UniAudio-Token: Empowering Semantic Speech Tokenizers with General Audio Perception

Anonymous ACL submission

Abstract

Semantic speech tokenizers have become a widely used interface for Audio-LLMs, owing to their compact single-codebook design and strong linguistic alignment. However, their focus on linguistic abstraction induces *acoustic blindness*, limiting their applicability beyond speech-centric tasks. We propose **UniAudio-Token**, a framework that empowers semantic tokenizers with general audio perception without compromising speech ability. Instead of altering the semantic paradigm, UniAudio-Token mitigates its information loss through two key innovations: (1) Semantic-Acoustic Primitives (SAP) provide structured supervision by decomposing audio into linguistic content, vocal attributes, and auditory-scene primitives; and (2) Semantic-Acoustic Equilibrium (SAE) introduces a content-aware gating mechanism that adaptively restores fine-grained acoustic details from shallow layers. Extensive evaluations show that UniAudio-Token learns comprehensive universal representations while preserving high-fidelity speech generation. When integrated with downstream LLMs, it outperforms all single-codebook baseline tokenizers on both understanding and generation tasks, effectively serving as a unified audio interface.

1 Introduction

Audio-LLMs aim to extend LLMs to spoken and auditory interaction, requiring audio representations that support both **understanding** and **generation**. While continuous features from pretrained audio encoders are effective for perception, they significantly struggle with audio generation (Yang et al., 2025b). In contrast, discrete audio tokens can be handled in the same modeling paradigm as text tokens, enabling a unified token-level interface for both audio input and output. This property has motivated recent Audio-LLMs to continue adopting discrete audio tokenizers (Zeng et al., 2024; KimiTeam et al., 2025; Zhang et al., 2025a), and makes improving audio tokenizers a critical problem rather than merely an architectural choice.

Model	Single Codebook	General Audio	Linguistic Alignment
EnCodec (Défossez et al., 2023)	✗	✓	✗
SpeechTokenizer (Zhang et al., 2024)	✗	✓	✓
CosyVoice2 (Du et al., 2024)	✓	✗	✓
GLM-4-Voice-Tokenizer (Zeng et al., 2025)	✓	✗	✓
StableToken (Song et al., 2026)	✓	✗	✓
WavTokenizer (Ji et al., 2025)	✓	✓	✗
UniAudio-Token (Ours)	✓	✓	✓

Table 1: Comparison of audio tokenizers. UniAudio-Token uniquely combines single-codebook modeling, general audio perception, and linguistic alignment.

Among discrete audio tokenizers, **semantic speech tokenizers** have been widely adopted in recent Audio-LLMs (Du et al., 2024; Zeng et al., 2025; Song et al., 2026) due to two compelling advantages: (1) single-codebook design, which enables direct integration into standard LLM architectures and compact sequences crucial for long-context processing; and (2) inherent linguistic alignment, as initialization from ASR encoders facilitates seamless text-audio interaction.

However, as Audio-LLMs expand from speech to universal auditory perception, including music, sound events, and complex acoustic scenes, semantic speech tokenizers lag behind. Optimized strictly for linguistic content extraction, deep ASR encoders actively suppress vocal cues and auditory scene details as *noise*. This induces **acoustic blindness**, fundamentally limiting the LLM's understanding of the full acoustic scene.

Alternatively, single-codebook acoustic-centric models (Ji et al., 2025) prioritize waveform reconstruction. While preserving acoustic details, they lack explicit semantic guidance. Consequently, their audio tokens fail to form distinct categorical clusters based on meaning. Semantically distinct but acoustically similar sounds (e.g. *rain* and *white noise*) may collapse into overlapping distributions.

This harsh semantic-acoustic trade-off forces a

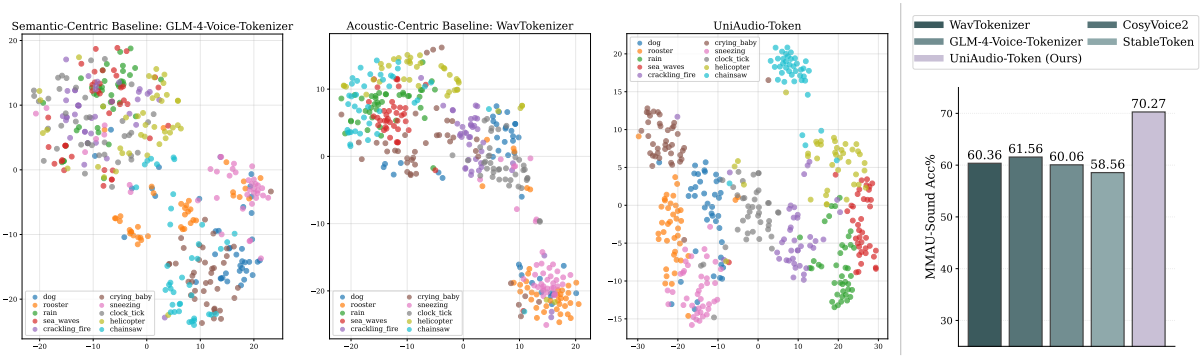


Figure 1: ESC-10 token sequence t-SNE Visualization. *(Left)* A semantic-centric baseline (GLM-4-Voice-Tokenizer) suffers from acoustic blindness, mapping distinct events to overlapping regions. *(Center Left)* An acoustic-centric baseline (WavTokenizer) exhibits insufficient semantic discrimination. *(Center Right)* UniAudio-Token resolves these issues via Semantic-Acoustic Equilibrium, forming well-separated clusters. *(Right)* When integrated with Qwen2.5-3B, UniAudio-Token shows superior performance on the MMAU benchmark.

fragmented paradigm: Audio-LLMs are either confined to speech-only understanding (Zeng et al., 2024), or rely on heterogeneous architectures that combine external continuous encoders and adapters for general audio perception with discrete tokenizers for speech generation (Huang et al., 2025; Xu et al., 2025a). In this work, we aim to unify this divide by using a single codebook to support both high-fidelity **speech generation** and high-level **general audio understanding** simultaneously.

Empowering semantic tokenizers with universal audio perception is non-trivial. It involves two fundamental conflicts: (1) *The supervision conflict*: ASR targets (Du et al., 2024; Zeng et al., 2025; Song et al., 2026) extract linguistics but ignore acoustics, whereas reconstruction targets (Ji et al., 2025) focus on raw acoustic nuances, hindering semantic extraction. Semantic distillation (Zhang et al., 2024) remains speech-centric and fails to generalize to general audio. Recent work has begun exploring supervision beyond transcription (Zhang et al., 2026), but only at the Audio-LLM level, without addressing acoustic blindness at the audio tokenizer level, which is the fundamental representational bottleneck. (2) *The architectural bottleneck*: Deep semantic encoders irreversibly lose fine-grained acoustic cues in higher layers, while naive feature fusion risks diluting the linguistic abstraction required for content-faithful speech generation. A mechanism that dynamically balances these competing information streams is required.

To address these challenges, we propose the **UniAudio-Token** framework to empower single-codebook semantic speech tokenizers with universal audio perception. Our core insight is that mitigating the semantic-acoustic tension requires dual rectification: explicitly disentangling linguistic content from vocal attributes and auditory scenes *at the supervision level*, and dynamically bridging the information bottleneck to recover lost acoustic details *at the architectural level*.

Specifically, we introduce two innovations: (1) **Semantic-Acoustic Primitives (SAP):** Resolving the supervision conflict, this structured supervision protocol decomposes raw audio into fundamental linguistic content, vocal attributes, and auditory-scene building blocks. It explicitly disentangles content from style, forcing the model to allocate capacity for vocal and acoustic details without interfering with the semantic backbone. (2) **Semantic-Acoustic Equilibrium (SAE):** Addressing the architectural bottleneck, this content-aware gating mechanism adaptively injects fine-grained acoustic details from shallow layers into deep semantic streams when needed, mitigating acoustic blindness without corrupting semantic representations.

Extensive evaluations demonstrate UniAudio-Token effectively bridges linguistic alignment and universal representation. At the tokenizer level, it achieves high Cluster Purity on ESC (Piczak, 2015), forming distinct clusters for diverse audio events where baselines struggle. Crucially, this acquisition of general audio perception does not compromise speech generation capabilities; instead, UniAudio-Token even surpasses specialized speech tokenizers in generation quality. At the Audio-LLM level, integrating this universal frontend with Qwen2.5 also yields superior performance on both understanding and generation. Further analysis validates the adaptive behavior of the SAE mechanism.

2 Related Work

Semantic Speech Tokenizers. The evolution of LLMs has pushed spoken dialogue systems from traditional cascaded pipelines towards end-to-end Audio-Language Models (Zhang and Wang, 2019; Zhang et al., 2020; Tang et al., 2024; Zhang et al., 2023; Gong et al., 2024; Hu et al., 2024; Fang et al., 2025; Défossez et al., 2024; Li et al., 2025; Wang et al., 2025; Bai et al., 2024; Ghosh et al., 2025; Zhang et al., 2025b), driving two distinct tokenizer paradigms. Early self-supervised learning (SSL) units (Hsu et al., 2021; Baevski et al., 2020; Huang et al., 2022) primarily encode phonetic information and suffer from high Gross Pitch Error in generation (Sicherman and Adi, 2023; Mousavi et al., 2026), making them unsuitable for high-fidelity end-to-end synthesis. Recent systems adopt supervised ASR-based tokenization (Du et al., 2024; KimiTeam et al., 2025; Song et al., 2026), quantizing intermediate representations of ASR encoders into compact flat tokens that represent linguistic units while implicitly retaining prosodic features learned from large-scale transcription. However, despite their success in speech understanding and synthesis, we find this paradigm fundamentally suffers from *acoustic blindness* in general audio tasks.

