UniCodec: Unified Audio Codec with Single Domain-Adaptive Codebook

Anonymous ACL submission

Abstract

The emergence of audio language models is empowered by neural audio codecs, which establish critical mappings between continuous waveforms and discrete tokens compatible with language model paradigms. The evolutionary trends from multi-layer residual vector quantizer to single-layer quantizer are beneficial for language-autoregressive decoding. However, the capability to handle multi-domain audio signals through a single codebook remains constrained by inter-domain distribution discrepancies. In this work, we introduce **UniCodec**, a unified audio codec with a single codebook to support multi-domain audio data, including *speech*, *music*, and *sound*. To achieve this, we propose a partitioned domain-adaptive codebook method based on domain Mixture-of-Experts strategy to capture the distinct characteristics of each audio domain. Furthermore, to enrich the semantic density of the codec without auxiliary modules, we propose a self-supervised mask prediction modeling approach. Comprehensive objective and subjective evaluations demonstrate that UniCodec achieves excellent audio reconstruction performance across the three audio domains, outperforming existing unified neural codecs with a single codebook, and even surpasses state-of-the-art *domain-specific* codecs on both acoustic and semantic representation capabilities[1].

1 Introduction

Many recent developments of speech language models (SLMs) (Bai et al., 2023; Défossez et al., 2024; Peng et al., 2024; Ji et al., 2024a) integrate the speech modality with text-based large language models (LLMs) and have led to significant advancements in speech understanding and generation tasks. This paradigm relies on discrete acoustic codec models, which convert high-rate speech signals into a finite set of discrete speech tokens, bridging the gap between continuous speech signals and discrete-token-based language models, thus enabling speech applications powered by LLMs.

Most existing neural audio codecs (NACs) (Zeghidour et al., 2022; Kumar et al., 2023; Ji et al., 2024b; Défossez et al., 2023; Défossez et al., 2024) employ a *multi-layer* Residual Vector Quantizer (RVQ), where each quantizer operates on the residual of the previous quantizer. This RVQ structure generates multiple parallel hierarchical token streams for downstream language models to decode, hence it increases the complexity and the generation latency of SLMs (Xie and Wu, 2024a,b; Défossez et al., 2024). To address this problem, several recent works, including WavTokenizer (Ji et al., 2024c), Single-Codec (Li et al., 2024), and BigCodec (Xin et al., 2024), focus on developing *single-layer* quantizer to streamline the process. Integrating a single-layer quantizer with LLMs facilitates rapid extraction of speech features on input audio while significantly reducing the burden of autoregressive modeling. These works demonstrate that using a single VQ to discretize speech could achieve competitive performance in both audio reconstruction and generation tasks. Therefore, our work follows this trend and focuses on developing high-performing single-layer quantizer codec.

An ideal codec should be able to perform well across various audio domains, such as speech, music, and sound, with distinct domain characteristics. Prior RVQ-based neural audio codecs using *multi-layer RVQ and hence multi-codebooks*, such as DAC (Kumar et al., 2023) and Encodec (Défossez et al., 2023), exhibit strong reconstruction capabilities for speech, music, and sound. However, previous studies such as Wavtokenizer (Ji et al., 2024c) show that using a *unified single-codebook codec* for speech, music, and sound still poses a great challenge: The unified codec suffers from notable per-

[1] We will make our code and model checkpoints publicly available to ensure reproducibility.

Table 1: Comparison of recent codec models based on single codebook, compatibility with speech, music, and sound domains, and whether they use *separate* models for different domains or a *unified* model.

Model	Single Codebook	Speech	Music&Sound	Separate/Unified model
DAC (Kumar et al., 2023)	✘	✔	✔	Unified
Encodec (Défossez et al., 2023)	✘	✔	✔	Unified
Mimi (Défossez et al., 2024)	✘	✔	✔	Unified
SemantiCodec (Liu et al., 2024)	✘	✔	✔	Unified
SpeechTokenizer (Zhang et al., 2023)	✘	✔	✘	-
BigCodec (Xin et al., 2024)	✔	✔	✘	-
TAAE (Parker et al., 2024)	✔	✔	✘	-
Wavtokenizer (Ji et al., 2024c)	✔	✔	✔	Separate&Unified
UniCodec	✔	✔	✔	**Unified**

formance degradation compared to domain-specific codec models, since the substantial distribution discrepancies between these domains make it difficult to effectively capture their distinct characteristics with a single codebook. To tackle this challenge, in this work, we develop a **unified audio codec with a single codebook, designed to support multiple audio domains—including speech, music, and sound—while achieving both low bitrate and high acoustic reconstruction quality**.

In addition to powerful acoustic reconstruction capabilities, strong semantic representation capabilities (that is, encapsulating rich semantic information) of NACs are crucial for effective integration of NACs with LLMs, since strong semantic capabilities can ease understanding of audio content and facilitate generation of semantically reasonable audio. There are two main challenges in enriching the semantic representations of NACs. (1) There is an inherent trade-off between semantic richness and reconstruction performance, since semantic features provide a higher-level, more abstract understanding, while reconstruction features emphasize fine-grained details of audio. (2) The majority of existing works enrich semantic capabilities through distillation from additional pretrained speech semantic encoders (Zhang et al., 2023; Défossez et al., 2024), separate semantic codebooks (Liu et al., 2024), or auxiliary semantic modules (Ye et al., 2024). However, methods using an additional pretrained semantic encoder are constrained by reliance on a pretrained speech encoder, are less elegant and not fully adaptable, and difficult to support unified modeling of speech, music, and sound. Moreover, an auxiliary semantic module introduces additional computation cost and degrades the efficiency of feature extraction. Since both reconstruction quality and efficiency are critical for NACs, we explore a more elegant approach by **directly learning semantic information through the codec itself, without additional modules, while preserving high reconstruction ability**.

