

BridgeCodec: Mamba Enhanced Neural Audio Codec with Schrödinger Bridge at Low Bitrate

Zijian Lin¹, Jing Yang², Jinghao Luo², Zhuo Wang², Fan Fan², Zhiyong Wu^{1,**}

¹ Shenzhen International Graduate School, Tsinghua University, China

² Central Media Technology Institute, Huawei

linzj24@mails.tsinghua.edu.cn, zywu@sz.tsinghua.edu.cn

Abstract

Neural audio codecs achieve remarkable performance but operate as closed ecosystems with rigidly co-trained encoder-decoder pairs, severely hindering interoperability. To resolve this, we propose BridgeCodec, a versatile framework that decouples these pairs by translating latent representations between mismatched, frozen endpoints. We formulate this cross-codec translation as a Schrödinger Bridge optimal transport problem, enabling robust mapping between disparate distributions without predefined noise priors. To efficiently capture long-range speech temporal dependencies, we integrate a Mamba-enhanced U-Net architecture. We validate BridgeCodec on an extreme deployment scenario: mapping a lightweight 8 kHz source encoder directly to a 48 kHz target decoder at an ultra-low bitrate of 1 kbps. Experimental results demonstrate that BridgeCodec achieves superior wideband reconstruction and high perceptual quality, successfully bridging heavily bottlenecked, heterogeneous codecs.¹

Index Terms: low-bitrate audio codec, Schrödinger bridge, mamba-enhanced model

1. Introduction

Neural audio codecs (NACs) have revolutionized audio processing by achieving a unified framework for efficient compression and high-fidelity reconstruction [1]. Beyond traditional transmission, these systems now serve as critical intermediaries in downstream speech synthesis systems. State-of-the-art NACs typically employ Residual Vector Quantization (RVQ) [1, 2, 3, 4] to capture fine-grained acoustic details. Recent advancements further branch into two paradigms: integrating semantic information via multi-codebook designs to enhance LLM prediction [5, 6, 7, 8], and exploring single-layer VQ to push the boundaries of ultra-low bitrates [9, 10, 11, 12].

Despite their remarkable performance, current NACs operate as closed ecosystems relying on strictly co-trained encoder-decoder pairs. This coupling creates a severe "Interoperability Gap" across heterogeneous networks, preventing communication between devices with codecs deployed independently for with asymmetric hardware capabilities or varying compute constraints. Since simultaneously updating these structurally frozen endpoints may be practically infeasible and risks systemic failures, direct communication is fundamentally broken. To resolve this, we propose *BridgeCodec*, a versatile framework that completely decouples codec pairs by directly translating highly compressed latent representations between any mismatched, frozen endpoints. To demonstrate its robustness in

extreme deployment scenarios, we rigorously validate BridgeCodec on a highly asymmetric testbed: mapping a lightweight 8 kHz encoder to a 48 kHz decoder at an ultra-low 1 kbps. This confirms its capability to seamlessly bridge structurally heterogeneous latents and fully activate the generative priors of the target decoder.

To decouple encoder and decoder, we draw inspiration from the recent success of the Schrödinger Bridge (SB) [13, 14] in diverse speech tasks, including generation [15], enhancement [16, 17], bandwidth extension [18, 19] and speech vocoder [20], where it has been extensively validated to outperform flow matching [21] and standard diffusions [22, 23]. Therefore, we formulate our latent transformation as an SB optimal transport problem. Crucially, unlike standard score-based and flow models that require predefined noise priors, SB natively maps highly compressed, disparate latent spaces by adaptively learning optimal transport between empirical distributions. Moreover, capturing the long-range temporal dependencies intrinsic to speech (e.g., prosody, intonation, and phonetic coherence) is critical for high-fidelity reconstruction. Extensive recent literature [24, 25, 26] has firmly established the superiority of Mamba [27] in modeling extended speech contexts. Based on these findings, we integrate Mamba into a U-Net backbone [28]. This tailored architecture enables BridgeCodec to capture global temporal dynamics with linear complexity, strategically bypassing the localized receptive fields of convolutions and the quadratic computational bottlenecks of Transformers. Finally, we maximize perceptual quality through a two-stage training strategy: structural latent-space alignment followed by audio-domain refinement. Experimental results demonstrate that BridgeCodec successfully bridges the mismatched codecs under stringent low-bitrate constraints. Remarkably, it sustains high perceptual quality even with a single function evaluation (1 NFE), empirically validating the structural directness of our chosen SB transport path and robustness of the framework.

