LoSATok: Low-Dimensional Semantic-Acoustic Tokenizer for Cross-Domain Audio Understanding and Generation

Zhisheng Zhang[1*†], Xiang Li[1†], Yixuan Zhou[1], Jing Peng[1], Guoyang Zeng[2], Zhiyong Wu[1‡]
[1]Shenzhen International Graduate School, Tsinghua University, China
[2]ModelBest Inc., China
zhangzs25@mails.tsinghua.edu.cn

Abstract

Audio tokenizers are fundamental to unifying audio understanding and generation. Understanding requires high-level semantics, while generation demands semantic and acoustic details. Existing unified tokenizers jointly encode both in high-dimensional continuous latents, which increases the modeling burden of Diffusion Transformers (DiTs) for generation. We propose **LoSATok**, a low-dimensional audio tokenizer for cross-domain audio understanding and generation. Motivated by the observation that 1280-dimensional semantic encoder features are compressible, we introduce a Semantic Bottleneck that compresses them into 128 dimensions, regularized by the proposed time-relation loss for temporal feature consistency. We further design a dual-level semantic supervision method that leverages both high- and low-dimensional semantic signals, enabling the tokenizer to jointly capture semantics and acoustic details within a compact latent space. Experiments on speech, music, and general audio show that SemBo preserves strong low-dimensional semantic capacity and LoSATok retains competitive understanding performance compared with several semantic representations, while consistently improving DiT modeling performance on speech, music, and audio generation. These results demonstrate that LoSATok's low-dimensional representations can effectively support audio understanding and generation. Our code is provided at https://github.com/wxzyd123/LoSATok.

1 Introduction

Understanding and generation are two important research directions in the audio domain (Yan et al., 2025). The former focuses on extracting high-level semantic information from audio signals to support tasks, e.g., automatic speech recognition (ASR) and audio event classification, while the latter requires models to produce natural and realistic audio signals, e.g., text-to-speech (TTS) (Zhou et al., 2026), text-to-music (TTM) (Lei et al., 2025), and text-to-audio (TTA) (Xu et al., 2025b). These two categories of tasks typically rely on distinct model architectures or representation designs. Although such separate designs have achieved promising results, they limit the model's ability to perform unified modeling across diverse tasks.

Recently, several works have explored joint modeling of understanding and generation, which can be categorized into model-level and representation-level approaches. At the **model level**, unified audio language models, e.g., UALM (Tian et al., 2026), enable a single model to handle both understanding and generation tasks via post-training, reasoning, etc. Such methods generally require more complex training pipelines, and their internal representations may still fail to simultaneously satisfy the requirements of semantic preservation, acoustic reconstruction, and efficient modeling. Another line of work aims to unify understanding and generation at the **representation level**, i.e., through a unified audio representation. DashengTokenizer (Dinkel et al., 2026) injects acoustic details into a semantic representation to jointly model semantics and acoustics, exhibiting stronger generality than purely acoustic representations across various understanding and generation tasks. RAE (Zheng et al., 2026) shows that semantically rich continuous representations can improve generative modeling with Diffusion Transformers (DiTs) (Peebles and Xie, 2023). However, these high-dimensional representations often increase the modeling burden on downstream generators, requiring wider DiT blocks or more parameters to converge effectively.

In this paper, we consider a research question: *Can we learn a low-dimensional, semantically rich, and acoustically reconstructable audio representation that supports cross-domain modeling*

*Work conducted while interning at ModelBest.
†Equal contribution.
‡Corresponding author: Zhiyong Wu.

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 21062–21079
October 24-29, 2026 ©2026 Association for Computational Linguistics

for both understanding and generation? Acoustic signals are usually low-dimensional (Xu et al., 2025b), but existing semantic representations are mostly high-dimensional, e.g., > 768 dimensions. Therefore, a direct low-dimensional semantic signal is lacking. To this end, we investigate the intrinsic structure of the high-dimensional semantic representations from MiDashengLM (Dinkel et al., 2025). We find that these semantic representations exhibit information redundancy: **their variance is not uniformly distributed across dimensions but instead concentrated in some principal directions**. Effective rank and Principal Component Analysis (PCA) reveal that a substantial portion of the semantic information is preserved even after compressing the high-dimensional representations to **128** dimensions. Training-free channel merging or PCA-based reduction retains rich semantics at 128 dimensions. However, these methods are not optimized for temporal structure, making them suboptimal as semantic supervision.

Building on this, we propose the **Sem**antic **Bo**ttleneck (**SemBo**) to learn a low-dimensional semantic signal from high-dimensional semantic representations. Specifically, we freeze a pre-trained semantic encoder and employ a compressor to project the high-dimensional representation into a low-dimensional space and a restorer to reconstruct it. In addition to the high-dimensional reconstruction loss, we expect the high-dimensional semantic signal to directly supervise the low-dimensional one, while the two have different dimensionalities. Inspired by the Gram loss (Jing et al., 2019), we propose a time-relation loss that preserves the semantic relationships across different temporal frames. Compared with training-free strategies, SemBo explicitly models the relationship between low- and high-dimensional semantic representations via a learnable non-linear mapping, yielding a low-dimensional semantic representation that is better suited as a supervision signal.

Based on this low-dimensional semantic signal, we propose the **Lo**w-dimensional **S**emantic-**A**coustic **Tok**enizer (**LoSATok**), an architecture for cross-domain audio representation modeling. To jointly encode semantic and acoustic features, we introduce a dual-level semantic supervision scheme. The high-dimensional semantic supervision provides a complete semantic target, while the low-dimensional semantic supervision serves as a direct and compact target that guides the tokenizer to also learn low-dimensional semantic features. Meanwhile, we inject acoustic signals into the tokenizer, enabling native joint modeling of semantic and acoustic properties.

We comprehensively assess the effectiveness of LoSATok across speech, music, and general audio domains. Experiments across speech, music, and general audio show that SemBo retains approximately 93% of the teacher encoder's average understanding performance at 10% of the feature dimension, even better than several Self-Supervised Learning (SSL) representations, e.g., HuBERT and WavLM. LoSATok provides a compact reconstructable latent whose sampled view preserves strong global and event-level understanding, while trading off fine-grained content probing for improved DiT learnability. Under a 208M-parameter DiT, LoSATok outperforms the acoustic baseline in both single-task and multi-task training, while reducing the DiT inference cost from 1,926 to 944 GFLOPs. These results position LoSATok as a generation-efficient semantic–acoustic tokenizer. The main contributions are summarized as follows:

- We analyze the high-dimensional semantic representations' redundancy, reveal their low-rank structure, and verify the feasibility of low-dimensional semantic supervision.
- We propose the Semantic Bottleneck, which compresses high-dimensional semantic signals with a learnable network via semantic reconstruction and time-relation preservation.
- We propose LoSATok with dual-level semantic supervision, enabling a 128-dimensional reconstructable semantic–acoustic latent and explicitly characterizing its trade-offs across semantic probing, acoustic reconstruction, and generation.
- We validate the efficiency and effectiveness of the proposed low-dimensional unified representation across speech, music, and general audio, covering 15 understanding tasks and three generation tasks (TTA, TTM, and TTS).

2 Related Work

Unified Understanding and Generation. Audio understanding and generation have previously relied on different models or representations. Some recent studies have explored unified modeling of the two. One line of work seeks unification at the model level. UALM (Tian et al., 2026) unifies understanding, generation, and reasoning within one model through different training strategies, such as SFT and DPO. Ming-UniAudio (Yan et al., 2025)

targets speech scenarios and uses high-dimensional general representations for LLM modeling and low-dimensional representations for DiT modeling, thereby unifying speech understanding, generation, and editing. These methods achieve model-level unification through more complex training pipelines and engineering architectures. Another line of work operates at the representation level, aiming to support both understanding and generation within a single general-purpose representation. JMAS-VAE (Cheng et al., 2026) utilizes joint-marginal distillation losses and adaptive weighting, improving the balance between speech reconstruction, understanding, and generation. DashengTokenizer (Dinkel et al., 2026) injects acoustic information into frozen semantic features to construct a continuous audio tokenizer, enabling joint semantic-acoustic modeling. However, such high-dimensional representations increase the burden on downstream DiT modeling and require wider DiT blocks. Concurrently, WavCube (Yang et al., 2026) models speech understanding and generation by directly compressing high-dimensional semantic representations into 128-dimensional semantic representations and constructing acoustic details from them. Our work differs in four respects: 1) we jointly and natively model semantics and acoustics, reducing the difficulty of modeling acoustics from semantic representations; 2) we analyze the compressibility of high-dimensional semantic representations and propose a time-relation loss to explicitly learn the mapping between high- and low-dimensional representations; 3) we achieve cross-domain understanding and generation beyond the speech domain; and 4) we investigate the efficiency and performance of acoustic and semantic tokenizers for downstream DiT-based generation.

Discrete and Continuous Audio Tokenizers. Audio tokenizers have been widely used across audio-related tasks, e.g., ASR and TTS. Neural audio codecs (NACs) typically use discrete token representations (Zhang et al., 2026). DAC (Kumar et al., 2023) and SNAC (Siuzdak et al., 2024) quantize encoder representations to achieve efficient audio compression. Existing discrete NACs can be easily combined with autoregressive language models (Ye et al., 2025), but they suffer from quantization loss (Zhou et al., 2026), which limits their ability to perform fine-grained reconstruction. Unlike discrete tokenizers, continuous audio tokenizers (CATs) model audio directly in the continuous latent space. They usually offer stronger reconstruction capabilities and are better suited to continuous generative architectures such as DiT. The UniFlow-Audio (Xu et al., 2025b) architecture, based on an acoustic VAE, achieves superior reconstruction. Recent work introduces semantic information into continuous tokenizers to bridge the gap between acoustic and semantic representations. SemanticVocoder (Xie et al., 2026) reconstructs audio from semantic latents and restores acoustic details from semantic information using a generative decoder. This paper focuses on CATs, aiming to preserve more acoustic details with rich semantics and better adapt to DiTs.

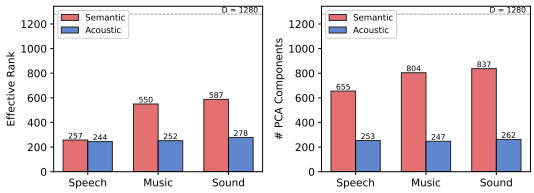


Figure 1: Effective rank and PCA components.

