EMPOWERING UNIFIED AUDIO LANGUAGE MODELING WITH HIERARCHICAL SEMANTIC REPRESENTATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

We present SL-Audio, an audio-language modeling framework that supports audio captioning, text-to-audio generation, and text-guided audio separation through hierarchical semantic representations. Unlike existing approaches that typically address these capabilities independently or rely on a single semantic granularity, SL-Audio organizes audio into multiple semantic layers, ranging from high-level semantic concepts to progressively finer acoustic representations. This hierarchical design enables different tasks to flexibly access representations at appropriate levels of abstraction, where higher-level representations provide semantic guidance and finer representations capture progressively richer acoustic details. Building on this hierarchy, SL-Audio enables a single LLM backbone to support audio captioning, text-to-audio generation, and text-guided audio separation within a unified framework, allowing different tasks to leverage different levels of the hierarchy according to their representation requirements. Experiments on these representative tasks demonstrate that the proposed hierarchical representations effectively support diverse audio-language tasks within a unified framework, providing an initial step toward unified audio-language modeling.

1 INTRODUCTION

Recent advancements in multimodal large language models (MLLMs) have made significant progress in integrating understanding and generation within a unified framework in visual domain (Ge et al., 2024; Pan et al., 2025; Chen et al., 2025b). However, extending this paradigm to audio domain remains significantly under-explored. While some pioneering works have explored unifying audio understanding and generation, they mainly focus on speech-centric scenarios rather than general audio (Xu et al., 2025; AI et al., 2025). In these speech-oriented audio language modeling frameworks, speech is typically decomposed into linguistic and acoustic tokens. The linguistic component, which carries most of the semantic information in speech, participates in the reasoning and generation process of LLMs, while fine-grained acoustic details are progressively restored in subsequent stages (Zhang et al., 2023; Borsos et al., 2023a). General audio, however, differs fundamentally from speech. It encompasses diverse sound events and acoustic patterns without explicit linguistic structure, making such linguistic-acoustic decomposition far less straightforward.

A central challenge therefore lies in how general audio should be represented for audio language modeling. Different audio-language tasks impose inherently different requirements on audio representations. High-level semantic abstraction is critical for audio understanding and open-ended generation, while precise acoustic details are essential for audio editing and manipulation tasks. Despite recent progress in hierarchical representations for general audio (e.g., SemantiCodec (Liu et al., 2024a)), existing approaches mainly focus on improving the tokenizer or representation itself, while paying less attention to how hierarchical audio representations should interact with large language models for unified reasoning and generation. Moreover, simply decomposing general audio into semantic and acoustic components is often insufficient to capture the multi-level structure required by diverse general audio tasks.

To address these challenges, we propose **SL-Audio**, a unified audio large language modeling framework built upon hierarchical semantic audio representations and a coarse-to-fine prediction paradigm. Instead of treating audio as a monolithic sequence, **SL-Audio** explicitly represents audio at multiple **S**emantic **L**ayers, ranging from high-level meaning and events to progressively finer acoustic real-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

izations. This hierarchical design enables task-dependent modeling across diverse audio-language scenarios. High-level representations support semantic audio understanding and open-ended generation, while finer-grained layers enable precise control over acoustic details for tasks such as audio editing and manipulation.

By exposing a shared semantic interface across layers, SL-Audio enables a single LLM backbone to support audio understanding, generation, and editing through next-layer prediction. We adopt audio captioning as a representative understanding task, text-to-audio generation as a generation task, and text-queried audio separation as a representative editing task. Together, these tasks demonstrate how the proposed hierarchy bridges semantic abstraction and acoustic fidelity while accommodating task-specific granularity requirements.

Our main contributions are summarized as follows:

- We propose a multi-layer audio tokenizer that produces coarse-to-fine representations, spanning global semantic content to fine-grained acoustic details, enabling both semantic understanding and high-fidelity reconstruction.
- We introduce SL-Audio, an LLM-based audio framework that leverages hierarchical representations and next-layer prediction to unify audio understanding, text-to-audio generation, and text-queried audio separation within a single framework. Extensive experiments demonstrate that SL-Audio achieves competitive or superior results compared to specialized expert models, marking a meaningful step toward unified and controllable audio-language modeling.