Our contributions are summarized as follows. First, we propose *BridgeCodec*, a SB based framework that decouples encoder-decoder to enable direct speech communication between mismatched endpoints. Second, we introduce a Mamba-enhanced U-Net architecture tailored to efficiently capture long-range speech dependencies within compressed representations. Third, leveraging a two-stage training strategy, we demonstrate that BridgeCodec achieves high-fidelity 48 kHz wideband reconstruction from 8 kHz bottlenecked priors at 1 kbps.

2. Methodology

2.1. Overall Architecture

Neural audio codecs (NACs) typically compress an audio waveform $X \in \mathbb{R}^{T \times x}$ via an encoder $\text{En}(\cdot)$ into a latent represen-

^{**} indicates the corresponding author.

¹ Demo: <https://thuhsai.github.io/interspeech2026-BridgeCodec/>

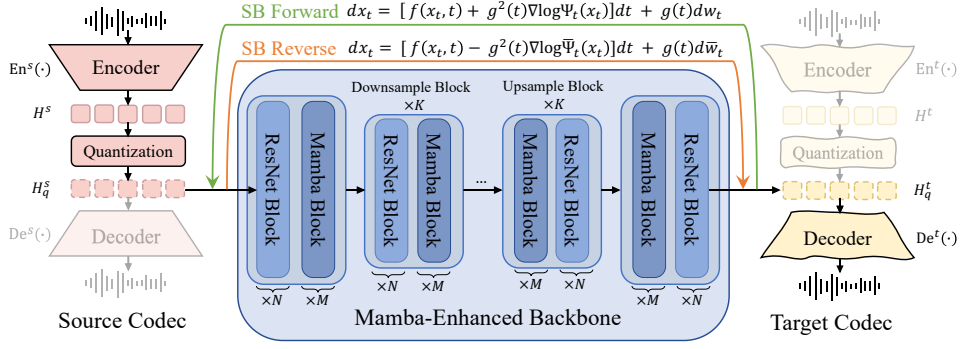


Figure 1: The BridgeCodec framework consists of $En^s(\cdot)$, Mamba-enhanced backbone, and $De^t(\cdot)$. The source codec is marked with straight lines and the target codec with wavy lines, indicating they can use different model architectures and quantization methods.

tation $H \in \mathbb{R}^{D_c \times T_H}$, where D_c and T_H denote the channel dimension and frame length, respectively. A quantizer with L vector quantization (VQ) layers then maps H into discrete tokens $Q \in \mathbb{R}^{L \times T_H}$. The embeddings corresponding to these tokens are summed across all L layers to reconstruct the quantized continuous feature $H_q \in \mathbb{R}^{D_c \times T_H}$. Finally, a decoder $De(\cdot)$ reconstructs the audio X_q from H_q .

To achieve interoperability between heterogeneous codecs, BridgeCodec fundamentally decouples the source encoder $En^s(\cdot)$ and the target decoder $De^t(\cdot)$. We formulate an intermediary latent translator using the Schrödinger Bridge (SB) [13, 14]. As illustrated in Fig. 1, a Mamba-enhanced backbone translates the source latent feature H_q^s directly into the target latent feature H_q^t . Because this transformation operates entirely in the latent space, the source and target codecs can maintain entirely different architectures, dimensions, and quantization paradigms. During deployment, frozen $En^s(\cdot)$, SB translator, and frozen $De^t(\cdot)$ seamlessly compose a unified and flexible asymmetric communication pipeline.

In this work, we validate BridgeCodec on a rigorous asymmetric testbed: connecting a lightweight, low-bitrate narrowband (8 kHz) source encoder to a high-fidelity wideband (48 kHz) target decoder. In this demanding setup, the SB-driven translator is tasked with not only aligning structurally mismatched latent dimensions but also performing latent-space bandwidth extension. By effectively modeling and compensating for the high-frequency semantics missing in the heavily bottlenecked H_q^s prior, BridgeCodec robustly recovers 48 kHz wideband audio without requiring any structural modifications to the endpoints.

