défossez, Laurent Mazaré, Manu Orsini, Amélie Royer, Patrick Pérez, Hervé Jégou, Edouard Grave, and Neil Zeghidour. 2024. Moshi: A speech-text foundation model for real-time dialogue. *Preprint*, arXiv:2410.00037.

Yitian Gong, Luozhijie Jin, Kuangwei Chen, Dong Zhang, Ruifan Deng, Xiaogui Yang, Xin Zhang, Zhaoye Fei, Qinyuan Cheng, Shimin Li, and Xipeng Qiu. 2026. XY-Tokenizer: Mitigating the semantic-acoustic conflict in low-bitrate speech codecs. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9350–9369, San Diego, California, United States. Association for Computational Linguistics.

Albert Gu and Tri Dao. 2024. Mamba: Linear-time sequence modeling with selective state spaces. In *Proceedings of the Conference on Language Modeling (COLM)*.

Albert Gu, Karan Goel, and Christopher Ré. 2022. Efficiently modeling long sequences with structured state spaces. In *Proceedings of the International Conference on Learning Representations*.

Yifeng Han, Lin Li, and Jianwei Zhang. 2020. A coordinated representation learning enhanced multimodal machine translation approach with multi-attention. In *Proceedings of the 2020 International Conference on Multimedia Retrieval, ICMR 2020*, pages 571–577. ACM.

Haorui He, Zengqiang Shang, Chaoren Wang, Xuyuan Li, Yicheng Gu, Hua Hua, Liwei Liu, Chen Yang, Jiaqi Li, Peiyang Shi, Yuancheng Wang, Kai Chen, Pengyuan Zhang, and Zhizheng Wu. 2025. Emilia: A large-scale, extensive, multilingual, and diverse dataset for speech generation. *IEEE Transactions on Audio, Speech and Language Processing*, 33:4044–4054.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE ACM Trans. Audio Speech Lang. Process.*, 29:3451–3460.

Hangrui Hu, Xinfa Zhu, Ting He, Dake Guo, Bin Zhang, Xiong Wang, Zhifang Guo, Ziyue Jiang, Hongkun Hao, Zishan Guo, Xinyu Zhang, Pei Zhang, Baosong Yang, Jin Xu, Jingren Zhou, and Junyang Lin. 2026. Qwen3-tts technical report. *Preprint*, arXiv:2601.15621.

Shengpeng Ji, Ziyue Jiang, Wen Wang, Yifu Chen, Minghui Fang, Jialong Zuo, Qian Yang, Xize Cheng, Zehan Wang, Ruiqi Li, Ziang Zhang, Xiaoda Yang, Rongjie Huang, Yidi Jiang, Qian Chen, Siqi Zheng, and Zhou Zhao. 2025. WavTokenizer: An efficient acoustic discrete codec tokenizer for audio language modeling. In *Proceedings of the International Conference on Learning Representations*.

Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. 2020. HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis. In *Advances in Neural Information Processing Systems*.

Rithesh Kumar, Prem Seetharaman, Alejandro Luebs, Ishaan Kumar, and Kundan Kumar. 2023. High-fidelity audio compression with improved RVQGAN. In *Advances in Neural Information Processing Systems 36*.

Moreno La Quatra, Alkis Koudounas, Lorenzo Vaiani, Elena Baralis, Luca Cagliero, Paolo Garza, and Sabato Marco Siniscalchi. 2024. Benchmarking representations for speech, music, and acoustic events. In *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2024 - Workshops*, pages 505–509. IEEE.

Kushal Lakhotia, Evgeny Kharitonov, Wei-Ning Hsu, Yossi Adi, Adam Polyak, Benjamin Bolte, Tu Anh Nguyen, Jade Copet, Alexei Baevski, Adelrahman Mohamed, and Emmanuel Dupoux. 2021. On generative spoken language modeling from raw audio. *Transactions of the Association for Computational Linguistics*, 9:1336–1354.

Junhyeok Lee, Xiluo He, Jihwan Lee, Helin Wang, Shrikanth Narayanan, Thomas Thebaud, Laureano Moro-Velázquez, Jesús Villalba, and Najim Dehak. 2026. Reconstruct! don't encode: Self-supervised representation reconstruction loss for high-intelligibility and low-latency streaming neural audio codec. *Preprint*, arXiv:2603.05887.

Jiaqi Li, Xiaolong Lin, Zhekai Li, Shixi Huang, Yuancheng Wang, Chaoren Wang, Zhenpeng Zhan, and Zhizheng Wu. 2025. DualCodec: A low-frame-rate, semantically-enhanced neural audio codec for speech generation. In *International Speech Communication Association, Interspeech 2025*. ISCA.

Lin Li, Kaixi Hu, Yunpei Zheng, Jianquan Liu, and Kong Aik Lee. 2021. COOPNet: Multi-modal cooperative gender prediction in social media user profiling. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2021*, pages 4310–4314. IEEE.