2 Related Work

Audio Large Language Models The success of large language models has stimulated extensive research on extending their capabilities beyond text to speech and audio, giving rise to audio large language models (Audio LLMs) that aim to process and reason over auditory information. Most existing Audio LLMs for general audio primarily focus on audio understanding tasks, such as audio captioning, audio question answering, and classification. Representative systems, including SALMONN (Tang et al., 2024), Qwen-Audio (Chu et al., 2023; 2024), and AudioFlamingo (Kong et al., 2024; Ghosh et al., 2025; Goel et al., 2025), typically rely on pretrained audio encoders (e.g., Whisper (Radford et al., 2023) or CLAP (Wu et al., 2023)) to extract semantic features aligned with text, followed by language modeling over these representations. Beyond understanding, only a limited number of works explore audio generation within LLM-based frameworks. UniAudio (Yang et al., 2024) takes an important step toward unifying multiple audio tasks within a single framework, encompassing generation, understanding, and transformation. However, its generation capability remains largely constrained by discrete codec-based representations and task-specific designs. Overall, compared to rapid progress in audio understanding, LLM-based audio generation and precise editing remain under-explored, especially in terms of controllability, fidelity, and multi-task flexibility.

Audio Generation Audio generation has generally been dominated by specialized expert models rather than general-purpose LLMs. Diffusion-based approaches, such as AudioLDM (Liu et al., 2023), AudioLDM2 (Liu et al., 2024b), Make-an-Audio 2 (Huang et al., 2023) and Stable Audio (Evans et al., 2025), achieve strong generation quality by modeling audio in continuous latent spaces conditioned on text embeddings. In parallel, autoregressive models over discrete audio tokens, including AudioLM (Borsos et al., 2023a) and SoundStorm (Borsos et al., 2023b), demonstrate impressive fidelity for speech and sound continuation by leveraging neural audio codecs. Despite their success, these models are typically designed for specific generation tasks and rely on carefully engineered architectures. Using LLMs as a general backbone for audio generation, capable of flexible conditioning, reasoning, and integration with language, remains a promising yet under-explored direction. Our work addresses this gap by leveraging hierarchical semantic representations that allow a single LLM to perform both planning and acoustic refinement within a unified framework.

Text-queried Audio Separation Text-queried audio separation aims to extract target sounds from a mixture given free-form textal descriptions, extending separation beyond fixed source categories. Most existing approaches condition discriminative separators on pretrained audio-text alignment models (e.g., CLAP (Wu et al., 2023)), enabling strong zero-shot generalization (**?**Ma et al., 2024), with some methods further supporting multimodal queries (Cheng et al., 2025). In parallel, generative

approaches reformulate separation as conditional latent-space generation using diffusion or rectified-flow models to improve robustness and perceptual quality (Scheibler et al., 2023; Yuan et al., 2024). Despite this progress, existing methods often struggle to balance semantic correctness with signal-level fidelity, particularly for fine-grained acoustic details. This limitation highlights the need for representations that simultaneously expose semantic structure and retain low-level acoustic information. Recent work shows that decoupling high-level semantic alignment from low-level reconstruction can improve both separation accuracy and perceptual quality (Yin et al., 2025), consistent with the design principles adopted in our layered representations.

Audio Representations in LLMs Audio representations play a critical role in bridging audio modality and text-based LLMs. A common line of work relies on discrete acoustic codecs that map continuous audio into sequences of tokens, allowing LLM architectures to operate on audio via token-based modeling (Chen et al., 2025a; Zhang et al., 2023; Ji et al., 2024; Liu et al., 2024a; Yang et al., 2025). Across both speech LLMs and general audio LLMs, existing tokenizers follow two dominant design trends: (1) **Compactness**, which compresses audio into as few tokens as possible to alleviate long-sequence modeling while preserving reconstruction quality; (2) **Semantic efficiency**, which exposes high-level semantic information that is easier for LLMs to model. In speech-oriented systems, semantic-efficient representations are often organized into linguistic/acoustic hierarchies, where linguistic tokens participate in language-level reasoning while acoustic tokens restore fine-grained details (Chen et al., 2025a; Zhang et al., 2023; Borsos et al., 2023a). However, such decomposition is less explored for general audio, which lacks explicit linguistic structure and contains significantly more diverse sound events and acoustic patterns. SemantiCodec(Liu et al., 2024a) extend semantic-acoustic representation learning to general audio, but mainly focus on improving the tokenizer or representation itself, with limited exploration of how hierarchical general audio representations should interact with large language models. In this work, we aim to design an audio tokenizer tailored for LLM-based modeling across multiple tasks. Rather than relying on discrete codecs, we adopt continuous representations to reduce information loss. More importantly, our tokenizer produces multi-layer representations spanning global semantic concepts and fine-grained acoustic structures, enabling a single LLM to support audio understanding, generation, and manipulation within a unified framework.