2.2. Schrödinger Bridges

The core challenge in BridgeCodec is establishing a direct mapping between the source latent H_q^s and the target latent H_q^t . Standard diffusions are mathematically ill-suited for this task, as they constrain the terminal distribution to a Gaussian noise prior ($p_T = \mathcal{N}(0, \mathbf{I})$), which destructively overwrites the structural speech priors preserved in H_q^s . Inspired by recent success of SB in speech tasks [15, 16, 17, 18, 19, 20], we bypass this limitation by formulating the latent translation as a SB optimal transport problem. Because SB adaptively learns transitions between arbitrary empirical distributions without predefined noise priors, it uniquely qualifies to directly and robustly bridge two structurally mismatched, compressed latent spaces.

Formally, SB finds an optimal path measure $p \in \mathcal{P}_{[0, T]}$ that minimizes the Kullback-Leibler divergence with a refer-

ence path p_{ref} , while satisfying boundary constraints $p_0 = p_{\text{target}}$ and $p_T = p_{\text{source}}$:

$$\min_{p \in \mathcal{P}_{[0, T]}} D_{\text{KL}}(p \parallel p_{\text{ref}}) \quad \text{s.t.} \quad p_0 = p_{\text{target}}, p_T = p_{\text{source}}, \quad (1)$$

where p_{source} and p_{target} are the two distributions to be connected. The path measure solution to Equation 1 is a pair of forward-backward stochastic differential equations (SDEs):

$$dx_t = [f(x_t, t) + g^2(t) \nabla \log \Psi_t(x_t)] dt + g(t) dw_t, \quad (2)$$

$$dx_t = [f(x_t, t) - g^2(t) \nabla \log \bar{\Psi}_t(x_t)] dt + g(t) d\bar{w}_t, \quad (3)$$

where $f(x_t, t)$ and $g(t)$ are inherited from the reference SDE defining p_{ref} ; dw_t and $d\bar{w}_t$ denote distinct Wiener processes capturing stochasticity. The forward and backward evolution are characterized by distinct nonlinear drift terms $+g^2(t) \nabla \log \Psi_t(x_t)$ and $-g^2(t) \nabla \log \bar{\Psi}_t(x_t)$.

The marginal distribution of x_t is given by $p_t = \Psi_t \bar{\Psi}_t$ for any $t \in [0, T]$.

SB has tractable closed-form solutions under linear drift $f(x_t) = f(t)x_t$ and Gaussian boundaries, avoiding general computational intractability [15]. Let $p_0 = \mathcal{N}(x_0, \epsilon_0^2 \mathbf{I})$ and $p_T = \mathcal{N}(x_T, \epsilon_T^2 \mathbf{I})$, with $\epsilon_T = e^{\int_0^T f(\tau) d\tau} \epsilon_0$. As $\epsilon_0 \rightarrow 0$, the forward/backward terms $\Psi_t, \bar{\Psi}_t$ can be formulated as $\bar{\Psi}_t = \mathcal{N}(\alpha_t x_0, \alpha_t^2 \sigma_t^2 \mathbf{I})$, $\Psi_t = \mathcal{N}(\bar{\alpha}_t x_T, \bar{\alpha}_t^2 \bar{\sigma}_t^2 \mathbf{I})$, where $\alpha_t = e^{\int_0^t f(\tau) d\tau}$, $\sigma_t^2 = \int_0^t g^2(\tau) / \alpha_\tau^2 d\tau$, $\bar{\alpha}_t = \alpha_t \alpha_T^{-1}$, $\bar{\sigma}_t^2 = \sigma_T^2 - \sigma_t^2$. The marginal $p_t = \Psi_t \bar{\Psi}_t$ is thus formulated as:

$$p_t = \mathcal{N}(w_0(t)x_0 + w_T(t)x_T, \alpha_t^2 \sigma_t^2 \bar{\sigma}_t^2 / \sigma_T^2 \mathbf{I}), \quad (4)$$

where, $w_0(t) = \alpha_t \bar{\sigma}_t^2 / \sigma_T^2$, $w_T(t) = \bar{\alpha}_t \sigma_t^2 / \sigma_T^2$ are time-varying weights balancing initial/terminal mean contributions.