Tzu-Quan Lin, Heng-Cheng Kuo, Tzu-Chieh Wei, Hsi-Chun Cheng, Chun Wei Chen, Hsien-Fu Hsiao, Yu Tsao, and Hung-yi Lee. 2026. An exploration of mamba for speech self-supervised models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2026*, pages 10337–10350. Association for Computational Linguistics.

Steven R. Livingstone and Frank A. Russo. 2018. The ryerson audio-visual database of emotional speech and song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in north american english. *PloS one*, 13(5):e0196391.

Ilya Loshchilov and Frank Hutter. 2017. SGDR: Stochastic gradient descent with warm restarts. In *Proceedings of the International Conference on Learning Representations*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *Proceedings of the*

International Conference on Learning Representations.

Xudong Mao, Qing Li, Haoran Xie, Raymond Y. K. Lau, Zhen Wang, and Stephen Paul Smolley. 2017. Least squares generative adversarial networks. In *International Conference on Computer Vision, ICCV 2017*, pages 2813–2821. IEEE Computer Society.

Fabian Mentzer, David Minnen, Eirikur Agustsson, and Michael Tschannen. 2024. Finite scalar quantization: VQ-VAE made simple. In *Proceedings of the International Conference on Learning Representations*.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. LibriSpeech: An ASR corpus based on public domain audio books. In *International Conference on Acoustics, Speech and Signal Processing, ICASSP 2015*, pages 5206–5210. IEEE.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning, ICML 2023*, pages 28492–28518. PMLR.

Antony W. Rix, John G. Beerends, Michael P. Hollier, and Andries P. Hekstra. 2001. Perceptual evaluation of speech quality (PESQ)–a new method for speech quality assessment of telephone networks and codecs. In *International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2001*, pages 749–752. IEEE.

Takaaki Saeki, Detai Xin, Wataru Nakata, Tomoki Koriyama, Shinnosuke Takamichi, and Hiroshi Saruwatari. 2022. UTMOS: UTokyo-SaruLab system for VoiceMOS challenge 2022. In *International Speech Communication Association, Interspeech 2022*, pages 4521–4525. ISCA.

Hubert Siuzdak. 2024. Vocos: Closing the gap between time-domain and fourier-based neural vocoders for high-quality audio synthesis. In *Proceedings of the International Conference on Learning Representations*.

Jianlin Su, Murtadha H. M. Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. 2024. RoFormer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063.

Cees H. Taal, Richard C. Hendriks, Richard Heusdens, and Jesper Jensen. 2010. A short-time objective intelligibility measure for time-frequency weighted noisy speech. In *International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, pages 4214–4217. IEEE.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. 2017. Neural discrete representation learning. In *Advances in Neural Information Processing Systems*, pages 6306–6315.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Junyu Wang, Zizhen Lin, Tianrui Wang, Meng Ge, Longbiao Wang, and Jianwu Dang. 2025a. Mamba-SEUNet: Mamba UNet for monaural speech enhancement. In *International Conference on Acoustics, Speech and Signal Processing, ICASSP 2025*, pages 1–5. IEEE.

Xinsheng Wang, Mingqi Jiang, Ziyang Ma, Ziyu Zhang, Songxiang Liu, Linqin Li, Zheng Liang, Qixi Zheng, Rui Wang, Xiaoqin Feng, Weizhen Bian, Zhen Ye, Sitong Cheng, Ruibin Yuan, Zhixian Zhao, Xinfa Zhu, Jiahao Pan, Liumeng Xue, Pengcheng Zhu, and 6 others. 2025b. Spark-tts: An efficient llm-based text-to-speech model with single-stream decoupled speech tokens. *arXiv preprint arXiv:2503.01710*.

Detai Xin, Xu Tan, Shinnosuke Takamichi, and Hiroshi Saruwatari. 2024. BigCodec: Pushing the limits of low-bitrate neural speech codec. *Preprint*, arXiv:2409.05377.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Zhen Ye, Peiwen Sun, Jiahe Lei, Hongzhan Lin, Xu Tan, Zheqi Dai, Qiuqiang Kong, Jianyi Chen, Jiahao Pan, Qifeng Liu, Yike Guo, and Wei Xue. 2025a. Codec does matter: Exploring the semantic shortcoming of codec for audio language model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 25697–25705. AAAI Press.

Zhen Ye, Xinfa Zhu, Chi-Min Chan, Xinsheng Wang, Xu Tan, Jiahe Lei, Yi Peng, Haohe Liu, Yizhu Jin, Zheqi Dai, Hongzhan Lin, Jianyi Chen, Xingjian Du, Liumeng Xue, Yunlin Chen, Zhifei Li, Lei Xie, Qiuqiang Kong, Yike Guo, and Wei Xue. 2025b. Llasa: Scaling train-time and inference-time compute for llama-based speech synthesis. *Preprint*, arXiv:2502.04128.