Acoustic Audio Tokenizers. In parallel, acoustic tokenizers target high-fidelity waveform reconstruction. Neural codecs (Zeghidour et al., 2022; Défossez et al., 2023; Yang et al., 2023; Kumar et al., 2023) typically employ multi-codebook RVQ to reduce distortion, which necessitates specialized architectural adaptations or flattening for LLM integration. Recent acoustic-centric models like WavTokenizer (Ji et al., 2025) instead employ single-codebook vector quantization to achieve extreme compression with a flat structure. While this facilitates direct LLM integration and preserves general audio details, it lacks explicit semantic alignment, limiting the performance on linguistic-intensive tasks. Although semantic distillation techniques (Défossez et al., 2024; Zhang et al., 2025a; Ye et al., 2025) can improve semantic awareness, they retain multi-codebook design and restrict semantic supervision to speech, leaving general audio events semantically entangled.

3 Methods

To resolve the architectural fragmentation of Audio-LLMs, **UniAudio-Token** establishes a *unified discrete interface* that projects speech and general audio into a single codebook. As shown in Figure 2, we reconcile linguistic generation and universal perception via two core innovations: Semantic-Acoustic Primitives (SAP) for disentangled supervision, and Semantic-Acoustic Equilibrium (SAE) for adaptive semantic-acoustic feature fusion.

3.1 Semantic-Acoustic Primitives (SAP)

Existing audio tokenizers face a fundamental trade-off: ASR-based supervision provides strong linguistic alignment but limited discrimination for non-speech signals, while reconstruction-based objectives preserve acoustic details but lack explicit semantic guidance. To address this limitation, we introduce a structured supervision strategy termed **Semantic-Acoustic Primitives (SAP)**. SAP serves as a protocol generated by an LLM to provide supervision that captures both semantic content and acoustic cues. Unlike traditional ASR corpora that focus solely on linguistic content, SAP explicitly separates and annotates the full spectrum of acoustic information, enabling the tokenizer to remain discriminative across diverse audio types.

Structure Design. SAP describes each audio clip using three complementary layers: (1) **Linguistic Content**, i.e., the verbatim transcript for speech; (2) **Vocal Attributes**, which characterize *how* speech is produced through six normalized fields: *Age, Gender, Emotion, Accent, Prosody,* and *Timbre*; and (3) **Auditory Scene**, which captures the acoustic environment, including *Transient Events* (e.g., door slams) and *Persistent Events* (e.g., engine rumble).

This structure separates semantic meaning from acoustic cues explicitly and provides low-entropy supervision targets, thereby stabilizing tokenizer optimization. Example JSON annotations are provided in Appendix A for better understanding.

Data Curation. Since manually annotating such fine-grained attributes is prohibitively costly, we develop an automated pipeline to derive SAP labels from large-scale ASR corpora: (1) **Acoustic Captioning.** An audio-language model generates rich, unstructured textual descriptions of the audio, capturing vocal style and auditory-scene information missing from ASR transcripts. (2) **Structured Synthesis.** An LLM teacher aggregates the ground-truth transcription and the generated acoustic captions, normalizes them into predefined SAP fields, and outputs a valid JSON object. (3) **Quality Validation.** We apply a multi-level validation

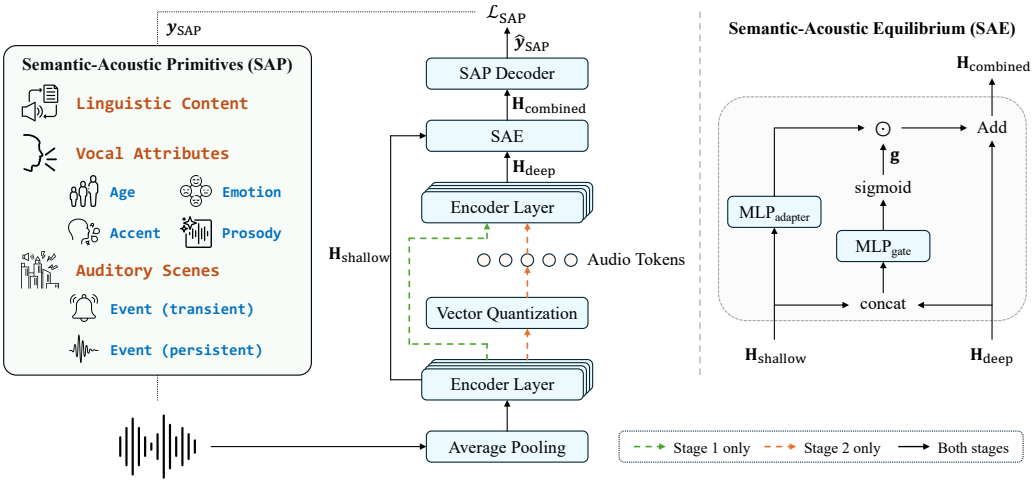


Figure 2: The framework of UniAudio-Token. *(Left)* The model is supervised by Semantic-Acoustic Primitives (SAP), which cover linguistic content, vocal attributes, and auditory scenes. *(Center)* Vector Quantization (VQ) converts hidden states into discrete audio tokens. *(Right)* Semantic-Acoustic Equilibrium (SAE) adaptively fuses shallow acoustic details with deep semantic features, mitigating the loss of fine-grained acoustic cues in deep layers.

mechanism to reduce hallucinations, including ontology constraints for categorical fields, logical consistency checks, and content-duration alignment. Only samples passing all checks are retained. Human evaluation further verifies the reliability of the SAP annotations, with details in Appendix B.

To facilitate interactive capabilities, we further derive an **SAP-Instruct** dataset from the structured annotations, including Direct QA, Multiple Choice, and True/False Verification pairs. These diverse formats encourage the model to attend to specific acoustic sub-features during training.

3.2 Model Architecture

As illustrated in Figure 2, UniAudio-Token consists of an audio encoder, an SAE module, a quantization layer, and an SAP decoder.

Semantic-Acoustic Equilibrium (SAE). ASR-centric speech encoders, such as Whisper (Radford et al., 2023), progressively abstract audio into high-level semantic representations. While beneficial for ASR, this process often discards low-level acoustic details required by SAP, such as vocal texture and auditory events. The **Semantic-Acoustic Equilibrium (SAE)** mechanism addresses this bottleneck by adaptively fusing semantically rich deep features with acoustically rich shallow features.

Let $\mathbf{H}_{\text{shallow}}$ denote the output from a shallow encoder layer, and $\mathbf{H}_{\text{deep}}$ denote the final-layer representation. First, we project the shallow features into the deep feature space:

$$\mathbf{H}_{\text{ada_shallow}} = \mathbf{MLP}_{\text{adapter}}\left(\mathbf{H}_{\text{shallow}}\right), \quad (1)$$

where $\mathbf{MLP}_{\text{adapter}}$ is learnable. Then the SAE computes a content-aware fusion gate $\mathbf{g}$:

$$\mathbf{g} = \sigma\left(\mathbf{MLP}_{\text{gate}}\left([\mathbf{H}_{\text{deep}}; \mathbf{H}_{\text{shallow}}]\right)\right), \quad (2)$$

where $[\cdot;\cdot]$ denotes concatenation and σ is the sigmoid function. $\mathbf{MLP}_{\text{gate}}$ is learnable. The final fused representation $\mathbf{H}_{\text{combined}}$ is then obtained via:

$$\mathbf{H}_{\text{combined}} = \mathbf{H}_{\text{deep}} + \mathbf{g} \odot \mathbf{H}_{\text{ada_shallow}}, \quad (3)$$

where $\odot$ denotes element-wise multiplication. The SAE mechanism allows the model to adaptively retain acoustic details necessary for SAP-supervised learning while preserving semantic abstraction.

Vector Quantization. We discretize the continuous hidden states using a standard Vector Quantization (VQ) layer (van den Oord et al., 2017). Given a learnable codebook $\mathcal{C} = \{\mathbf{e}_k\}_{k=1}^{K} \subset \mathbb{R}^D$, where K is the codebook size, the input vector $\mathbf{h}_t$ at time step t is mapped to its nearest code vector:

$$\mathbf{h}_t^q = \mathbf{e}_k, \quad \text{where } k = \underset{j}{\operatorname{argmin}} \|\mathbf{h}_t - \mathbf{e}_j\|_2^2. \quad (4)$$

The sequence of indices k forms the audio tokens.

3.3 Training Strategy

We initialize the encoder and decoder from whisper-large-v3 (Radford et al., 2023) and train the full model end-to-end with a mixture of SAP generation and SAP-Instruct QA tasks. This initialization preserves strong linguistic alignment, while SAP supervision then expands the representation toward vocal attributes and auditory scenes.

