EntangleCodec: A Unified Discrete Audio Tokenizer via Semantic-Acoustic Entanglement

Hui Li[*], **Yangfan Gao**[*], **Junlin Shang**, **Changhao Jiang**, **Tao Gui**, **Qi Zhang**[†], **Xuanjing Huang**
Fudan University
hui_li25@m.fudan.edu.cn, qzhang@fudan.edu.cn

Abstract

Audio tokenizers bridge continuous audio and Audio Language Models (ALMs), yet existing methods often trade off understanding and generation. Reconstruction-oriented codecs lack rich semantics, while semantic-aware tokenizers typically use separate semantic and acoustic streams, introducing redundancy or misalignment.

We propose **EntangleCodec**, a unified discrete audio tokenizer that learns caption-aligned semantic-acoustic representations before quantization. Rich audio captions enable a compact token stream to capture linguistic, speaker, prosodic, emotional, and acoustic information beyond ASR transcripts. A flow-matching diffusion decoder further supports high-quality reconstruction across speech, music, and general audio.

EntangleCodec achieves reconstruction quality competitive with specialized codecs, outperforms all codec-based baselines on audio understanding by up to **+7.4%** on MMAR, and supports both TTS and TTA generation in a unified framework. Furthermore, EntangleCodec-based audio language models demonstrate strong scaling behavior: even at *0.6B* parameters, the model surpasses specialized continuous-representation LLMs with over *13B* parameters across three benchmarks using **22**× fewer parameters; scaling to *8B* further establishes new state-of-the-art results on MMAR, highlighting that representation quality is as critical as model scale in audio language modeling. Code and model weights are available at https://github.com/luckyerr/EntangleCodec.

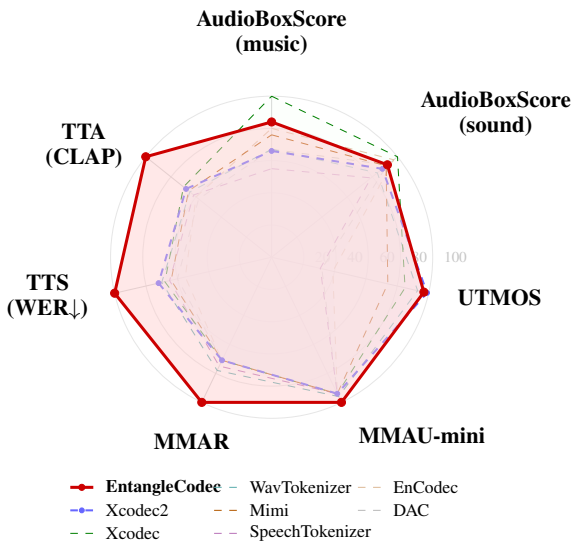


Figure 1: Comprehensive tokenizer comparison across seven dimensions (normalized 0–100; TTS WER inverted so higher = better).Reconstruction: AudioBoxScore (music/sound), UTMOS; Understanding: MMAU-mini, MMAR; Generation: TTS (WER↓), TTA (CLAP).

1 Introduction

The recent success of large language models has motivated growing interest in Audio Language Models (ALMs), spanning speech, music, and general audio understanding and generation. A central challenge is how to represent audio as discrete tokens that are both semantically informative and acoustically faithful. Existing audio tokenizers often favor one side of this trade-off: reconstruction-oriented codecs preserve fine acoustic detail but provide limited semantic information, while semantic-aware tokenizers introduce additional semantic streams that require fusion with acoustic representations. Moreover, semantic supervision is often limited to ASR transcripts, ignoring non-textual cues such as speaker identity, emotion, prosody, acoustic scenes, and musical structure.

We propose **EntangleCodec**, a unified discrete audio tokenizer that learns caption-aligned semantic-acoustic representations before quantization. Instead of separating semantic and acous-

[*] Equal contribution.
[†] Corresponding authors.

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 18015–18031
October 24-29, 2026 ©2026 Association for Computational Linguistics

tic information into independent streams, EntangleCodec integrates them within a shared pre-quantization representation, reducing the need for late fusion. To enrich semantic supervision, we align audio with rich captions rather than ASR transcripts through contrastive learning, enabling the resulting tokens to capture linguistic content as well as speaker, affective, prosodic, and environmental cues. A flow-matching diffusion decoder then reconstructs high-quality audio from the discrete tokens, allowing the same tokenizer to support TTS, TTA, and audio QA without task-specific architectural changes.

Key empirical findings. EntangleCodec achieves reconstruction quality comparable to specialized codecs despite its unified design (UTMOS: 3.96 vs. 4.02). For audio understanding, it outperforms all codec-based baselines by up to **+7.4%** on MMAR. Furthermore, a *0.6B* EntangleCodec-based language model surpasses specialized continuous-representation models with over *13B* parameters across MMAR, MMAU-mini, and MMAU, using **22**× fewer parameters; scaling to *8B* further establishes new state-of-the-art results on MMAR (**42.6%**). These results highlight that tokenizer representation quality is as critical as model scale in audio language modeling.

Our main contributions are:

- **EntangleCodec**, a unified codec-based tokenizer that learns semantic-acoustic entangled representations before quantization, reducing the need for dual-stream separation and late fusion while maintaining competitive reconstruction quality.
- **Caption-aligned discrete representation learning**, which goes beyond ASR-only supervision by contrastively aligning audio representations with rich captions covering linguistic content, speaker identity, emotion, prosody, and acoustic scenes, leading to more expressive discrete audio tokens.
- **A unified audio language modeling framework** for TTS, TTA, and audio QA, requiring no task-specific tokenizer or architectural modification and demonstrating the broad applicability of EntangleCodec across both generation and understanding tasks.

2 Related Work

2.1 Audio Codec and Tokenization

Neural audio codecs provide the foundation for discrete audio language modeling. Early codecs such as SoundStream (Zeghidour et al., 2021) and EnCodec (Défossez et al., 2023) utilized residual vector quantization (RVQ) for high-fidelity reconstruction at low bitrates, while DAC (Kumar et al., 2023) further improved codebook utilization and reconstruction quality. Recent tokenizers focus more explicitly on language-model compatibility: WavTokenizer (Ji et al., 2025) reduces token rates while preserving perceptual quality, Mimi (Défossez et al., 2024) supports streaming-friendly tokenization, and SpeechTokenizer (Zhang et al., 2024), XCodec (Ye et al., 2025a), and XCodec2 (Ye et al., 2025b) introduce semantic-aware designs for speech generation and audio modeling.

Most existing codecs follow one of two routes. Reconstruction-oriented codecs use a single acoustic encoder but provide limited semantic information for downstream reasoning. Semantic-aware codecs often introduce separate semantic and acoustic streams, which improves task relevance but requires additional fusion between independently learned representations. In contrast, EntangleCodec learns a caption-aligned semantic-acoustic representation before quantization, aiming to preserve both reconstruction fidelity and semantic richness within a unified discrete token stream.

2.2 Audio-Language Models for Understanding

Audio-language models extend LLMs to speech, music, and general audio understanding. LTU (Gong et al., 2024) and Audio Flamingo (Kong et al., 2024) connect audio encoders with language models for audio question answering, while MU-LLaMA (Liu et al., 2024) and GAMA (Ghosh et al., 2024) target music understanding and general audio reasoning. SALMONN (Tang et al., 2024) combines speech and audio encoders to handle diverse audio tasks, and Qwen2-Audio (Chu et al., 2024) further improves performance through stronger audio encoders.