Lijun Yu, Jose Lezama, Nitesh B. Gundavarapu, Luca Versari, Kihyuk Sohn, David Minnen, Yong Cheng, Vighnesh Birodkar, Agrim Gupta, Xiuye Gu, Alexander G. Hauptmann, Boqing Gong, Ming-Hsuan Yang, Irfan Essa, David A. Ross, and Lu Jiang. 2024. Language model beats diffusion—tokenizer is key to visual generation. In *Proceedings of the International Conference on Learning Representations*.

Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi. 2022. SoundStream: An end-to-end neural audio codec. *IEEE ACM Trans. Audio Speech Lang. Process.*, 30:495–507.

Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. 2023. SpeechGPT: Empowering large language models with intrinsic cross-modal conversational abilities. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15757–15773. Association for Computational Linguistics.

Xiangyu Zhang, Jianbo Ma, Mostafa Shahin, Beena Ahmed, and Julien Epps. 2025a. Rethinking mamba in speech processing by self-supervised models. In *International Conference on Acoustics, Speech and Signal Processing, ICASSP 2025*, pages 1–5. IEEE.

Xiangyu Zhang, Qiquan Zhang, Hexin Liu, Tianyi Xiao, Xinyuan Qian, Beena Ahmed, Eliathamby Ambikairajah, Haizhou Li, and Julien Epps. 2025b. Mamba in speech: Towards an alternative to self-attention. *IEEE ACM Trans. Audio Speech Lang. Process.*, 33:1933–1948.

Xin Zhang, Lin Li, Xiangni Lu, Jianquan Liu, and Kong Aik Lee. 2026. Speaking clearly: A simplified whisper-based codec for low-bitrate speech coding. In *International Conference on Acoustics, Speech and Signal Processing, ICASSP 2026*, pages 17037–17041. IEEE.

Xin Zhang, Dong Zhang, Shimin Li, Yaqian Zhou, and Xipeng Qiu. 2024. SpeechTokenizer: Unified speech tokenizer for speech language models. In *Proceedings of the International Conference on Learning Representations*.

Yue Zhao, Hanwen Jiang, Zhenlin Xu, Chutong Yang, Ehsan Adeli, and Philipp Krähenbühl. 2026. Spherical leech quantization for visual tokenization and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12913–12923.

Yue Zhao, Yuanjun Xiong, and Philipp Krähenbühl. 2025. Image and video tokenization with binary spherical quantization. In *Proceedings of the International Conference on Learning Representations*.

A Details of Architecture

A.1 Encoder

The input waveform is resampled to 16 kHz and converted to an 80-channel mel-spectrogram using a 25 ms window and a 10 ms hop. The trainable codec encoder follows a Whisper-style convolutional front-end (Radford et al., 2023): two 1D convolutional layers with kernel size 3 project the input to 768 dimensions and reduce the frame rate from 100 Hz to 50 Hz. The resulting sequence is processed by 8 bidirectional Mamba-1 blocks (Gu and Dao, 2024; Zhang et al., 2025b) with hidden dimension 768, state dimension 16, convolution width 4, expansion factor 2, RMSNorm, and a 1536-dimensional SwiGLU feed-forward layer.

A.2 Down/Upsampling

The quantizer operates at 12.5 Hz. The downsampler stacks every four adjacent 50 Hz encoder frames along the channel dimension and maps the resulting representation to a 128-dimensional latent sequence z_e through weight-normalized 1D projections and residual convolutional units with dilations 1, 3, and 9, following common neural audio codec designs (Zeghidour et al., 2022; Défossez et al., 2023).

The upsampler mirrors this module. It projects the quantized 128-dimensional latent back to 768×4 channels, applies the same residual convolutional units, and unstacks frames to recover a 50 Hz sequence before decoding. No attention-based projection is used in either sampling module.

A.3 Quantizer

The bottleneck uses 5 RSLQ levels with a fixed spherical Leech codebook of 196,560 entries per level (Zhao et al., 2026; Conway and Sloane, 1988). The first level is used as the semantic SLQ level, and the remaining four levels form the residual acoustic path. The main text gives the RSLQ computation in Sec. 3.3.

A.4 Decoder

The decoder mirrors the encoder. It applies 8 bidirectional Mamba-1 blocks to the upsampled 50 Hz sequence and reconstructs an 80-channel mel-spectrogram with two transposed 1D convolutional layers. A jointly trained Vocos vocoder (Siuzdak, 2024) with 12 layers, hidden dimension 512, intermediate dimension 4096, FFT size 640, and hop size 160 converts the reconstructed mel-spectrogram into a 16 kHz waveform.