Our training pipeline consists of two stages. In **Stage 1**, we bypass the VQ layer and train the SAE module together with the decoder using only SAP prediction loss ($\mathcal{L}_{\text{SAP}}$). The goal is to adapt the pre-trained ASR decoder into an SAP decoder, aligning the continuous hidden space with structured SAP. In **Stage 2**, we insert the VQ layer and primarily optimize the codebook to produce discrete audio tokens, while preserving the SAP-aligned representation learned in the previous stage. Following the framework of VQ (van den Oord et al., 2017), the objective function combines the SAP prediction loss with quantization and commitment losses:

$$\mathcal{L} = \mathcal{L}_{\text{SAP}} + \lambda_1 \underbrace{\|\text{sg}[\mathbf{h}] - \mathbf{h}^q\|_2^2}_{\mathcal{L}_{\text{quantization}}} + \lambda_2 \underbrace{\|\mathbf{h} - \text{sg}[\mathbf{h}^q]\|_2^2}_{\mathcal{L}_{\text{commitment}}}, \quad (5)$$

where λ_1, λ_2 are hyperparameters, and $\text{sg}[\cdot]$ the stop-gradient operator. The decoder is optimized by $\mathcal{L}_{\text{SAP}}$, the encoder and SAE module by $\mathcal{L}_{\text{SAP}}$ and $\mathcal{L}_{\text{commitment}}$, and the codebook by $\mathcal{L}_{\text{quantization}}$.

4 Experimental Setup

Implementation Details. For SAP data curation, we utilize Qwen3-Omni-Captioner (Xu et al., 2025b) to produce detailed acoustic captions and Qwen3-30B-A3B-Instruct-2507 (Yang et al., 2025a) to perform structured synthesis. SAP-Instruct is curated using the more powerful Qwen3-235B-A22B-Instruct-2507 (Yang et al., 2025a) to ensure high-quality instruction following.

UniAudio-Token uses a single codebook with a vocabulary size of 8,192 and a token frame rate of 25Hz. Full training details, including datasets and hyperparameters, are listed in Appendix C.

For downstream Audio-LLMs, we integrate all audio tokenizers with the same Qwen2.5 (Qwen et al., 2025) LLM backbone and train them under the same settings for fair comparison.

Baselines. We compare UniAudio-Token against representative single-codebook acoustic and semantic audio tokenizers, including (1) **WavTokenizer** (Large, 75Hz) (Ji et al., 2025); (2) **CosyVoice2** (Du et al., 2024); (3) **GLM-4-Voice-Tokenizer** (Zeng et al., 2025); and (4) **StableToken** (Song et al., 2026). An additional brief introduction of these baselines is provided in Appendix D.

Evaluation & Benchmarks. We use t-SNE visualization (van der Maaten and Hinton, 2008) and clustering metrics (Silhouette Score (Rousseeuw, 1987) and Cluster Purity (Manning et al., 2008)) on ESC (Piczak, 2015) to measure UniAudio-Token's discriminability across diverse sound events. We evaluate speech reconstruction and text-to-speech (TTS) synthesis on LibriSpeech (Panayotov et al., 2015) and SEED-TTS (Anastassiou et al., 2024) using content faithfulness (WER) and speech quality (MOS predicted by MOSNet (Lo et al., 2019)).

We evaluate the tokenizer-LLM systems on MMAU (Sakshi et al., 2025), MMAR (Ma et al., 2025), and MMSU (Wang et al., 2026). These comprehensive audio understanding tasks cover diverse audio inputs, including speech, music, and sounds.

5 Results

We evaluate UniAudio-Token from multiple perspectives. First, we examine its intrinsic quality through latent space analysis and speech reconstruction (§5.1). Second, we assess its effectiveness as an interface for Audio-LLMs on both understanding and generation tasks (§5.2). Finally, we analyze the contribution of SAE through ablation studies and mechanism visualizations (§5.3).

5.1 Tokenizer-Level Performance

Latent Space Disentanglement. To examine how discrete tokens capture granular acoustic characteristics, we visualize token sequences on ESC-50 (Piczak, 2015), which is not included in our training data. ESC-50 covers a broad taxonomic range, from transient human sounds to stationary environmental textures and continuous mechanical noises. This comparison against baselines validates whether UniAudio-Token can perceive and disentangle diverse non-speech sounds.

Since standard dimensionality reduction techniques cannot be directly applied to discrete token sequences, we adopt a **Bag-of-Tokens** approach. For an audio clip with token sequence $T = [t_0, t_1, \ldots, t_n]$, we compute a token histogram vector:

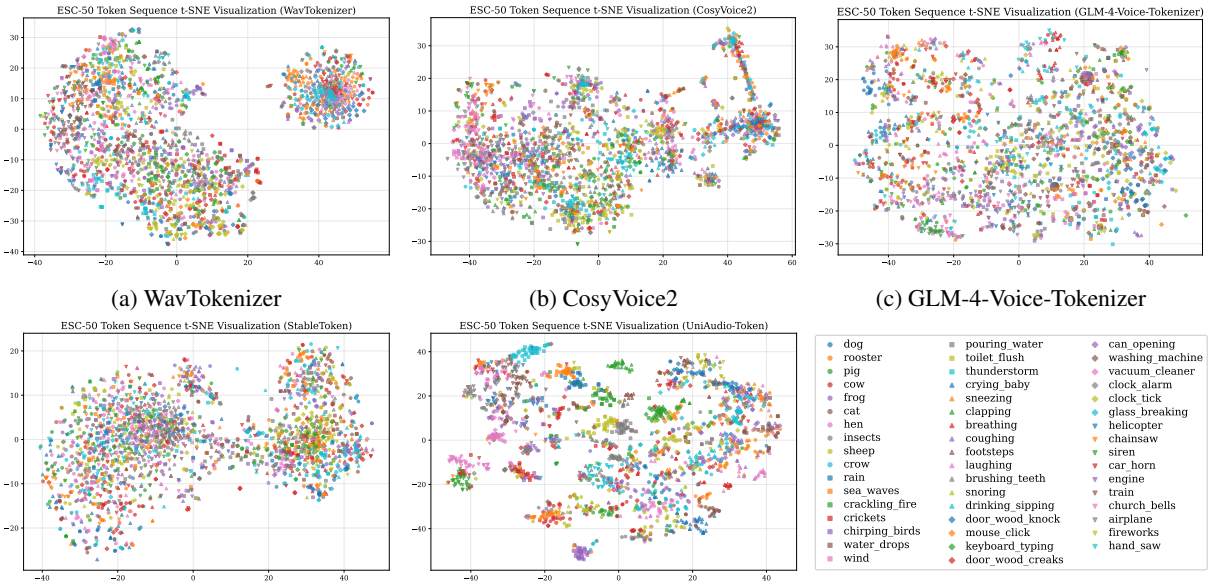
$$H = [h_0, h_1, \ldots, h_{V-1}] \in \mathbb{N}^V, \quad (6)$$

where V is the codebook size, and the i-th element

$$h_i = \sum_{k=0}^{n} \mathbb{I}(t_k = i), \quad i = 0, 1, \ldots, V-1, \quad (7)$$

denotes the frequency of token ID i in T. We subsequently apply t-SNE (van der Maaten and Hinton, 2008) to project these high-dimensional histogram vectors into two dimensions for visualization.

Figure 3 compares UniAudio-Token with baseline tokenizers. Baselines exhibit severe feature



(a) WavTokenizer (b) CosyVoice2 (c) GLM-4-Voice-Tokenizer

(d) StableToken (e) UniAudio-Token (Ours) (f) Legend

Figure 3: t-SNE visualization of token sequences on ESC-50. UniAudio-Token (Figure 3e) exhibits the most clear and semantically meaningful clusters, whereas the baselines show significant feature fragmentation and overlap.

Model	**ESC-10**		**ESC-50**	
	Sil. ↑	Purity ↑	Sil. ↑	Purity ↑
WavTokenizer	-0.030	0.450	-0.108	0.215
GLM-4-Voice-Tokenizer	-0.182	0.373	-0.304	0.133
CosyVoice2	-0.016	0.413	-0.100	0.216
StableToken	-0.035	0.468	-0.096	0.174
UniAudio-Token (Ours)	**0.091**	**0.730**	**0.023**	**0.390**

Table 2: Clustering analysis on ESC-10 and ESC-50. UniAudio-Token is the only one achieving positive Silhouette Scores, indicating valid cluster separation.

entanglement and fragmentation, while UniAudio-Token forms compact and well-separated clusters. This demonstrates that our method effectively captures acoustic characteristics of general audio, forming globally coherent and class-consistent representations. Additional visualization results on the ESC-10 subset are provided in Appendix E.

To complement the qualitative visualization, we compute the Silhouette Score (Rousseeuw, 1987) and Cluster Purity (Manning et al., 2008) directly on the high-dimensional token histogram vectors to avoid information loss from dimensionality reduction. As shown in Table 2, UniAudio-Token is the only model achieving positive Silhouette Scores on both ESC-10 and ESC-50, while all baselines exhibit negative scores, indicating that their token distributions fail to form valid clusters aligned with acoustic categories. In terms of Cluster Purity, our model achieves huge improvements over baselines on both ESC-10 and ESC-50. These quantitative results support UniAudio-Token's meaningful discriminability across diverse acoustic events.