3 Proposed Scheme

3.1 SL-Audio-Tokenizer

In this work, we design a hierarchical audio tokenizer that produces a sequence of layered audio representations, thus facilitating the distinct requirements of different audio tasks. Specifically, for tasks like text-to-audio generation or audio-to-text understanding, high-level semantic representations are crucial for achieving effective text-audio semantic alignment. Meanwhile, for tasks like text-queried audio separation not only require semantic alignment but also demand the preservation of low-level acoustic details of the original audio, necessitating fine-grained acoustic representations. Motivated by these considerations, we propose a novel hierarchical tokenizer that delivers multiple layers of audio representations within a single unified framework, see Fig. 1. The layered tokens produced by our SL-Audio-Tokenizer captures both global semantic information and fine-grained acoustic details, enabling high-fidelity audio reconstruction and achieving strong performance across diverse audio tasks.

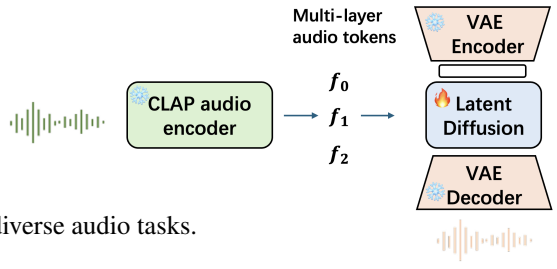


Figure 1: SL-Audio-Tokenizer

3.1.1 Feature Encoding Stage

Our unified hierarchial paradigm leverages a semantic audio encoder to extract three layers of audio representations, forming a versatile feature foundation for diverse audio tasks.

We employ a pre-trained CLAP (Wu et al., 2023) audio encoder as the backbone of our semantic encoder. The CLAP encoder is a Transformer-based architecture consisting of four groups of Swin Transformer blocks, each operating on 2D time-frequency audio patches. Within each group,

windowed attention models local relationships across contiguous time frames and frequency bins. As the network goes deeper, patch-merge layers progressively aggregate neighboring windows, expanding the attention receptive field and naturally forms a hierarchical representation from local time-frequency cues to global structures (see (Wu et al., 2023) for details).Within this hierarchical framework, we extract three layered audio representations from different depths of the CLAP encoder. Specifically, we leverage the second and fourth Transformer groups of the CLAP encoder to construct a three-layer audio token hierarchy with progressively increasing resolution.

The **coarsest layer**, denoted as f_0, is obtained from the global pooled audio embedding of the fourth Transformer group. This layer produces a single global token that captures high-level semantic information of the entire audio clip. The **middle layer** f_1 is also derived from the output of the fourth Transformer group, but instead of global pooling, we apply a 2×2 time-frequency pooling to obtain a compact set of 16 audio tokens, preserving coarse acoustic structure while retaining limited temporal-spectral variation beyond global semantics. The **finest layer** f_2 is extracted from the output feature map of the second Transformer group, which retains substantially higher time-frequency resolution, yielding 256 audio tokens, capturing detailed local acoustic information.

The three feature maps are flattened into 1D sequences f_0, f_1, and f_2, containing 1, 16, and 256 audio tokens, respectively. Together, they form a compact multi-scale token hierarchy ranging from global semantic concepts (f_0), through intermediate acoustic structures (f_1), to fine-grained time-frequency details (f_2).