These systems typically rely on continuous audio representations produced by dedicated encoders. Such representations are effective for understanding, but they are not naturally suited to autoregres-

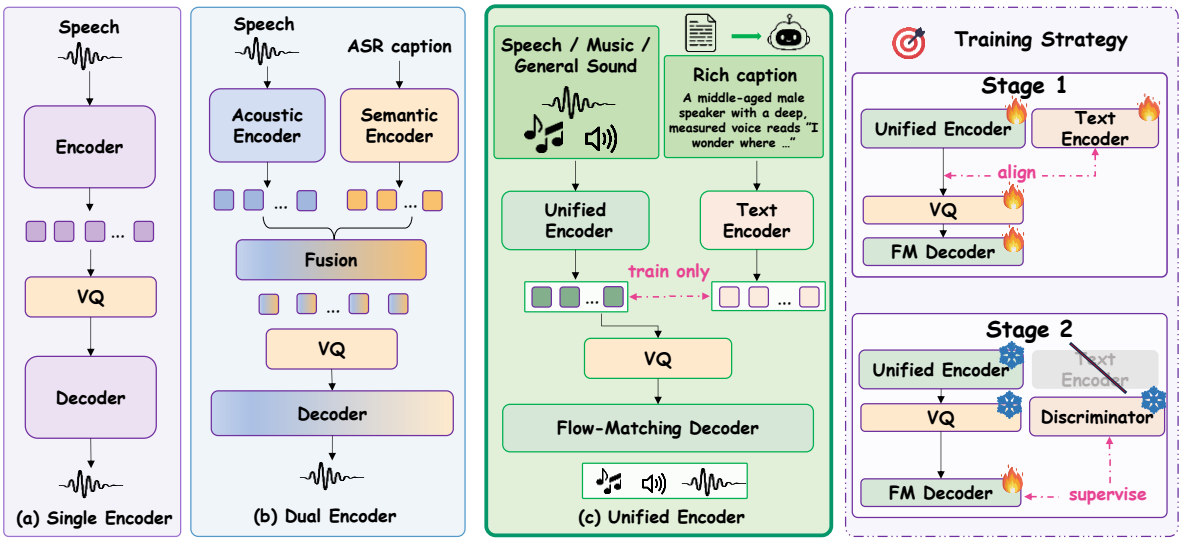


Figure 2: **Overview of EntangleCodec.** Compared with single-encoder acoustic codecs and dual-encoder semantic-acoustic codecs, EntangleCodec learns caption-aligned unified tokens that support both reconstruction and downstream audio-language tasks. The right panel shows the two-stage training strategy: joint codec learning with rich captions, followed by decoder refinement.

sive audio generation. EntangleCodec instead provides discrete audio tokens that can be consumed by LLMs, enabling audio understanding and generation to share the same token interface.

2.3 Text-Conditioned Audio Generation

Text-conditioned audio generation includes text-to-speech (TTS) and text-to-audio (TTA). Autoregressive models such as AudioLM (Borsos et al., 2023) and VALL-E (Wang et al., 2023) demonstrate the effectiveness of predicting discrete audio tokens for speech generation, while TTA models often rely on text-audio alignment signals such as CLAP (Elizalde et al., 2023) to synthesize audio from natural language descriptions. Strong TTS systems such as NaturalSpeech3 (Ju et al., 2024) and CosyVoice (Du et al., 2024) achieve high-quality speech synthesis, whereas general TTA systems focus on environmental sound, music, and acoustic scenes.

Despite this progress, TTS and TTA are usually treated with separate modeling pipelines, tokenizers, or conditioning mechanisms. EntangleCodec provides a unified discrete representation for both speech and general audio, allowing TTS and TTA to be formulated as text-conditioned token prediction tasks within the same audio language modeling framework.

3 Methods

3.1 Overview

EntangleCodec aims to build a unified discrete audio representation that simultaneously supports both understanding and generation tasks. As illustrated in Figure 2, the overall architecture consists of three core components: a unified encoder, a discrete quantizer, and a diffusion-based decoder. The fundamental distinction from existing audio tokenizers lies in two key design decisions: first, employing a single shared encoder rather than dual encoders to jointly model semantic and acoustic information in the same representation space; second, achieving semantic alignment through rich audio captions rather than ASR transcripts, encoding multimodal semantics beyond text.

3.2 Unified Encoder

Unified Modeling. To obtain a discrete representation suitable for both audio understanding and generation, EntangleCodec integrates semantic and acoustic cues within the same pre-quantization representation. Unlike dual-encoder tokenizers that fuse separately extracted semantic and acoustic features, our encoder directly maps the input audio into a unified semantic-acoustic feature sequence, which is subsequently quantized into discrete tokens.

Given a Mel-spectrogram $\mathbf{M} \in \mathbb{R}^{128 \times T_m}$, we

first project it into a D_{enc}-dimensional sequence:

$$\mathbf{M}' = \mathrm{Linear}(\mathbf{M}^\top) \in \mathbb{R}^{T_m \times D_{enc}}. \quad (1)$$

After prepending a learnable [CLS] token $\mathbf{c} \in \mathbb{R}^D$, the encoder input becomes

$$\mathbf{E} = [\mathbf{c}; \mathbf{M}'] \in \mathbb{R}^{(T_m+1) \times D_{enc}}. \quad (2)$$

This sequence is processed by the Transformer-based encoder:

$$\mathbf{H}' = \mathrm{Encoder}(\mathbf{E}) \in \mathbb{R}^{(T_m+1) \times D_{enc}}. \quad (3)$$

We discard the [CLS] token and use the remaining frame-level outputs $\mathbf{H} \in \mathbb{R}^{T_m \times D_{enc}}$ as the unified semantic-acoustic representation for subsequent quantization. By performing semantic-acoustic integration before quantization, EntangleCodec avoids an explicit late-fusion module and provides a shared representation space for both reconstruction and downstream audio understanding.

Rich Semantic Alignment. To enrich the semantic content of the unified audio representation, EntangleCodec aligns audio with rich captions rather than ASR transcripts. Unlike transcripts that mainly preserve linguistic content, rich captions additionally describe speaker attributes, emotion, prosody, acoustic scenes, and sound events, providing a broader supervision signal for semantic-acoustic representation learning. Notably, these rich captions are automatically generated by a large language model (LLM), which comprehensively characterizes the multi-dimensional semantic attributes of audio in natural language, thereby providing the encoder with a substantially richer supervision signal than conventional ASR transcripts, details in Appendix C.2.

We implement this alignment through audio-text contrastive learning. Given the encoder output $\mathbf{H}'$, we take the class-token representation $\mathbf{h}_{\text{cls}} = \mathbf{H}'_0$ as the global audio embedding and project it into a D_{align}-dimensional space:

$$\mathbf{a} = \mathrm{Proj}_{\text{audio}}(\mathbf{h}_{\text{cls}}) \in \mathbb{R}^{D_{align}}, \quad (4)$$

$$\mathbf{t}_{\text{emb}} = \mathrm{TextEncoder}(\mathbf{t}) \in \mathbb{R}^{D_{align}}, \quad (5)$$

where $\mathbf{t}$ denotes the rich caption generated by the LLM, and $\mathrm{TextEncoder}(\cdot)$ is a 12-layer Transformer text encoder. The audio and text embeddings are then aligned using a CLIP-style contrastive loss, encouraging the unified encoder to encode semantic cues beyond transcript-level content while retaining the acoustic information required for reconstruction.

Pre-quantization Processing. The unified representation $\mathbf{H}$ is projected to D_{quant} dimensions via a two-layer MLP and is L2-normalized to satisfy the constraints of single-codebook vector quantization:

$$\mathbf{h} = \mathrm{MLP}_{\mathrm{VQ}}(\mathbf{H}) \in \mathbb{R}^{T_m \times D_{quant}}, \quad \tilde{\mathbf{h}} = \frac{\mathbf{h}}{\|\mathbf{h}\|_2}. \quad (6)$$

The normalized representation $\tilde{\mathbf{h}}$ is then fed into a single-codebook vector quantizer for discretization, producing the final token sequence.