Speech Reconstruction Fidelity. A universal tokenizer should improve general audio understanding without sacrificing speech generation. Following prior work (Du et al., 2024; Zeng et al., 2024; Song et al., 2026), we train a flow matching model to reconstruct speech from discrete tokens. We evaluate on LibriSpeech (Panayotov et al., 2015) and SEED (Anastassiou et al., 2024), using WER and MOS (predicted by MOSNet (Lo et al., 2019)).

Table 3 shows that UniAudio-Token not only preserves, but also **further improves** speech reconstruction fidelity. It achieves a significantly lower WER and the highest average MOS. This indicates that retaining fine-grained acoustic cues can instead improve speech reconstruction capability.

We attribute this improvement to two factors: (1) *Linguistic content and vocal attributes are not fully separable* (Fant, 1971; Polyak et al., 2020). Overly aggressive semantic compression may remove acoustic details that are important for phonetic realization, such as aspiration and consonant transitions. SAE helps retain such cues, enabling clearer phoneme reconstruction. (2) *UniAudio-Token better preserves accent-specific features*. In our analysis, baselines tend to normalize pronunciation, which can introduce transcription mismatches (e.g., "colour" vs. "color", and "centre" vs. "cen-

Model	Frame Rate	BPS	WER ↓					MOS ↑				
			LS-clean	LS-other	SEED en	SEED zh	Average	LS-clean	LS-other	SEED en	SEED zh	Average
WavTokenizer	75Hz	900	5.07	13.09	5.60	4.02	6.95	3.37	3.09	3.01	3.13	3.15
GLM-4-Voice-Tokenizer	12.5Hz	175	4.04	9.33	3.54	3.23	5.04	4.07	3.99	**4.16**	4.10	4.08
CosyVoice2	25Hz	325	4.25	9.68	4.34	2.75	5.26	3.36	3.25	3.31	3.58	3.38
StableToken	25Hz	325	3.84	7.99	3.44	2.62	4.47	4.09	3.83	4.01	4.18	4.03
UniAudio-Token (Ours)	25Hz	325	**3.47**	**6.79**	**2.55**	**1.90**	**3.68** (-0.79)	**4.19**	**4.18**	4.13	**4.25**	**4.19** (+0.11)

Table 3: Speech reconstruction results measured via WER (↓) and MOS (↑).

Tokenizer	MMAU				MMAR				MMSU		
	Speech	Sound	Music	Overall	Speech	Sound	Music	Overall	Perception	Reasoning	Overall
WavTokenizer	36.94	60.36	57.78	51.70	39.80	31.52	29.61	36.30	32.83	45.37	38.90
CosyVoice2	39.94	61.56	62.57	54.70	41.50	35.76	30.58	38.10	27.44	45.83	36.34
GLM-4-Voice-Tokenizer	43.24	60.06	62.28	55.20	39.46	40.00	36.89	40.10	32.40	47.64	39.78
StableToken	**45.05**	58.56	55.99	53.20	42.18	39.39	31.07	39.10	31.98	49.71	40.56
UniAudio-Token (Ours)	**45.05**	**70.27**	**67.96**	**61.10** (+5.90)	**45.24**	**43.64**	**40.29**	**45.80** (+5.70)	**35.54**	**52.07**	**43.54** (+2.98)

Table 4: Downstream Audio-LLMs audio understanding performance comparison, measured via accuracy (%).

ter") after reconstruction. By retaining accent characteristics, UniAudio-Token better matches original speech and reduces recognition errors.

Overall, the latent-space and reconstruction results show that UniAudio-Token provides a discrete representation both discriminative for general audio events and faithful for speech reconstruction.

5.2 Downstream Audio-LLM Performance

A suitable audio tokenizer for Audio-LLMs should provide effective discrete representations for downstream **understanding** and **generation**. Following this paradigm (Song et al., 2026; Du et al., 2024), we integrate each tokenizer with the same Qwen2.5-3B backbone to evaluate understanding, and the same Qwen2.5-0.5B backbone for generation. All systems are tuned identically for fair comparison.

Universal Audio Understanding. As shown in Table 4, UniAudio-Token yields the best performance across all three benchmarks. On **speech** tasks, UniAudio-Token matches or outperforms semantic tokenizers, confirming that its universal design does not compromise linguistic information. The largest improvements appear in **sound** and **music** categories, where baseline semantic tokenizers are limited by acoustic blindness and acoustic-centric tokenizers lack sufficient semantic structure. In contrast, SAP supervision encourages the codebook to encode vocal attributes and auditory-scene cues, while SAE adaptively restores shallow acoustic details. Together, these components provide the LLM with richer evidence for reasoning over complex sound events and musical content.

Tokenizer	SIM ↑	WER ↓	MOS ↑
CosyVoice2	.758 | **.762** | .760	2.71 | 1.39 | 2.05	3.75 | 3.37 | 3.56
UniAudio-Token	**.792** | .742 | **.767**	**1.78** | **1.29** | **1.54**	**4.07** | **3.68** | **3.88**

Table 5: TTS results measured via SIM (↑), WER (↓), and MOS (↑) on SEED-TTS benchmark (en | zh | avg.).

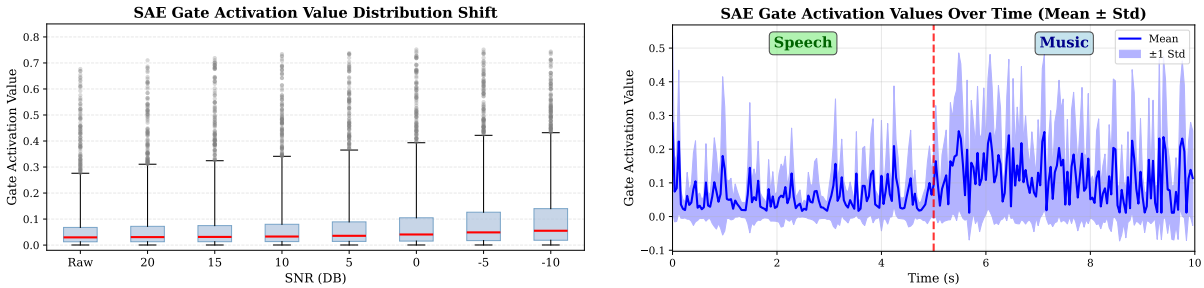
Controllable TTS Synthesis. We further assess UniAudio-Token on controllable text-to-speech (TTS) synthesis tasks. As other tokenizers do not support speaker embedding conditioning, we compare with CosyVoice2 on SEED-TTS. We condition on CAM++ (Wang et al., 2023) and use ERes2Net (Chen et al., 2023) for speaker similarity (SIM) evaluation to avoid bias. Table 5 shows that UniAudio-Token yields significantly better WER and MOS, with slightly higher average SIM. These results prove that UniAudio-Token is also effective for autoregressive LLM-based speech generation.

5.3 Analysis of SAE

We next analyze SAE, a key mechanism of our framework. We first study the effect of fusion depth (§5.3.1), and then visualize the gate to verify its behavior as a content-aware adaptive mechanism rather than a static residual connection (§5.3.2).

5.3.1 Impact of Fusion Depth

We evaluate the effect of injecting acoustic features from different encoder layers, using two complementary metrics: WER on LibriSpeech for phonetic preservation, and Non-Linguistic Score (NLS) on AudioSet (Gemmeke et al., 2017) for acoustics. NLS evaluation details are provided in Appendix F.



(a) Noise-Adaptive Gating

(b) Modality-Aware Gating

Figure 4: Visualization of the SAE gate activation **g**. The gate increases under lower SNR (Figure 4a), and activates more strongly for music than speech (Figure 4b), demonstrating its content-aware dynamic behavior.

Configuration	WER (%) ↓		NLS ↑
	LS-clean	LS-other	
Baseline (w/o SAE)	2.47	5.71	2.93
+ SAE (L_1)	**2.41**	5.62	3.08
+ SAE (L_3)	2.43	**5.58**	**3.16**
+ SAE (L_5)	2.46	5.64	2.95

Table 6: Impact of fusion depth in SAE. L_k denotes using acoustic features from the k-th encoder layer.

As shown in Table 6, SAE consistently outperforms the baseline without SAE, especially on the challenging LS-other and NLS. This confirms that supplementing deep semantic features with shallow acoustic cues improves non-linguistic discriminability without harming phonetic preservation.

The choice of fusion depth reveals a **semantic-acoustic trade-off**. L_1 features contain rich low-level details and yield the best WER on clean speech, but they are less structurally aligned with deep semantic representations. Conversely, L_5 features have undergone substantial semantic abstraction and therefore retain fewer fine-grained acoustic cues, resulting in only marginal NLS improvement.

Fusion from L_3 provides the **optimal equilibrium**, achieving a peak NLS of 3.16 while maintaining competitive WER. This suggests that L_3 retains sufficient acoustic cues (e.g., timbral patterns and transient events) while remaining compatible with deep semantic features for effective fusion.