3.1.2 Feature Decoding Stage

The audio de-tokenizer reconstructs audio in a coarse-to-fine manner by consuming the three layers of audio tokens. When only the coarsest layer is available, the model produces audio that is semantically aligned with the original signal but lacks detailed acoustic structure. As progressively finer-layer tokens become accessible, the model recovers increasingly acoustic details, achieving high-fidelity reconstruction.

We employ a diffusion-based model as the de-tokenizer. Specifically, the three tokens f_0,f_1 and f_2 are first projected to a shared channel dimension via linear projections, and then concatenated along the sequence dimension to serve as the conditioning input to the diffusion model. By modulating the cross-attention mask, the de-tokenizer can selectively access different subsets of the layered tokens, thereby controlling the decoding granularity. The decoding at granularity level i could be formulated as:

$$\hat{x}_i = D_\theta(z \mid \mathbf{M}_i), \quad \mathbf{M}_i : \{f_0, f_1, f_2\}_{\leq i} \; i \in \{0, 1, 2\}, \tag{1}$$

where z is the random noise for the diffusion process and $\mathbf{M}_i$ controls feature visibility through cross-attention mask such that only $\{f_0, f_1, f_2\}_{\leq i}$ could be attended to in the cross-attention module.

We employ the pre-trained DiT (Evans et al., 2025) as our de-tokenizer and fine-tune it to decode audios conditioned on the layered tokens. During training, the audio tokenizer is frozen and only de-tokenizer is fine-tuned. We sum up the loss across all three layers, enabling joint optimization of coarse-to-fine decoding within a single model. This design allows us to keep both the tokenizer and de-tokenizer structures unchanged while supporting multi-granularity decoding within one unified model. During inference, audio can be decoded at any desired granularity simply by selecting the corresponding cross-attention mask. Training details of the tokenizer are provided in the Appendix A.3.

3.2 SL-Audio

Given the proposed audio tokenizer, we build SL-Audio, a unified audio large language model that supports audio understanding, text-to-audio generation, and text-queried audio separation within a single framework.

3.2.1 Progressive Multi-Layer Audio Modeling in LLM

Next-Layer Prediction Given the multi-layer audio tokens (f_0, f_1, f_2), which encode audio information from high-level semantics to fine-grained acoustic details, SL-Audio adopts a layer-wise coarse-to-fine progressive modeling strategy. The key idea is to organize audio token generation as a

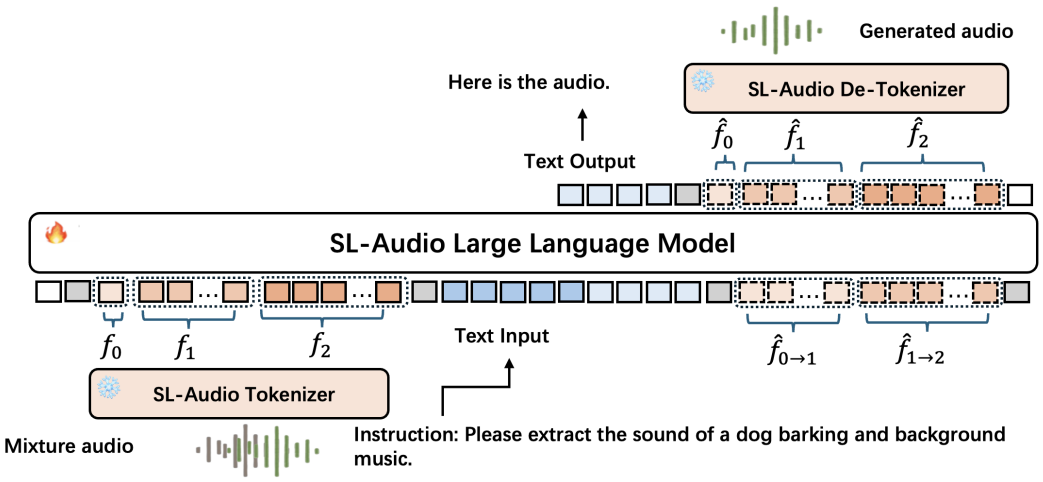


Figure 2: Overview of the SL-Audio architecture: a unified multi-layer autoregressive model for joint text-audio processing across multiple feature levels (f_0, f_1, f_2).

layer-wise prediction process, where each finer layer is generated conditioned on all previous coarser ones, rather than predicting all audio tokens one by one. This hierarchical generation process aligns with human auditory perception, where semantic understanding precedes structural reasoning and detailed acoustic perception, enabling early global planning and more stable fine-grained generation.