Our contributions can be summarized as follows:

- We introduce UniCodec, a unified audio codec with a single quantizer, designed to support various audio types, including speech, music, and sound, with a single codebook. To achieve this, we propose a partitioned domain-adaptive codebook method based on domain Mixture-of-Experts (MoE) strategy to effectively capture the distinct characteristics of each audio domain.
- We propose a self-supervised, masked modeling approach to enrich semantic information without extra modules.
- Comprehensive objective and subjective evaluations show that UniCodec achieves better reconstruction and semantic performance compared to existing unified codecs with a single codebook, and even outperforms domain-specific codecs.

2 Related Work

Neural Audio Codecs Neural Audio Codecs (NACs) aim to compress audio signals into highly compressed discrete tokens while preserving high reconstruction quality. The predominant paradigm of NACs utilizes the Vector Quantized Variational Autoencoder (VQ-VAE) (van den Oord et al., 2017; Gârbacea et al., 2019) architecture, where an encoder transforms the audio signal into a latent representation, a quantizer discretizes this representation, and a decoder reconstructs the signal. SoundStream (Zeghidour et al., 2022) enhances this approach by incorporating Residual Vector Quantization (RVQ), and improves both modeling and reconstruction capabilities for NACs. Encodec (Défossez et al., 2023) further refines

SoundStream by introducing multi-scale discriminators and a loss-balancing strategy to optimize reconstruction performance. Numerous works such as DAC (also named RVQGAN) (Kumar et al., 2023) and Mimi (Défossez et al., 2024) continue enhancing RVQ-based NACs. While multi-codebook residual modeling boosts reconstruction quality, it complicates the autoregressive process in SLMs and suffers from unacceptable latency. In contrast, single-layer quantizer codecs, such as Single-Codec (Li et al., 2024), WavTokenizer (Ji et al., 2024c), BigCodec (Xin et al., 2024), and TAAE (Parker et al., 2024), show promising potentials due to their ability to seamlessly integrate into SLMs with low latency and reduced computational overhead. However, there is still much room to improve the performance of single-layer low-bitrate codecs; hence, this work focuses on enhancing single-layer low-bitrate codecs.

Unified Audio Signal Modeling A unified NAC capable of processing various audio types, such as speech, music, and sound, will be greatly beneficial for constructing universal audio language models (ALMs) that are generalizable to various audio types. RVQ-based audio codec models, such as SoundStream (Zeghidour et al., 2022), Encodec (Défossez et al., 2023), and DAC (Kumar et al., 2023), are trained on a combination of speech, music, and sound datasets. While these codecs achieve high reconstruction quality, their performance significantly degrades in low-bitrate scenarios, particularly when restricted to the first codebook. Although existing single-layer codecs (Ji et al., 2024c) perform well in one or two audio domains, they struggle to simultaneously maintain superior performance on speech, music, and sound domains while operating at a low bitrate.

Semantic Audio Representation Learning Discrete tokens compressed by acoustic NACs lack high-level semantic information, which is essential for effective SLMs. To address this issue, models such as SpeechTokenizer (Zhang et al., 2023) and Mimi (Défossez et al., 2024) leverage self-supervised-learning (SSL) based speech representation models to distill semantic information into the first-layer codebook. XCodec (Ye et al., 2024) concatenates acoustic tokens with semantic tokens produced by SSL models before the RVQ stage and introduces a semantic reconstruction loss. FunCodec (Du et al., 2024) offers various methods to integrate SSL-based semantic tokens with RVQ-based acoustic tokens. However, these approaches rely on SSL encoders, which complicate the training process and constrain the semantic capabilities of NACs. SemantiCodec (Liu et al., 2024) combines quantized semantic tokens with acoustic tokens and introduces a diffusion process to enhance reconstruction quality, but the diffusion process introduces additionally training cost. In contrast, UniCodec requires neither additional SSL encoders nor complex diffusion process, hence simplifying the training process while encapsulating rich semantic information.

3 Methodology

UniCodec is built upon the highly competitive single-layer encoder-VQ-decoder codec, Wavtokenizer (Ji et al., 2024c). The left part of Figure 1 provides an overview of the architecture of UniCodec, which comprises three modules: an encoder that processes the input audio to generate a latent feature representation, a quantizer that discretizes the feature into tokens through a single codebook, and a decoder that reconstructs the audio signal from the compressed, discrete tokens. We first make the following enhancements over Wavtokenizer (Section 3.1). We enhance the encoder by incorporating transformer layers, which can better capture and represent complex patterns. We also enhance the codebook utilization rate to maximize the use of codebook and improve efficiency. More importantly, to build a unified tokenizer capable of supporting multi-domain audio reconstruction, we propose two novel strategies: a partitioned domain-adaptive codebook (Section 3.2), and a domain mixture-of-experts (MoE) encoder structure (Section 3.3), which is detailed in the upper-right part of Figure 1. UniCodec is trained end-to-end through two stages. In the first acoustic training stage, the model is trained by optimizing a reconstruction loss applied over both time and frequency domains, along with a perceptual loss using discriminators operating at different resolutions, the same as Wavtokenizer. In the following semantic training stage (Section 3.4), which is depicted in the lower-right part of Figure 1), a contrastive loss is added into the training objective.

3.1 Enhanced Encoder and Quantizer

The encoder of Wavtokenizer (Ji et al., 2024c) consists of convolutional blocks followed by a two-layer LSTM and a final 1D convolution layer,

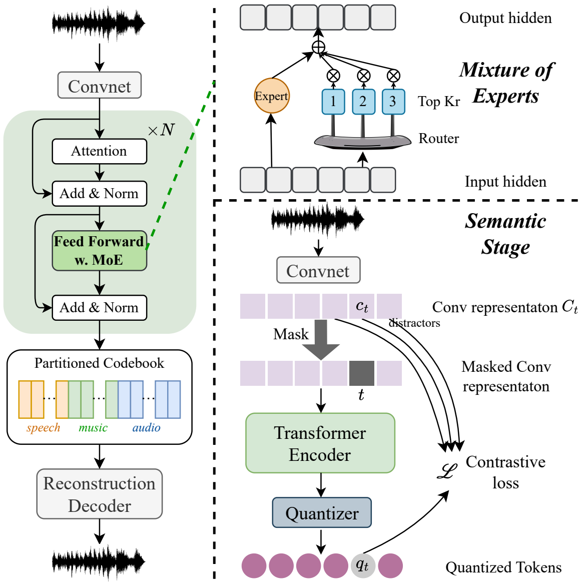

Figure 1: Left: Overview of the proposed UniCodec. Upper-right: the domain MoE encoder structure. Lower-right: the semantic training stage.