5.3.2 Adaptive Gating Behavior

We further examine whether the learned gate acts as a content-aware adaptive controller. To this end, we statistically analyze and visualize the gate activations under controlled acoustic conditions.

Noise-Adaptive Gating. We mix clean speech in LibriSpeech with music from MusicBench (Melechovsky et al., 2024) at different Signal-to-Noise Ratios (SNRs). As shown in Figure 4a, gate activation increases as SNR decreases, indicating that when background noise (music) becomes more prominent, SAE injects more acoustic information to compensate for increased acoustic complexity.

Modality-Aware Gating. We further examine the temporal dynamics of the gate on a concatenated clip containing 5 seconds of speech followed by 5 seconds of music. Figure 4b reveals strikingly distinct activation patterns across modalities. During speech, the gate remains relatively suppressed, making the model rely more on deep semantic abstraction. During music, the gate becomes stronger and more variable, actively capturing acoustic textures essential for non-linguistic perception.

These results support our motivation: rather than applying a fixed fusion strategy, SAE dynamically regulates the flow of shallow acoustic information according to the input content. This adaptive behavior helps UniAudio-Token balance linguistic abstraction with acoustic detail, yielding a unified representation for both speech and general audio.

6 Conclusion

In this paper, we address the critical limitation of *acoustic blindness* in current semantic speech tokenizers by introducing UniAudio-Token, a novel framework empowering semantic speech tokenizers with general audio perception. It leverages Semantic-Acoustic Primitives (SAP) as a supervision protocol and a Semantic-Acoustic Equilibrium (SAE) mechanism to adaptively rectify the acoustic information loss inherent in traditional semantic-centric paradigms. Extensive experiments validate its effectiveness in improving general audio discriminability, speech reconstruction fidelity, and downstream Audio-LLM performance.

Limitations

UniAudio-Token is designed as a compact single-codebook tokenizer for Audio-LLMs, with an emphasis on balancing linguistic alignment, general audio perception, and speech generation. Due to its compact low-bitrate design and text-based supervision, its waveform-level reconstruction quality for complex non-speech audio still falls short of specialized high-bitrate acoustic codecs optimized for high-fidelity waveform reconstruction. Moreover, while the UniAudio-Token framework itself is architecturally language-agnostic and can in principle support multilingual speech, our current training and evaluation mainly cover English and Chinese, constrained by the availability of audio-text data resources. Future work may explore further improving non-speech audio reconstruction fidelity and more diverse language coverage, while preserving the compactness and semantic alignment required by LLM-based modeling.

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. *Preprint*, arXiv:2406.02430.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4218–4222, Marseille, France. European Language Resources Association.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems*, volume 33, pages 12449–12460. Curran Associates, Inc.

Jisheng Bai, Haohe Liu, Mou Wang, Dongyuan Shi, Wenwu Wang, Mark D. Plumbley, Woon-Seng Gan, and Jianfeng Chen. 2024. Audiosetcaps: An enriched audio-caption dataset using automated generation pipeline with large audio and language models. *Preprint*, arXiv:2411.18953.

Evelina Bakhturina, Vitaly Lavrukhin, Boris Ginsburg, and Yang Zhang. 2021. Hi-Fi Multi-Speaker English TTS Dataset. In *Interspeech 2021*, pages 2776–2780.

Hui Bu, Jiayu Du, Xingyu Na, Bengu Wu, and Hao Zheng. 2017. Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline. In *2017 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA)*, pages 1–5.

Guoguo Chen, Shuzhou Chai, Guan-Bo Wang, Jiayu Du, Wei-Qiang Zhang, Chao Weng, Dan Su, Daniel Povey, Jan Trmal, Junbo Zhang, Mingjie Jin, Sanjeev Khudanpur, Shinji Watanabe, Shuaijiang Zhao, Wei Zou, Xiangang Li, Xuchen Yao, Yongqing Wang, Zhao You, and Zhiyong Yan. 2021. GigaSpeech: An Evolving, Multi-Domain ASR Corpus with 10,000 Hours of Transcribed Audio. In *Interspeech 2021*, pages 3670–3674.

Yafeng Chen, Siqi Zheng, Hui Wang, Luyao Cheng, Qian Chen, and Jiajun Qi. 2023. An Enhanced Res2Net with Local and Global Feature Fusion for Speaker Verification. In *Interspeech 2023*, pages 2228–2232.

Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. 2023. High fidelity neural audio compression. *Transactions on Machine Learning Research*. Featured Certification, Reproducibility Certification.

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

Zhihao Du, Yuxuan Wang, Qian Chen, Xian Shi, Xiang Lv, Tianyu Zhao, Zhifu Gao, Yexin Yang, Changfeng Gao, Hui Wang, Fan Yu, Huadai Liu, Zhengyan Sheng, Yue Gu, Chong Deng, Wen Wang, Shiliang Zhang, Zhijie Yan, and Jingren Zhou. 2024. Cosyvoice 2: Scalable streaming speech synthesis with large language models. *Preprint*, arXiv:2412.10117.

Qingkai Fang, Shoutao Guo, Yan Zhou, Zhengrui Ma, Shaolei Zhang, and Yang Feng. 2025. Llama-omni: Seamless speech interaction with large language models. In *International Conference on Learning Representations*, volume 2025, pages 57607–57624.

Gunnar Fant. 1971. *Acoustic theory of speech production: with calculations based on X-ray studies of Russian articulations*. Walter de Gruyter.

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)*, pages 776–780.

Sreyan Ghosh, Arushi Goel, Jaehyeon Kim, Sonal Kumar, Zhifeng Kong, Sang-gil Lee, Chao-Han Yang, Ramani Duraiswami, Dinesh Manocha, Rafael Valle,

and Bryan Catanzaro. 2025. Audio flamingo 3: Advancing audio intelligence with fully open large audio language models. In *Advances in Neural Information Processing Systems*, volume 38, pages 41819–41886. Curran Associates, Inc.

Yuan Gong, Hongyin Luo, Alexander Liu, Leonid Karlinsky, and James R Glass. 2024. Listen, think, and understand. In *International Conference on Learning Representations*, volume 2024, pages 18516–18545.

Haorui He, Zengqiang Shang, Chaoren Wang, Xuyuan Li, Yicheng Gu, Hua Hua, Liwei Liu, Chen Yang, Jiaqi Li, Peiyang Shi, Yuancheng Wang, Kai Chen, Pengyuan Zhang, and Zhizheng Wu. 2024. Emilia: An extensive, multilingual, and diverse speech dataset for large-scale speech generation. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pages 885–890.

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

Shujie Hu, Long Zhou, Shujie Liu, Sanyuan Chen, Lingwei Meng, Hongkun Hao, Jing Pan, Xunying Liu, Jinyu Li, Sunit Sivasankaran, Linquan Liu, and Furu Wei. 2024. WavLLM: Towards robust and adaptive speech large language model. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4552–4572, Miami, Florida, USA. Association for Computational Linguistics.

Ailin Huang, Boyong Wu, Bruce Wang, Chao Yan, Chen Hu, Chengli Feng, Fei Tian, Feiyu Shen, Jingbei Li, Mingrui Chen, Peng Liu, Ruihang Miao, Wang You, Xi Chen, Xuerui Yang, Yechang Huang, Yuxiang Zhang, Zheng Gong, Zixin Zhang, and 126 others. 2025. Step-audio: Unified understanding and generation in intelligent speech interaction. *Preprint*, arXiv:2502.11946.

Wenyong Huang, Zhenhe Zhang, Yu Ting Yeung, Xin Jiang, and Qun Liu. 2022. SPIRAL: Self-supervised perturbation-invariant representation learning for speech pre-training. In *International Conference on Learning Representations*.

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 *International Conference on Learning Representations*, volume 2025, pages 93809–93826.

KimiTeam, Ding Ding, Zeqian Ju, Yichong Leng, Songxiang Liu, Tong Liu, Zeyu Shang, Kai Shen, Wei Song, Xu Tan, Heyi Tang, Zhengtao Wang, Chu Wei, Yifei Xin, Xinran Xu, Jianwei Yu, Yutao Zhang, Xinyu Zhou, Y. Charles, and 21 others. 2025. Kimi-audio technical report. *Preprint*, arXiv:2504.18425.

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*, volume 36, pages 27980–27993. Curran Associates, Inc.

Tianpeng Li, Jun Liu, Tao Zhang, Yuanbo Fang, Da Pan, Mingrui Wang, Zheng Liang, Zehuan Li, Mingan Lin, Guosheng Dong, Jianhua Xu, Haoze Sun, Zenan Zhou, and Weipeng Chen. 2025. Baichuan-audio: A unified framework for end-to-end speech interaction. *Preprint*, arXiv:2502.17239.

Xinjian Li, Shinnosuke Takamichi, Takaaki Saeki, William Chen, Sayaka Shiota, and Shinji Watanabe. 2023. Yodas: Youtube-oriented dataset for audio and speech. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8.

Chen-Chou Lo, Szu-Wei Fu, Wen-Chin Huang, Xin Wang, Junichi Yamagishi, Yu Tsao, and Hsin-Min Wang. 2019. MOSNet: Deep Learning-Based Objective Assessment for Voice Conversion. In *Interspeech 2019*, pages 1541–1545.