3.3 Diffusion-based Decoder

The diffusion decoder reconstructs Mel-spectrograms from the quantized representation $\mathbf{Z}_q$ using Flow Matching (Lipman et al., 2023). The decoder is implemented as a Llama-style Transformer flow predictor, taking the noisy Mel-spectrogram $\mathbf{x}_t$, a time-step embedding, and $\mathbf{Z}_q$ as a conditioning prefix as input, and predicting the velocity field:

$$\mathbf{v}_\theta(\mathbf{x}_t, t, \mathbf{Z}_q) \in \mathbb{R}^{T_m \times 128}. \quad (7)$$

3.4 Training Objectives and Inference

EntangleCodec adopts a two-stage training strategy. The first stage jointly trains the audio encoder, text encoder, quantizer, and diffusion decoder, enabling the encoder to capture rich multimodal semantics through contrastive learning. The second stage freezes the encoder and quantizer, discards the text encoder, and trains only the decoder to refine reconstruction quality.

Stage 1: Semantic Learning and Joint Optimization. The total loss consists of three components:

$$\mathcal{L}_{\text{stage1}} = \mathcal{L}_{\text{flow}} + \mathcal{L}_{\text{contrast}} + \mathcal{L}_{\text{vq}} \quad (8)$$

Flow Matching Loss $\mathcal{L}_{\text{flow}}$ is based on the Rectified Flow objective. Given an interpolated sample $\mathbf{x}_t = (1-(1-\sigma_{\min})t)\mathbf{z} + t\mathbf{x}$, the decoder is trained to predict the velocity field:

$$\mathcal{L}_{\text{flow}} = \mathbb{E}_{t,\mathbf{x},\mathbf{z}}\Big[\big\|\mathbf{v}_\theta(\mathbf{x}_t, t, \mathbf{Z}_q) - (\mathbf{x} - (1-\sigma_{\min})\mathbf{z})\big\|_1\Big] \quad (9)$$

Contrastive Loss $\mathcal{L}_{\text{contrast}}$ aligns audio with LLM-generated captions via a CLIP-style bidirectional cross-entropy loss, enabling the encoder to learn

rich semantics including speaker characteristics, emotion, and prosody:

$$\mathcal{L}_{\text{contrast}} = \frac{1}{2}\left(\mathcal{L}_{\text{a2t}} + \mathcal{L}_{\text{t2a}}\right) \tag{10}$$

VQ Commitment Loss $\mathcal{L}_{\text{vq}}$ stabilizes quantization and prevents codebook collapse:

$$\mathcal{L}_{\text{vq}} = \|\text{sg}(\tilde{\mathbf{h}}) - \mathbf{z}_q\|_2^2 + \beta\|\tilde{\mathbf{h}} - \text{sg}(\mathbf{z}_q)\|_2^2 \tag{11}$$

Stage 2: Refining Reconstruction Quality. The encoder and quantizer are frozen, and only the decoder is updated using the reconstruction-oriented flow and VQ objectives together with an adversarial loss (Appendix B.3). This stage improves perceptual reconstruction quality while preserving the semantic-acoustic representation learned in Stage 1.

The two stages are trained for 500k and 200k steps respectively, with 10% conditional dropout applied throughout to support classifier-free guidance.

Inference. For **audio reconstruction**, $\mathbf{Z}_q$ is obtained by encoding and quantizing the input audio. For **conditional generation** (e.g., TTS, TTA), an autoregressive LLM predicts token indices from the text condition, and the corresponding codebook embeddings serve as $\mathbf{Z}_q$. Both modes share the same decoder and differ only in the source of discrete tokens. Sampling is performed with a single Euler step, with optional classifier-free guidance:

$$\tilde{\mathbf{v}}_\theta = (1+\gamma)\,\mathbf{v}_\theta(\mathbf{x}_t, t, \mathbf{Z}_q) - \gamma\,\mathbf{v}_\theta(\mathbf{x}_t, t, \varnothing), \tag{12}$$

where $\varnothing$ denotes the null condition and $\gamma = 1.0$ by default. The resulting Mel-spectrogram is converted to a waveform using the Vocos vocoder.

4 Experiments

4.1 Experimental Setup

4.1.1 Datasets

Training Data. We train EntangleCodec on a diverse corpus covering speech, music, and general audio, including LibriSpeech (Panayotov et al., 2015), MusicBench (Melechovský et al., 2024), AudioSet (Gemmeke et al., 2017), AudioCaps (Kim et al., 2019), and WavCaps (Mei et al., 2023). Downstream audio language models are trained on task-specific instruction and generation data for audio understanding, TTS, and TTA. Detailed dataset statistics are provided in Appendix C.

Evaluation Data. We evaluate reconstruction on held-out speech, music, and general-audio test sets, audio understanding on MMAR (Ma et al., 2026), MMAU-mini (Sakshi et al., 2025), and MMAU (Sakshi et al., 2025), and audio generation on standard TTS and TTA benchmarks. Detailed evaluation sets are listed in Appendix C.

4.1.2 Implementation Details

EntangleCodec is trained in two stages: Stage 1 jointly learns the tokenizer and decoder with reconstruction, contrastive, and VQ losses; Stage 2 freezes the tokenizer and refines the decoder for reconstruction. After training, EntangleCodec is frozen and used to train Qwen3-series audio language models (Team, 2025) for understanding and generation. Detailed optimization settings and hyperparameters are provided in Appendix B.2.

4.1.3 Baselines

We compare EntangleCodec along three dimensions. For reconstruction, we evaluate against representative neural audio codecs under matched compression settings using official pretrained checkpoints. For audio understanding, we use a codec-controlled setting where identical Qwen3-0.6B (Team, 2025) language models are trained on tokens from different codecs, isolating the effect of tokenization. We additionally compare with specialized continuous-representation audio LLMs using official checkpoints or reported results when available. For generation, we follow the same codec-controlled protocol for TTS and TTA. The full list of baselines is provided in Appendix D.

4.1.4 Evaluation Metrics

We evaluate reconstruction with UTMOS (Saeki et al., 2022), PESQ (Rix et al., 2001), voiced/unvoiced F1, and the 4 metrics in AudioBox Aesthetics (Tjandra et al., 2025) (AudioBoxScore); audio understanding with accuracy; TTS with WER and UTMOS; and TTA with CLAP Score (Elizalde et al., 2023). Lower WER is better, while higher values are better for all other metrics. Detailed metric definitions are provided in Appendix F.

4.2 Results and Analysis

4.2.1 Finding 1: EntangleCodec Maintains Strong Reconstruction Quality

▷ *EntangleCodec achieves reconstruction quality comparable to specialized neural codecs while pro-*

Model	T./L.	Speech				Sound	Music
		UTMOS↑	F1↑	STOI↑	SIM↑	AudioBoxScore	AudioBoxScore
GT	/	4.08	0.98	1.00	1.00	3.64/5.16/4.30/5.84	5.93/6.43/5.24/6.59
DAC-Codec	50/1	1.30	0.97	0.62	0.25	3.16/3.99/4.25/5.07	3.58/3.97/5.18/4.61
EnCodec	75/1	1.57	0.92	0.77	0.25	3.39/4.86/3.46/5.63	4.92/5.32/4.74/5.52
WavTokenizer	75/1	3.79	0.98	**0.90**	0.65	2.84/3.99/3.04/5.07	3.59/4.39/4.18/5.16
SpeechTokenizer	50/1	1.27	0.97	0.64	0.17	2.48/4.10/2.43/5.20	2.34/3.85/2.88/5.00
Xcodec	50/1	3.42	0.97	0.85	0.48	**3.48/4.66/4.11/5.62**	**6.25/6.75/5.82/6.92**
Mimi	50/4	3.03	0.97	0.85	0.50	3.14/4.25/3.67/5.24	4.12/4.81/5.26/5.37
Xcodec2	50/1	**4.02**	0.98	0.88	**0.76**	3.13/4.14/3.08/5.27	3.26/4.26/4.08/5.45
WavTokenizer	40/1	3.58	0.97	0.85	0.48	3.11/4.30/3.58/5.19	4.78/5.34/5.29/5.61
EntangleCodec	**50/1**	3.96	**0.98**	0.87	0.66	3.12/4.50/3.36/5.45	4.98/5.70/5.10/5.88

Table 1: Tokenizer reconstruction quality comparison. **Bold** = best result; underline = second best (GT excluded). Δ denotes improvement of EntangleCodec over the second-best baseline.