which limits its capacity for effective feature extraction. To enhance the ability to encode audio into compact representations while ensuring high-quality audio reconstruction, inspired by Mimi Codec in Moshi (Défossez et al., 2024), we replace the LSTM sequence modeling in the encoder with a contextual Transformer architecture following the convolutional blocks. Consistent with Mimi, the Transformer consists of 8 layers, 8 attention heads, RoPE position encodings, GELU activations (Hendrycks and Gimpel, 2016), with a hidden size of 512 and an MLP dimension of 2048.

Scaling the training data to cover multiple audio domains necessitates scaling the codebook concurrently, which introduces the challenge of optimizing codebook utilization during the vector quantization process. To improve codebook utilization and improve efficiency, we adopt the SimVQ algorithm (Zhu et al., 2024), which effectively and efficiently mitigates the issue of representation collapse in vector-quantized model by using a simple linear layer.

3.2 Domain-adaptive Codebook

To achieve seamless integration of data from three distinct domains—speech, music, and sound—into a unified audio tokenizer, we propose a novel partitioned domain-adaptive codebook. In this framework, the codebook is divided into three specialized regions: the first region, spanning indices 0 to 4095, is dedicated to the speech domain; the second, from 4096 to 8191, is for the music domain; and the remaining indices from 8191 to 16383 are allocated for the sound domain. This design is inspired by the hypothesis in Semanticodec (Liu et al., 2024) that general sound tends to encompass a broader range of sounds than speech and music, hence we allocate a larger region for sound. During the training process, the model only updates the codebook entries corresponding to the domain of the input sample, ensuring that domain-specific features are accurately captured and learned. This partitioned codebook approach facilitates the construction of a unified audio tokenizer that can effectively handle the unique characteristics of each domain, providing a flexible solution for multi-domain audio representation. The ablation experimental results in Table 6 of Section 5.3 validate this strategy achieves performance improvement when scaling up the amount of training data covering different audio types and also codebook size.

3.3 Domain MoE

For training the codec on data from multiple audio domains, we employ a domain Mixture-of-Experts (MoE) strategy for the Feed-Forward Networks (FFNs) in our Transformer encoder, inspired by the DeepSeekMoE architecture (Dai et al., 2024). Different from traditional MoE architectures, such as GShard (Lepikhin et al., 2020), DeepSeekMoE utilizes finer-grained experts, designates some as *shared experts* and the rest as *routed experts* This architectural design is well-suited to capture domain-specific features while maintaining high performance and computational efficiency. For the FFN input u_t of the t-th token, the computation of the FFN hidden output h_t can be formulated as follow:

$$h_t = u_t + \sum_{i=1}^{N_s} FFN_i^s(u_t) + \sum_{i=1}^{N_r} g_{i,t} FFN_i^r(u_t) \tag{1}$$

$$g_{i,t} = \frac{g'_{i,t}}{\sum_{j=1}^{N_r} g'_{j,t}} \tag{2}$$

$$g'_{i,t} = \begin{cases} s_{i,t}, & s_{i,t} \in Topk(s_{j,t} | 1 \leq j \leq N_r, K_r) \\ 0, & otherwise \end{cases} \tag{3}$$

$$s_{i,t} = Sigmoid(u_t^T e_i) \tag{4}$$

where N_s and N_r denote the numbers of shared experts and routed experts, respectively. $FFN_i^s(\cdot)$ and $FFN_i^r(\cdot)$ denote the i-th shared expert and the i-th routed expert, respectively. $g(i,t)$ is the gating value for the i-th expert. K_r is the number of activated routed experts. si,t is the token-to-expert affinity. e_i is the centroid vector of the i-th routed expert, and $Topk(\cdot, K)$ denotes the set comprising K highest scores among the affinity scores calculated for the t-th token and all routed experts. Considering the trade-off between computational cost and performance on all three audio domains, we set $N_s = 1$, $N_r = 3$, and $K_r = 1$.

3.4 Semantic Training Stage

To simultaneously enhance semantic representation capabilities while preserving high reconstruction ability, we introduce a domain-agnostic masked modeling approach for UniCodec, inspired by Wav2Vec 2.0 (Baevski et al., 2020). Notably, our approach does not add any extra modules. Specifically, we mask a proportion of the features output from the convolution layers in the encoder before passing them into the contextual Transformer layers. Following the masking strategy of Wav2Vec 2.0 (Baevski et al., 2020), we randomly sample a proportion p of all time steps to serve as starting indices and then mask the subsequent M consecutive time steps from each sampled index, allowing overlapping spans.

After the contextual Transformer layers and the quantizer, the quantized output q_t, centered over the masked time step t, requires the model to identify the unmasked convolutional latent representation c_t from a set of $K + 1$ convolutional latent representations $\hat{c} \in C_t$, which includes c_t and K distractors (Gutmann and Hyvärinen, 2010; Oord et al., 2018). These distractors are uniformly sampled from other masked time steps within the same utterance. The contrastive loss is computed as:

$$L_m = -log\frac{exp(sim(q_t, c_t)/K)}{\sum_{\hat{c} \in C_t} exp(sim(q_t, \hat{c})/K)} \tag{5}$$

where we compute the cosine similarity $sim(a, b) = a^T b/||a||||b||$ between quantized tokens and unmasked convolutional latent representations (He et al., 2020; Chen et al., 2020).