A.5 Semantic Teacher

The semantic branch is only used during training. We instantiate the frozen teacher as either Whisper-small (Radford et al., 2023) or SenseVoice-small (An et al., 2024) and use its representations for semantic distillation and reconstruction alignment, as described in Sec. 3.4. The teacher encoder and the auxiliary projection/alignment paths are removed at inference, leaving the acoustic codec as a single-tower encoder–RSLQ–decoder system.

B RSLQ Codebook Behavior

We analyze the validation-time codebook behavior of the five RSLQ levels to verify that the fixed Leech codebook (Zhao et al., 2026; Conway and Sloane, 1988) is effectively used under the proposed split-residual design. Figure 5 reports the codebook usage rate and normalized entropy during training. The first level, which receives semantic supervision, maintains a lower but stable usage rate and a normalized entropy of about 0.81. This pattern is expected because semantic distillation encourages a more concentrated token distribution over linguistically relevant regions of the codebook, rather than uniformly spreading probability mass across all entries. In contrast, the four residual acoustic levels reach substantially higher usage rates and normalized entropy values above 0.94, indicating broad utilization of the fixed lattice codebook for speaker, prosodic, and fine acoustic residual information.

These results support the intended division of labor in RSLQ. The semantic level forms a compact supervised allocation, while the acoustic residual levels preserve high-entropy code usage. Importantly, all levels remain stable after convergence, suggesting that the fixed Leech lattice avoids the severe under-utilization and collapse commonly associated with learnable VQ/RVQ codebooks (van den Oord et al., 2017; Zeghidour et al., 2022; Défossez et al., 2023; Kumar et al., 2023).

C Training Objective Details

To accelerate training, we pre-extract semantic teacher representations for the full LibriSpeech training corpus (Panayotov et al., 2015) and store them offline. For the Whisper-small variant (Radford et al., 2023), we store the 50 Hz encoder features directly. For the SenseVoice-small variant (An et al., 2024), the teacher produces 16.67

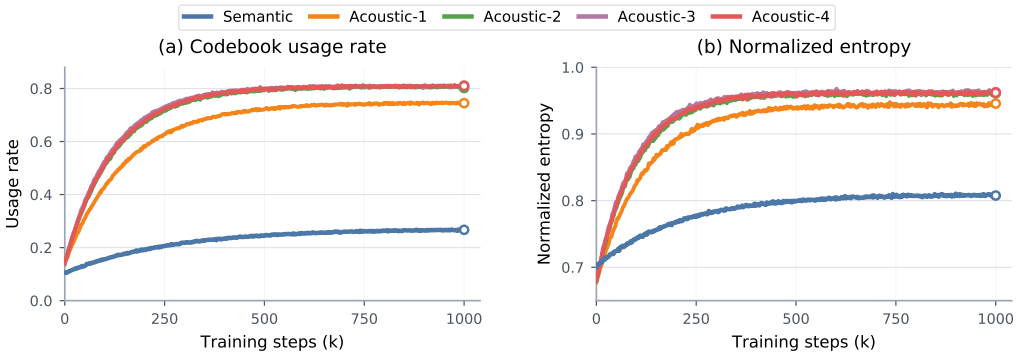

Figure 5: Validation codebook behavior of the five RSLQ levels. The semantic level shows lower but stable usage and normalized entropy under semantic supervision, while the four residual acoustic levels maintain higher utilization and entropy throughout training.

Hz features; we linearly interpolate them to the 12.5 Hz codec token rate before storage, so that semantic distillation can be applied at the RSLQ bottleneck resolution. During generator training, both teacher-supervised variants use the same loss coefficients: $\lambda_{\text{sem}} = 15$, $\lambda_{\text{align}} = 1$, $\lambda_{\text{rec}} = 15$, $\lambda_{\text{adv}} = 1$, and $\lambda_{\text{feat}} = 1$. The frozen teacher features are used only for auxiliary semantic supervision and reconstruction alignment, and no teacher branch is used at inference.

C.1 Semantic Projector

The lightweight projector P_ψ depends on the frozen semantic teacher. For the Whisper-small variant (Radford et al., 2023), the teacher encoder produces 50 Hz features, while the semantic RSLQ level operates at 12.5 Hz. We therefore implement P_ψ with the same linear upsampling module used in the codec, converting the 12.5 Hz semantic sequence to 50 Hz before aligning it with Whisper encoder features. For the SenseVoice-small variant (An et al., 2024), P_ψ is a single linear layer that maps the semantic RSLQ embedding to the SenseVoice feature dimension.