Model	TPS/VQ Layers	MMAU-mini↑	MMAU↑	MMAR↑	Average↑
WavTokenizer	75/1	32.7	32.3	26.9	30.6
SpeechTokenizer	400/8	31.9	31.6	25.9	29.8
Xcodec	50/1	32.0	32.6	24.2	29.6
Xcodec2	50/1	26.8	25.7	24.6	25.7
XY-Tokenizer	100/8	31.6	30.1	24.9	28.9
EntangleCodec (Ours)	**50/1**	**34.2**	**35.1**	**34.3**	**34.5**
Δ vs. 2nd best		+1.5	+2.5	+7.4	+3.9

Table 2: Audio understanding performance. All discrete-tokenizer models share the same Qwen3-0.6B backbone. **Bold** = best; underline = second best. Δ denotes improvement over the second-best baseline.

viding a unified semantic-acoustic representation.

Speech Reconstruction. Our model achieves **3.96** UTMOS and **0.98** F1, approaching the best baselines of 4.02 and 0.98, respectively., and remains competitive on STOI and SIM. Learning a semantic-aware representation does not come at the cost of speech reconstruction quality.

Sound and Music Reconstruction. Our model ranks second on both sound (**4.11**) and music (**5.42**) AudioBoxScore, trailing only XCodec which lacks semantic alignment. EntangleCodec is the only tokenizer that consistently ranks among the top performers across all three audio domains, confirming that semantic-acoustic entanglement and reconstruction fidelity are complementary rather than conflicting objectives.

4.2.2 Finding 2: Representation Quality Improves Audio Understanding

▷ *Under controlled comparisons, EntangleCodec produces more effective discrete tokens for audio understanding; the resulting small audio LLM also remains competitive with much larger specialized models.*

Codec-controlled comparison. Table 2 compares different tokenizers under the same backbone and training setup, isolating the effect of the codec itself. EntangleCodec achieves **34.2%** on MMAU-mini, **35.11%** on MMAU, and **34.3%** on MMAR, outperforming the best discrete codec baseline by +**1.5%**, +**2.51%**, and +**7.4%**, respectively, establishing a new state-of-the-art among discrete tokenizer-based audio LLMs. The consistent improvements across all three benchmarks indicate that EntangleCodec tokens provide stronger semantic cues for downstream reasoning than reconstruction-oriented codec tokens.

Model	TPS/VQ Layers	TTS (Speech)		TTA (Sound)	
		WER↓	UTMOS↑	AudioBox Score↑	CLAP↑
WavTokenizer	75/1	17.2	1.29	2.85/3.39/2.48/5.20	0.01
Xcodec	50/1	25.6	2.95	**3.58/5.06/4.72/6.01**	0.03
Xcodec2	50/1	23.2	1.53	3.07/5.11/2.51/4.37	0.02
XY-Tokenizer	100/8	<u>12.1</u>	<u>2.95</u>	<u>3.17/4.81/3.02/6.01</u>	<u>0.04</u>
EntangleCodec	50/1	**9.8**	**3.89**	3.29/4.85/3.19/5.75	**0.17**
Δ vs. 2nd best		−2.3	+0.94	−0.57	+0.13

Table 3: Audio generation performance. All discrete-tokenizer models share the same Qwen3-0.6B backbone. **Bold** = best; <u>underline</u> = second best. Δ denotes improvement over the second-best baseline. TTS WER: lower is better (shown inverted for Δ).

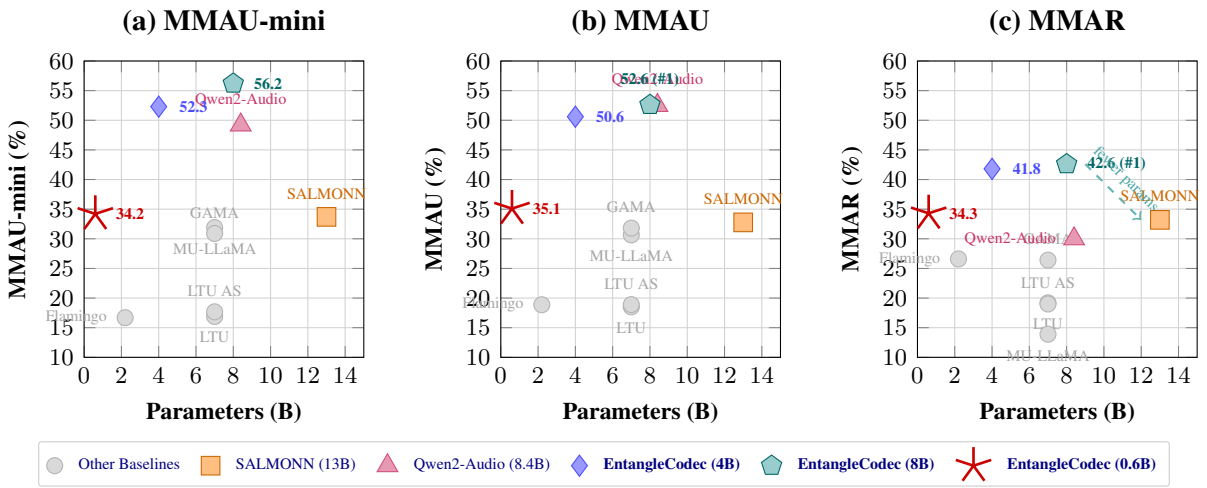


Figure 3: Parameter efficiency across audio understanding benchmarks. EntangleCodec scales consistently from 0.6B to 8B, achieving the best results on MMAU at 8B and on MMAR at both 4B and 8B, while outperforming larger specialized audio LLMs with fewer parameters.

Comparison with continuous-representation audio LLMs. Figure 3 shows EntangleCodec-LLM against continuous-representation audio LLMs. At only **0.6B** parameters, EntangleCodec-LLM already surpasses all discrete-tokenizer baselines and outperforms several much larger continuous-representation models, including SALMONN-13B by **+0.3%**, **+0.6%**, and **+1.1%** on MMAU-mini, MMAU, and MMAR, and GAMA-7B by **+3.1%**, **+1.6%**, and **+7.9%**, despite using **22**× fewer parameters than SALMONN. These results demonstrate that a well-designed discrete codec is not inherently a bottleneck for audio understanding — with sufficiently rich token representations, discrete-tokenizer LLMs can match or exceed continuous-representation systems at a fraction of the model size.

Effect of scaling. As shown in Figure 3, EntangleCodec-based LLMs scale effectively from **0.6B** to **4B** and **8B** parameters, with consistent and substantial gains across all benchmarks. At 4B, the model achieves **52.3%**, **50.6%**, and **41.8%** on MMAU-mini, MMAU, and MMAR, surpassing Qwen2-Audio-8.4B on MMAU-mini by **+3.1%** despite fewer parameters. Scaling to 8B further improves performance to **56.2%**, **52.6%**, and **42.6%**, achieving state-of-the-art results on MMAU-mini and matching Qwen2-Audio on MMAU while surpassing all baselines on MMAR. The strong scaling behavior indicates that EntangleCodec provides a solid representational foundation that continues to benefit from increased language model capacity, suggesting that the performance gap with continuous-representation systems will narrow further as model scale grows.