Our preliminary experiments show that training from scratch with reconstruction, masked modeling, and contrastive loss is challenging, as the single-quantizer codec struggles to simultaneously perform reconstruction and mask prediction. Therefore, we first train the codec model with reconstruction-related loss following Wavtokenizer in the **initial acoustic training stage**, omitting the masking strategy. Then we introduce this **semantic training stage** with a more difficult mask prediction goal, allowing the codec to encapsulate high-level semantic information after acquiring initial reconstruction ability.

4 Experimental Setup

Datasets. We train UniCodec on approximately 80,000 hours of data spanning speech, music, and audio domains. For the speech domain, we use Librilight (Kahn et al., 2020), LibriTTS (Zen et al., 2019), VCTK (Veaux et al., 2016), and CommonVoice (Ardila et al., 2019). For the music domain, we use Jamendo (Bogdanov et al., 2019) and MusicDB (Rafii et al., 2017) datasets. For the audio domain, we use AudioSet (Gemmeke et al., 2017). We evaluate the speech reconstruction performance on LibriTTS test-clean. We evaluate the audio and music reconstruction performance on the AudioSet eval and MusicDB test sets, respectively.

Training details. Throughout the entire training process, all input speech, music, and audio samples are resampled to 24 kHz. The batch size is 10×32 on 32 NVIDIA A800 80G GPUs. We uniformly truncate excessively long segments in the training data to a fixed duration of 10 seconds and feed them into the model. We use the AdamW optimizer (Kingma and Ba, 2015; Loshchilov and Hutter, 2019) with an initial learning rate of 2e-4 and betas set to (0.9, 0.999). The learning rate is decayed based on a cosine scheduler (Loshchilov and Hutter, 2017).

During training, we provide a domain ID for each sample to allow the model to use partitioned domain-adaptive codebook to capture the distinct characteristics of each domain. However, for fair comparisons during evaluation, we do not provide domain IDs; instead, we rely on the codebook to autonomously learn the distinct paradigms of each domain and rely on the quantizer to select the nearest token from the entire codebook. As explained in Section 3, we design initial acoustic training and semantic training stages for UniCodec to balance acoustic and semantic capabilities. We follow the Wav2vec 2.0 (Baevski et al., 2020) mask strategy and configuration. The mask ratio p and mask length M is set to 0.1 and 5.

Training with large-scale and diverse dataset in both acoustic and semantic stages ensure generalization ability of UniCodec. However, our preliminary experiments indicate that large-scale data

Table 2: **Objective reconstruction results** of UniCodec and baselines on **speech, music and audio** domains on LibriTTS test-clean, MusicDB test set, and Audioset eval set, in terms of Mel Distance and STFT Distance. **TPS** denotes token per second. We **bold** the best results in all the models, and **bold and underline** the best results in single-codebook codec models.

Model	Unified	TPS↓	LibriTTS test-clean		MusicDB test		AudioSet eval	
			Mel Dist↓	STFT Dist↓	Mel Dist↓	STFT Dist↓	Mel Dist↓	STFT Dist↓
DAC	✔	600	0.3697	1.5525	**0.3578**	**1.9621**	0.4581	2.1378
Encodec	✔	600	0.5367	1.8271	0.5565	2.1678	0.7601	2.6273
Mimi	✔	100	0.6709	1.9859	0.6714	2.2526	0.8406	2.6639
TAAE	✘	50	0.7508	2.2426	1.4067	4.1340	1.9335	5.2897
DAC	✘	75	0.7217	2.1662	1.8894	6.2476	1.7063	5.2923
BigCodec	✘	80	0.4427	1.7385	1.3803	4.2366	1.8632	5.6171
Wavtokenizer (speech)	✘	75	0.5001	1.7879	0.6586	3.0335	0.5990	2.5479
Wavtokenizer (music/audio)	✘	75	0.5451	1.8649	0.4516	2.2450	0.4536	2.1871
Wavtokenizer (unified)	✔	75	0.5308	1.8614	0.5435	2.5451	0.5193	2.3727
UniCodec **(Ours)**	✔	75	**0.3442**	**1.5147**	**0.3959**	**2.1822**	**0.3820**	**2.1065**

training performs worse compared to training on only LibriTTS dataset. Upon analysis, we find that diverse and noisy data significantly hinders codec reconstruction learning. To further improve the reconstruction ability, we select high-quality data for a further **fine-tuning stage**. More details about the fine-tuning stage are in Appendix C.

Evaluation Metrics. We adopt a comprehensive set of evaluation metrics, as follows.

Tokens Per Frame (TPF): The number of parallel tokens per timestep of encoded audio, affecting ease of modeling token sequences in generative models.

Tokens Per Second (TPS): The number of tokens per second. It determines the context length required by a generative model, especially when residual tokens are used in flattened form.

Downsample Rate (DR): The token compression rate. It is calculated by dividing the input audio sample rate by TPS, indicating the difficulty of compressing audio waveforms into tokens.

Mel Distance (Reconstruction): L1 distance between the mel-scaled magnitude spectrograms of the ground truth and the generated sample.

STFT Distance (Reconstruction): L1 distance between time-frequency representations of the ground truth and the prediction, computed using multi-scale Short-Time Fourier Transform (STFT).

More details about the metrics for speech reconstruction evaluation can be found in Appendix E.

Baselines. We select both state-of-the-art (SOTA) *multi-layer* quantizer codec models and *single-layer* quantizer codec models as the baselines. For multi-layer codecs, we compare against DAC (Kumar et al., 2023), Encodec (Défossez et al., 2023), SpeechTokenizer (Zhang et al., 2023), and Mimi (Défossez et al., 2024). For single-layer codecs, we compare with the official checkpoints provided by Wavtokenizer (speech) [2], Wavtokenizer (music and audio) [3], BigCodec (Xin et al., 2024) [4], and TAAE (Parker et al., 2024) [5].