C.2 Reconstruction Loss

We use a multi-scale mel-spectrogram reconstruction loss, following common practice in neural speech generation and codec training (Kong et al., 2020; Défossez et al., 2023). For each STFT scale $k \in \{5, \ldots, 11\}$, let $M_k(\cdot)$ denote the mel-spectrogram computed with FFT size 2^k. Given the original audio x and reconstructed audio $\hat{x}$, the reconstruction loss is

$$\mathcal{L}_{\text{rec}} = \sum_{k=5}^{11} \|M_k(x) - M_k(\hat{x})\|_1. \tag{15}$$

This objective focuses reconstruction learning on perceptually relevant spectral structure. We do not use an additional time-domain waveform L1 term.

C.3 Adversarial and Feature-Matching Losses

We use a multi-period discriminator (MPD) (Kong et al., 2020) and a multi-scale short-time Fourier transform discriminator (MS-STFTD) (Défossez et al., 2023). Let $D_i(\cdot)$ denote the output of the i-th discriminator and let N be the number of discriminators. The discriminator is optimized with the least-squares GAN objective (Mao et al., 2017):

$$\mathcal{L}_D = \frac{1}{N} \sum_{i=1}^{N} \left[(D_i(x) - 1)^2 + D_i(\hat{x})^2 \right]. \tag{16}$$

The generator adversarial loss is

$$\mathcal{L}_{\text{adv}} = \frac{1}{N} \sum_{i=1}^{N} (D_i(\hat{x}) - 1)^2. \tag{17}$$

We also apply feature matching on intermediate discriminator activations. Let $D_i^j(\cdot)$ denote the j-th layer feature map from discriminator D_i, let K be the number of feature layers, and let ϵ be a small constant for numerical stability. The feature-matching loss is

$$\mathcal{L}_{\text{feat}} = \frac{1}{NK} \sum_{i=1}^{N} \sum_{j=1}^{K} \frac{\|D_i^j(x) - D_i^j(\hat{x})\|_1}{\|D_i^j(x)\|_1 + \epsilon}, \tag{18}$$

which stabilizes adversarial training by matching real and reconstructed audio in discriminator feature space. Since RSLQ has no learnable codebook, no VQ codebook loss, commitment loss, or entropy regularization is added.

Model	Codebook Size	bps	Frame Rate	N_q	Single Tower	Param	SIM↑	STOI↑	PESQ NB↑	PESQ WB↑	UTMOS↑	WER(%)↓
Ground Truth	–	–	–	–	–	–	1.00	1.00	4.55	4.64	3.50	4.59
EnCodec (24 kHz)	1024	1500	75	2	✓	15M	0.59	0.83	1.96	1.58	1.48	17.19
DAC (16 kHz)	1024	1500	50	3	✓	74M	0.45	0.79	1.61	1.26	1.42	22.31
SpeechTokenizer (16 kHz)	1024	1000	50	2	✓	103.7M	0.32	0.74	1.51	1.23	1.93	14.71
BigCodec (16 kHz)	8192	1040	80	1	✓	159M	0.82	0.91	3.03	2.44	3.55	8.29
Mimi (24 kHz)	2048	1100	12.5	8	✓	79M	0.71	0.88	2.64	2.12	3.07	9.41
XCodec (16 kHz)	1024	1000	50	2	×	160M	0.67	0.85	2.50	1.99	3.66	7.02
XCodec2.0 (16 kHz)	65536	800	50	1	×	820M	0.81	0.90	2.83	2.26	3.64	6.85
DualCodec (24 kHz)	16384/4096	1075	12.5	1/6	×	664M	0.83	0.91	3.07	2.54	3.65	6.26
XY-Tokenizer (16 kHz)	1024	1000	12.5	8	×	520M	0.84	0.90	2.88	2.28	3.46	6.19
SimWhisper-Codec (16 kHz)	2016	1100	12.5	8	✓	291M	0.81	0.92	3.12	2.56	3.49	7.61
BiMTokenizer-Whisper (Ours, 16 kHz)	196560	1100	12.5	5	✓	253M	**0.85**	**0.93**	**3.38**	**2.83**	3.75	**6.10**
BiMTokenizer-SenseVoice (Ours, 16 kHz)	196560	1100	12.5	5	✓	253M	0.83	0.92	3.25	2.63	**3.76**	6.32

Table 5: Speech reconstruction comparison on the noisier LibriSpeech *test-other* set. Best codec results are in bold, and second-best results are underlined.