Configuration	T./L.	Speech				AudioBox Score	
		UTMOS↑	F1↑	STOI↑	SIM↑	Sound	Music
EntangleCodec	**50/1**	**3.96**	**0.98**	**0.87**	**0.66**	**3.12/4.50/3.36/5.45**	**4.98/5.70/5.10/5.88**
w/o Contrastive Loss	50/1	$3.04^{\downarrow 0.92}$	$0.97^{\downarrow 0.01}$	$0.82^{\downarrow 0.05}$	$0.50^{\downarrow 0.16}$	2.89/4.21/2.77/5.34	4.10/5.29/4.31/5.63
w/o Rich Caption	50/1	$3.24^{\downarrow 0.72}$	$0.97^{\downarrow 0.01}$	$0.83^{\downarrow 0.04}$	$0.51^{\downarrow 0.15}$	2.64/3.92/2.69/5.13	3.65/4.71/3.45/5.34
w/o Stage 2	50/1	$3.40^{\downarrow 0.56}$	$0.98^{=}$	$0.85^{\downarrow 0.02}$	$0.54^{\downarrow 0.12}$	2.84/4.22/2.76/5.34	4.35/5.53/4.35/5.79

Table 4: Ablation study on key components of EntangleCodec. **Bold** = best (full model); superscript ↓ indicates degradation from full model.

4.2.3 Finding 3: Unified Tokens Support Audio Generation

▷ *The same EntangleCodec tokens can be used for TTS and TTA generation without introducing task-specific tokenizers or separate acoustic representations.*

Text-to-speech. Table 3 shows that EntangleCodec achieves a WER of **9.8%** and a UTMOS of **3.89** on TTS, ranking **first** among all codec-controlled baselines on both metrics. Compared to the second-best baseline, WER is reduced by 2.3 points and UTMOS improves by $+0.94$, demonstrating that the proposed tokenizer preserves sufficient acoustic and prosodic information for intelligible and natural speech synthesis.

Text-to-audio. For TTA, EntangleCodec achieves a CLAP Score of **0.17**, outperforming the second-best baseline by $+0.13$—more than $4\times$ the score of XY-Tokenizer (0.04). While AudioBoxScore is slightly below Xcodec (-0.57 on average), Xcodec exhibits substantially degraded TTS performance (WER 25.6%), suggesting that its acoustic fidelity comes at the cost of semantic alignment. The strong CLAP improvement is consistent with the understanding results: caption-aligned semantic-acoustic tokens provide richer conditioning signals for generating audio that matches textual descriptions.

4.3 Ablation Studies

We isolate the contribution of three core design choices in EntangleCodec: the audio-text contrastive learning objective, the richness of semantic supervision, and the two-stage training strategy. All ablated variants use the same architecture (Table 4).

Effect of Contrastive Learning. Removing the contrastive loss (*w/o contrast*) leads to the most pronounced degradation among the ablated variants: UTMOS drops by **0.92** (3.96 → 3.04) and SIM drops by **0.16** (0.66 → 0.50), with consistent declines in AudioBoxScore for both sound and music. These results show that contrastive alignment is not merely an auxiliary semantic objective; it substantially shapes the encoder representation. By aligning audio features with text-level semantics, the contrastive objective encourages more structured representations that remain compatible with high-quality reconstruction.

Effect of Rich Captions vs. ASR-Only Supervision. Replacing rich audio captions with ASR transcripts (*w/o rich semantic*) yields a smaller but consistent degradation across all metrics. UTMOS drops by **0.72** and SIM by **0.15**, while AudioBoxScore for both sound and music also declines. The gap between *w/o contrast* and *w/o rich semantic* suggests that ASR-only contrastive alignment still provides useful semantic supervision, but rich captions offer additional information beyond transcript content, including speaker attributes, emotion, prosody, and acoustic scenes. This supports our hypothesis that ASR-only supervision provides an incomplete semantic signal for unified audio tokenization.

5 Conclusion

We presented **EntangleCodec**, a unified discrete audio tokenizer that learns caption-aligned semantic-acoustic representations for both audio understanding and generation. By combining rich caption alignment with a flow-matching decoder, EntangleCodec supports reconstruction, TTS, TTA, and audio QA without task-specific tokenizers or architectural changes. Experiments show EntangleCodec achieves competitive reconstruction quality, improves codec-based audio understanding by up to **+7.4%** on MMAR, and enables a *0.6B* audio LLM to surpass specialized models exceeding *13B* parameters. These results highlight the importance of representation quality in audio language modeling.

Limitations

While EntangleCodec demonstrates strong performance across multiple tasks, several limitations warrant further exploration and improvement.

Fine-grained Semantic Modeling. Although EntangleCodec uses rich audio captions to encode multimodal semantics, the granularity of semantics remains limited by annotation quality. For subtle semantic features—such as nuanced emotional shifts, complex musical harmonies, or specific acoustic scene details—current captions may not fully capture them. Finer-grained semantic modeling may require more detailed annotations or self-supervised learning methods.

Limited Exploration of Larger Scales. Due to computational resource and data constraints, our scaling experiments are limited to 8B parameters and a moderate amount of training data. While the consistent gains from 0.6B to 8B suggest that EntangleCodec would continue to benefit from further scaling, experiments at larger scales (e.g., 30B or 70B) remain unexplored. Furthermore, as parameters scale up, reconstruction and understanding tasks may have different demands on model capacity, and the optimal parameter allocation and architectural balance between the two objectives may shift. We leave a more comprehensive scaling analysis, along with strategies for balancing reconstruction and understanding under larger-scale training, to future work as resources permit.

Acknowledgements

The authors wish to thank the anonymous reviewers for their helpful comments. This work was partially funded by National Key R&D Program of China No.2025ZD0215702, National Natural Science Foundation of China (No. 62576106, 62476061, 62376061).

References

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. *CoRR*, abs/2406.02430.

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.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou. 2024. Qwen2-audio technical report. *CoRR*, abs/2407.10759.

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. *CoRR*, abs/2410.00037.

Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen. 2019. Clotho: An audio captioning dataset. *CoRR*, abs/1910.09387.

Zhihao Du, Qian Chen, Shiliang Zhang, Kai Hu, Heng Lu, Yexin Yang, Hangrui Hu, Siqi Zheng, Yue Gu, Ziyang Ma, Zhifu Gao, and Zhijie Yan. 2024. Cosyvoice: A scalable multilingual zero-shot text-to-speech synthesizer based on supervised semantic tokens. *CoRR*, abs/2407.05407.

Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. 2023. CLAP learning audio concepts from natural language supervision. In *IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2023, Rhodes Island, Greece, June 4-10, 2023*, pages 1–5. IEEE.

Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2017, New Orleans, LA, USA, March 5-9, 2017*, pages 776–780. IEEE.

Sreyan Ghosh, Sonal Kumar, Ashish Seth, Chandra Kiran Reddy Evuru, Utkarsh Tyagi, S. Sakshi, Oriol Nieto, Ramani Duraiswami, and Dinesh Manocha. 2024. GAMA: A large audio-language model with advanced audio understanding and complex reasoning abilities. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 6288–6313. Association for Computational Linguistics.

Yuan Gong, Alexander H. Liu, Hongyin Luo, Leonid Karlinsky, and James R. Glass. 2023. Joint audio and speech understanding. In *IEEE Automatic Speech Recognition and Understanding Workshop, ASRU 2023, Taipei, Taiwan, December 16-20, 2023*, pages 1–8. IEEE.

Yuan Gong, Hongyin Luo, Alexander H. Liu, Leonid Karlinsky, and James R. Glass. 2024. Listen, think, and understand. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

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 *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Zeqian Ju, Yuancheng Wang, Kai Shen, Xu Tan, Detai Xin, Dongchao Yang, Eric Liu, Yichong Leng, Kaitao Song, Siliang Tang, Zhizheng Wu, Tao Qin, Xiangyang Li, Wei Ye, Shikun Zhang, Jiang Bian, Lei He, Jinyu Li, and Sheng Zhao. 2024. Naturalspeech 3: Zero-shot speech synthesis with factorized codec and diffusion models. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, Proceedings of Machine Learning Research, pages 22605–22623. PMLR / OpenReview.net.

Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. 2019. Audiocaps: Generating captions for audios in the wild. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 119–132. Association for Computational Linguistics.