5 Results and Discussions

5.1 Reconstruction Evaluation

We compare the reconstruction performance of UniCodec against a broad selection of SOTA and competitive codec models as baselines. Table 2 presents the results of UniCodec and baselines on speech (LibriTTS test-clean), music (MusicDB test), and audio (AudioSet eval) domains, in terms of Mel Distance and STFT Distance. As shown in Table 2, UniCodec demonstrates excellent reconstruction performance on all three domains, outperforming the unified single-codebook baseline Wavtokenizer (unified) and also speech-specific single-codec baselines such as BigCodec, TAAE, and Wavtokenizer (speech). In the music and audio domains, UniCodec also outperforms the music/audio-specific baseline Wavtokenizer (music/audio) on both MusicDB test set and AudioSet eval set. Even when compared to multi-layer RVQ-based unified baselines such as Encodec and Mimi, the single-layer unified UniCodec shows superior performance across all three domains, except for slightly lower performance compared to DAC (which has a much larger tokens-per-second rate) in the music domain. The Real-Time Factors (RTF) and comparisons of the number of parameters can be found in Appendix B.

[2] wavtokenizer_medium_speech_320_24k_v2.ckpt
[3] wavtokenizer_medium_music_audio_320_24k_v2.ckpt
[4] huggingface.co/Alethia/BigCodec/resolve/main/bigcodec.pt
[5] huggingface.co/stabilityai/stable-codec-speech-16k

Table 3: **Objective reconstruction results** on the **Speech** domain from UniCodec and baselines on LibriTTS test-clean, in terms of naturalness, distortion, and intelligibility. **DR** denotes the Downsample Rate (the input audio sample rate division by Tokens Per Second (TPS)). **Unified** denotes the codec model can support all three domains of speech, music, and sound. The results of models marked by † are cited from the Wavtokenizer paper (Ji et al., 2024c) and others are reproduced by us based on the checkpoints released by the corresponding work.

Model	Unified	DR (↑)	TPF (↓)	TPS (↓)	PESQ (↑)	STOI (↑)	F1 (↑)	UTMOS (↑)
Ground Truth†	-	-	-	-	-	-	-	4.0562
DAC	✔	40	8	600	3.5197	0.9709	0.9546	3.6905
Encodec†	✔	40	8	600	2.7202	0.9391	0.9527	3.0399
SpeechTokenizer†	✘	40	8	600	2.6121	0.9165	0.9495	3.8794
Mimi	✔	240	8	100	2.2695	0.9118	0.912	3.5731
TAAE	✘	320	2	50	1.8955	0.8816	0.9260	4.1389
DAC	✘	320	1	75	1.1763	0.7739	0.7560	1.3531
BigCodec	✘	200	1	80	2.6872	0.9293	0.9480	4.0367
Wavtokenizer (speech)†	✘	320	1	75	2.3730	0.9139	0.9382	**4.0486**
Wavtokenizer (unified)	✔	320	1	75	1.8379	0.8718	0.9175	3.6115
UniCodec **(Ours)**	✔	320	1	75	**3.0266**	**0.9493**	**0.9486**	3.9873

Table 4: **Subjective MUSHRA test reconstruction results** from codec models on **speech, music and audio** domains, on LibriTTS test-clean, MusicDB test set and AudioSet eval set. We report mean and standard deviation.

Model	Unified	LibriTTS test-clean (↑)	MusicDB test (↑)	AudioSet eval (↑)
Ground Truth	-	93.52 ± 1.99	96.18 ± 1.47	95.28 ± 2.18
Wavtokenizer (speech)	✘	85.44 ± 2.29	-	-
Wavtokenizer (music & audio)	✘	-	75.24 ± 2.38	80.19 ± 2.43
Wavtokenizer (unified)	✔	80.40 ± 2.54	56.10 ± 3.74	62.21 ± 3.42
UniCodec **(Ours)**	✔	**90.74 ± 2.06**	**77.77 ± 2.45**	**82.43 ± 2.56**

Table 3 further compares the speech domain reconstruction performance of different codec models on **LibriTTS test-clean**, using PESQ, STOI, F1 and UTMOS, assessing the codecs in terms of naturalness, distortion, and intelligibility. The unified UniCodec significantly outperforms WavTokenizer (unified) across all metrics. Even compared to WavTokenizer (speech) and BigCodec, which are SOTA speech-specific models with single-layer quantizers, UniCodec achieves better PESQ and STOI, demonstrating superior reconstruction quality. Furthermore, despite having a much higher downsampling rate (DR), UniCodec remains competitive with multi-layer quantizer models such as Encodec, Mimi, and SpeechTokenizer, which have higher tokens per second (TPS). Appendix A also reports the reconstruction performance on **LibriTTS test-other**.

The reconstruction results of the MUSHRA subjective test are shown in Table 4. UniCodec outperforms WavTokenizer (unified) markedly in reconstruction quality across speech, music, and audio domains. Even when compared to domain-specific codecs, UniCodec performs slightly better than WavTokenizer (speech) in the speech domain, and WavTokenizer (music/audio) in the music and audio domains. These results further demonstrate that **in all three domains, UniCodec achieves superior subjective reconstruction performance while maintaining a high compression rate**.

5.2 Semantic Evaluation

We evaluate the semantic richness of different codec models on several speech, music, and audio domain datasets of the ARCH benchmark (La Quatra et al., 2024). The speech domain includes the RAVDESS (Livingstone and Russo, 2018) and Audio-MNIST (Becker et al., 2024) datasets, the music domain includes the MTT (Law et al., 2009) and MS-DB (Rafii et al., 2017) datasets, and the audio domain includes the ESC50 (Piczak, 2015) and VIVAE (Holz et al., 2022) datasets. We extract embeddings corresponding to the discrete codebooks of each acoustic codec model as its respective representations and evaluate the classification accuracy of the codec models on the ARCH datasets using these representations. The experimental results, as shown in Table 5, demonstrate that our UniCodec outperforms WavTokenizer, DAC (configured with a single quantizer) and Encodec (configured with two-layer quantizers), in terms of classification accuracy. Furthermore, performance comparison

Table 5: **Semantic representation evaluation results** on the ARCH benchmark, in terms of classification accuracy. The results of models marked by † are cited from the Wavtokenizer paper (Ji et al., 2024c).