3 Semantic Bottleneck

Effective Rank and PCA Analysis. To obtain low-dimensional semantic supervision, we explore whether high-dimensional semantic representations are compressible. If semantic representations concentrate along a few principal directions, they contain redundancy. We analyze acoustic and semantic representations of DashengTokenizer using effective rank and Principal Component Analysis (PCA). The semantic representations are extracted from MiDashengLM (Dinkel et al., 2025).

As shown in Figure 1, although the original feature dimension is 1280, the effective rank is only 257 on speech and remains below 50% of the original dimension in the music and audio domains of semantic representations, suggesting that semantic energy concentrates along dominant directions. PCA also shows that fewer than 1280 components are sufficient to preserve 90% of the variance. These results indicate that semantic representations are more complex than acoustic features but still contain substantial dimensional redundancy, which motivates subsequent compression. Further details are provided in Appendix E.1.

Training-Free Dimensionality Reduction. Based on the above low-rank property, we test whether training-free dimensionality reduction preserves semantic information. The first method is *channel merging* (CM). For a 1280-dimensional semantic

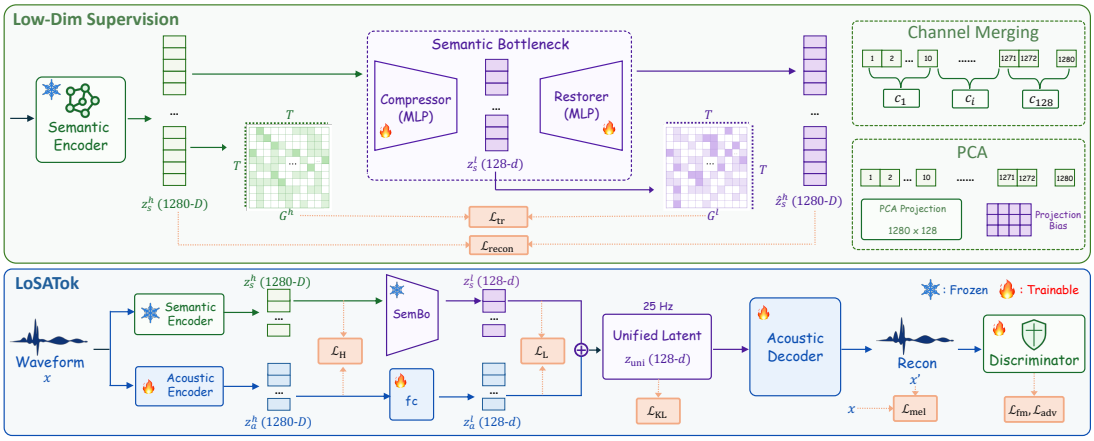


Figure 2: The architecture and training components of SemBo and LoSATok.

Table 1: Understanding on the XARES benchmark.

Method	Dimension	ESC(↑)	FSC(↑)	GTZAN(↑)
MiDashengLM	1280	96.95	98.26	91.19
CM	128	92.80	86.11	89.39
PCA	128	**94.95**	78.06	90.49
SemBo (Ours)	128	93.75	**89.11**	**89.29**

representation, we average every 10 channels and obtain a 128-dimensional representation. The second method is *PCA reduction*. We randomly sample 6000 examples from the cross-domain data in Table 12, fit PCA using extracted semantic features, and retain the first 128 principal components.

Table 1 shows that both methods preserve semantic information on downstream understanding tasks, confirming that the semantic representation has a compressible redundant structure. However, these methods do not explicitly model the mapping between high- and low-dimensional semantic latents, and thus provide weak supervision for unified representations, as discussed in Section 5.6.

Semantic Bottleneck. To obtain a more effective low-dimensional semantic signal, we propose **Sem**antic **Bo**ttleneck (**SemBo**). SemBo consists of a compressor and a restorer, as shown in Figure 2. Given the high-dimensional output z_s^h from a frozen semantic encoder, the compressor C maps it to $z_s^l = C(z_s^h), \quad z_s^l \in \mathbb{R}^{T\times d}$, where $d = 128$. The restorer R then reconstructs the high-dimensional representation $\hat{z}_s^h = R(z_s^l)$.

Both modules are implemented with lightweight 2-layer multilayer perceptrons (MLPs). The reconstruction loss is given in Eq. (1).

$$\mathcal{L}_{\text{recon}} = \left\| \text{norm}\left(\hat{z}_s^h\right) - \text{sg}\left(\text{norm}(z_s^h)\right) \right\|_2, \quad (1)$$

where $\text{sg}(\cdot)$ denotes stop-gradient and $\text{norm}(\cdot)$ denotes normalization. Since reconstruction alone does not directly constrain the low-dimensional representation, we introduce a time-relation loss to achieve direct high-dimensional supervision for z_s^l. First, we compute temporal similarity matrices with input representation normalization $\mathbf{G}^h = z_s^h(z_s^h)^\top \in \mathbb{R}^{T\times T}, \mathbf{G}^l = z_s^l(z_s^l)^\top \in \mathbb{R}^{T\times T}$ for high- and low-dimensional representations. Then, we align the temporal relation of $\mathbf{G}^h$ and $\mathbf{G}^l$:

$$\mathcal{L}_{\text{tr}} = \left\| \mathbf{G}^l - \text{sg}(\mathbf{G}^h) \right\|_2. \quad (2)$$

Intuition. Direct feature-wise alignment between z_s^h and z_s^l is infeasible because they have different channel dimensions. However, their normalized temporal Gram matrices are both $T \times T$ and describe the same dimension-independent temporal geometry: which frames are semantically similar and which are dissimilar. Matching these matrices therefore transfers the temporal-semantic relations of the teacher space into the bottleneck. Moreover, $\mathcal{L}_{\text{recon}}$ constrains only the composite mapping $R \circ C$, leaving the bottleneck under-determined under many invertible reparameterizations. In contrast, $\mathcal{L}_{\text{tr}}$ directly anchors the geometry of z_s^l, explaining why removing it causes the larger downstream degradation in Fig. 4.

The final objective for SemBo is in Eq. (3).

$$\mathcal{L}_{\text{SemBo}} = \lambda_{\text{recon}}\mathcal{L}_{\text{recon}} + \mathcal{L}_{\text{tr}}, \quad (3)$$

where $\lambda_{\text{recon}} = 10^3$ with hyperparameter sensitivity analysis in Section 5.6. Compared with CM and PCA, SemBo performs well on FSC and GTZAN tasks in Table 1 and explicitly learns a nonlinear mapping between high- and low-dimensional semantic spaces, making it better suited as the low-dimensional semantic supervision for LoSATok.

4 LoSATok

Problem Definition. Given an input waveform x, LoSATok maps it with an encoder into a continuous latent representation $z_{\text{uni}} \in \mathbb{R}^{T \times d}$, where d denotes the latent dimension and is set to 128. A unified (i.e., semantic-acoustic) representation should:

1. *Semantic Richness:* contain rich semantic information for downstream understanding models.
2. *Acoustic Reconstruction:* preserve basic acoustic details for reconstruction.
3. *Generation Efficiency:* provide a compact and effective latent space for generative models.

Therefore, the objective of LoSATok is to learn a low-dimensional semantic-acoustic representation.

Dual-Level Semantic Supervision. Let the high-dimensional acoustic representation be $z_a^h = \mathrm{E}_a(x) \in \mathbb{R}^{T \times 1280}$ and the low-dimensional acoustic representation be $z_a^l = \mathrm{fc}\left(\mathrm{E}_a(x)\right)$, where E_a is the acoustic encoder that encodes the waveform into a high-dimensional acoustic representation and $\mathrm{fc}(\cdot)$ represents a linear layer for acoustic channel compression. For semantic-acoustic joint modeling, we aim to obtain the unified representation $z_{\text{pre}} = z_a^l + z_s^l$. Then, we obtain $\mu, \log \sigma^2 = f_{\text{KL}}(z_{\text{pre}}), \quad z_{\text{uni}} = \mu + \sigma \odot \epsilon$.

To enable z_{uni} to capture both semantic and acoustic features, we introduce dual-level semantic supervision. Specifically, we use both high- and low-dimensional semantic signals to supervise the latent space. The high-dimensional semantic representation provides global semantic information, while the low-dimensional semantic representation directly constrains the acoustic features. These two levels complement each other, allowing LoSATok to preserve generation-friendly representations:

$$\mathcal{L}_{\text{H}} = \left\| z_a^h - \mathrm{sg}(z_s^h) \right\|_2, \ \mathcal{L}_{\text{L}} = \left\| z_a^l - \mathrm{sg}(z_s^l) \right\|_2 \tag{4}$$

Architecture. LoSATok utilizes an encoder-decoder architecture. The encoder consists of a semantic encoder and an acoustic encoder. For the semantic encoder, we use the MiDashengLM (Dinkel et al., 2025) audio encoder, which is pretrained on large-scale data. For the acoustic encoder, following (Dinkel et al., 2026), we adopt a non-overlapping patch embedding scheme to align the acoustic and semantic representations. The decoder is based on Vocos (Siuzdak, 2024) and generates audio from the unified representation. The frame rate of LoSATok is 25Hz, and detailed architecture information is shown in Appendix E.3.

Training Objectives. In addition to semantic capabilities, LoSATok must also ensure audio reconstruction capability. Following (Kumar et al., 2023), we improve the naturalness and fine-grained quality of reconstructed audio using a mel reconstruction loss $\mathcal{L}_{\text{mel}}$ and adversarial training. We adopt a Multi-Frequency Discriminator (MFD) (Ye et al., 2025) with a hinge loss objective. The training objective of LoSATok is given in Eq. (5).

$$\begin{aligned} \mathcal{L} = \lambda_{\text{mel}} \mathcal{L}_{\text{mel}} + \lambda_{\text{sem}} \left(\mathcal{L}_{\text{H}} + \mathcal{L}_{\text{L}}\right) + \\ \lambda_{\text{KL}} \mathcal{L}_{\text{KL}} + \lambda_{\text{fm}} \mathcal{L}_{\text{fm}} + \lambda_{\text{adv}} \mathcal{L}_{\text{adv}}, \end{aligned} \tag{5}$$

where $\mathcal{L}_{\text{fm}}$ is the feature-matching loss, and $\mathcal{L}_{\text{adv}}$ is the adversarial loss. $\mathcal{L}_{\text{KL}}$ is the Kullback-Leibler (KL) divergence loss. The loss weights are set to $\{\lambda_{\text{mel}}, \lambda_{\text{sem}}, \lambda_{\text{KL}}, \lambda_{\text{fm}}, \lambda_{\text{adv}}\} = \{45, 45, 10^{-2}, 1, 1\}$. Details of these objectives are provided in Appendix E.3.