Zhifeng Kong, Arushi Goel, Rohan Badlani, Wei Ping, Rafael Valle, and Bryan Catanzaro. 2024. Audio flamingo: A novel audio language model with few-shot learning and dialogue abilities. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, Proceedings of Machine Learning Research, pages 25125–25148. PMLR / OpenReview.net.

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: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. 2023. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Samuel Lipping, Parthasaarathy Sudarsanam, Konstantinos Drossos, and Tuomas Virtanen. 2022. Clotho-aqa: A crowdsourced dataset for audio question answering. In *30th European Signal Processing Conference, EUSIPCO 2022, Belgrade, Serbia, August 29 - Sept. 2, 2022*, pages 1140–1144. IEEE.

Shansong Liu, Atin Sakkeer Hussain, Chenshuo Sun, and Ying Shan. 2024. Music understanding llama: Advancing text-to-music generation with question answering and captioning. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2024, Seoul, Republic of Korea, April 14-19, 2024*, pages 286–290. IEEE.

Ziyang Ma, Yinghao Ma, Yanqiao Zhu, Chen Yang, Yi-Wen Chao, Ruiyang Xu, Wenxi Chen, Yuanzhe Chen, Zhuo Chen, Jian Cong, and 1 others. 2026. Mmar: A challenging benchmark for deep reasoning in speech, audio, music, and their mix. *Advances in Neural Information Processing Systems*, 38.

Xinhao Mei, Chutong Meng, Haohe Liu, Qiuqiang Kong, Tom Ko, Chengqi Zhao, Mark D. Plumbley, Yuexian Zou, and Wenwu Wang. 2023. Wavcaps: A chatgpt-assisted weakly-labelled audio captioning dataset for audio-language multimodal research. *CoRR*, abs/2303.17395.

Jan Melechovský, Zixun Guo, Deepanway Ghosal, Navonil Majumder, Dorien Herremans, and Soujanya Poria. 2024. Mustango: Toward controllable text-to-music generation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 8293–8316. Association for Computational Linguistics.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An ASR corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2015, South Brisbane, Queensland, Australia, April 19-24, 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, 23-29 July 2023, Honolulu, Hawaii, USA*, Proceedings of Machine Learning Research, 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 *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2001, 7-11 May, 2001, Salt Palace Convention Center, Salt Lake City, Utah, USA, Proceedings*, 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 *23rd Annual Conference of the International Speech Communication Association, Interspeech 2022, Incheon, Korea, September 18-22, 2022*, pages 4521–4525. ISCA.

S. Sakshi, Utkarsh Tyagi, Sonal Kumar, Ashish Seth, Ramaneswaran Selvakumar, Oriol Nieto, Ramani Duraiswami, Sreyan Ghosh, and Dinesh Manocha. 2025. MMAU: A massive multi-task audio understanding and reasoning benchmark. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. 2024. SALMONN: towards generic hearing abilities for large language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Qwen Team. 2025. Qwen3 technical report. *CoRR*, abs/2505.09388.

Zeyue Tian, Yizhu Jin, Zhaoyang Liu, Ruibin Yuan, Xu Tan, Qifeng Chen, Wei Xue, and Yike Guo. 2025. Audiox: Diffusion transformer for anything-to-audio generation. *CoRR*, abs/2503.10522.

Andros Tjandra, Yi-Chiao Wu, Baishan Guo, John Hoffman, Brian Ellis, Apoorv Vyas, Bowen Shi, Sanyuan Chen, Matt Le, Nick Zacharov, Carleigh Wood, Ann Lee, and Wei-Ning Hsu. 2025. Meta audiobox aesthetics: Unified automatic quality assessment for speech, music, and sound. *CoRR*, abs/2502.05139.

Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, Lei He, Sheng Zhao, and Furu Wei. 2023. Neural codec language models are zero-shot text to speech synthesizers. *CoRR*, abs/2301.02111.

LLM-Core Xiaomi. 2025. Mimo-audio: Audio language models are few-shot learners. *CoRR*, abs/2512.23808.

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 *Thirty-Ninth AAAI Conference on Artificial Intelligence, Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence, Fifteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2025, Philadelphia, PA, USA, February 25 - March 4, 2025*, 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. *CoRR*, abs/2502.04128.

Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi. 2021. Soundstream: An end-to-end neural audio codec. *CoRR*, abs/2107.03312.

Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J. Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu. 2019. Libritts: A corpus derived from librispeech for text-to-speech. In *20th Annual Conference of the International Speech Communication Association, Interspeech 2019, Graz, Austria, September 15-19, 2019*, pages 1526–1530. ISCA.

Xin Zhang, Dong Zhang, Shimin Li, Yaqian Zhou, and Xipeng Qiu. 2024. Speechtokenizer: Unified speech tokenizer for speech language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Hyperparameter	Value
D_{enc}	768
D_{align}	512
D_{quant}	14
GPUs	8 NVIDIA A100 80GB
Optimizer	AdamW
AdamW β_1, β_2	0.9, 0.999
Weight decay	10^{-2}
Gradient clipping	1.0
Stage 1 learning rate	2×10^{-4}
LR scheduler	ReduceLROnPlateau
Scheduler patience	5 epochs
LR reduction factor	0.5
Stage 2 learning rate	1×10^{-4}
Per-GPU batch size	32
Gradient accumulation	4
Global batch size	1024
Stage 1 steps	500k
Stage 2 steps	200k
Conditional dropout	10%

Table 5: Training hyperparameters of EntangleCodec.

A Claim of the Usage of AI Assistants

In preparing this manuscript, AI assistants were used only to improve the clarity, style, and readability of selected passages. They did not contribute to the study design, methodological development or implementation, data collection or analysis, or the generation of the manuscript's primary scientific contributions. All substantive research decisions, analyses, interpretations, and conclusions remain the sole responsibility of the authors.

B Implementation Details

B.1 Audio Preprocessing

All audio is resampled to 24 kHz. Log-Mel spectrograms are extracted with 128 Mel bins, a Hann window of 1024 samples (42.7 ms), a hop size of 480 samples (20 ms), and an FFT size of 1024, yielding a frame rate of 50 Hz. The log-Mel values are computed from power spectrograms and normalized to zero mean and unit variance per utterance using statistics estimated on the training set. During training, we randomly crop 10-second segments; utterances shorter than 10 seconds are padded with silence. At inference, arbitrary-length audio is processed without padding or chunking.

B.2 Hyperparameters

Table 5 and Table 6 summarize the hyperparameters used for EntangleCodec and the downstream audio language models.

Component	Hyperparameter	Value
Audio LM Understanding	Base model	Qwen3-0.6B
	Pretraining steps	50k
	Pretraining LR	3×10^{-5}
	SFT steps	10k
	SFT LR	1×10^{-5}
Audio LM Generation	Base model	Qwen3-0.6B
	Tasks	TTS and TTA
	Training epochs	3
	Batch size	256
	Learning rate	3×10^{-5}
Inference	CFG scale γ	1.0
	Vocoder	Vocos

Table 6: Training and inference hyperparameters of downstream audio language models.

B.3 Adversarial Decoder Refinement

In Stage 2, we introduce an adversarial objective to improve the perceptual quality of reconstructed Mel-spectrograms. The unified encoder, vector quantizer, and a pre-trained discriminator are kept frozen, and only the flow-matching decoder is updated. The discriminator acts as a fixed perceptual critic, providing adversarial supervision by comparing reconstructed Mel-spectrograms with ground-truth ones. This encourages the decoder to produce reconstructions that are both close to the target and perceptually realistic.

The objective of Stage 2 is

$$\mathcal{L}_{\text{stage2}} = \mathcal{L}_{\text{flow}} + \mathcal{L}_{\text{adv}}, \tag{13}$$

where all loss weights are set to 1.0. Since the encoder and quantizer remain frozen, the adversarial signal only affects the decoder and does not modify the learned semantic-acoustic token representation.