Model	TPS (↓)	Speech		Music		Audio	
		RAVDESS (↑)	AM (↑)	MTT (↑)	MS-DB (↑)	ESC50 (↑)	VIVAE (↑)
Encodec†	150	27.43	36.49	19.00	32.45	16.99	26.30
DAC†	100	25.00	62.87	25.02	51.37	20.65	29.91
Wavtokenizer (speech)†	75	32.55	69.57	-	-	-	-
Wavtokenizer (music&audio)†	75	-	-	28.35	57.64	25.50	**35.63**
UniCodec	75	**40.28**	**70.94**	**29.55**	**59.29**	**26.00**	34.17
w/o semantic stage	75	36.81	69.84	28.09	54.05	20.80	30.21

Table 6: Ablation study of UniCodec by evaluating the effects of domain ID during evaluation, the domain MoE module, domain-adaptive codebook, and the semantic training stage and the fine-tuning stage.

Model	**LibriTTS test-clean**		**MusicDB test**		**AudioSet eval**	
	Mel Dist ↓	STFT Dist ↓	Mel Dist ↓	STFT Dist ↓	Mel Dist ↓	STFT Dist ↓
UniCodec	**0.3442**	**1.5147**	0.3959	2.1822	**0.3820**	2.1065
w. domain id	0.3474	1.5151	**0.3912**	**2.1818**	0.3824	**2.1061**
w/o finetune stage	0.4476	1.7005	0.4490	2.2505	0.4366	2.1659
w/o semantic&finetune stage	0.4481	1.6978	0.4534	2.2690	0.4380	2.1723
w/o MoE	0.4883	1.8024	0.4592	2.3153	0.4548	2.2633
w/o partitioned codebook	0.4873	1.7742	0.5064	2.3031	0.5135	2.2382

against the counterpart that excludes the semantic stage training (w/o semantic stage) verifies the effectiveness of the proposed semantic training using mask prediction and contrastive loss. In future work, we plan to explore UniCodec-based ALM on downstream audio tasks such as audio continuation and generation.

5.3 Ablation study

We conduct ablation study by evaluating the effect of proposed methods and modules on the LibriTTS test-clean, MusicDB test, and AudioSet eval sets. As shown in Table 6, providing the domain ID for the partitioned domain-adaptive codebook during evaluation performs comparably to the default setting without providing domain ID. The only exception is the music domain, where performance improves slightly due to the inherent mixed nature of songs, which contain both speech and music elements. These results demonstrate that the partitioned domain-adaptive codebook can autonomously capture distinct domain-specific features. The third row shows that without the fine-tuning stage, a significant performance degradation is observed when trained on large but noisy data. This highlights the critical role of high-quality data in codec training. The fourth row reports results without both semantic training and fine-tuning stages. Comparison between the third and fourth rows shows that our proposed semantic stage enhances semantic information while preserving reconstruction ability. Furthermore, removing the MoE module from UniCodec without the semantic and fine-tuning stages (i.e., only the initial acoustic training stage) results in an additional performance degradation. Removing the partitioned domain-adaptive codebook (i.e. naive single codebook) leads to even greater degradation than removing the MoE module. These results confirm the effectiveness of the proposed domain MoE and partitioned domain-adaptive codebook strategy in achieving a unified codec with superior reconstruction ability.

6 Conclusions

In this work, we introduce UniCodec, a low-bitrate unified audio tokenizer designed to support multi-domain audio data, including speech, music, and sound, using a single quantizer. To achieve this goal of unified modeling, we propose the partitioned domain-adaptive codebook and the domain MoE strategy to capture the distinct characteristics of each domain. To enrich the semantic information without introducing additional modules, we propose a self-supervised mask prediction modeling algorithm during codec training. Comprehensive objective and subjective evaluations demonstrate that, as a unified audio codec with a single codebook, UniCodec achieves excellent performance in both acoustic and semantic capabilities.

7 Limitations

Our experiments reveal that UniCodec training will be disrupted by noisy or low-quality inputs. Modeling speech in complex environments, such as noisy settings or with overlapped speech, remains a challenge. We anticipate that future work will address these issues, improving model robustness for such scenarios.

Although our experiments demonstrate that the proposed semantic training stage with mask prediction and contrastive loss effectively captures semantic information, it remains challenging for a unified single-codebook codec to balance both acoustic and semantic density across diverse domain data. We believe that it is a promising research direction to focus on enhancing semantic capabilities while preserving reconstruction performance, without introducing additional modules.

We have evaluated the model in streaming use cases but have observed some performance degradation. Future work should aim to improve streaming capabilities while maintaining high reconstruction quality.

Due to space limit and computational constraints, we have focused on demonstrating UniCodec's reconstruction capabilities and have not yet explored training UniCodec with LLM to function as an Audio Language Model (ALM). In future work, we plan to investigate the performance of UniCodec-based ALM on downstream audio tasks.

References

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M Tyers, and Gregor Weber. 2019. Common voice: A massively-multilingual speech corpus. *arXiv preprint arXiv:1912.06670*.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Sören Becker, Johanna Vielhaben, Marcel Ackermann, Klaus-Robert Müller, Sebastian Lapuschkin, and Wojciech Samek. 2024. Audiomnist: Exploring explainable artificial intelligence for audio analysis on a simple benchmark. *Journal of the Franklin Institute*, 361(1):418–428.

Dmitry Bogdanov, Minz Won, Philip Tovstogan, Alastair Porter, and Xavier Serra. 2019. The mtg-jamendo dataset for automatic music tagging. ICML.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR.

Damai Dai, Chengqi Deng, Chenggang Zhao, RX Xu, Huazuo Gao, Deli Chen, Jiashi Li, Wangding Zeng, Xingkai Yu, Y Wu, et al. 2024. Deepseekmoe: Towards ultimate expert specialization in mixture-of-experts language models. *arXiv preprint arXiv:2401.06066*.

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. *arXiv preprint arXiv:2410.00037*.