In the context of our work, the terminal distribution mean x_T and the initial distribution mean x_0 correspond to the source codec feature H_q^s and the target codec feature H_q^t , respectively. Optimization uses a bridge loss predicting x_0 from sampled $x_t \sim p_t$:

$$\mathcal{L}_{\text{bridge}} = \mathbb{E}_{x_T \sim p_T, x_0 \sim p_0} \mathbb{E}_t [\|x_\theta(x_t, t, x_T) - x_0\|_2^2], \quad (5)$$

where $x_\theta(\cdot)$ denotes the parameterized prediction model.

2.3. Mamba-Enhanced Backbone

While standard U-Nets excel at extracting localized structural patterns from acoustic latents, they inherently lack the receptive field to capture long-range temporal correlations critical

to speech, such as prosodic fluctuations and phonetic coherence. To overcome this localized bias, we incorporate Mamba [27], a selective State Space Model, into our latent translation backbone. Using input-dependent discrete recurrences, Mamba effectively models global sequential dynamics in linear time. Crucially, recent speech processing literature [24, 25, 26] has demonstrated that Mamba matches or even exceeds the generative fidelity of Transformers while maintaining significantly faster inference speeds. This avoids the quadratic computational bottlenecks of attention mechanisms while perfectly complementing the spatial feature processing capabilities of U-Nets.

Architecturally, we interleave Mamba modules within the K resolution levels of the U-Net, as shown in Fig. 1. At each encoder downsampling stage, the input latent $H_q^s \in \mathbb{R}^{B \times 1 \times D_c \times T_H}$ is first processed by N convolutional residual blocks. This local feature extraction expands the representation to an intermediate tensor $H_{\text{conv}} \in \mathbb{R}^{B \times C \times D_r \times T_H}$, where C and D_r denote the channel and hidden dimensions, respectively. To establish a global temporal context, H_{conv} is flattened along its spatial dimensions into a sequence $\mathbb{R}^{B \times T_H \times D_m}$ (with $D_m = C \times D_r$) and routed through M consecutive Mamba layers. The globally-enriched output is subsequently reshaped back to its original topology $\mathbb{R}^{B \times C \times D_r \times T_H}$ to ensure seamless propagation to the next resolution level. The decoder strictly mirrors this Mamba-enhanced hierarchy, ensuring synchronized local-global feature alignment throughout the Schrödinger Bridge transport path.

2.4. Two-Stage Training Strategy

While optimizing the SB objective ($\mathcal{L}_{\text{bridge}}$) ensures strict optimal transport in the latent space, relying solely on latent alignment can lead to sub-optimal perceptual quality, as the model remains blind to the acoustic characteristics of the actual decoded waveforms. To align latent structural accuracy with audio-domain fidelity, we employ a two-stage training strategy. First, the SB model is trained exclusively in the highly compressed latent domain using Equation 5 to establish a robust transport path between the heterogeneous codecs. Then, at the second stage, to inject real-world acoustic constraints, the pre-trained SB translator is cascaded with the frozen target decoder $\text{De}^t(\cdot)$ to reconstruct the time-domain waveform $\hat{X}$. The translator is then fine-tuned against the ground-truth wideband audio X using a joint objective:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{bridge}} \mathcal{L}_{\text{bridge}} + \lambda_t \mathcal{L}_{\text{time}} + \lambda_{\text{mel}} \mathcal{L}_{\text{mel}}, \quad (6)$$

where $\mathcal{L}_{\text{time}} = \|X - \hat{X}\|_1$ is the time-domain L1 loss, and $\mathcal{L}_{\text{mel}} = \|S(X) - S(\hat{X})\|_1$ is the frequency-domain mel-spectrogram L1 loss [3], with $S(\cdot)$ denoting the mel-extraction function. The scalars λ_{bridge} , λ_t , and λ_{mel} balance the respective loss terms.

3. Experiments

3.1. Dataset

We conducted experiments on the clean speech dataset from the ICASSP 2022 Deep Noise Suppression (DNS) Challenge [29]. This dataset supports fullband 48 kHz audio processing, with a total duration of approximately 2,425 hours and around 1.09 million clips covering six languages. 10,000 clips (around 1% of the dataset) are designated as the validation set and another 10,000 clips as the test set. The remaining majority of clips are used as the training set.