C Dataset Details

C.1 Training Corpus Statistics

Table 7 summarizes the datasets used to train EntangleCodec, covering three audio domains. In total, the training corpus comprises approximately 3,200 hours of audio.

C.2 Rich Audio Caption Generation

A central contribution of EntangleCodec is the use of rich audio captions rather than ASR transcripts for semantic alignment. We generate captions using MIMO-Audio (Xiaomi, 2025) with a structured prompt designed to elicit descriptions across four semantic dimensions:

(1) **Speaker attributes** — gender, estimated age, accent, speaking style, emotional state;

(2) **Acoustic environment** — indoor/outdoor, reverberation level, presence and type of background noise;

(3) **Musical attributes** — tempo, key, instrumentation, mood, dynamics;

(4) **Sound events** — event type, approximate onset and offset, spatial characteristics. For music-only clips, speaker attributes are omitted; for speech-only clips, musical attributes are omitted.

For LibriSpeech and LibriTTS, the original ASR transcript is appended to the caption as an additional semantic anchor.

The following examples illustrate the qualitative difference between ASR transcripts and our rich captions across all three audio domains.

C.3 LLM Training Data

Table 8 summarizes the datasets used for audio language model training across understanding and generation tasks.

D Baselines

We establish baselines along three evaluation dimensions: tokenizer reconstruction quality, audio understanding, and audio generation.

Tokenizer Reconstruction Quality. We compare EntangleCodec with seven state-of-the-art neural audio codecs: DAC (Kumar et al., 2023), EnCodec (Défossez et al., 2023), WavTokenizer (Ji et al., 2025), SpeechTokenizer (Zhang et al., 2024), XCodec (Ye et al., 2025a), Mimi (Défossez et al., 2024), and XCodec2 (Ye et al., 2025b). All codec baselines are evaluated using their official pretrained checkpoints under matched compression settings whenever applicable.

Audio Understanding. We evaluate EntangleCodec under a codec-controlled setting by training identical Qwen3-0.6B (Team, 2025) language models on discrete tokens produced by different codecs, while keeping the model architecture, training data, and optimization setup fixed. This isolates the effect of tokenizer quality on downstream understanding performance. In addition, we compare with specialized continuous-representation audio LLMs equipped with dedicated audio encoders, including LTU (Gong et al., 2024), LTU-AS (Gong et al., 2023), Audio Flamingo (Kong et al., 2024), MU-LLaMA (Liu et al., 2024), GAMA (Ghosh et al., 2024), SALMONN (Tang et al., 2024), and Qwen2-Audio (Chu et al., 2024), using official checkpoints or reported results when available.

Audio Generation. For generation tasks, we follow the same codec-controlled protocol as in audio understanding: identical LLM architectures are trained on discrete tokens produced by different codecs, with the training and decoding setup kept fixed. This allows us to directly assess how tokenizer quality affects downstream generation performance.

E Benchmark Details

E.1 Audio Understanding Benchmarks

MMAR. MMAR (Ma et al., 2026) is a multi-domain audio reasoning benchmark containing 1,000 single-choice questions across three domains: speech (covering speaker characteristics, emotion, prosody, and dialogue understanding), music (covering music theory, instrument recognition, and mood), and environmental sound (covering event recognition, scene classification, and causal reasoning). Each question has four answer choices. We report top-1 accuracy following the standard zero-shot evaluation protocol.

MMAU-mini and MMAU. MMAU (Sakshi et al., 2025) is a large-scale audio understanding benchmark with 9,000 questions in the test split spanning sound events, music, and speech. MMAU-mini is a stratified subset of 1,000 questions designed for rapid evaluation. All models are evaluated zero-shot with top-1 accuracy as the primary metric.

Example Questions. Figures 7, 8 and 9 illustrate the type of reasoning required by each benchmark and contrast the responses of EntangleCodec-LLM with a strong baseline.

E.2 Generation Benchmarks

SEED-TTS. SEED-TTS (Anastassiou et al., 2024) evaluates TTS systems on a diverse set of 3,000 utterances spanning multiple speakers and speaking styles. We report WER measured by Whisper-large-v3 (Radford et al., 2023) as the primary intelligibility metric, and UTMOS for naturalness.

AudioCaps (TTA). The AudioCaps (Kim et al., 2019) test set contains 979 clips, each paired with five human-written captions. We use the first caption as the generation condition. CLAP Score is computed as the cosine similarity between the CLAP text embedding of the condition and the CLAP audio embedding of the generated output.

Domain	Dataset	Hours	Clips	Avg. Dur. (s)
Speech	LibriSpeech (Panayotov et al., 2015)	960.0	281,241	12.3
	LibriTTS (Zen et al., 2019)	585.0	147,616	14.3
Music	MusicBench (Melechovský et al., 2024)	327.0	52,768	22.3
Sound	AudioSet (Gemmeke et al., 2017)	1,154.0	415,440	10.0
	AudioCaps (Kim et al., 2019)	105.8	38,118	10.0
	WavCaps (Mei et al., 2023)	127.0	45,720	10.0
Total		**3258.8**	**980,903**	

Table 7: Training data statistics for EntangleCodec.

Speech Caption Example — LibriSpeech, male speaker, read speech

ASR Transcript:
"He looked out across the grey expanse of water and said nothing for a long time."

Rich Caption (Ours):
"A middle-aged male speaker with a deep, measured voice reads **"He looked out across the grey expanse of water and said nothing for a long time"** aloud at a slow, deliberate pace. The tone is calm and somewhat solemn, conveying a reflective mood. The speaker has a mild British Received Pronunciation accent. The recording is clean and studio-quality, with very low background noise and minimal room reverberation. No music or non-speech sound events are present."

Figure 4: Example of a rich speech caption generated by our captioning pipeline.

Task	Dataset / Type	Samples
Understanding	MusicQA (Liu et al., 2024) / Music QA	70,011
	ClothoAQA (Lipping et al., 2022) / Sound QA	21,114
	COMPA-R (Ghosh et al., 2024) / Audio reasoning	200,234
TTS	LibriSpeech (Panayotov et al., 2015) / (text, speech)	281,241
	LibriTTS (Zen et al., 2019) / (text, speech)	354,780
TTA	IFCaps (Tian et al., 2025) / (caption, audio)	24,000
	AudioCaps (Kim et al., 2019) / (caption, audio)	38,118

Table 8: Audio language model training data.

Model	UTMOS↑	F1↑	STOI↑	SIM↑
GT	4.06	0.98	1.00	1.00
DAC	1.28	0.97	0.61	0.24
EnCodec	1.54	0.92	0.75	0.23
WavTokenizer	3.76	0.98	**0.89**	0.63
SpeechTokenizer	1.25	0.97	0.63	0.16
Xcodec	3.40	0.97	0.84	0.46
Mimi	3.09	0.99	0.85	0.50
Xcodec2	**4.00**	0.99	0.87	**0.75**
EntangleCodec	3.94	**0.99**	0.86	0.65

Table 9: Reconstruction quality on LibriTTS *test-clean*. **Bold** = best; underline = second best (GT excluded).

Clotho (TTA). The Clotho (Drossos et al., 2019) evaluation split contains 1,045 clips with five captions each. We follow the same evaluation protocol as AudioCaps. Clotho covers a broader range of acoustic scenes and environmental sounds with longer, more descriptive captions, providing a complementary evaluation to AudioCaps.

LibriTTS Reconstruction. To complement the LibriSpeech reconstruction results in Table 1, we additionally evaluate all tokenizers on LibriTTS *test-clean* (Zen et al., 2019). LibriTTS features longer utterances (average 14.3 s vs. 12.3 s for LibriSpeech) and more diverse speaking styles including expressive and conversational speech, providing a more challenging reconstruction testbed. Results are reported in Table 9.