Downstream DiT Generation. The unified representation z_{uni} can serve as the target representation for downstream generative models. For TTA, TTM, and TTS tasks, given a text condition c, DiT learns to generate LoSATok latents from noise, modeled as $p_\theta(z_{\text{uni}} \mid c)$. The generated latents are then converted back into audio by the LoSATok decoder.

5 Experiments

5.1 LoSATok Training Setup

Datasets. We select a diverse 13.2K-hour cross-domain dataset, consisting of 34.6% speech data (LibriSpeech (Panayotov et al., 2015), VCTK (Yamagishi et al., 2019), and the Common Voice-en subset (Ardila et al., 2020)), 28.6% music data (MTG-Jamendo (Bogdanov et al., 2019) and MUSDB (Rafii et al., 2017)), and 36.8% general audio data (AudioSet (Gemmeke et al., 2017)).

Training Details. We train the model for one million steps on 8 NVIDIA H100 GPUs. The initial learning rate is set to 1×10^{-4}, and optimization is performed using AdamW. The global batch size is set to 64 during training.

5.2 Audio Understanding

Evaluation Setup. For understanding evaluation, we use the XARES benchmark (Zhang et al., 2025) with a linear probing protocol across 15 cross-domain tasks. XARES normalizes all task results to a $0 \sim 1$ scale, with higher values indicating better performance. We report percentages (%).

Speech Understanding. We evaluate 7 speech understanding tasks: ASR (LS100h), emotion recog-

Table 2: Understanding evaluation of representations on speech, music, and general audio domains. "Latent Dim." denotes the representation dimension of tokenizers, and "Avg." denotes the average scores of understanding tasks.

Model	Latent Dim.	Speech Understanding(↑)							Music Understanding(↑)				Audio Understanding(↑)				Avg.
		LS100h	CD	FSC	LibCnt	LSMF	RAV	VocS	FMA	GTZAN	MT	NSynth	Clo	DES	ESC	Urb8	
Low-dimensional Tokenizers																	
EnCodec	128	0.00	39.16	3.48	33.78	88.44	27.71	43.06	28.80	35.14	23.37	32.59	0.57	3.82	16.80	40.34	27.80
EzAudio	64	0.00	38.80	1.63	37.19	84.90	25.56	35.26	30.74	40.54	17.57	32.18	0.40	17.01	18.05	34.97	27.65
UniFlow-Audio	128	0.00	39.78	1.32	37.15	74.57	23.26	34.73	30.17	40.74	16.44	31.73	0.57	16.90	19.85	35.16	26.82
LoSATok																	
z_s^l	128	89.16	72.85	89.11	75.28	96.86	71.25	90.59	65.37	89.29	22.86	63.16	6.11	46.72	93.75	84.94	**70.49**
z_{pre}	128	82.74	70.34	57.13	71.66	97.08	59.79	90.18	63.43	87.09	44.25	59.62	5.46	38.47	90.95	84.47	66.84
μ	128	61.67	67.65	32.48	68.01	95.51	65.28	88.51	59.54	85.89	44.27	51.76	4.78	39.75	89.55	83.04	62.51
z_{uni}	128	39.00	65.41	25.10	66.63	95.14	60.42	88.03	58.06	85.39	41.87	51.39	4.86	36.67	88.90	82.66	59.30
Generalization of SemBo																	
Qwen3-Omni	2048	85.03	67.65	95.54	83.39	98.80	59.93	86.86	55.66	78.68	17.68	32.67	1.76	32.67	74.85	77.29	63.23
w/ SemBo	128	83.79	61.65	86.16	70.19	98.92	43.06	77.76	57.71	70.07	13.49	33.69	1.73	27.61	63.00	73.79	57.51
High-dimensional Tokenizer																	
DAC	1024	0.00	43.90	2.55	43.93	91.58	36.45	42.17	35.43	54.06	12.78	38.57	0.73	20.20	31.20	50.26	33.59
SemantiCodec	768	0.00	60.30	46.40	66.05	98.14	55.00	84.28	57.94	66.67	8.62	68.14	1.89	48.57	81.75	84.54	55.22
Whisper	1280	90.00	71.32	97.78	64.40	94.93	68.47	91.48	58.90	71.77	0.00	63.50	3.10	22.64	62.45	75.74	62.43
HuBERT	1024	82.45	58.06	98.73	49.72	85.94	47.71	81.71	42.81	51.85	0.00	48.22	1.74	3.64	37.35	57.40	49.82
WavLM	1024	67.06	45.16	96.07	50.37	75.95	32.57	71.19	36.11	48.14	0.00	40.14	0.79	16.76	31.45	53.13	44.33
MingTok-Audio	896	93.45	66.85	98.58	62.92	97.10	57.08	88.59	54.45	71.17	25.00	57.00	2.00	39.80	62.25	72.87	63.27
DashengTok	1280	68.71	80.56	83.39	79.98	98.06	83.06	93.15	66.29	89.99	57.65	77.91	5.64	55.40	96.40	83.80	74.67
MiDashengLM	1280	80.32	80.10	98.26	80.79	98.47	87.15	93.15	65.14	91.19	38.65	77.59	6.12	52.02	96.95	86.23	75.48

nition (Crema-D, RAV), intent classification (FSC), speaker counting (LibCnt), gender classification (LSMF), and vocal sound classification (VocS). For ASR, we connect the representation to Qwen2.5-0.5B for evaluation, while for the other tasks, we use the MLP classifier as in XARES.

Music Understanding. We evaluate the representation capability on music understanding tasks, including music genre classification (FMA, GTZAN), music transcription (MAESTRO, MT), and musical timbre/source recognition (Nsynth).

Audio Understanding. We select 4 types of general audio understanding tasks: audio captioning (Clo), sound/environment classification (ESC, Urb8), and domestic sound event detection (DES).

Results. Table 2 reports the understanding performance of the different LoSATok views. The 128-dimensional semantic view z_s^l achieves an average score of 70.49 across 15 tasks, retaining 93.4% of the performance of the 1280-dimensional MiDashengLM (75.48). Following the XARES ASR protocol, which connects the representation to Qwen2.5-0.5B, z_s^l obtains 89.16 on LS100h, approaching the performance of the ASR-oriented Whisper representation. The deterministic pre-KL unified embedding z_{pre} achieves an average score of 66.84, including 82.74 on LS100h, indicating that substantial linguistic information remains accessible after acoustic-detail injection. The posterior mean μ obtains 62.51 on average, while the reparameterized latent z_{uni} obtains 59.30, still outperforming several high-dimensional SSL representations, including *HuBERT (Hsu et al., 2021) and WavLM (Chen et al., 2022)*. The effect of KL regularization and sampling is concentrated on fine-grained content tasks: from z_{pre} to z_{uni}, LS100h decreases from 82.74 to 39.00 and FSC from 57.13 to 25.10. In contrast, global and event-level tasks are considerably more stable, with LSMF changing from 97.08 to 95.14 and ESC-50 from 90.95 to 88.90. Moreover, acoustic-detail injection benefits tasks whose relevant information is primarily acoustic: on MT, the unified latent improves the score from 22.86 to 41.87 over the semantic view. Overall, these results reveal a structured trade-off: deterministic views provide stronger accessibility for fine-grained understanding, whereas the sampled unified latent retains competitive broad-domain understanding while offering the regularized, reconstructable representation required for downstream generation.

Generalization of SemBo on Qwen3-Omni. To test whether the observed redundancy is specific to MiDashengLM, we adapt the analysis using the 2048-dimensional Qwen3-Omni (Xu et al., 2025a) audio encoder. Its effective rank is only approximately 100∼125 across domains, and PCA-128 already explains 84∼90% of the variance. We further train SemBo to compress the Qwen3-Omni features from 2048 to 128 dimensions, corresponding to a 16× reduction. As shown in Table 2, SemBo retains 57.51/63.23, or approximately 91%, of the average understanding score, while preserving ASR performance (83.79 vs. 85.03). These results suggest that both the redundancy pattern and semantic bottleneck generalize beyond the MiDashengLM.

5.3 Audio Generation

Evaluation Pipelines. To evaluate the downstream generation performance of LoSATok on TTA, TTM,

Table 3: Downstream generation results of *general-purpose* continuous audio tokenizers (CATs). "GFLOPs" represents the inference efficiency of DiT sampling process. The best results are highlighted in **bold**.

Model	Latent Dim.	DiT Dim.	GFLOPs(↓)	Text-To-Speech			Text-To-Music				Text-To-Audio			
				WER(↓)	SIM(↑)	UTMOS(↑)	FAD(↓)	FD(↓)	KL(↓)	CLAP(↑)	FAD(↓)	FD(↓)	KL(↓)	CLAP(↑)
Single-task Training														
UniFlow-Audio	128	512	1926	3.589	0.408	2.768	6.147	36.098	1.807	0.250	4.925	40.017	2.613	0.243
DashengTok	1280	512	1028	100.0	0.015	1.251	23.257	131.417	5.068	0.027	34.681	115.557	4.109	0.002
	1280	1536	1910	75.469	0.103	1.322	8.460	51.246	1.832	0.237	7.238	43.916	2.395	0.245
	1280	1536	7855	3.652	0.287	3.144	**3.780**	25.848	**1.751**	0.268	4.138	**24.605**	**1.738**	0.379
LoSATok	128	512	**944**	**3.030**	**0.548**	**3.367**	4.156	**25.089**	1.788	**0.282**	**2.760**	25.743	1.844	**0.381**
Multi-task Joint Training														
UniFlow-Audio	128	512	1926	4.529	0.344	2.259	4.752	28.173	1.724	0.272	4.066	37.217	2.554	0.252
DashengTok	1280	1536	1910	84.761	0.074	1.296	5.780	36.227	1.797	0.268	4.482	37.146	2.137	0.283
LoSATok	128	512	**944**	**3.667**	**0.507**	**3.310**	**3.366**	**20.127**	**1.628**	**0.300**	**1.987**	**21.363**	**1.700**	**0.396**

and TTS tasks, we follow the evaluation protocol of DashengTokenizer and adopt UniFlow-Audio (Xu et al., 2025b), a flow-based DiT paradigm for cross-domain generation. Specifically, we replace the VAE in UniFlow-Audio with the tokenizer under evaluation. During training, the tokenizer is frozen, and the DiT architecture is trained, while other settings remain unchanged for a fair comparison.