Model	kbps	SeedTTS-ZH					SeedTTS-EN				
		PESQ-NB↑	PESQ-WB↑	SIM↑	STOI↑	UTMOS↑	PESQ-NB↑	PESQ-WB↑	SIM↑	STOI↑	UTMOS↑
Ground Truth	–	4.55	4.64	1.00	1.00	2.78	4.55	4.64	1.00	1.00	3.53
BigCodec (16 kHz)	1.04	2.88	2.26	0.80	0.91	2.90	2.80	2.21	0.80	0.91	3.59
XCodec2.0 (16 kHz)	0.80	2.85	2.23	0.85	0.90	2.89	2.67	2.09	0.81	0.90	3.69
XY-Tokenizer (16 kHz)	1.00	2.98	2.31	**0.88**	0.91	2.78	2.75	2.18	**0.84**	0.90	3.59
SimWhisper-Codec (16 kHz)	1.10	3.08	2.44	0.82	0.92	2.99	2.89	2.28	0.80	0.92	3.61
BiMTokenizer-Whisper (Ours, 16 kHz)	1.10	**3.30**	**2.66**	0.85	**0.93**	3.20	**3.13**	**2.51**	**0.84**	**0.93**	3.89
BiMTokenizer-SenseVoice (Ours, 16 kHz)	1.10	3.15	2.49	0.82	0.92	**3.23**	2.93	2.28	0.80	0.92	**3.91**

Table 6: Speech reconstruction comparison on Seed-TTS-Eval dataset. ZH/EN split results are reported in the same row for each model. Best codec results are in bold.

D Evaluation Details

For model-level comparison on LibriSpeech *test-clean* (Panayotov et al., 2015), we report codebook size, bitrate in bits per second (bps), frame rate, number of quantizers (N_q), single-tower architecture status, and parameter count. Speech intelligibility is measured by Short-Time Objective Intelligibility (STOI) (Taal et al., 2010) and word error rate (WER). WER transcriptions are obtained with a HuBERT-based ASR model (Hsu et al., 2021).[1]

Acoustic quality is measured by narrow-band and wide-band Perceptual Evaluation of Speech Quality (PESQ-NB/PESQ-WB) (Rix et al., 2001), UTMOS (Saeki et al., 2022), and the Virtual Speech Quality Objective Listener (ViSQOL) (Chinen et al., 2020). We also report speaker similarity (SIM), computed as the cosine similarity between speaker embeddings extracted from the original and reconstructed speech using a pretrained speaker verification model.[2]

[1] https://huggingface.co/facebook/hubert-large-ls960-ft

[2] https://github.com/microsoft/UniSpeech/tree/main/downstreams/speaker_verification

E Details of LLM-Based Speech Generation

E.1 Speech Generation Model and Training Details

For autoregressive speech generation, we retrain a TTS-oriented variant of BiMTokenizer on the English and Chinese subsets of Emilia (He et al., 2025), totaling approximately 96.7K hours of speech (46.8K English and 49.9K Chinese). To make token-level language-model training tractable, we subsample the original spherical Leech codebook to 2,048 entries at every quantization level. The tokenizer uses eight quantization levels: one semantic level followed by seven acoustic residual levels. At a frame rate of 12.5 Hz, each level contributes 11 bits per frame, yielding a total bitrate of $12.5 \times 8 \times 11 = 1{,}100$ bps. We adopt Qwen3-0.6B (Yang et al., 2025) as the language-model backbone and follow the autoregressive training formulation of Qwen3-TTS (Hu et al., 2026).

During training, text tokens and speech tokens are concatenated along the temporal dimension, and the autoregressive loss is computed only on the speech tokens. The model is trained on the VoxBox dataset proposed in Spark-TTS (Wang

et al., 2025b). We use a maximum token budget of 131,072 per training step, train the model for 400k steps, and set the maximum learning rate to 2×10^{-4}.

E.2 Evaluation Protocol

We evaluate zero-shot speech generation on the Seed-TTS-Eval benchmark (Anastassiou et al., 2024), including both the English (*test-en*) and Chinese (*test-zh*) subsets. We report word error rate (WER), speaker similarity (SIM), and UTMOS. During inference, the model is conditioned on the prompt text, reference speech, and target text, and autoregressively generates the target speech tokens, which are then decoded into waveforms by the BiMTokenizer decoder.

F Reconstruction under Noisy and Out-of-Distribution Conditions

Noisy Conditions. To further evaluate robustness under more challenging acoustic conditions, we conduct additional reconstruction experiments on the full LibriSpeech *test-other* set (Panayotov et al., 2015). Compared with *test-clean*, this subset contains noisier and more variable speech, making it a useful setting for testing whether a codec can preserve intelligibility, speaker information, and perceptual quality beyond clean read speech.

The ground-truth recordings already show a clear degradation from *test-clean*: UTMOS (Saeki et al., 2022) decreases from 4.09 to 3.50, and WER increases from 2.16% to 4.59%. Most codecs therefore exhibit lower STOI (Taal et al., 2010) and PESQ (Rix et al., 2001) scores and higher WER under this setting. Nevertheless, BiMTokenizer remains strong on noisy speech. The Whisper-supervised variant achieves the best codec result on SIM, STOI, PESQ-NB, PESQ-WB, and WER, showing stronger reconstruction quality and intelligibility than the competing codecs. The SenseVoice-supervised variant obtains the highest UTMOS, although its gains over the Whisper-supervised variant are limited to this metric.