3.2. Implementation Details and Baselines

We trained a source codec and a target codec for 8 kHz and 48 kHz speech separately using open-source AudioCraft.² The source codec uses ratios [2, 4, 5, 8] to produce 128-dimensional latents at 25 Hz, with 4-layer RVQ and 1024 codebooks per layer, achieving 1 kbps with 2.8 MB parameters. The target codec uses ratios [2, 4, 5, 6, 8] to produce 128-dimensional latents at 25 Hz, with 24-layer RVQ and 1024 codebooks per layer, achieving 6 kbps with 134 MB parameters. The SB model in BridgeCodec uses a U-Net backbone with 4 resolution levels and 2× downsampling per level, each followed by 4 residual blocks and 5 Mamba layers whose hidden size is 128 for long-range temporal modeling. It was trained in two stages, first with learning rate 5e-5, batch size 32 for 100k steps, and then with learning rate 5e-6, batch size 8 for 10k steps. We used VP noise schedule following the setting of Bridge-TTS [15] and SDE sampler with 10 NFE for inference. The number of function evaluations (NFE) refers to the count of forward passes of the neural network during sampling, directly related to computational cost. For comparative evaluation, we trained an SB model based on NCSN++ [23] with the same parameters and configuration as the baseline but excluding the Mamba layers, named BridgeCodec_{UNet}.

3.3. Evaluation Metrics

We adopt several metrics to evaluate the performance of BridgeCodec. STOI [30] is an intrusive metric that assesses speech intelligibility. Word Error Rate (WER) quantifies audio intelligibility by calculating the discrepancies between transcriptions and original texts. We only calculate WER on the English subset of the test set and transcribe the speeches through a Whisper-large V3 [31]. Speaker similarity (SIM) reflects the ability of the codec to preserve speaker characteristics. We use WavLM-base-sv [32] to extract the speaker embeddings of the reconstructed and ground-truth waveforms, and calculate the cosine similarity between them to measure SIM. DNSMOS [33] is a widely adopted metric for evaluating objective speech quality. Mel distance (MEL) [3] calculates the distance between log mel spectrograms of the reconstructed and ground-truth waveforms. The multiscale STFT distance (STFT) [3] computes the distance between log magnitude spectrograms of the reconstructed and ground-truth waveforms. To evaluate perceptual fidelity from a human perspective, we conduct a subjective Mean Opinion Score (MOS) test on the naturalness of the reconstruction speech. A panel of 30 fluent listeners evaluates the reconstructed speech quality of 15 randomly sampled test-set clips on a standard 5-point scale with 1-point intervals.

4. Results

Table 1 presents the comprehensive evaluation of BridgeCodec against the mismatched baselines. The target codec (48 kHz, 6 kbps) serves as the upper bound for high-fidelity reconstruction, while the source codec (8 kHz, 1 kbps) acts as the lower bound, exhibiting severely degraded performance across all metrics due to its narrowband constraint and ultra-low bitrate.

4.1. Objective Results

BridgeCodec successfully closes the massive interoperability gap between these endpoints. Without altering the 1 kbps transmission bitrate, it dramatically improves upon the source

²<https://github.com/facebookresearch/audiocraft>

Table 1: *Main results of BridgeCodec on the testset. All audio is resampled to 48 kHz for STOI, STFT and MEL. And due to evaluation model constraints, all audio is resampled to 16 kHz for DNSMOS, WER and SIM. MOS represents the subjective Mean Opinion Score for naturalness and speech quality with 95% confidence intervals.*