The LibriTTS results are consistent with those reported on LibriSpeech in the main paper: EntangleCodec achieves an UTMOS of **3.94**, within 0.06 points of the top-performing Xcodec2 (4.00) and substantially ahead of reconstruction-focused codecs such as DAC (1.28) and EnCodec (1.54). The slight overall drop compared to LibriSpeech results (e.g., WavTokenizer: 3.79 → 3.76) is consistent across all methods, attributable to the greater prosodic variability and longer utterance lengths in LibriTTS.

Music Caption Example — MusicBench, piano solo

Rich Caption (Ours):
"A solo piano piece in a minor key, played at a slow to moderate tempo of approximately 72 BPM. The melody is introspective and melancholic, with a sparse left-hand accompaniment using sustained bass notes. Dynamic range is moderate, with soft passages punctuated by occasional crescendos. The recording has a natural concert-hall reverb. No vocals or other instruments are present. The overall mood is contemplative and slightly sorrowful."

Figure 5: Example of a rich music caption generated by our captioning pipeline.

Sound Caption Example — AudioSet, urban environment

Rich Caption (Ours):
"A busy urban street soundscape recorded outdoors. Continuous traffic noise forms the background, with cars passing at irregular intervals. A car horn honks twice in quick succession approximately 2 seconds into the clip. Faint footsteps on pavement are audible in the right channel. The acoustic environment has moderate urban reverberation, consistent with a narrow street between medium-height buildings. No speech or music is present."

Figure 6: Example of a rich sound caption generated by our captioning pipeline.

F Evaluation Metrics

Reconstruction Quality. We evaluate tokenizer reconstruction using complementary objective and perceptual metrics. **UTMOS** (Saeki et al., 2022) is a deep learning-based predictor of mean opinion score (MOS) for speech quality, ranging from 1 to 5, where higher is better. **STOI** measures the intelligibility of speech signals, ranging from 0 to 1, where higher values indicate better intelligibility. **F1 Score** evaluates voiced/unvoiced frame classification accuracy, where higher is better. For sound and music, we additionally report **AudioBoxScore** (Tjandra et al., 2025), which averages four sub-scores to assess overall audio quality across domains.

Audio Understanding. We report **Accuracy** for audio understanding tasks, following the standard evaluation protocol of each benchmark.

Audio Generation. For TTS, we report **WER** (Word Error Rate), measured by an automatic speech recognition system, where lower is better. We also report UTMOS for naturalness assessment. For TTA, we report **CLAP Score** (Elizalde et al., 2023), which measures semantic alignment between generated audio and the conditioning text description.

G Token Space UMAP Visualization

We visualize EntangleCodec representations with UMAP to examine whether the learned tokens capture meaningful semantic structure. For each audio clip, we compute the mean token embedding and project it to two dimensions. As shown in Figure 10, EntangleCodec forms clear domain-level clusters for speech, music, and general sound, while also exhibiting meaningful intra-domain organization, such as acoustic/electronic music and slow tempo patterns.

We further visualize the joint audio–text embedding space used for caption alignment. The audio and caption clusters are closely aligned across domains, indicating that rich caption supervision encourages the tokenizer to encode semantic information beyond reconstruction-oriented acoustic detail. These visualizations support the quantitative results, showing that EntangleCodec learns discrete representations that are both acoustically grounded and semantically structured.

Example 1 — Speech: Behavior Recognition

Audio A clip of a male speaker performing a vocal imitation, with exaggerated tonal patterns and rhythmic speech, recorded in an indoor environment.

Question Why did the speaker sing a song at the end?
(A) To express awe for the lion
(B) To showcase his musical talent
(C) To imitate Africans

Ground Truth: (C) To imitate Africans
EntangleCodec-LLM (0.6B): **(C) To imitate Africans** ✓
XCodec2-LLM (0.6B): (B) To showcase his musical talent ×
Speechtokenizer-LLM (0.6B): (B) To express awe for the lion ×

The rhythmic and tonal characteristics of the vocal performance carry cultural imitation cues. EntangleCodec's rich caption supervision enables the model to distinguish intentional imitation from musical performance.

Figure 7: MMAR Example 1 — Speech: Behavior Recognition.

Example 2 — Music: Humor Reduction

Audio A clip of a musical piece with a comedic tone, featuring a notable pitch pattern that contributes to its humorous effect.

Question How to make this song less funny?
(A) Change the note F# to F
(B) Change the key from F major to C major
(C) Change the note D to C#
(D) Change the note F# to D

Ground Truth: (D) Change the note F# to D
EntangleCodec-LLM (0.6B): **(D) Change the note F# to D** ✓
XCodec2-LLM (0.6B): (A) Change the note F# to F ×

The humorous effect stems from the unexpected use of F#, which creates a comedic tension. Replacing it with D resolves this tension naturally. EntangleCodec's fine-grained pitch supervision enables the model to identify the specific note responsible for the comedic character of the piece.

Figure 8: MMAR Example 2 — Music: Humor Reduction.

Example 3 — Sound: Fine-grained Audio Discrimination

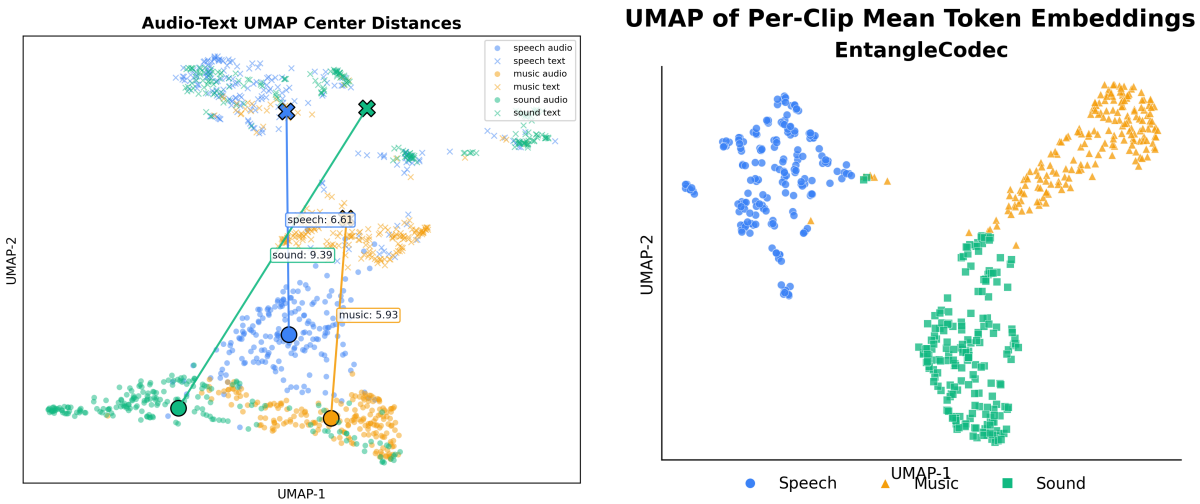
Audio A clip ending with a brief water-related sound, requiring fine-grained discrimination between two acoustically similar sound events.

Question Is the sound at the end of the audio pouring water or spitting water?
(A) Pouring water sound
(B) Spitting water sound

Ground Truth: (B) Spitting water sound
EntangleCodec-LLM (0.6B): **(B) Spitting water sound** ✓
XCodec2-LLM (0.6B): (A) Pouring water sound ×

The two sounds share similar acoustic properties but differ in subtle spectral and temporal patterns. EntangleCodec's fine-grained audio caption supervision enables the model to capture these nuanced distinctions, while XCodec2-LLM conflates the two acoustically similar sound events.

Figure 9: MMAR Example — Sound: Fine-grained Audio Discrimination.



(a) Audio–text UMAP center distances.

(b) Per-clip mean token embeddings.

Figure 10: UMAP visualization of EntangleCodec representations. Left: audio and caption embeddings are projected into a shared space, where smaller center distances indicate closer audio–text alignment. Right: per-clip mean token embeddings form clear domain-level clusters for speech, music, and general sound, with additional intra-domain structure.