In the main experiments, we use the same version of the generation architecture, where the hidden dimension of DiT is set to 512 and the number of layers is set to 12. This standard trainable DiT configuration contains approximately 208M parameters. For DashengTokenizer, since its representation dimension is 1280, we follow its original settings (Dinkel et al., 2026) and consider three configurations to support training for comparison:

1. *Align with the original VAE setting:* DiT dimension is 512, the layer number is 12, and the number of trainable parameters is 215M.
2. *Wider DiT block with comparable parameter:* DiT dimension is 1536, the layer number is 2, and the number of trainable parameters is 322M.
3. *Larger trainable parameter scale:* DiT dimension is 1536, the layer number is 12, and the number of trainable parameters is 975M.

Moreover, the GFLOPs denote the inference efficiency of DiT. We evaluate this metric on LibriTTS with an average duration of 5.647 s per utterance.

Training and Inference. During training, we consider both single-task training and multi-task joint training. For single-task training, all experiments are conducted on 4 NVIDIA H100 GPUs. Specifically, TTA is trained on the WavCaps (Mei et al., 2024) dataset for 100K steps, TTM is trained on the LP-MusicCaps-MTT (Doh et al., 2023) subset for 50K steps, and TTS is trained on the LibriTTS (Zen et al., 2019) training set for 50K steps. For multi-task joint training, we merge the TTA, TTM, and TTS training data based on UniFlow-Audio, constructing a larger joint training setting with a more complex task composition, and train the model on 8 NVIDIA GPUs for 150K steps. In all training settings, the batch size on each GPU is set to 24.

During inference, the classifier-free guidance scale is 3.0, and the number of inference steps is 20. During evaluation, we evaluate the models on the AudioCaps (Kim et al., 2019) test set, the MusicCaps (Agostinelli et al., 2023) test set, and the LibriTTS test set, respectively. For TTA and TTM tasks, we use Fréchet Audio Distance (FAD) (Kilgour et al., 2019), Fréchet Distance (FD), KL divergence, and CLAPScore (Elizalde et al., 2023) as evaluation metrics. For the TTS task, we use Word Error Rate (WER), Speaker Similarity, and UTMOS (Saeki et al., 2022) for evaluation.

Results. Table 3 demonstrates our strong performance on cross-domain generation tasks. Compared with the acoustic tokenizer from UniFlow-Audio, our semantically rich representation is better suited for DiT learning under the same parameter settings, leading to superior generation quality. For example, when trained only on TTS, LoSATok achieves SIM and UTMOS scores of 0.548 and 3.367, respectively, exceeding UniFlow-Audio's 0.408 and 2.768. For the high-dimensional unified representation DashengTokenizer, LoSATok achieves clearly better performance despite using fewer trainable parameters (208M vs. 322M). For the 1280-dimensional representation, substantially more parameters are required to achieve performance comparable to LoSATok (208M vs. 975M), and DashengTokenizer performs poorly on the TTS task. Moreover, LoSATok achieves better or competitive single-task performance with substantially fewer trainable parameters, and is especially strong on TTS and TTA semantic alignment. Moreover, downstream DiT models achieve the lowest infer-

Table 4: Reconstruction performance of discrete and continuous audio tokenizers.

Model	Frame Rate	RTF(↓)	AudioSet		MUSDB18		SeedTTS-EN		SeedTTS-ZH	
			Mel-16k(↓)	STFT-16k(↓)	Mel-16k(↓)	STFT-16k(↓)	PESQ(↑)	STOI(↑)	PESQ(↑)	STOI(↑)
DAC	50	0.0043	0.615	1.884	0.578	1.755	3.786	0.969	3.877	0.967
SNAC	–	**0.0019**	1.156	2.700	1.172	2.600	1.817	0.872	1.879	0.867
XY-Tokenizer	12.5	0.0099	1.096	2.795	1.138	2.714	2.173	0.901	2.319	0.909
GLM-4-Voice	12.5	0.0754	–	–	–	–	1.110	0.600	1.110	0.610
X-Codec 2	50	–	1.263	3.000	1.275	2.852	2.086	0.897	2.250	0.901
EzAudio	50	0.0054	0.298	1.226	0.259	1.168	3.649	0.989	3.850	0.987
UniFlow-Audio	50	0.0167	**0.268**	**1.185**	**0.232**	**1.129**	3.833	**0.992**	4.040	**0.991**
MingTok-Audio	50	0.0050	0.500	1.949	0.472	1.792	3.976	0.983	**4.199**	0.983
DashengTokenizer	25	0.0034	0.370	1.736	0.329	1.584	**4.122**	0.987	4.164	0.988
LoSATok	25	0.0033	0.760	2.296	0.712	2.159	3.051	0.947	3.157	0.947

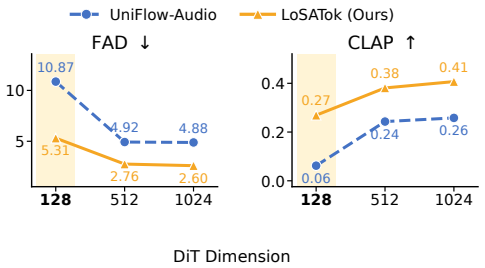


Figure 3: DiT dimensions for downstream generation.

ence cost with LoSATok. These results demonstrate the benefits of our low-dimensional unified representation for cross-domain generation.

5.4 Impact of DiT Dimensions

Dimension Setup. Prior studies show the hidden dimension of DiT also affects generation performance (Zheng et al., 2026). We therefore aim to investigate how different DiT dimensions influence acoustic and unified tokenizers. Specifically, we set the DiT dimension to 128, 512, and 1024, respectively, while keeping the number of DiT layers fixed at 12 and all other settings unchanged.

Results. From Fig. 3, we find that when the DiT dimension is 128, i.e., equal to the representation dimension, the acoustic representation has almost no generation capability. For example, its CLAP score is only 0.06, while the FAD reaches 10.87, indicating a large distribution gap between the generated and real audio and almost no semantic alignment with the text prompt. In contrast, LoSATok still maintains non-trivial generation capability at 128 dimensions, achieving performance comparable to UniFlow-Audio with DiT dimensions of 512 or 1024. This experiment provides intuitive evidence of the significant difference between the semantically rich LoSATok and acoustic tokenizers across different generation settings.

5.5 Audio Reconstruction

Baselines. For discrete NACs, we compare against DAC (Kumar et al., 2023), SNAC (Siuzdak et al., 2024), XY-Tokenizer (Gong et al., 2025), GLM-4-Voice-Tokenizer (Zeng et al., 2024), and X-Codec 2 (Ye et al., 2025). For CATs, we utilize audio tokenizers from EzAudio (Hai et al., 2025), UniFlow-Audio (Xu et al., 2025b), MingTok-Audio (Yan et al., 2025), and DashengTok (Dinkel et al., 2026).

Evaluation. We measure the Mel-spectrogram (Mel-16k) and Short-Time Fourier Transform (STFT-16k) distances on MUSDB18 (Rafii et al., 2017) and AudioSet-eval (Gemmeke et al., 2017). We use the wide-band Perceptual Evaluation of Speech Quality (PESQ) and Short-Time Objective Intelligibility (STOI) as evaluation metrics on SeedTTS (EN/ZH) (Anastassiou et al., 2024).

Results. Table 4 shows the reconstruction performance. We find that although LoSATok outperforms some NACs, e.g., SNAC and XY-Tokenizer, in reconstruction quality, it still lags behind advanced acoustic tokenizers. For a representation designed for generation and understanding, reconstruction metrics mainly reflect the degree of distortion. The reconstruction score may not be directly related to generation. For example, UniFlow-Audio achieves the best reconstruction performance, but its understanding and generation capabilities are inferior to those of LoSATok. We also conduct subjective evaluations on the generated audio in Appendix D, and the results suggest LoSATok-based generation has good perceptual quality.

5.6 Ablation Study

Components of Semantic Bottleneck. In this section, we evaluate the effects of $\mathcal{L}_{\text{recon}}$ and $\mathcal{L}_{\text{tr}}$ in SemBo, as well as the sensitivity of λ_{recon}. Fig. 4 shows the results of the ablation study. We find that removing $\mathcal{L}_{\text{tr}}$ leads to a large performance drop on the ESC and FSC understanding tasks, with scores

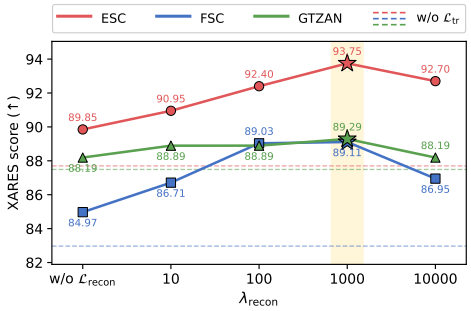


Figure 4: Ablation results of Semantic Bottleneck.

Table 5: Ablation results of the LoSATok.

Method	Reconstruction						Understanding		
	AudioSet		MUSDB		SeedTTS-EN		ESC	FSC	GTZAN
	Mel-16k	STFT-16k	Mel-16k	STFT-16k	PESQ	STOI			
w/o $\mathcal{L}_{\mathrm{H}}$	0.759	2.315	0.700	2.161	2.952	0.944	91.10	54.79	**86.99**
w/o $\mathcal{L}_{\mathrm{L}}$	0.585	2.111	0.558	1.998	3.121	0.953	47.25	6.30	53.76
w/ CM	**0.575**	**2.076**	**0.538**	**1.955**	**3.178**	**0.956**	52.45	5.11	56.26
AE	0.776	2.340	0.705	2.167	2.909	0.942	**91.25**	**59.87**	86.49

of 87.70 and 82.97, respectively. Removing $\mathcal{L}_{\mathrm{recon}}$ also degrades understanding performance, but the results are slightly better than those obtained without ("w/o") $\mathcal{L}_{\mathrm{tr}}$, which verifies the effectiveness of our proposed time-relation loss. Combining $\mathcal{L}_{\mathrm{recon}}$ and $\mathcal{L}_{\mathrm{tr}}$ yields better semantic performance. We further vary λ_{recon} from 10 to 10^4 on the understanding tasks and find that when $\lambda_{\mathrm{recon}} = 10^3$, the model achieves the best average cross-domain understanding performance, with scores of 93.75 on ESC, 89.11 on FSC, and 89.29 on GTZAN. Therefore, we set $\lambda_{\mathrm{recon}} = 10^3$ in Eq. (3).