Out-of-Distribution To further assess generalization beyond the training distribution, we evaluate BiMTokenizer on Seed-TTS-Eval (Anastassiou et al., 2024), which covers both English (*seedtts-test-en*) and Mandarin Chinese (*seedtts-test-zh*). Since our codec is trained exclusively on English LibriSpeech (Panayotov et al., 2015), seedtts-test-en probes generalization across speakers, recording conditions, and content style, whereas seedtts-test-zh additionally probes *cross-lingual* generalization to a language entirely unseen during training. Following common practice in speech codec evaluation on this benchmark, we report PESQ-NB/PESQ-WB (Rix et al., 2001), SIM, STOI (Taal et al., 2010), and UTMOS (Saeki et al., 2022).

As shown in Table 6, BiMTokenizer attains the strongest results on the majority of metrics on both subsets. BiMTokenizer-Whisper leads on PESQ-NB, PESQ-WB, and STOI on both seedtts-test-en (3.13, 2.51, 0.93) and seedtts-test-zh (3.30, 2.66, 0.93), and ties XY-Tokenizer (Gong et al., 2026) for the best SIM on seedtts-test-en (0.84). BiMTokenizer-SenseVoice obtains the highest UTMOS on both subsets (3.91 on en, 3.23 on zh). The only metric on which BiMTokenizer does not lead is SIM on seedtts-test-zh, where XY-Tokenizer reaches 0.88 versus our 0.85; this is consistent with XY-Tokenizer being trained on a corpus that includes Mandarin speakers, whereas our model has had no exposure to Mandarin speaker characteristics during training, and the gap is small in absolute terms. Importantly, the codec-internal ranking is preserved across the two languages, indicating that the learned representations encode acoustic and phonetic regularities that transfer beyond the training language rather than fitting language-specific patterns of LibriSpeech English.

These results indicate that the proposed single-tower architecture is not only effective on clean speech but also robust under more difficult acoustic conditions and generalizable to out-of-distribution and even cross-lingual data. Compared with recent dual-tower codecs (Li et al., 2025; Gong et al., 2026; Chen et al., 2026), BiMTokenizer achieves stronger results on most reconstruction metrics while retaining a single-tower architecture. This further supports our main claim that the single-tower route has not been exhausted, and that improving the temporal backbone and quantization bottleneck can substantially strengthen low-bitrate speech coding.

G Computational Efficiency Details

We provide two complementary efficiency analyses. The first isolates the sequence backbone by comparing BiMamba (Gu and Dao, 2024; Zhang et al., 2025b) with a bidirectional Transformer counterpart (Vaswani et al., 2017), and the second measures the complete encoding and decoding cost

Group	Teacher	λ_{sem}	λ_{align}	Backbone	Quantizer	Reconstruction				Understanding	
						SIM↑	UTMOS↑	PESQ↑	WER↓	SLURP↑	AM↑
Core Design	–	0	0	BiMamba	RSLQ	**0.88**	**4.12**	**3.56/2.99**	**2.55**	**7.94**	**78.26**
	–	0	0	Bi-Self-Attn	RSLQ	0.85	3.90	3.28/2.77	4.15	7.62	75.18
	–	0	0	BiMamba	GroupFSQ	0.84	4.03	3.42/2.84	2.69	7.75	75.31
	–	0	0	BiMamba	RVQ	0.84	4.00	3.42/2.75	2.75	7.81	76.26
SenseVoice	SenseVoice	0	1	BiMamba	RSLQ	**0.86**	4.15	**3.50/2.95**	2.56	8.58	81.25
	SenseVoice	15	0	BiMamba	RSLQ	0.84	**4.18**	3.46/2.89	2.58	18.22	97.79
	SenseVoice	15	1	BiMamba	RSLQ	0.85	**4.18**	3.45/2.85	**2.53**	18.48	**98.02**
	SenseVoice	20	1	BiMamba	RSLQ	0.82	4.15	3.42/2.78	2.55	**19.90**	97.46
Whisper	Whisper	5	1	BiMamba	RSLQ	0.86	4.07	3.46/2.86	2.51	10.89	84.65
	Whisper	10	1	BiMamba	RSLQ	0.86	4.10	3.48/2.92	2.51	12.24	86.28
	Whisper	15	1	BiMamba	RSLQ	**0.87**	**4.21**	**3.56/3.03**	**2.44**	12.51	86.20
	Whisper	20	1	BiMamba	RSLQ	**0.87**	4.18	3.52/2.99	2.52	**13.68**	**87.11**