Zhihao Du, Shiliang Zhang, Kai Hu, and Siqi Zheng. 2024. Funcodec: A fundamental, reproducible and integrable open-source toolkit for neural speech codec. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2024, Seoul, Republic of Korea, April 14-19, 2024*, pages 591–595. IEEE.

Cristina Gârbacea, Aäron van den Oord, Yazhe Li, Felicia SC Lim, Alejandro Luebs, Oriol Vinyals, and Thomas C Walters. 2019. Low bit-rate speech coding with vq-vae and a wavenet decoder. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 735–739. IEEE.

Jort F Gemmeke, Daniel PW 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)*, pages 776–780. IEEE.

Michael Gutmann and Aapo Hyvärinen. 2010. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 297–304. JMLR Workshop and Conference Proceedings.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738.

Dan Hendrycks and Kevin Gimpel. 2016. Bridging nonlinearities and stochastic regularizers with gaussian error linear units. corr abs/1606.08415 (2016). *arXiv preprint arXiv:1606.08415*.

Natalie Holz, Pauline Larrouy-Maestri, and David Poeppel. 2022. The variably intense vocalizations of affect and emotion (vivae) corpus prompts new perspective on nonspeech perception. *Emotion*, 22(1):213.

Keith Ito. 2017. The lj speech dataset. `https://keithito.com/LJ-Speech-Dataset/`.

Shengpeng Ji, Yifu Chen, Minghui Fang, Jialong Zuo, Jingyu Lu, Hanting Wang, Ziyue Jiang, Long Zhou, Shujie Liu, Xize Cheng, et al. 2024a. Wavchat: A survey of spoken dialogue models. *arXiv preprint arXiv:2411.13577*.

Shengpeng Ji, Minghui Fang, Ziyue Jiang, Siqi Zheng, Qian Chen, Rongjie Huang, Jialung Zuo, Shulei Wang, and Zhou Zhao. 2024b. Language-codec: Reducing the gaps between discrete codec representation and speech language models. *arXiv preprint arXiv:2402.12208*.

Shengpeng Ji, Ziyue Jiang, Wen Wang, Yifu Chen, Minghui Fang, Jialong Zuo, Qian Yang, Xize Cheng, Zehan Wang, Ruiqi Li, et al. 2024c. Wavtokenizer: an efficient acoustic discrete codec tokenizer for audio language modeling. *arXiv preprint arXiv:2408.16532*.

Jacob Kahn, Morgane Riviere, Weiyi Zheng, Evgeny Kharitonov, Qiantong Xu, Pierre-Emmanuel Mazaré, Julien Karadayi, Vitaliy Liptchinsky, Ronan Collobert, Christian Fuegen, et al. 2020. Libri-light: A benchmark for asr with limited or no supervision. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7669–7673. IEEE.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *ICLR*.

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*.

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

Edith Law, Kris West, Michael I Mandel, Mert Bay, and J Stephen Downie. 2009. Evaluation of algorithms using games: The case of music tagging. In *ISMIR*, pages 387–392. Citeseer.

Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. 2020. Gshard: Scaling giant models with conditional computation and automatic sharding. *arXiv preprint arXiv:2006.16668*.

Hanzhao Li, Liumeng Xue, Haohan Guo, Xinfa Zhu, Yuanjun Lv, Lei Xie, Yunlin Chen, Hao Yin, and Zhifei Li. 2024. Single-codec: Single-codebook speech codec towards high-performance speech generation. In *Interspeech 2024*, pages 3390–3394.

Haohe Liu, Xuenan Xu, Yi Yuan, Mengyue Wu, Wenwu Wang, and Mark D Plumbley. 2024. Semanticodec: An ultra low bitrate semantic audio codec for general sound. *arXiv preprint arXiv:2405.00233*.

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

Ilya Loshchilov and Frank Hutter. 2017. SGDR: stochastic gradient descent with warm restarts. In *ICLR*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *ICLR*.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Julian D. Parker, Anton Smirnov, Jordi Pons, CJ Carr, Zack Zukowski, Zach Evans, and Xubo Liu. 2024. Scaling transformers for low-bitrate high-quality speech coding. *CoRR*, abs/2411.19842.

Jing Peng, Yucheng Wang, Yu Xi, Xu Li, Xizhuo Zhang, and Kai Yu. 2024. A survey on speech large language models. *arXiv preprint arXiv:2410.18908*.

Karol J Piczak. 2015. Esc: Dataset for environmental sound classification. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 1015–1018.

Zafar Rafii, Antoine Liutkus, Fabian-Robert Stöter, Stylianos Ioannis Mimilakis, and Rachel Bittner. 2017. The musdb18 corpus for music separation.

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 *2001 IEEE international conference on acoustics, speech, and signal processing. Proceedings (Cat. No. 01CH37221)*, volume 2, 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. *arXiv preprint arXiv:2204.02152*.

Aäron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. 2017. Neural discrete representation learning. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 6306–6315.

Christophe Veaux, Junichi Yamagishi, Kirsten MacDonald, et al. 2016. Superseded-cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit.

Zhifei Xie and Changqiao Wu. 2024a. Mini-omni: Language models can hear, talk while thinking in streaming. *arXiv preprint arXiv:2408.16725*.

Zhifei Xie and Changqiao Wu. 2024b. Mini-omni2: Towards open-source gpt-4o with vision, speech and duplex capabilities. *arXiv preprint arXiv:2410.11190*.

Detai Xin, Xu Tan, Shinnosuke Takamichi, and Hiroshi Saruwatari. 2024. Bigcodec: Pushing the limits of low-bitrate neural speech codec. *CoRR*, abs/2409.05377.

Zhen Ye, Peiwen Sun, Jiahe Lei, Hongzhan Lin, Xu Tan, Zheqi Dai, Qiuqiang Kong, Jianyi Chen, Jiahao Pan, Qifeng Liu, et al. 2024. Codec does matter: Exploring the semantic shortcoming of codec for audio language model. *arXiv preprint arXiv:2408.17175*.

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

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. *arXiv preprint arXiv:1904.02882*.