Components of LoSATok. We explore the effectiveness of LoSATok's objectives. All models are trained with a global batch size of 16 for 100K steps. We exclude the KL divergence to remove its potential influence on the results. Thus, this section uses an autoencoder (AE). KL-divergence experiments are provided in Appendix B. Table 5 reports the results. We find that both $\mathcal{L}_{\mathrm{H}}$ and $\mathcal{L}_{\mathrm{L}}$ in LoSATok's dual-level semantic supervision affect understanding tasks. Removing $\mathcal{L}_{\mathrm{L}}$ almost eliminates the representation's understanding ability, while removing $\mathcal{L}_{\mathrm{H}}$ also leads to a certain performance drop on the FSC task. Although Section 3 shows that a training-free approach can also preserve high-dimensional understanding ability, this experiment shows that the training-free CM strategy is unsuitable as a semantic supervision signal, as it causes a substantial drop in understanding performance. These results validate the effectiveness of SemBo and LoSATok's design.

Projection-Adapter Baseline. A possible alternative to learning an explicitly supervised low-dimensional latent is to compress an existing high-dimensional tokenizer before DiT modeling. We therefore freeze DashengTokenizer and train a $1280 \rightarrow 128 \rightarrow 1280$ linear adapter for 200K steps on the same data used by LoSATok. The downstream DiT is trained on the 128-dimensional bottleneck using the setup in Section 5.3. As shown in Table 6, the adapter retains considerably more reconstruction fidelity than semantic accessibility: STOI remains 0.952, but ESC-50 decreases from 96.40 to 37.05, and GTZAN from 89.99 to 54.96. Its TTS performance is also consistently below LoSATok in three metrics, e.g., SIM 0.476 vs. 0.548. These results show that post-hoc dimensionality reduction does not preserve the organization of the semantic space; explicit low-dimensional semantic supervision is necessary.

Table 6: Comparison of LoSATok with a linear projection on DashengTokenizer.

Model	Latent Dim.	STOI	Understanding		TTS		
			ESC-50	GTZAN	WER↓	SIM↑	UTMOS↑
DashengTok	1280	0.987	96.40	89.99	100.0	0.015	1.251
+ adapter	128	**0.952**	37.05	54.96	3.456	0.476	2.623
LoSATok	128	0.947	**88.90**	**85.39**	**3.030**	**0.548**	**3.367**

6 Conclusion

In this paper, we propose LoSATok, a unified low-dimensional tokenizer for cross-domain understanding and generation. We introduce low-dimensional semantic representations based on time-relation loss and dual-level semantic supervision to enrich semantics. On understanding tasks, LoSATok matches several SSL models. On generation tasks, LoSATok outperforms acoustic and high-dimensional unified audio tokenizers.

Acknowledgment

This work is supported by National Natural Science Foundation of China (62076144) and Shenzhen Science and Technology Program (JCYJ20220818101014030).

Limitations

LoSATok aims to propose a tokenizer for semantic-acoustic modeling that can support both understanding and generation tasks. Its advantage lies in introducing a low-dimensional unified representation that is easier for DiT to learn, while still

preserving useful semantic and acoustic information for speech, music, and general audio tasks.

However, LoSATok is not intended to replace tokenizers designed for high-fidelity reconstruction. Compared with acoustic tokenizers, LoSATok sacrifices a certain degree of reconstruction fidelity in order to obtain a low-dimensional representation that has a stronger semantic structure and is more suitable for generative modeling. Nevertheless, both subjective and objective experiments demonstrate its advantages in generation tasks.

In addition, although LoSATok retains a certain level of audio understanding capability and outperforms some SSL representations, e.g., HuBERT and WavLM, its semantic ability has not yet fully reached or approached the level of high-dimensional semantic representations, e.g., MiDashengLM. In our framework, SemBo is mainly used to demonstrate that high-dimensional semantic representations can be effectively compressed, while LoSATok introduces acoustic information and KL regularization to support downstream generation tasks. This design involves trade-offs among semantics, acoustics, and generation. Further optimizing the balance among these three aspects within a low-dimensional representation is a promising direction for future research.

References

Andrea Agostinelli, Timo I Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, et al. 2023. Musiclm: Generating music from text. *arXiv preprint arXiv:2301.11325*.

Philip Anastassiou, Jiawei Chen, Jitong Chen, Yuanzhe Chen, Zhuo Chen, Ziyi Chen, Jian Cong, Lelai Deng, Chuang Ding, Lu Gao, et al. 2024. Seed-tts: A family of high-quality versatile speech generation models. *arXiv 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.

Dmitry Bogdanov, Minz Won, Philip Tovstogan, Alastair Porter, and Xavier Serra. 2019. The mtg-jamendo dataset for automatic music tagging. In *Machine learning for music discovery workshop, international conference on machine learning (ICML 2019)*, pages 1–3. Long Beach, CA, United States.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. 2022. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.

Changhao Cheng, Wei Wang, Wangyou Zhang, Dongya Jia, Jian Wu, Zhuo Chen, and Yanmin Qian. 2026. On the distillation loss functions of speech vae for unified reconstruction, understanding, and generation. *arXiv preprint arXiv:2604.12383*.

Heinrich Dinkel, Gang Li, Jizhong Liu, Jian Luan, Yadong Niu, Xingwei Sun, Tianzi Wang, Qiyang Xiao, Junbo Zhang, and Jiahao Zhou. 2025. Midashenglm: Efficient audio understanding with general audio captions. *arXiv preprint arXiv:2508.03983*.

Heinrich Dinkel, Xingwei Sun, Gang Li, Jiahao Mei, Yadong Niu, Jizhong Liu, Xiyang Li, Yifan Liao, Jiahao Zhou, Junbo Zhang, et al. 2026. Dasheng-tokenizer: One layer is enough for unified audio understanding and generation. *arXiv preprint arXiv:2602.23765*.

SeungHeon Doh, Keunwoo Choi, Jongpil Lee, and Juhan Nam. 2023. Lp-musiccaps: Llm-based pseudo music captioning. *arXiv preprint arXiv:2307.16372*.

Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. 2023. Clap learning audio concepts from natural language supervision. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. 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.

Yitian Gong, Luozhijie Jin, Ruifan Deng, Dong Zhang, Xin Zhang, Qinyuan Cheng, Zhaoye Fei, Shimin Li, and Xipeng Qiu. 2025. Xy-tokenizer: Mitigating the semantic-acoustic conflict in low-bitrate speech codecs. *arXiv preprint arXiv:2506.23325*.

Jiarui Hai, Yong Xu, Hao Zhang, Chenxing Li, Helin Wang, Mounya Elhilali, and Dong Yu. 2025. EzAudio: Enhancing Text-to-Audio Generation with Efficient Diffusion Transformer. In *Interspeech 2025*, pages 4233–4237.

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.

Yongcheng Jing, Yezhou Yang, Zunlei Feng, Jingwen Ye, Yizhou Yu, and Mingli Song. 2019. Neural style transfer: A review. *IEEE transactions on visualization and computer graphics*, 26(11):3365–3385.

Kevin Kilgour, Mauricio Zuluaga, Dominik Roblek, and Matthew Sharifi. 2019. Fréchet Audio Distance: A Reference-Free Metric for Evaluating Music Enhancement Algorithms. In *Interspeech 2019*, pages 2350–2354.

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

Rithesh Kumar, Prem Seetharaman, Alejandro Luebs, Ishaan Kumar, and Kundan Kumar. 2023. High-fidelity audio compression with improved rvqgan. *Advances in Neural Information Processing Systems*, 36:27980–27993.

Shun Lei, Yaoxun Xu, Zhiwei Lin, Huaicheng Zhang, Wei Tan, Hangting Chen, Yixuan Zhang, Chenyu Yang, Haina Zhu, Shuai Wang, Zhiyong Wu, and Dong Yu. 2025. Levo: High-quality song generation with multi-preference alignment. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Xinhao Mei, Chutong Meng, Haohe Liu, Qiuqiang Kong, Tom Ko, Chengqi Zhao, Mark D Plumbley, Yuexian Zou, and Wenwu Wang. 2024. Wavcaps: A chatgpt-assisted weakly-labelled audio captioning dataset for audio-language multimodal research. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:3339–3354.

Zhikang Niu, Shujie Hu, Jeongsoo Choi, Yushen Chen, Peining Chen, Pengcheng Zhu, Yunting Yang, Bowen Zhang, Jian Zhao, Chunhui Wang, et al. 2025. Semantic-vae: Semantic-alignment latent representation for better speech synthesis. *arXiv preprint arXiv:2509.22167*.

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. IEEE.

William Peebles and Saining Xie. 2023. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205.

Zafar Rafii, Antoine Liutkus, Fabian-Robert Stöter, Stylianos Ioannis Mimilakis, and Rachel Bittner. 2017. Musdb18 - a corpus for music separation.

Takaaki Saeki, Detai Xin, Wataru Nakata, Tomoki Koriyama, Shinnosuke Takamichi, and Hiroshi Saruwatari. 2022. UTMOS: UTokyo-SaruLab System for VoiceMOS Challenge 2022. In *Interspeech 2022*, pages 4521–4525.

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

Hubert Siuzdak, Florian Grötschla, and Luca A Lanzendörfer. 2024. SNAC: Multi-scale neural audio codec. In *Audio Imagination: NeurIPS 2024 Workshop AI-Driven Speech, Music, and Sound Generation*.

Jinchuan Tian, Sang gil Lee, Zhifeng Kong, Sreyan Ghosh, Arushi Goel, Chao-Han Huck Yang, Wenliang Dai, Zihan Liu, Hanrong Ye, Shinji Watanabe, Mohammad Shoeybi, Bryan Catanzaro, Rafael Valle, and Wei Ping. 2026. UALM: Unified audio language model for understanding, generation and reasoning. In *The Fourteenth International Conference on Learning Representations (ICLR)*.

Zeyu Xie, Chenxing Li, Qiao Jin, Xuenan Xu, Guanrou Yang, Wenfu Wang, Mengyue Wu, Dong Yu, and Yuexian Zou. 2026. Semanticvocoder: Bridging audio generation and audio understanding via semantic latents. *arXiv preprint arXiv:2602.23333*.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, et al. 2025a. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*.

Xuenan Xu, Jiahao Mei, Zihao Zheng, Ye Tao, Zeyu Xie, Yaoyun Zhang, Haohe Liu, Yuning Wu, Ming Yan, Wen Wu, et al. 2025b. Uniflow-audio: Unified flow matching for audio generation from omni-modalities. *arXiv preprint arXiv:2509.24391*.