Ziyang Ma, Yinghao Ma, Yanqiao Zhu, Chen Yang, Yi-Wen Chao, Ruiyang Xu, Wenxi Chen, Yuanzhe Chen, Zhuo Chen, Jian Cong, Kai Li, Keliang Li, Siyou Li, Xinfeng Li, Xiquan Li, Zheng Lian, Yuzhe Liang, Minghao Liu, Zhikang Niu, and 15 others. 2025. Mmar: A challenging benchmark for deep reasoning in speech, audio, music, and their mix. In *Advances in Neural Information Processing Systems*, volume 38. Curran Associates, Inc.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press.

Jan Melechovsky, 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)*, pages 8293–8316, Mexico City, Mexico. Association for Computational Linguistics.

Pooneh Mousavi, Jarod Duret, Darius Petermann, Artem Ploujnikov, Luca Della Libera, Anastasia Kuznetsova, Cem Subakan, and Mirco Ravanelli. 2026. Dasb - discrete audio and speech benchmark. *Preprint*, arXiv:2406.14294.

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)*, pages 5206–5210.

Karol J. Piczak. 2015. Esc: Dataset for environmental sound classification. In *Proceedings of the 23rd ACM International Conference on Multimedia*, MM '15, page 1015–1018, New York, NY, USA. Association for Computing Machinery.

Adam Polyak, Lior Wolf, and Yaniv Taigman. 2020. TTS Skins: Speaker Conversion via ASR. In *Interspeech 2020*, pages 786–790.

Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert. 2020. MLS: A Large-Scale Multilingual Dataset for Speech Research. In *Interspeech 2020*, pages 2757–2761.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR.

Peter J. Rousseeuw. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65.

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 *International Conference on Learning Representations*, volume 2025, pages 84929–84964.

Amitay Sicherman and Yossi Adi. 2023. Analysing discrete self supervised speech representation for spoken language modeling. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Yuhan Song, Linhao Zhang, Chuhan Wu, Aiwei Liu, Wei Jia, Houfeng Wang, and Zhou Xiao. 2026. Stabletoken: A noise-robust semantic speech tokenizer for resilient speechLLMs. In *The Fourteenth International Conference on Learning Representations*.

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 *International Conference on Learning Representations*, volume 2024, pages 16607–16629.

Aaron van den Oord, Oriol Vinyals, and koray kavukcuoglu. 2017. Neural discrete representation learning. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605.

Dingdong Wang, Junan Li, Jincenzi Wu, Dongchao Yang, Xueyuan Chen, Tianhua Zhang, and Helen M. Meng. 2026. MMSU: A massive multi-task spoken language understanding and reasoning benchmark. In *The Fourteenth International Conference on Learning Representations*.

Hui Wang, Siqi Zheng, Yafeng Chen, Luyao Cheng, and Qian Chen. 2023. CAM++: A Fast and Efficient Network for Speaker Verification Using Context-Aware Masking. In *Interspeech 2023*, pages 5301–5305.

Xiong Wang, Yangze Li, Chaoyou Fu, Yike Zhang, Yunhang Shen, Lei Xie, Ke Li, Xing Sun, and Long Ma. 2025. Freeze-omni: A smart and low latency speech-to-speech dialogue model with frozen LLM. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 63345–63354. PMLR.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. 2025a. Qwen2.5-omni technical report. *Preprint*, arXiv:2503.20215.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, Yuanjun Lv, Yongqi Wang, Dake Guo, He Wang, Linhan Ma, Pei Zhang, Xinyu Zhang, Hongkun Hao, Zishan Guo, and 19 others. 2025b. Qwen3-omni technical report. *Preprint*, arXiv:2509.17765.

Junichi Yamagishi, Christophe Veaux, and Kirsten MacDonald. 2019. CSTR VCTK Corpus: English multi-speaker corpus for CSTR voice cloning toolkit (version 0.92). *University of Edinburgh. The Centre for Speech Technology Research (CSTR)*.

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

Dongchao Yang, Songxiang Liu, Rongjie Huang, Jinchuan Tian, Chao Weng, and Yuexian Zou. 2023. Hifi-codec: Group-residual vector quantization for high fidelity audio codec. *Preprint*, arXiv:2305.02765.

Zhengdong Yang, Shuichiro Shimizu, Yahan Yu, and Chenhui Chu. 2025b. When large language models meet speech: A survey on integration approaches. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 20298–20315, Vienna, Austria. Association for Computational Linguistics.

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. 2025. Codec does matter: Exploring the semantic shortcoming of codec for audio language model. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(24):25697–25705.

Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi. 2022. Soundstream: An end-to-end neural audio codec. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 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. In *Interspeech 2019*, pages 1526–1530.

Aohan Zeng, Zhengxiao Du, Mingdao Liu, Kedong Wang, Shengmin Jiang, Lei Zhao, Yuxiao Dong, and Jie Tang. 2024. Glm-4-voice: Towards intelligent and human-like end-to-end spoken chatbot. *Preprint*, arXiv:2412.02612.

Aohan Zeng, Zhengxiao Du, Mingdao Liu, Lei Zhang, Shengmin Jiang, Yuxiao Dong, and Jie Tang. 2025. Scaling speech-text pre-training with synthetic interleaved data. In *International Conference on Learning Representations*, volume 2025, pages 49396–49419.

Binbin Zhang, Hang Lv, Pengcheng Guo, Qijie Shao, Chao Yang, Lei Xie, Xin Xu, Hui Bu, Xiaoyu Chen, Chenchen Zeng, Di Wu, and Zhendong Peng. 2022. Wenetspeech: A 10000+ hours multi-domain mandarin corpus for speech recognition. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6182–6186.

Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. 2023. Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities. *Preprint*, arXiv:2305.11000.

Dong Zhang, Gang Wang, Jinlong Xue, Kai Fang, Liang Zhao, Rui Ma, Shuhuai Ren, Shuo Liu, Tao Guo, Weiji Zhuang, Xin Zhang, Xingchen Song, Yihan Yan, Yongzhe He, Cici, Bowen Shen, Chengxuan Zhu, Chong Ma, Chun Chen, and 80 others. 2025a. Mimo-audio: Audio language models are few-shot learners. *Preprint*, arXiv:2512.23808.

Linhao Zhang, Dehong Ma, Xiaodong Zhang, Xiaohui Yan, and Houfeng Wang. 2020. Graph lstm with context-gated mechanism for spoken language understanding. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):9539–9546.

Linhao Zhang, Yuhan Song, Aiwei Liu, Chuhan Wu, Sijun Zhang, Wei Jia, Yuan Liu, Houfeng Wang, and Xiao Zhou. 2026. Beyond transcription: Unified audio schema for perception-aware audiollms. *Preprint*, arXiv:2604.12506.

Linhao Zhang and Houfeng Wang. 2019. Using bidirectional transformer-crf for spoken language understanding. In *Natural Language Processing and Chinese Computing*, pages 130–141, Cham. Springer International Publishing.

Linhao Zhang, Jian Zhang, Bokai Lei, Chuhan Wu, Aiwei Liu, Wei Jia, and Xiao Zhou. 2025b. Wildspeechbench: Benchmarking end-to-end speechllms in the wild. *Preprint*, arXiv:2506.21875.

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*.

A Samples of Semantic-Acoustic Primitives (SAP)

In this section, we present some concrete examples of the Semantic-Acoustic Primitives (SAP) that are used as model supervision targets ($\mathbf{y}_{\text{SAP}}$) in our framework. Specifically, Figure 5, Figure 6, and Figure 7 illustrate examples of SAP data annotations associated with speech, music, and environmental sound audio clips, respectively.

Speech. Figure 5 illustrates the fine-grained annotation of a speech clip within a complex auditory environment. Beyond the verbatim transcript (the "linguistic_content" field), the SAP schema also includes vocal attributes such as the speaker's identity and the expressive prosody. It also details a high-level audio clip summary alongside discrete auditory events, such as impulsive applause and hall reverberation. By encoding these as part of supervision targets, the framework facilitates the model to perceive the acoustic environment as a complement to speech components.

Music. Figure 6 demonstrates an example of the SAP schema when encountering an instrumental music audio clip without a human voice. Due to the absence of speech, the "linguistic_content" and "vocal_attributes" fields are assigned `null` values. This sparsity prevents the model from hallucinating false linguistic and paralinguistic features in pure music audio clips. Moreover, this music instance prioritizes the "events" list to capture the concurrent instrumental layers. By separately supervising targets for each music component (such as lead guitar, rhythm guitar, drums, and bass), the framework encourages the model to learn disentangled representations of diverse instrument timbres and rhythmic patterns.

```
{
  "linguistic_content": "I'd like to thank my husband for supporting me and
      going on this journey with me. Thank you.",
  "vocal_attributes": {
    "age": "Adult",
    "gender": "Female",
    "emotion": "Happiness",
    "accent": "General American English",
    "prosody": "Slightly breathy, with a touch of vocal tremor and a natural
        rise and fall in pitch that conveys sincerity and emotion.",
    "timbre": "Gentle"
  },
  "auditory_scenes": {
    "summary": "The audio captures a heartfelt personal acknowledgment in a
        large, reverberant hall, followed by a dense, energetic, and sustained
        wave of applause from a large audience, with natural reverberation
        matching the acoustics of the space.",
    "events": [
      {
        "class": "Audience applause",
        "temporal_type": "impulsive",
        "properties": "Dense, energetic, sustained"
      },
      {
        "class": "Reverberation in hall",
        "temporal_type": "persistent",
        "properties": "Natural and sustained"
      }
    ]
  }
}
```


Figure 5: Example of SAP data annotation associated with a speech audio clip.