Formally, the layer-wise audio generation process is defined as

$$p(f_0, f_1, f_2 \mid \mathcal{C}) = \prod_{k=0}^{2} p\big(f_k \mid f_{<k}, \mathcal{C}\big), \tag{2}$$

where $\mathcal{C}$ denotes a task-dependent multimodal context, which may include textual instructions and optional audio inputs depending on the task setting. During the k-th autoregressive step, all tokens in f_k will be generated in parallel, conditioned on previous coarser layers $f_{<k}$ and context $\mathcal{C}$. This is realized by a hierarchical causal attention mechanism, which will be discussed later.

Hierarchical Causal Attention Mechanism To implement the proposed next-layer prediction and flexibly model interactions between text tokens and layered audio tokens under different task settings, SL-Audio adopts a hierarchial causal attention mechanism. Concretely, text tokens are always modeled with standard causal attention to preserve autoregressive language modeling. For audio tokens, tokens within the same layer attend to each other through full self-attention, whereas cross-layer interactions are restricted by a causal mask, allowing each layer to attend only to text tokens and coarser audio layers. This design explicitly enforces a coarse-to-fine generation dependency across audio layers, while allowing all tokens within the same layer to be generated in parallel.

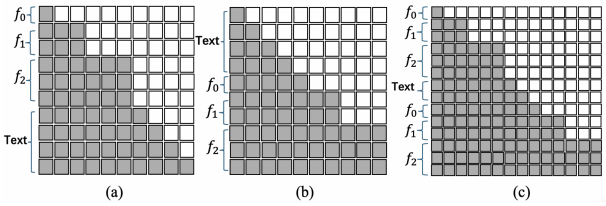


Figure 3: Hierarchical causal attention mechanism. (a) Audio-to-text understanding (b) Text-to-audio generation (c) Text-queried audio separation.

Given a formatted multimodal sequence, this attention mechanism naturally adapts to different audio-text tasks, see Fig.3 for details. In Text-to-audio understanding, text tokens can attend to all audio layers, allowing the model to aggregate multi-granular audio representations for semantic understanding (Fig.3(a)). In text-to-audio generation, audio layers are generated progressively, where each layer attends to all preceding text tokens and previously generated coarser audio layers (Fig.3(b)). In text-queried audio separation, generated audio representations attend to both instruction text and mixture audio layers provided earlier in the sequence (Fig.3(c)), allowing the model to generate edited audio conditioned on multi-layer acoustic context and textual instructions. Each audio layer is associated with a pair of dedicated start and end tokens, which explicitly mark the boundaries of

layer-specific audio representations.The concrete sequence construction under different task settings is detailed in Appendix A.1.

Input and Output Feature Spaces When audio tokens are provided as input, layered audio representations (f_0, f_1, f_2) are mapped into the LLM embedding space via layer-specific linear projections, allowing them to be processed jointly with text tokens in a unified sequence. During audio generation, SL-Audio reconstructs layered audio representations in a coarse-to-fine manner. For each layer, representations from the previous layer are up-sampled and projected to initialize the conditioning inputs for the current layer corresponding to $\hat{f}_{0\to1}, \hat{f}_{1\to2}$ in Fig. 2. The resulting hidden states are then projected back to the corresponding audio feature spaces to recover layered audio representations. This design provides a structured prior that facilitates stable next-layer prediction.

Training Objectives SL-Audio is jointly optimized with text token prediction and audio feature regression objectives. The overall training objective is formulated as

$$\mathcal{L} = -\sum_{i} \log P(t_i \mid t_{<i}, \mathcal{C}) + \sum_{i=0}^{K} \alpha_i \sum_{j=1}^{l_i} \left\| f_{i,j} - \hat{f}_{i,j} \right\|_1, \tag{3}$$

where the first term denotes the cross-entropy loss for text token prediction, and the second term corresponds to the regression loss for audio feature prediction. Here, $f_{i,j}$ and $\hat{f}_{i,j}$ are the ground-truth and predicted audio features at the j-th token of layer i, respectively, and l_i is the number of audio tokens at layer i. The weighting coefficient α_i controls the contribution of each layer during training.