Junichi Yamagishi, Christophe Veaux, and Kirsten MacDonald. 2019. Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit (version 0.92). *The Rainbow Passage which the speakers read out can be found in the International Dialects of English Archive:(http://web. ku. edu/˜ idea/readings/rainbow. htm)*.

Canxiang Yan, Chunxiang Jin, Dawei Huang, Haibing Yu, Han Peng, Hui Zhan, Jie Gao, Jing Peng, Jingdong Chen, Jun Zhou, et al. 2025. Ming-uniaudio: Speech llm for joint understanding, generation and editing with unified representation. *arXiv preprint arXiv:2511.05516*.

Guanrou Yang, Tian Tan, Qian Chen, Zhikang Niu, Yakun Song, Ziyang Ma, Yushen Chen, Zeyu Xie, Tianrui Wang, Yifan Yang, et al. 2026. Wavcube: Unifying speech representation for understanding and generation via semantic-acoustic joint modeling. *arXiv preprint arXiv:2605.06407*.

Zhen Ye, Xinfa Zhu, Chi-Min Chan, Xinsheng Wang, Xu Tan, Jiahe Lei, Yi Peng, Haohe Liu, Yizhu Jin,

Zheqi Dai, et al. 2025. Llasa: Scaling train-time and inference-time compute for llama-based speech synthesis. *arXiv preprint arXiv:2502.04128*.

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*.

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. *arXiv preprint arXiv:2412.02612*.

Junbo Zhang, Heinrich Dinkel, Yadong Niu, Chenyu Liu, Si Cheng, Anbei Zhao, and Jian Luan. 2025. X-ARES: A Comprehensive Framework for Assessing Audio Encoder Performance. In *Interspeech 2025*, pages 4868–4872.

Zhisheng Zhang, Xiang Li, Yixuan Zhou, Jing Peng, Shengbo Cai, Guoyang Zeng, and Zhiyong Wu. 2026. Unisrcodec: Unified and low-bitrate single codebook codec with sub-band reconstruction. *arXiv preprint arXiv:2601.02776*.

Boyang Zheng, Nanye Ma, Shengbang Tong, and Saining Xie. 2026. Diffusion transformers with representation autoencoders. In *The Fourteenth International Conference on Learning Representations (ICLR)*.

Yixuan Zhou, Guoyang Zeng, Xin Liu, Xiang Li, Renjie Yu, Ziyang Wang, Runchuan Ye, Weiyue Sun, Jiancheng Gui, Kehan Li, Zhiyong Wu, and Zhiyuan Liu. 2026. Hierarchical semantic-acoustic modeling via semi-discrete residual representations for expressive end-to-end speech synthesis. In *The Fourteenth International Conference on Learning Representations (ICLR)*.

A Discussion

In this section, we discuss some important aspects.

Ethical Considerations. To evaluate the subjective perceptual quality of the generated audio, we perform a subjective listening test in Appendix D. These experiments have received approval from the local Human Ethics Research Committee. Before the experiment, we obtain informed consent from all participants. We do not collect any personal information, and all responses are used solely for academic research and kept strictly confidential. After completing the test, each participant receives approximately $1.47 as compensation.

Reconstruction and Generation. Prior studies (Xie et al., 2026) have found that although models do not perform best on reconstruction tasks, they can achieve strong performance on downstream generation tasks. From a representation perspective, high-quality reconstruction requires representations with sufficient acoustic details, whereas generation tasks require representations that are easier for downstream models, such as DiTs, to learn. In the XARES benchmark, these tasks are typically evaluated by directly feeding the representations into the MLP or LLM. LoSATok's unified representations achieve comparable performance on understanding tasks. Compared with acoustically rich representations, semantically rich representations are easier for downstream models to learn, which explains why our model performs better in generation. Reconstruction evaluation mainly measures whether the tokenizer introduces audio distortion. The subjective experiments in Appendix D further show that audio generated based on LoSATok has good perceptual quality and outperforms acoustic tokenizers in generation tasks. Overall, LoSATok is not designed to be the best reconstruction tokenizer. It sacrifices part of reconstruction fidelity to obtain a lower-dimensional, more semantically organized latent space that is easier for text-conditioned DiT generation.

Term Explanation. In this paper, we use "unified" to refer to the integration of acoustic and semantic representations (Dinkel et al., 2026) for both generation and understanding. We use "cross-domain/general-purpose" to refer to joint modeling across the three domains of speech, music, and general audio (sound). The paper uses different terms accordingly.

Downstream Fairness. To ensure fairness in evaluating generation tasks, we follow the multi-task

Table 7: Generation, understanding, and reconstruction performance under different KL loss weights.

λ_{KL}	Generation			Understanding		Reconstruction					
	TTS			ESC↑	GTZAN↑	AudioSet		MUSDB		SeedTTS-EN	
	WER↓	SIM↑	UTMOS↑			Mel-16k↓	STFT-16k↓	Mel-16k↓	STFT-16k↓	PESQ↑	STOI↑
w/o $\mathcal{L}_{\text{KL}}$	3.338	0.463	3.170	**91.40**	86.99	**0.688**	**2.195**	**0.626**	**2.027**	**3.447**	**0.964**
10^{-4}	3.395	0.449	3.330	90.35	86.29	0.694	2.203	0.633	2.043	3.424	0.963
10^{-3}	3.158	0.491	3.284	91.10	**88.39**	0.694	2.203	0.636	2.044	3.405	0.962
10^{-2}	**3.030**	**0.548**	**3.367**	88.90	85.39	0.760	2.296	0.712	2.159	3.051	0.947

Table 8: Dimension reduction via FC layers.

Method	Dim.	ESC(↑)	FSC(↑)	GTZAN(↑)
SemBo	128	**93.75**	**89.11**	**89.29**
FC-Semantic	128	41.15	13.60	67.87
FC-Unified	128	42.55	4.75	56.96

training framework of UniFlow-Audio and replace only the VAE component, while keeping all other architectural components unchanged. We also keep the number of training GPUs, batch size, and training steps consistent, freeze the respective audio tokenizers, and train the DiT blocks and some connection layers. For understanding tasks, we feed the reparameterized representations from LoSATok into the XARES benchmark for evaluation.

Why Are Low-dimensional Semantic Representations Important? We train SemBo separately as the low-dimensional semantic signals. This is because, without an explicit low-dimensional semantic supervision target, learning semantic dimensionality reduction jointly with tokenizer reconstruction can cause the low-dimensional representation to be dominated by the acoustic reconstruction objective, making it difficult to preserve semantics.

To verify this, we construct a direct-projection baseline using DashengTokenizer. This baseline uses two trainable fully connected (FC) layers to reduce the dimensionality of the high-dimensional semantic and acoustic representations, respectively, and retains dual-level semantic supervision to obtain the low-dimensional unified representation. Specifically, we replace the low-dimensional semantic representation from SemBo with the semantic representation directly projected by the fully connected layer, so that semantic dimensionality reduction is optimized jointly with tokenizer reconstruction. The training architecture and implementation details are kept consistent with Section 5.6.

We further evaluate understanding capabilities of the resulting low-dimensional unified representation on the XARES benchmark in Table 8. Compared with SemBo, semantic representations obtained via FC-based dimensionality reduction have significantly weaker comprehension ability. Consequently, the unified representations learned from such weak semantic supervision are also much worse than those from LoSATok. This indicates that applying dimensionality reduction through an FC layer together with reconstruction training, without using directly frozen semantic signals as teacher supervision, causes the trainable parameters to focus more on acoustic details and leads to severe loss of semantic information. In contrast, independently trained SemBo can provide more reliable low-dimensional semantic supervision, making it better suited to guide the learning of low-dimensional unified representations.

B KL Divergence in LoSATok

LoSATok improves downstream generation by utilizing the KL divergence. Prior studies have shown that KL divergence can regularize the representation distribution by mapping it to a more structured latent space (Niu et al., 2025), e.g., a Gaussian distribution, which facilitates modeling by downstream DiT models and further improves generation quality. However, a strong KL constraint may over-smooth the representations (Yang et al., 2026), while a weak KL constraint may provide limited benefits for generation. Therefore, in this section, we analyze the impact of KL divergence and its weight λ_{KL} on LoSATok. All experiments follow the same training setup as LoSATok, using a global batch size of 64 and training for one million steps.

We evaluate the model on generation, understanding, and reconstruction tasks under four settings: without the KL loss and reparameterization layers, and with KL weights of 10^{-4}, 10^{-3}, and 10^{-2}. The results are shown in Table 7. When $\lambda_{\text{KL}} = 10^{-2}$, the model achieves the best performance on the TTS generation task, greatly outperforming both the setting without KL divergence and the setting with $\lambda_{\text{KL}} = 10^{-4}$. Specifically, WER decreases to 3.030%, SIM increases to 0.548, and

Table 9: Data scaling experiments on TTA.

Model	# Param	Text-to-Audio			
		FAD(↓)	FD(↓)	KL(↓)	CLAP(↑)
Large-scale Dataset, WavCaps, with ∼ 7,558 hours.					
UniFlow-Audio	208M	4.925	40.017	2.613	0.243
LoSATok	208M	**2.760**	**25.743**	**1.844**	**0.381**
Small-scale Dataset, AudioCaps, with ∼ 123 hours.					
UniFlow-Audio	208M	2.425	23.101	1.688	0.428
LoSATok	208M	**1.813**	**15.931**	**1.273**	**0.507**

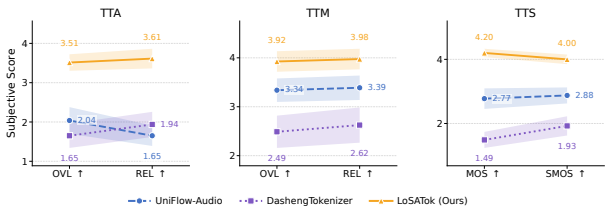


Figure 5: Subjective results on generation tasks with 95% confidence intervals.

UTMOS increases to 3.367, indicating improvements in intelligibility, speaker similarity, and perceptual audio quality. Meanwhile, performance on the ESC and GTZAN understanding tasks does not degrade substantially. Although reconstruction performance under this setting drops notably compared with the AE architecture without KL divergence, we select $\lambda_{\text{KL}} = 10^{-2}$ as the default configuration due to its benefits for generation.