Model	STOI $\uparrow$	STFT $\downarrow$	MEL $\downarrow$	DNSMOS $\uparrow$	WER $\downarrow$	SIM $\uparrow$	MOS $\uparrow$
target codec	0.96	1.63	1.05	3.26	1.70	0.99	4.24 ± 0.10
source codec	0.23	12.24	10.01	3.11	1.92	0.81	3.80 ± 0.11
BridgeCodec _{UNet}	0.86	2.27	1.74	3.09	2.16	0.93	-
BridgeCodec	0.88	2.22	1.65	3.17	2.11	0.96	4.12 ± 0.11
w/o Two stage training	0.87	2.19	1.68	3.15	2.04	0.95	-

codec’s baseline. Notably, STOI leaps from 0.23 to 0.88, and MEL drops from 10.01 to 1.65, indicating the improvement of speech intelligibility. Most impressively, BridgeCodec effectively reconstructs speaker characteristics during the extreme bandwidth extension, elevating SIM from 0.81 to 0.96, closely approaching the target codec’s upper bound of 0.99. While the WER sees a marginal increase compared to the bounds (2.11 vs. 1.92), this is an expected outcome given the massive generative hallucination required to map 8 kHz latents to 48 kHz wideband audio. The source encoder captures only narrowband, and the bridge performs generative extension. Generated high-frequency content may slightly perturb Whisper, leading to increased WER. Despite this, the semantic content and speech quality remains highly well-preserved.

4.2. Subjective Results

The objective improvements are further strongly corroborated by our subjective MOS evaluation. BridgeCodec achieves a remarkable MOS of 4.12 ± 0.11 , significantly outperforming the source codec (3.80 ± 0.11) and demonstrating that the generated 48 kHz audio sounds highly natural to human listeners. This confirms that BridgeCodec does not merely interpolate signals, but successfully activates the generative priors of the target decoder to produce perceptually pleasing wideband speech.

4.3. Ablation Studies

To validate our architectural and methodological designs, we conduct ablation studies on the Mamba backbone and the training strategy, as shown in Table 1.

Effectiveness of Mamba. The proposed model consistently outperforms the baseline BridgeCodec_{UNet} across all metrics. This confirms that while standard convolutions capture localized spatial correlations in the latent space, Mamba’s selective state space mechanism is essential for modeling long-range global temporal dependencies (such as prosody and pitch variations), thereby reducing semantic distortions.

Impact of Two-Stage Training. Comparing BridgeCodec with its "w/o Two-stage training" variant reveals that involving the two-stage training strategy enhances performance across most metrics, with only minor trade-offs observed in STFT and WER, suggesting that latent domain training alone may more effectively preserve semantic information. However, second-stage training is critical for improving perceptual quality and speaker similarity, allowing BridgeCodec to achieve higher DNSMOS and SIM.

Sensitivity to Inference Steps. To investigate the impact of inference step on reconstruction performance, we conduct an experiment on the number of function evaluations (NFE) in the Schrödinger Bridge sampler, as shown in Table 2. Typically, fewer NFEs translate to lower inference latency and higher efficiency, but this compromises reconstruction performance due to insufficient iterative refinement of latent features [18]. How-

ever, as illustrated in Table 2, BridgeCodec demonstrates a remarkable robustness to reduced NFE. Specifically, when NFE is minimized to 1, the model achieves the best performance in STOI, STFT, MEL, and DNSMOS.

This one-step capability suggests that the Schrödinger Bridge in our framework learns an exceptionally straight and efficient transport path between the source and target latents. While increasing NFE to 10 provides a marginal gain in speaker similarity (SIM) through iterative refinement, the 1-NFE configuration provides a superior balance of acoustic fidelity and structural alignment. This makes BridgeCodec a highly promising candidate for low-latency communication where minimal inference iterations are critical.

Table 2: *The results with different number of function evaluations.*

NFE	STOI	STFT	MEL	DNSMOS	WER	SIM
10	0.88	2.22	1.65	3.17	2.11	0.96
5	0.88	2.16	1.63	3.17	2.37	0.95
1	0.89	2.13	1.58	3.18	2.39	0.95

4.4. Deployment Efficiency and Universal Interoperability

Beyond individual reconstruction metrics, BridgeCodec establishes a new paradigm for codec interoperability. While our experiments utilize the challenging 8 kHz to 48 kHz transformation as a primary testbed, the core value of BridgeCodec lies in its ability to perform heterogeneous latent translation between structurally mismatched manifolds. By decoupling the source encoder from the specific requirements of target decoder, BridgeCodec effectively serves as a universal translation framework. This allows diverse audio codecs—regardless of their sampling rates, bitrates, or internal architectures—to communicate within a unified framework. Such a capability is crucial for breaking the closed-loop constraints of current NACs, enabling seamless high-fidelity audio flow across fragmented, resource-constrained network ecosystems.