Environmental Sounds. Figure 7 shows the SAP's capability to categorize and supervise diverse acoustic phenomena in an environmental sound recording. It highlights the distinction between different temporal granularities, such as *impulsive* wheel clatter, *transient* brake hisses, and *persistent* engine rumbles and background hiss. Therefore, the framework enables the model to learn representations that are sensitive to the start time and duration of acoustic events. Furthermore, the SAP captures not only the primary sound sources, but also the acoustic transformations (e.g., reverberation in Figure 5) imposed by the environment, thereby providing a foundation for downstream auditory reasoning tasks.

B Human Evaluation of SAP Data Annotation Quality

The reliability of the SAP data annotations is further verified through a manual quality assessment on a randomly sampled subset of the data by three human experts. Specifically, we conducted a manual audit on a uniformly random sample of 500 audio segments, covering speech, music, and environmental sounds. Three expert volunteers in audio processing performed the evaluation. The study adhered to ethical guidelines, with all participants providing informed consent and being notified of their right to withdraw without penalty. No financial incentives were involved in this process.

We validate the automated SAP generation by measuring the alignment between the audio content and the JSON outputs through the consensus of annotators. An attribute was considered valid only if at least two out of three experts confirmed its precision. Results are reported by the accuracy calculated over the eight fine-grained fields that comprise the vocal attributes and auditory scenes. We also calculated the 95% Confidence Intervals (CI) using the Wilson score interval method.

As detailed in Figure 8, the automated SAP generation pipeline demonstrates great precision across the board, with the majority of attributes surpassing 95% accuracy. It can be observed that objective vocal characteristics (e.g., *Age* 98.4%, *Accent* 96.0%) and high-level environmental descriptions (e.g., *Summary* 96.4%) exhibit high reliability, with the lower bounds of their 95% confidence intervals consistently exceeding 92%. The performance drop in *Emotion* is likely attributable to the inherent subjec-

```
{
  "linguistic_content": null,
  "vocal_attributes": {
    "age": null,
    "gender": null,
    "emotion": null,
    "accent": null,
    "prosody": null,
    "timbre": null
  },
  "auditory_scenes": {
    "summary": "The audio clip is a high-fidelity, instrumental funk track
        characterized by a classic 1970s studio production style, with no vocals
        or extraneous sounds, and concludes with an abrupt, unresolved ending.",
    "events": [
      {
        "class": "Lead guitar riff",
        "temporal_type": "persistent",
        "properties": "Bright, slightly overdriven, rapid descending melodic
            phrase"
      },
      {
        "class": "Rhythm guitar chords",
        "temporal_type": "persistent",
        "properties": "Sharp and staccato"
      },
      {
        "class": "Drum pattern",
        "temporal_type": "persistent",
        "properties": "Punchy and syncopated"
      },
      {
        "class": "Bass guitar line",
        "temporal_type": "persistent",
        "properties": "Melodic, syncopated, played with a pick"
      }
    ]
  }
}
```


Figure 6: Example of SAP data annotation associated with a music audio clip.

tivity of perceiving emotional states, and the complexity of identifying all the transient and persistent acoustic cues without error or omission accounts for the lower accuracy in *Events* list. Nonetheless, even for these nuanced fields, the lower confidence bound of accuracy remains above 85%, confirming that the pipeline provides reliable supervision across vocal attributes and auditory scenes.

C Training Details of UniAudio-Token

C.1 Training Datasets for UniAudio-Token

To develop a robust and versatile audio tokenizer, we trained UniAudio-Token on a massive, diverse corpus. Our training collection spans multiple domains, including high-quality speech, multi-lingual recordings, and diverse environmental sounds and music. Specifically, we incorporated major speech corpora such as Emilia (96,750 hours) and Yodas (29,155 hours) for broad linguistic coverage, alongside AudioSet (4,922 hours) to enhance the model's perception of non-linguistic acoustic events. A comprehensive summary of these open-source datasets and their respective durations is provided in Table 7.

C.2 Training Hyperparameters of UniAudio-Token

Table 8 summarizes the detailed hyperparameter configurations used during the training process of UniAudio-Token. To achieve a stable yet efficient optimization, we implement a multi-level learning rate strategy tailored to the specific functional roles of different model modules. Specifically, we apply a **conservative learning rate** (1e-5) to the pre-trained audio Encoder. This cautious approach is critical for preserving the rich semantic features and universal representations acquired during the

```
{
  "linguistic_content": null,
  "vocal_attributes": {
    "age": null,
    "gender": null,
    "emotion": null,
    "accent": null,
    "prosody": null,
    "timbre": null
  },
  "auditory_scenes": {
    "summary": "The audio is a low-fidelity field recording of a train passing
        through a tunnel or enclosed space, characterized by loud, rhythmic
        metallic clatter, deep rumble, and mechanical noise, followed by a sharp
        hiss from pneumatic brakes. The recording exhibits a reverberant, boomy
        acoustic environment, indicating an enclosed space. No human or animal
        presence is detected.",
    "events": [
      {
        "class": "Metallic clatter from train wheels",
        "temporal_type": "impulsive",
        "properties": "Loud, rhythmic"
      },
      {
        "class": "Pneumatic brake hiss",
        "temporal_type": "transient",
        "properties": "Sharp, brief"
      },
      {
        "class": "Train engine rumble",
        "temporal_type": "persistent",
        "properties": "Deep, continuous"
      },
      {
        "class": "Background recording hiss",
        "temporal_type": "persistent",
        "properties": "faint, persistent"
      }
    ]
  }
}
```


Figure 7: Example of SAP data annotation associated with an environmental sounds audio clip.

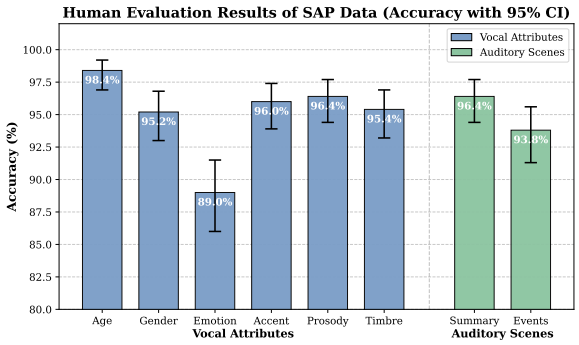


Figure 8: Human evaluation results of SAP data.

large-scale foundational pre-training phase, thereby avoiding catastrophic forgetting. In contrast, the **Decoder utilizes a significantly higher learning rate** (6e-4) to facilitate the rapid acquisition of the novel Semantic-Acoustic Primitives (SAP) generation task. Other training configurations include:

- **Learning Rate Schedule**: A Cosine schedule with 1,000 warmup iterations is adopted to stabilize the initial gradient updates and provide a smooth transition towards convergence.
- **Optimization Strategy**: We employ the AdamW optimizer ($\beta_1 = 0.9, \beta_2 = 0.999$) with a weight decay of 1×10^{-2} to ensure robust generalization. **Gradient clipping** with a threshold of 1.0 is applied to prevent gradient explosion during training.
- **Loss Balancing**: To maintain a high-quality discrete codebook while ensuring reconstruction fidelity, we set the **Quantization Loss Weight** (λ_1) to 10.0 and the **Commitment Loss Weight** (λ_2) to 2.5.

Dataset	Duration (# Hours)
LibriSpeech (Panayotov et al., 2015)	960
Multilingual LibriSpeech (Pratap et al., 2020)	27,322
GigaSpeech (Chen et al., 2021)	10,000
Yodas (Li et al., 2023)	29,155
Hi-Fi TTS (Bakhturina et al., 2021)	292
VCTK (Yamagishi et al., 2019)	44
LibriTTS (Zen et al., 2019)	586
AISHELL-1 (Bu et al., 2017)	150
WenetSpeech (Zhang et al., 2022)	10,005
Common Voice (Ardila et al., 2020)	2,133
Emilia (He et al., 2024)	96,750
AudioSet (Gemmeke et al., 2017)	4,922

Table 7: Overview of public datasets included in the training of UniAudio-Token.

Hyperparameter	Value
Learning rate	
Encoder max LR	1×10^{-5}
Decoder max LR	6×10^{-4}
Other max LR	2×10^{-4}
LR schedule	Cosine (with linear warmup)
Warmup iterations	1,000
Optimization	
Optimizer	AdamW ($\beta_1 = 0.9, \beta_2 = 0.999$)
Weight decay	1×10^{-2}
Gradient clipping	1.0
Loss weights	
Quantization loss (λ_1)	10.0
Commitment loss (λ_2)	2.5

Table 8: Hyperparameter configurations shared across the two training stages of UniAudio-Token.