We further analyze the effects of different KL weights across tasks. KL divergence affects both understanding and reconstruction performance. The AE architecture without KL divergence performs better on the ESC understanding task, but its SIM score on TTS generation is only 0.463, which is relatively low. When the KL weight is too small, e.g., $\lambda_{\text{KL}} = 10^{-4}$, overall performance decreases across understanding, generation, and reconstruction tasks. In contrast, $\lambda_{\text{KL}} = 10^{-3}$ achieves a more balanced trade-off among the three types of tasks. To enhance generation performance without substantially compromising understanding performance, we adopt $\lambda_{\text{KL}} = 10^{-2}$ as the main experimental setting. This choice increases the SIM score of the TTS task from 0.491 to 0.548 compared with $\lambda_{\text{KL}} = 10^{-3}$, and further improves audio quality and intelligibility. To support different use cases, we plan to release the pre-trained weights for both $\lambda_{\text{KL}} = 10^{-2}$ and $\lambda_{\text{KL}} = 10^{-3}$.

C Effect of Data Source and Scale

We aim to evaluate the performance of acoustic and unified tokenizers under different training data scales. In this section, we conduct experiments on the large-scale WavCaps dataset (∼7,558 hours) and the small-scale AudioCaps dataset (∼123 hours). Both models are trained for 100K steps on 4 NVIDIA H100 GPUs.

Table 9 presents the TTA results under different data scales. On the small-scale AudioCaps dataset, the acoustic tokenizer also achieves competitive performance, with an FAD of 2.425 and a CLAP score of 0.428. This indicates that the distribution of the generated audio is close to that of real audio and that the generated content follows the text prompts reasonably well. However, on the large-scale WavCaps dataset, the limitations of the acoustic tokenizer become more evident. Its FAD and CLAP scores are only 4.925 and 0.243, respectively, indicating degraded overall audio quality and text-audio alignment. In contrast, LoSATok achieves stable and superior performance across both data scales. Specifically, it obtains an FAD of 1.813 and a CLAP score of 0.507 when trained on AudioCaps, and an FAD of 2.760 and a CLAP score of 0.381 when trained on WavCaps.

This experiment highlights the gap between tokenizers that rely solely on acoustic information and those with rich acoustic and semantic representations in generation tasks. On large-scale datasets, LoSATok converges faster and achieves better generation quality than the acoustic tokenizer. On small-scale datasets, although both tokenizers perform well, LoSATok still demonstrates a stronger downstream generation advantage.

D Human Study

We test the objective metrics of the generation task in Section 5.3. In this section, we evaluate the subjective perception of the generated audio.

Participant Profile. We design a subjective evaluation questionnaire on the Credamo platform and recruit 20 participants for the experiment. All participants are over 18 years old, and their informed consent is obtained. Their average response time is 397.15 seconds. Each participant receives approximately $1.47 as compensation.

Metrics and Questionnaire Design. We design the subjective experiment following UniFlow-Audio (Xu et al., 2025b). For the TTA and TTM tasks, we evaluate the generated audio in terms of overall audio quality (OVL) and text relevance (REL). OVL measures the naturalness, clarity, and

overall listening quality of the generated audio, while REL measures its consistency with the text description. For the TTS task, we evaluate speech naturalness (MOS) and speaker similarity (SMOS). All metrics are rated on a 0–5 scale, where 0 indicates very poor quality or a complete mismatch, and 5 indicates excellent quality or a high degree of matching. Participants are instructed to wear headphones in a quiet environment and rate each sample only after listening to the entire audio.

Results and Analysis. Fig. 5 presents the results of our subjective evaluation. We report the mean scores and 95% confidence intervals (CI), shown as the shaded regions in the figure. For the TTA and TTM tasks, LoSATok outperforms UniFlow-Audio and DashengTokenizer in both OVL and REL, with large improvements on TTA. Specifically, LoSATok achieves a REL score of 3.61 ± 0.25, while UniFlow-Audio and DashengTokenizer achieve only 1.65 ± 0.26 and 1.94 ± 0.32, respectively. For the TTS task, LoSATok achieves a MOS score of 4.20 ± 0.13, outperforming the baselines. These results demonstrate that LoSATok achieves better perceptual quality and text relevance in generation tasks. Although LoSATok does not achieve the best reconstruction performance, its unified representation is better suited for downstream DiT-based generation tasks.

E Detailed Information

E.1 Details of Effective Rank and PCA Analysis

In Section 3, we analyze the effective rank and PCA spectra of the semantic and acoustic representations of DashengTokenizer to study the compressibility of high-dimensional semantic representations. In this section, we provide a detailed explanation of the relevant concepts.

Given a frozen semantic encoder E_s, it outputs a semantic latent $z_s^h = \mathrm{E}_s(x) \in \mathbb{R}^{T \times D}$, where T is the number of temporal frames and $D = 1280$ is the representation dimension. We collect frame-level representations over the dataset and compute the covariance matrix along the feature dimension.

First, we compute the effective rank of the high-dimensional semantic representations. Let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_D$ be the eigenvalues of the feature covariance matrix $\Sigma \in \mathbb{R}^{D \times D}$ computed across all frame-level representations in the dataset. We normalize them into a probability distribution:

$$p_i = \frac{\lambda_i}{\sum_{j=1}^{D} \lambda_j}. \tag{6}$$

The effective rank is then defined in Eq. (7).

$$r_{\text{eff}} = \exp\left(-\sum_{i=1}^{D} p_i \log p_i\right) \in [1, D]. \tag{7}$$

The effective rank measures the number of dimensions over which the feature energy is effectively distributed.

Second, we perform PCA analysis and compute the number of principal components required to preserve a fraction α of the total variance:

$$k_\alpha = \min\left\{k : \frac{\sum_{i=1}^{k} \lambda_i}{\sum_{j=1}^{D} \lambda_j} \geq \alpha\right\}. \tag{8}$$

In our analysis, we set $\alpha = 0.9$.

Fig. 1 shows that the effective dimensionality of the 1280-dimensional semantic representations can be substantially lower than their original dimensionality, indicating that high-dimensional semantic representations contain considerable redundancy and are therefore potentially compressible.

Table 10: Training hyperparameters for SemBo.

Parameter	Value
optimizer	AdamW
$[\beta_1, \beta_2]$	$[0.8, 0.99]$
lr	1×10^{-4}
lr_scheduler	CosineLR
warmup_steps	1,000
max_steps	100,000

E.2 Details of SemBo

The SemBo architecture consists of a two-layer MLP-based compressor and restorer, each with 0.72M parameters. For the semantic encoder, we use the audio encoder of MiDashengLM (Dinkel et al., 2025), considering its strong performance on the audio understanding leaderboard on the XARES benchmark. We train SemBo for 100K steps using the same 13K-hour training dataset as LoSATok. The training details are summarized in Table 10. We use AdamW as the optimizer, with β_1 and β_2 set to 0.8 and 0.99, respectively. The initial learning rate is 1×10^{-4}, and a cosine learning-rate scheduler is adopted with 1,000 warmup steps.

Table 11: Architecture details of LoSATok.

Module	Parameters	Architecture	Dim.
Semantic Encoder	632 M	32-layer Transformer + SemBo	128
Acoustic Encoder	0.82 M	2D Conv + Linear	128
KL divergence	0.03 M	$2 \times$ Linear	128
Decoder	171 M	Vocos	128

E.3 Details of LoSATok

LoSATok Architecture. LoSATok is designed based on DashengTokenizer. The encoder consists of a semantic encoder and an acoustic encoder, with detailed parameter information summarized in Table 11. The semantic encoder is composed of a pretrained MiDashengLM encoder followed by SemBo, and its output dimension is 128. The acoustic encoder applies a 2D convolutional layer to perform non-overlapping compression, and then uses a linear layer to project the feature dimension to 128. Two linear layers are used to compute the KL divergence based on the unified embedding z_{uni}, enabling adaptation to downstream DiT-based generation tasks. The acoustic decoder is based on Vocos (Siuzdak, 2024) and reconstructs audio details from the unified representation.

For the 25Hz tokenizer, given an input audio signal x, we extract its Mel-spectrogram $x_{\text{mel}} \in \mathbb{R}^{T,F}$, where F denotes the number of Mel bins. The acoustic encoder uses 2D convolutions to compress the frequency dimension to 1 and downsample the time dimension by a factor of 4, producing a 25Hz representation. Since the semantic representation is also at 25Hz, it can be directly added to the generated representation to obtain a unified representation. In the decoder, it is first upsampled to 50 Hz, then passed through Vocos blocks, followed by an Inverse Short-Time Fourier Transform (ISTFT) head to reconstruct the audio.

Details of LoSATok's Objectives. We describe LoSATok's training objectives in Eq. (5).

- $\mathcal{L}_{\text{mel}}$: The multi-scale Mel-spectrogram L_1 loss. We compare the reconstructed waveform $\hat{x}$ with the original waveform x using multiple window sizes $[32, 64, \ldots, 2048]$ and different numbers of Mel bins $[5, 10, \ldots, 320]$, thereby ensuring the perceptual quality of the reconstructed audio.
- $\mathcal{L}_{\text{sem}} = \mathcal{L}_{\text{H}} + \mathcal{L}_{\text{L}}$: The semantic alignment loss. We perform ℓ_2 alignment between acoustic representations and their corresponding high- and low-dimensional semantic targets, thereby improving the understanding capability of the general-purpose representation.
- $\mathcal{L}_{\text{KL}}$: The KL divergence loss. At the low-dimensional bottleneck, we compute the closed-form KL divergence between the posterior distribution $q(z \mid x) = \mathcal{N}(\mu, \sigma^2 I)$ and the prior distribution $\mathcal{N}(0, I)$:

$$\mathcal{L}_{\text{KL}} = -\frac{1}{2} \sum_{d} \left(1 + \log \sigma_d^2 - \mu_d^2 - \sigma_d^2\right), \quad (9)$$

  where μ denotes the mean and σ^2 denotes the variance. The unified representation is then reparameterized and fed into the decoder.
- $\mathcal{L}_{\text{fm}}$: The feature matching loss based on a multi-frequency discriminator. Given real and generated audio as inputs, we apply an L_1 matching loss to each intermediate feature, which helps stabilize adversarial training.
- $\mathcal{L}_{\text{adv}}$: The hinge-style adversarial loss, which improves the generation quality through the discriminator.

LoSATok Training Details. LoSATok uses the same training dataset as SemBo in Table 12. Compared to SemBo in Table 10, the number of training steps is increased to one million, and the minimum learning rate is 1×10^{-5} for sufficient training.