5. Conclusion

We presented BridgeCodec, a framework for universal audio codec interoperability that decouples rigidly paired encoders and decoders. By integrating Schrödinger Bridge for latent transport and a Mamba-enhanced backbone for temporal modeling, BridgeCodec achieves high-fidelity 48 kHz reconstruction from 8 kHz inputs at an ultra-low 1 kbps. Our two-stage training strategy effectively balances structural alignment with perceptual naturalness, supporting robust single-step inference. This work provides a scalable solution for high-quality and seamless audio communication across heterogeneous and resource-constrained network ecosystems.

6. Generative AI Use Disclosure

During the preparation of this manuscript, the authors used generative AI tools exclusively for the purpose of language editing and manuscript polishing to improve readability. These tools were not used to generate any core scientific ideas, experimental data, or technical contributions. All authors have thoroughly reviewed and approved the final version of the manuscript, and assume full responsibility for the integrity and entirety of its content.

7. Acknowledgement

This work was supported by National Natural Science Foundation of China (62076144).

8. References

- [1] N. Zeghidour, A. Luebs, A. Omran, J. Skoglund, and M. Tagliasacchi, “SoundStream: An end-to-end neural audio codec,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 495–507, 2021.
- [2] A. Défossez, J. Copet, G. Synnaeve, and Y. Adi, “High Fidelity Neural Audio Compression,” *Transactions on Machine Learning Research*, 2023, featured Certification, Reproducibility Certification.
- [3] R. Kumar, P. Seetharaman, A. Luebs, I. Kumar, and K. Kumar, “High-fidelity audio compression with improved RVQGAN,” *Advances in Neural Information Processing Systems*, vol. 36, pp. 27980–27993, 2023.
- [4] H. Liu, X. Xu, Y. Yuan, M. Wu, W. Wang, and M. D. Plumbley, “SemantiCodec: An Ultra Low Bitrate Semantic Audio Codec for General Sound,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 18, no. 8, pp. 1448–1461, 2024.
- [5] X. Zhang, D. Zhang, S. Li, Y. Zhou, and X. Qiu, “SpeechTokenizer: Unified Speech Tokenizer for Speech Language Models,” in *International Conference on Learning Representations (ICLR)*, 2024.
- [6] A. Défossez, L. Mazaré, M. Orsini, A. Royer, P. Pérez, H. Jégou, E. Grave, and N. Zeghidour, “Moshi: a speech-text foundation model for real-time dialogue,” *arXiv preprint arXiv:2410.00037*, 2024.
- [7] Z. Ye, P. Sun, J. Lei, H. Lin, X. Tan, Z. Dai, Q. Kong, J. Chen, J. Pan, Q. Liu *et al.*, “Codec does matter: Exploring the semantic shortcoming of codec for audio language model,” in *AAAI Conference on Artificial Intelligence (AAAI)*, 2025, pp. 25 697–25 705.
- [8] Y. Gong, L. Jin, R. Deng, D. Zhang, X. F. Zhang, Q. Cheng, Z. Fei, S. Li, and X. Qiu, “Xy-tokenizer: Mitigating the semantic-acoustic conflict in low-bitrate speech codecs,” *arXiv preprint arXiv:2506.23325*, 2025.
- [9] A. van den Oord, O. Vinyals, and K. Kavukcuoglu, “Neural Discrete Representation Learning,” in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017, pp. 6306–6315.
- [10] S. Ji, Z. Jiang, W. Wang, Y. Chen, M. Fang, J. Zuo, Q. Yang, X. Cheng, Z. Wang, R. Li, Z. Zhang, X. Yang, R. Huang, Y. Jiang, Q. Chen, S. Zheng, and Z. Zhao, “WavTokenizer: an Efficient Acoustic Discrete Codec Tokenizer for Audio Language Modeling,” in *International Conference on Learning Representations (ICLR)*, 2025.
- [11] D. Xin, X. Tan, S. Takamichi, and H. Saruwatari, “BigCodec: Pushing the limits of low-bitrate neural speech codec,” *arXiv preprint arXiv:2409.05377*, 2024.
- [12] H. Yang, I. Jang, and M. Kim, “Generative De-Quantization for Neural Speech Codec Via Latent Diffusion,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 1251–1255.