Table 7: Component analysis on reconstruction and semantic representation. The core-design block removes semantic supervision and first reports the full BiMamba+RSLQ reconstruction-only setting, followed by backbone and quantizer replacements. The SenseVoice block studies semantic distillation, reconstruction alignment, and distillation weights under the SenseVoice teacher. The Whisper block varies the distillation weight with reconstruction alignment enabled. PESQ reports PESQ-NB/PESQ-WB. Best and second-best values are marked within each block for each metric. AM denotes AudioMNIST.

of each codec. We report Multiply-Accumulate operations (MACs) for arithmetic complexity and Real-Time Factor (RTF) for wall-clock inference speed relative to audio duration. Following common neural audio codec evaluation practice (Zeghidour et al., 2022; Défossez et al., 2023; Kumar et al., 2023), MACs are computed for a 1-second audio input and reported in G/s. We use PyFlops for supported modules and manually calculate modules not supported by PyFlops, such as state-space model (SSM) blocks (Gu and Dao, 2024). All measurements are conducted with batch size 1 on a single NVIDIA H100 GPU. For the codec-level comparison in Table 4, both MACs and RTF are measured on LibriSpeech test-clean (Panayotov et al., 2015).

Figure 4 isolates the computational effect of the sequence backbone. The BiMamba (Zhang et al., 2025b) and Transformer (Vaswani et al., 2017) variants share the same codec encoder front-end, linear downsampling and upsampling modules, RSLQ bottleneck, decoder, and vocoder. The only architectural change is that each bidirectional Mamba block is replaced by a bidirectional self-attention block with RoPE (Su et al., 2024). This controlled setting ensures that the comparison reflects the backbone design rather than differences in temporal context, quantization, or decoder structure. MACs and RTF are measured for audio inputs of 5, 20, 40, 80, 160, and 320 seconds, with all non-backbone codec components kept unchanged.

Table 4 reports codec-level efficiency. Parameter count is listed once per model, while encoding and decoding costs are reported separately. Encoding corresponds to waveform-to-token conversion, and decoding corresponds to waveform reconstruction from the discrete tokens. Total MACs and total RTF are obtained by summing the corresponding encoding and decoding values, which reflects the full inference cost of tokenization followed by reconstruction.

MACs and RTF provide complementary measures. MACs quantify arithmetic complexity, whereas RTF also depends on hardware and operator implementation. Reporting both therefore separates the model's computational footprint from its realized wall-clock speed. Section 5 discusses the quantitative comparisons and their implications.

H Component Analysis Details

Table 7 is organized to isolate component choices from supervision choices. The core-design block removes semantic supervision and evaluates every variant at the same 1.1 kbps bitrate. In the backbone ablation, only the BiMamba blocks are replaced by bidirectional RoPE-based self-attention (Vaswani et al., 2017; Su et al., 2024); all other codec components remain unchanged. In the quantizer ablations, GroupFSQ (Casanova et al., 2025) uses 8 groups with per-group levels $[8, 7, 6, 6]$, while RVQ uses 8 quantizer layers with a codebook size of 2048 per layer, following standard neural codec practice (Zeghidour et al., 2022; Défossez et al., 2023; Kumar et al., 2023). These controlled settings en-

sure that the observed differences can be attributed to the backbone or quantizer rather than to bitrate or semantic supervision.

Within the reconstruction-only block, BiMamba with RSLQ achieves the strongest result on every reconstruction metric. Replacing BiMamba with self-attention primarily affects intelligibility and spectral fidelity, while replacing RSLQ with GroupFSQ or RVQ produces smaller but consistent degradations across SIM, UTMOS, PESQ, and WER. The downstream understanding scores in this block are reported for completeness, but they should not be interpreted as measures of semantic supervision because none of these variants uses a semantic teacher.

The teacher-specific blocks retain the BiMamba+RSLQ architecture and vary only the semantic objectives. In the SenseVoice block (An et al., 2024), introducing semantic distillation at $\lambda_{\text{sem}} = 15$ increases SLURP accuracy from 8.58 to 18.48 and AudioMNIST accuracy from 81.25 to 98.02, while preserving competitive reconstruction quality. Comparing the two settings with $\lambda_{\text{sem}} = 15$ further isolates the contribution of reconstruction alignment. Enabling this objective improves WER from 2.58% to 2.53% and AudioMNIST accuracy from 97.79 to 98.02. Increasing λ_{sem} to 20 yields the highest SLURP accuracy, but reduces SIM and PESQ, indicating that excessive semantic supervision compromises acoustic preservation.

The Whisper block (Radford et al., 2023) exhibits a similar trend. Increasing λ_{sem} from 5 to 15 improves reconstruction, intelligibility, and SLURP performance. A further increase to 20 produces the best SLURP and AudioMNIST scores in this block, but slightly degrades UTMOS, PESQ, and WER. We therefore use $\lambda_{\text{sem}} = 15$ and $\lambda_{\text{align}} = 1$ for both final variants, which provides a consistent operating point with a favorable balance between semantic retention and acoustic fidelity.