E.4 Details of Training Datasets

To train SemBo and LoSATok, we select approximately 13K hours of training data, as shown in Table 12. To achieve general-purpose capabilities, we keep the proportions of different domains roughly balanced. In addition, to evaluate the generation capability of LoSATok on downstream tasks, we follow previous studies (Xu et al., 2025b; Dinkel et al., 2026) and select multiple datasets. For the LP-MusicCaps dataset, because MusicCaps is used as the test set, we only use the MagnaTagATune (MTT) subset for training. The joint multi-task training in Section 5.3 uses LibriTTS, LP-MusicCaps, and WavCaps as large-scale training datasets. AudioCaps is used for comparison in the data scaling experiments.

E.5 Details of Downstream Generation

Training-Curve Visualization. Figure 6 shows the comparison curves of LoSATok, UniFlow-Audio, and DashengTokenizer in terms of convergence speed and performance during multi-task joint training, with evaluation conducted every 10K steps. The experimental results show that LoSATok converges faster. In terms of the CLAP

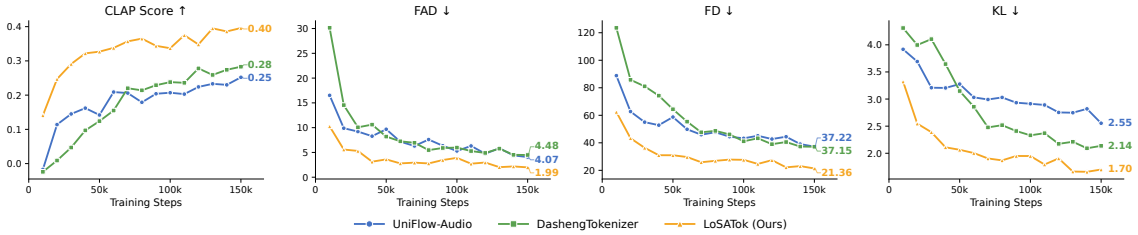


Figure 6: Text-to-audio training process of different audio tokenizers.

Table 12: Datasets for LoSATok training and generation.

Domain	Dataset	Subset	Hours (h)
Dataset for LoSATok			
Speech	VCTK	train	82
	Common Voice	train (EN)	3,507
	LibriSpeech	train	961
Music	MTG–Jamendo	train	3,763
	MUSDB	train	3
General Audio	AudioSet	bal & unbal	4,836
Dataset for Generation			
Speech	LibriTTS	train	523
Music	LP-MusicCaps	MTT–train	125
General Audio	WavCaps	train	7,558
	AudioCaps	train	123

metric, LoSATok reaches performance comparable to that achieved by UniFlow-Audio and DashengTokenizer at 150K training steps after only 21K and 28K steps, respectively, saving approximately 120K training steps. This result verifies the effectiveness of LoSATok in improving training convergence efficiency.

Evaluation Metrics. For different generation tasks, we follow (Xu et al., 2025b) and adopt the following evaluation strategies:

- TTA & TTM: For the FD metric, we compute the distributional distance between generated and reference audio based on features extracted by PANNs CNN14. The CLAP metric is used to measure the semantic alignment between the input prompt text and the generated audio, and we conduct the evaluation using the pretrained weights [1] released by LAION-AI.
- TTS: For the WER metric, we adopt a pretrained model [2] from the NeMo toolkit to evaluate the intelligibility of generated speech. For the SIM metric, we use pretrained weights based on WavLM-large fine-tuned on the speaker verification task [3] to measure the speaker similarity between synthesized and real speech. For the UTMOS metric, we use it as an objective measure of the perceptual quality of synthesized audio.

E.6 Real Time Factor

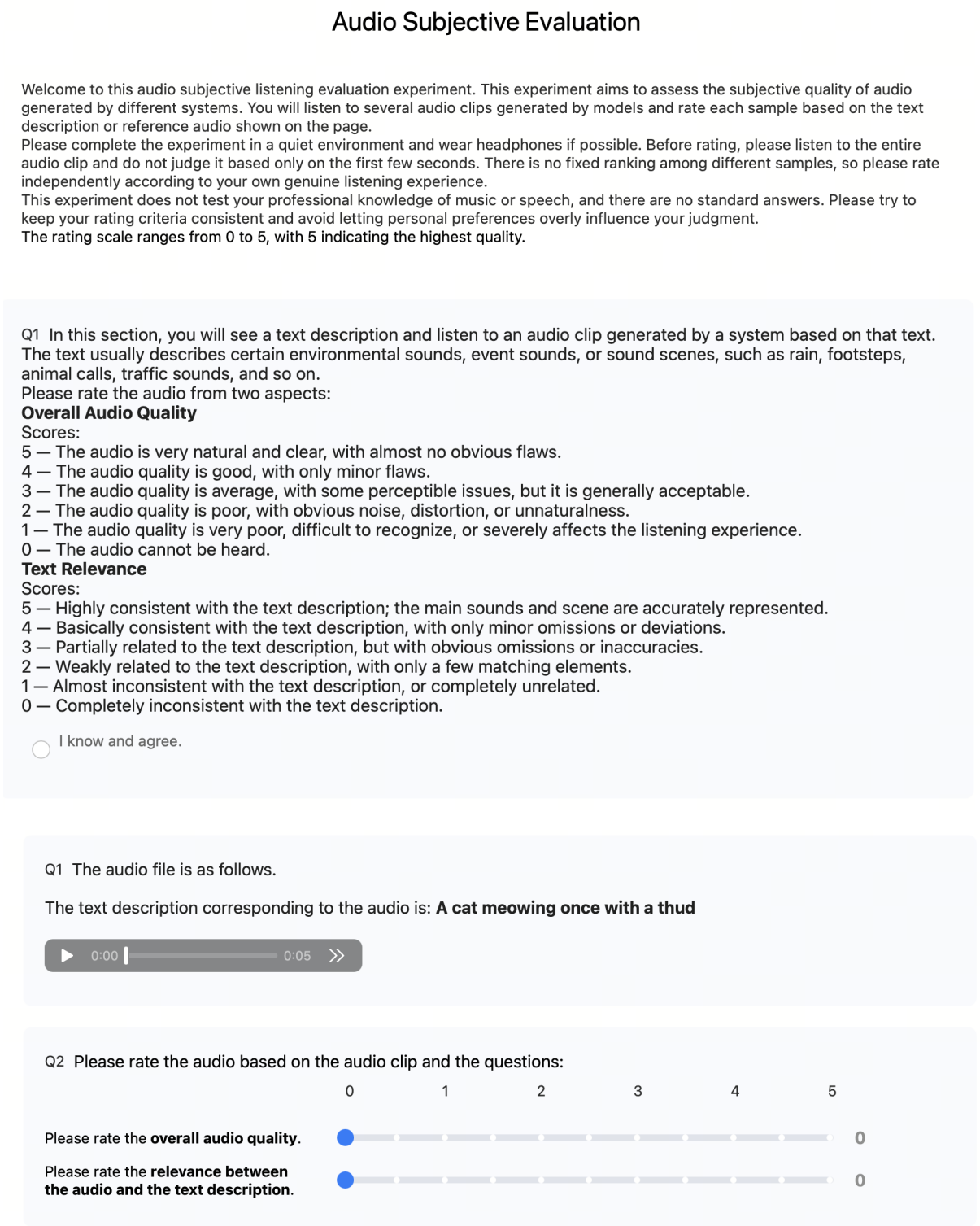
In Table 4, we compute the real-time factor (RTF) to evaluate the efficiency of audio tokenizers. RTF is defined as the total time required by the tokenizer to reconstruct an audio sample divided by the duration of the audio. For each model, we conduct the evaluation on a single NVIDIA H100 GPU, with no additional GPU processes running during testing. We measure the RTF on the four datasets listed in Table 4 and report the average value.

The results show that LoSATok achieves an RTF of only 0.0033, outperforming several CATs, e.g., UniFlow-Audio, and is therefore able to meet practical real-time requirements.

E.7 Full-Text of Human Study

In Figure 7, we provide screenshots of the instructions of the subjective experiment questionnaire and the TTA question settings.

[1] https://github.com/LAION-AI/CLAP
[2] https://huggingface.co/nvidia/stt_en_conformer_transducer_xlarge
[3] https://github.com/BytedanceSpeech/seed-tts-eval



Audio Subjective Evaluation

Welcome to this audio subjective listening evaluation experiment. This experiment aims to assess the subjective quality of audio generated by different systems. You will listen to several audio clips generated by models and rate each sample based on the text description or reference audio shown on the page.
Please complete the experiment in a quiet environment and wear headphones if possible. Before rating, please listen to the entire audio clip and do not judge it based only on the first few seconds. There is no fixed ranking among different samples, so please rate independently according to your own genuine listening experience.
This experiment does not test your professional knowledge of music or speech, and there are no standard answers. Please try to keep your rating criteria consistent and avoid letting personal preferences overly influence your judgment.
The rating scale ranges from 0 to 5, with 5 indicating the highest quality.

Q1 In this section, you will see a text description and listen to an audio clip generated by a system based on that text. The text usually describes certain environmental sounds, event sounds, or sound scenes, such as rain, footsteps, animal calls, traffic sounds, and so on.
Please rate the audio from two aspects:
Overall Audio Quality
Scores:
5 — The audio is very natural and clear, with almost no obvious flaws.
4 — The audio quality is good, with only minor flaws.
3 — The audio quality is average, with some perceptible issues, but it is generally acceptable.
2 — The audio quality is poor, with obvious noise, distortion, or unnaturalness.
1 — The audio quality is very poor, difficult to recognize, or severely affects the listening experience.
0 — The audio cannot be heard.
Text Relevance
Scores:
5 — Highly consistent with the text description; the main sounds and scene are accurately represented.
4 — Basically consistent with the text description, with only minor omissions or deviations.
3 — Partially related to the text description, but with obvious omissions or inaccuracies.
2 — Weakly related to the text description, with only a few matching elements.
1 — Almost inconsistent with the text description, or completely unrelated.
0 — Completely inconsistent with the text description.

I know and agree.

Q1 The audio file is as follows.

The text description corresponding to the audio is: **A cat meowing once with a thud**

0:00 0:05

Q2 Please rate the audio based on the audio clip and the questions:

0 1 2 3 4 5

Please rate the **overall audio quality**. 0

Please rate the **relevance between the audio and the text description**. 0

Figure 7: The screenshots of the subjective questionnaire.