- [13] E. Schrödinger, “Sur la théorie relativiste de l’électron et l’interprétation de la mécanique quantique,” in *Annales de l’Institut Henri Poincaré (AIHP)*, 1932, pp. 269–310.
- [14] V. De Bortoli, J. Thornton, J. Heng, and A. Doucet, “Diffusion Schrödinger bridge with applications to score-based generative modeling,” *Advances in Neural Information Processing Systems*, vol. 34, pp. 17 695–17 709, 2021.
- [15] Z. Chen, G. He, K. Zheng, X. Tan, and J. Zhu, “Schrödinger bridges beat diffusion models on text-to-speech synthesis,” *arXiv preprint arXiv:2312.03491*, 2023.
- [16] S. Han, S. Lee, J. Lee, and K. Lee, “Few-step Adversarial Schrödinger Bridge for Generative Speech Enhancement,” in *Conference of the International Speech Communication Association (INTERSPEECH)*, 2025, pp. 2380–2384.
- [17] J. Yang, S. Wang, C. Wu, L. Guo, and F. Fan, “Schrödinger Bridge Mamba for One-Step Speech Enhancement,” *arXiv preprint arXiv:2510.16834*, 2025.
- [18] C. Li, Z. Chen, F. Bao, and J. Zhu, “Bridge-SR: Schrödinger bridge for efficient SR,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025, pp. 1–5.
- [19] C. Li, Z. Chen, L. Wang, and J. Zhu, “Audio Super-Resolution with Latent Bridge Models,” in *Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2025.
- [20] T. Lei, A. Li, R. Chen, D. Yu, M. Yu, J. Lu, and C. Zheng, “BridgeVoC: Insights into Using Schrödinger Bridge for Neural Vocoders,” in *ICLR Workshop on Deep Generative Model in Machine Learning: Theory, Principle and Efficacy (ICLRW)*, 2025.
- [21] Y. Lipman, R. T. Q. Chen, H. Ben-Hamu, M. Nickel, and M. Le, “Flow Matching for Generative Modeling,” in *International Conference on Learning Representations (ICLR)*, 2023.
- [22] Y. Song and S. Ermon, “Generative modeling by estimating gradients of the data distribution,” *Advances in Neural Information Processing Systems*, vol. 32, 2019.
- [23] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, “Score-Based Generative Modeling through Stochastic Differential Equations,” in *International Conference on Learning Representations (ICLR)*, 2021.
- [24] K. Miyazaki, Y. Masuyama, and M. Murata, “Exploring the Capability of Mamba in Speech Applications,” in *Conference of the International Speech Communication Association (INTERSPEECH)*, 2024, pp. 237–241.
- [25] X. Zhang, Q. Zhang, H. Liu, T. Xiao, X. Qian, B. Ahmed, E. Ambikairajah, H. Li, and J. Epps, “Mamba in Speech: Towards an Alternative to Self-Attention,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 33, pp. 1933–1948, 2025.
- [26] X. Zhang, J. Ma, M. Shahin, B. Ahmed, and J. Epps, “Rethinking Mamba in Speech Processing by Self-Supervised Models,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025, pp. 1–5.
- [27] A. Gu and T. Dao, “Mamba: Linear-time sequence modeling with selective state spaces,” in *First Conference on Language Modeling (COLM)*, 2024.
- [28] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional networks for biomedical image segmentation,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Springer, 2015, pp. 234–241.
- [29] H. Dubey, V. Gopal, R. Cutler, A. Aazami, S. Matushevych, S. Braun, S. E. Eskimez, M. Thakker, T. Yoshioka, H. Gamper, and R. Aichner, “ICASSP 2022 Deep Noise Suppression Challenge,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 9271–9275.
- [30] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, “An algorithm for intelligibility prediction of time-frequency weighted noisy speech,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.

- [31] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International Conference on Machine Learning (ICML)*, 2023, pp. 28 492–28 518.
- [32] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao *et al.*, "WavLM: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.
- [33] C. K. Reddy, V. Gopal, and R. Cutler, "DNSMOS P. 835: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 886–890.