Xin Zhang, Dong Zhang, Shimin Li, Yaqian Zhou, and Xipeng Qiu. 2023. Speechtokenizer: Unified speech tokenizer for speech large language models. *arXiv preprint arXiv:2308.16692*.

Yongxin Zhu, Bocheng Li, Yifei Xin, and Linli Xu. 2024. Addressing representation collapse in vector quantized models with one linear layer. *arXiv preprint arXiv:2411.02038*.

A Speech Reconstruction Evaluation

We further evaluate UniCodec on the LibriTTS test-other set to assess its reconstruction ability on noisy data. The results in Table 7 show that the reconstructed speech from our model achieves a higher UTMOS score than the ground truth on the LibriTTS test-other noisy dataset. This indicates that UniCodec reconstructs speech with greater naturalness and quality, even in the presence of noise. As a unified codec with a single codebook, UniCodec outperforms Wavtokenizer (unified) across all metrics. Even when compared with other state-of-the-art speech-specific codecs with a single codebook, UniCodec maintains competitive performance.

B Real-Time Factor

To evaluate the real-time performance of different audio codec models, we compute the Real-Time Factor (RTF) for audio durations of 5, 10, 30, and 60 seconds. The evaluation is conducted on a test set of 1,000 audio clips to ensure a robust and fair comparison. All experiments are performed on an NVIDIA A100 GPU. RTF measures the processing speed relative to real-time feature extraction, a critical metric for NACs to minimize latency. Lower RTF values indicate faster processing. As shown in Table 8, UniCodec has more parameters than Wavtokenizer due to the incorporation of transformer layers and the MoE structure. This results in a higher RTF for UniCodec with 5-second inputs compared to Wavtokenizer. However, for 10, 30, and 60-second inputs, UniCodec exhibits better RTF performance, and benefits from the superior parallel processing capabilities of its transformer layers, compared to the LSTM module in Wavtokenizer. Semanticodec has a much larger RTF, making it unsuitable for real-time applications. For DAC, we do not report results for 30s and 60s due to out-of-memory issues.

C Fine-tuning Stage

In the finetune stage, we select high-quality speech data with a high UTMOS, including LibriTTS train-clean, VCTK, and LJSpeech (Ito, 2017). Additionally, the learning rate and mel loss coefficient are set to 5e-5 and 450, respectively. These training strategies in the finetune stage significantly enhance the model's ability to better learn reconstruction ability.

Table 7: **Objective reconstruction results** on the **Speech** domain from UniCodec and baselines on LibriTTS test-other, in terms of naturalness, distortion, and intelligibility. **DR** denotes the Downsample Rate (the input audio sample rate division by Tokens Per Second (TPS)). **Unified** denotes the codec model can support all three domains of speech, music, and sound. The results of models marked by † are cited from the Wavtokenizer paper (Ji et al., 2024c) and others are reproduced by us based on the checkpoints released by the corresponding work.

Model	**Unified**	**DR** (↑)	**TPF** (↓)	**TPS** (↓)	**PESQ** (↑)	**STOI** (↑)	**F1** (↑)	**UTMOS** (↑)
Ground Truth†	-	-	-	-	-	-	-	3.4831
DAC†	✔	48.9	9	900	3.7595	0.9576	0.9696	3.3566
Encodec†	✔	40	8	600	2.6818	0.9241	0.9338	2.6568
SpeechTokenizer†	✘	40	8	600	2.3269	0.8811	0.9205	3.2851
Mimi	✔	240	8	100	2.0952	0.8816	0.8875	3.0608
TAAE	✘	320	2	50	1.7539	0.8380	0.8994	3.7136
DAC†	✘	440	1	100	1.2454	0.7505	0.7775	1.4986
BigCodec	✘	200	1	80	**2.3817**	0.9094	**0.9237**	3.5453
Wavtokenizer (speech)†	✘	320	1	75	2.2614	0.8907	0.9172	3.4312
Wavtokenizer (unified)	✔	320	1	75	1.6649	0.8312	0.8874	3.0820
UniCodec	✔	320	1	75	2.2749	**0.9095**	0.9109	**3.5800**

Table 8: Real-Time Factors (RTFs) for audio codec models on test audio clips of 5s, 10s, 30s and 60s duration using an A100 GPU.

Model	Parameter (M)	RTF (5s)↓	RTF (10s)↓	RTF (30s)↓	RTF (60s)↓
DAC	76	0.01021	0.00771	-	-
SemantiCodec	507	1.10905	0.54455	0.69320	0.61164
Wavtokenizer	77	**0.00377**	0.00321	0.00286	0.00280
UniCodec	274	0.00467	**0.00287**	**0.00196**	**0.00187**

Table 9: Codebook utilization rate of the whole codebook and three domain-partitioned codebook in the condition of with and without domain id provided.

	Whole	Speech	Music	Audio
w/o domain id	99.63%	98.54%	100%	99.95%
w. domain id	99.62%	98.54%	100%	99.96%

D Codebook Utilization

We further evaluate the codebook utilization rate for both the entire codebook and the partitioned codebook across each domain. The results are evaluated on the LibriTTS test-clean, MusicDB test, and AudioSet eval sets. As shown in Table 9, the utilization rates for each domain-partitioned codebook are nearly fully exploited, demonstrating that our UniCodec's domain-adaptive codebook is both well-trained and effectively utilized.

E Speech Reconstruction Metrics

PESQ (Rix et al., 2001) (Distortion): A speech quality assessment metric that compares reconstructed speech with reference speech, with scores ranging from 1 to 5, and correlates with human judgment.

STOI (Intelligibility): A metric measuring speech intelligibility by comparing short-time spectral envelopes between reconstructed and ground truth speech, with scores ranging from 0 to 1.

F1 Score (Voiced/Unvoiced Classification): It balances precision and recall for voiced/unvoiced classification.

UTMOS (Saeki et al., 2022) (Naturalness): An automatic speech MOS (Mean Opinion Score) predictor evaluates the naturalness of generated speech, reflecting overall auditory quality.