D Baseline Audio Tokenizers

We compare UniAudio-Token against the following single-codebook audio tokenizers:

1. **WavTokenizer** (Ji et al., 2025), a high-compression single-codebook acoustic codecs. We use the officially released most powerful *large, 75Hz* variant in our experiments.

2. **CosyVoice2** (Du et al., 2024), a leading speech tokenization and generation model, which introduces Finite-Scalar Quantization (FSQ) to replace traditional Vector Quantization (VQ) in its audio tokenizer for enhanced codebook utilization and representation efficiency.

3. **GLM-4-Voice-Tokenizer** (Zeng et al., 2025), a representative semantic tokenizer tailored for Speech Large Language Models. It can compress speech into highly efficient discrete tokens at a significantly lower frame rate while ensuring robust semantic preservation. We use the officially released checkpoint which has a frame rate of 12.5Hz and a codebook size of 16,384 in our experiments.

4. **StableToken** (Song et al., 2026), a novel semantic speech tokenizer with superior noise robustness. It employs a multi-branch Voting-LFQ architecture and adopts a bit-wise voting mechanism and a noise-aware training strategy to extract noise-irrelevant semantic speech tokens.

E ESC-10 Token Sequence t-SNE Visualization Results

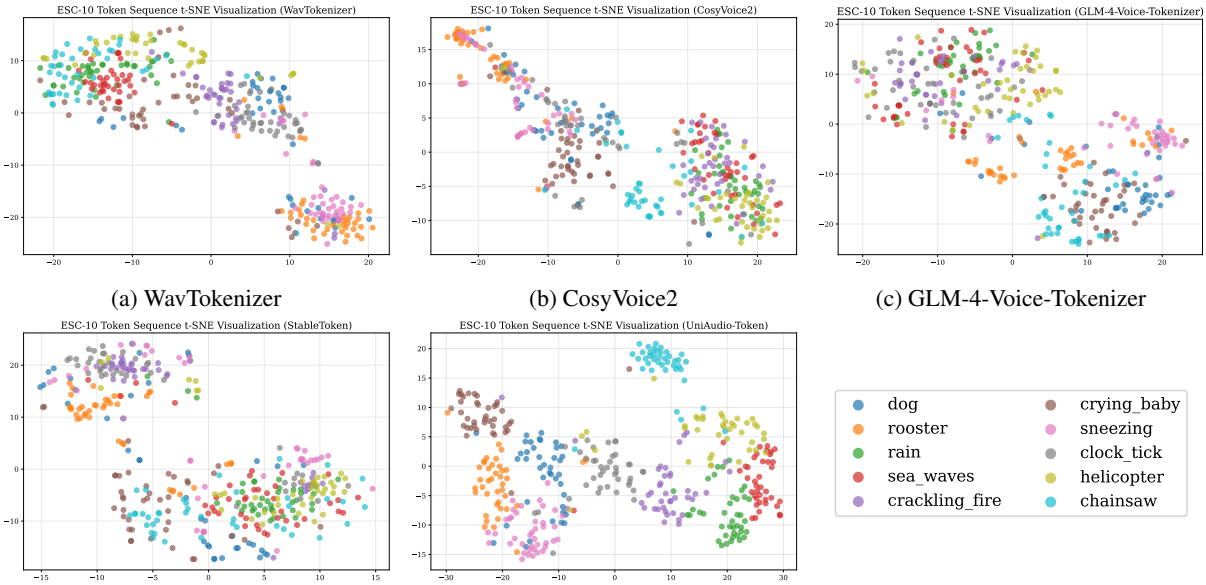
To further investigate the semantic representation capabilities of different audio tokenizers, we provide a t-SNE visualization of the token histograms on the ESC-10 dataset in Figure 9. This visualization maps high-dimensional token distributions into a 2D space to illustrate how well each model captures the underlying category information.

As shown in the figure, our UniAudio-Token (Figure 9e) exhibits the most distinct and semantically meaningful clusters. Samples belonging to the same sound category (e.g., *chainsaw*, *rooster*, or *sea waves*) are tightly grouped together with clear boundaries. In contrast, the baseline models, including WavTokenizer, CosyVoice2, GLM-4-Voice-Tokenizer, and StableToken (Figures 9a–9d), show significant feature fragmentation and overlap. In these baselines, different classes are often intermingled, suggesting that their token sequences lack sufficient discriminative power for environmental sound classification. These results demonstrate that UniAudio-Token's discrete representations are more effective at capturing high-level semantic features compared to existing speech-centric or general-purpose audio tokenizers.

F Non-Linguistic Score Evaluation Setup

In this section, we detail the methodology for calculating the Non-Linguistic Score (NLS), a metric designed to quantify the truthfulness of model-generated audio descriptions. Unlike traditional n-gram based metrics (e.g., BLEU or METEOR), the NLS focuses on the high-level alignment between generated content and the annotations provided in the SAP dataset.

To ensure a robust and nuanced evaluation, we employ an LLM-based judge framework. Specifi-



(a) WavTokenizer (b) CosyVoice2 (c) GLM-4-Voice-Tokenizer

(d) StableToken (e) UniAudio-Token (Ours) (f) Legend

Figure 9: t-SNE visualization of token histograms on the ESC-10 dataset. Our UniAudio-Token (Figure 9e) exhibits the most clear and semantically meaningful clusters, whereas the baselines (Figure 9a, 9b, 9c, 9d) show significant feature fragmentation and overlap.

cally, we utilize the Qwen3-235B-A22B-Instruct-2507 (Yang et al., 2025a) model to perform zero-shot scoring. This approach leverages the model's advanced reasoning capabilities to move beyond simple keyword matching, allowing for an assessment of complex audio attributes such as sound event sources, environmental context (e.g., spatial hints or reverberation), and audio characteristics (e.g., fidelity and mix).

The evaluation process is standardized through a carefully constructed prompt, as illustrated in Figure 10. The judge is tasked with assigning a score on a scale from 1 to 5, where a score of 5 represents "Perfect Consistency" and a score of 1 indicates a fundamental divergence in content. To minimize variance and maintain consistency across evaluations, the prompt provides explicit scoring criteria for each level, focusing on the preservation of principal information and the presence of contradictions. By enforcing a strict output format (score only), we ensure the results are directly parsable for large-scale quantitative analysis.

G LLM Usage Statement

In accordance with ACL policy on Generative AI tools and technologies, the authors hereby disclose the following: After the authors completed the initial draft of this paper, LLMs were utilized to enhance grammar and polish the writing of this manuscript. No new research ideas, experimental designs, or scientific content were generated by the LLMs. All scientific contributions, analyses, and conclusions presented in this work are solely those of the authors. We take full responsibility for the content of this paper, including all sections that have been revised or improved with LLM assistance. The LLMs are not authors and did not contribute to the research ideation or substantive scientific writing.

This statement is provided to ensure transparency and compliance with the ACL Guidelines on Generative Assistance in Authorship.

Prompt for Evaluating Non-Linguistic Score (NLS)

```
You are given two descriptions of the same audio clip:

- **Reference Description:** The ground-truth expert annotation describing the audio in detail.
- **Generated Description:** The generated description produced by a model.

Your task is to judge how well the Generated Description matches the Reference Description in
    its description of the audio clip.
**Audio can be any type**-music, singing, speech, spoken dialogue, environmental sounds, animal
    noises, or other non-musical audio.

**Please consider these aspects as relevant (if present):*230*
- **Sound events & sources:** Describe what sounds are present (e.g. music, speech,
    environmental sounds, specific objects, animals, vehicles, etc.)
- **Environmental context:** Indoors/outdoors, location hints, crowd noise, weather,
    reverberation, etc.
- **Audio characteristics:** Quality, fidelity, mix, spatial placement, background/foreground,
    clarity or distortion

**Scoring Criteria:**
Assign a score from **1** to **5** based on these standards:

- **5 - Perfect Consistency:**
  The Generated Description accurately covers almost all important information given in the
    Reference Description, including types of sound, main events, vocal or non-vocal details,
    environmental context, and tone. Any differences are minor and inconsequential.

- **4 - Mostly Consistent:**
  The Generated Description includes most major elements from the Reference Description. Some
    small details or less important aspects may be missing or stated differently, but all
    principal information matches closely. Only minor mistakes or omissions are present.

- **3 - Partially Consistent:**
  The Generated Description captures some main aspects correctly, but omits important details or
     contains noticeable errors regarding sound types, vocal or environmental features, or
    context. There is reasonable overlap, but inconsistency in parts.

- **2 - Mostly Inconsistent:**
  The Generated Description reflects very few features from the Reference Description. Most
    elements are missing, inaccurate, or contradicted (e.g., confuses speech for music; wrong
    language; misses major sound sources; incorrect context). There is substantial disagreement.

- **1 - Completely Inconsistent:**
  The Generated Description does not match the Reference Description in content or meaning. Major
     sound events/types, vocal or environmental information, and context differ fundamentally.

Below are the Generated Description and the Reference Description:

### [Generated Description]: ${generated}

### [Reference Description]: ${reference}

After evaluating, please output the score only without anything else. You dont need to provide
    any explanations.
```

Figure 10: Prompt template used by Qwen3-235B-A22B-Instruct-2507 for Non-Linguistic Score (NLS) evaluation.