3.2.2 Multimodal Pre-training and Instruct Tuning

We adopt a two-stage training pipeline for SL-Audio, consisting of large-scale multimodal pre-training followed by supervised instruction fine-tuning.

Multimodal Pre-training We first conduct large-scale multimodal pre-training to equip SL-Audio with unified audio-text understanding and coarse-to-fine generation capabilities. SL-Audio is pre-trained with audio-to-text understanding and text-to-audio generation tasks. We construct the pre-training dataset by combining multiple audio-text datasets including AudioCaps (Kim et al., 2019), WavCaps (Mei et al., 2024), AudioSetCaps (Bai et al., 2025), VGGSound (Bai et al., 2025) and YouTube-4M (Bai et al., 2025). These datasets jointly provide approximately 6.2 million audio-text pairs, corresponding to over 17,000 hours of audio. The details of datasets can be found in Appendix A.2. This stage encourages the model to learn (i) semantic alignment between audio and language (ii) the correspondence between multi-granularity layered audio representations and linguistic concepts (iii) the coarse-to-fine dependency required by next-layer audio prediction. Through large-scale pre-training, SL-Audio acquires general-purpose audio understanding and generation abilities that serve as a foundation for downstream instruction-following tasks. To enhance the model's bidirectional audio-language modeling capability, we adopt a dynamic training strategy in which, within each training batch, individual audio-text pairs are randomly assigned to either an audio-to-text understanding task or a text-to-audio generation task.

Multimodal Instruct Tuning Following multimodal pre-training, we conduct a second-stage multimodal instruction tuning to adapt SL-Audio to various audio-language tasks, including audio-to-text caption (SL-Audio-cap), text-to-audio generation (SL-Audio-gen), and text-queried audio separation (SL-Audio-sep).

For text-queried audio separation, we construct approximately 1.4 million triplets (mixture audio, target audio, text instruction) from AudioCaps, ESC-50 (Piczak, 2015), Clotho v2 (Drossos et al., 2020), FSD50K (Fonseca et al., 2021) and WavCaps, containing approximately 4000 hours of audios. The text instructions include operations such as sound extraction and removal, similar to prior work (Yin et al., 2025). The model is trained to generate multi-layer target audio tokens conditioned on both instruction text and mixture audio tokens, enabling precise audio manipulation that follows the textual instruction, while preserving acoustic fidelity via next-layer prediction.

In addition to separation, we optionally fine-tune SL-Audio on audio captioning and text-to-audio generation using the same large-scale audio-text datasets as in the pre-training stage, improving task-specific performance.

Table 1: Tokenizer performance comparison on the AudioCaps test set. Bold numbers denote the **best** result, and underlining denotes the second-best.

Model	Token/10s	Reconstruction Quality				Semantic Score	
		MEL ↓	STFT ↓	ViSQOL ↑	LPAPS ↓	CLAP-A ↑	AFSim ↑
DAC (3 RVQ, 1.5 kbps)	1500	5.54	4.44	2.36	3.86	0.42	0.61
SemanticCodec (2 VQ, 1.4 kbps)	1000	4.50	4.04	3.28	2.95	0.55	0.80
Encodec (1 RVQ, 0.75 kbps)	750	6.08	4.42	2.23	4.14	0.37	0.58
WavTokenizer	400	5.49	4.37	2.44	4.09	0.59	0.70
AudioLDM2-Tokenizer	8 + Text	7.05	4.85	2.12	4.02	0.47	0.78
SL-Audio-Tokenizer (Ours)							
SL-Audio-Tokenizer-f_0	1	10.80	6.43	1.69	4.63	0.72	0.79
SL-Audio-Tokenizer-f_0, f_1	17	7.40	4.52	1.96	3.79	**0.73**	0.79
SL-Audio-Tokenizer-f_0, f_1, f_2	273	**4.34**	**3.44**	**3.40**	**2.47**	**0.73**	**0.81**

4 EXPERIMENTS

We present our experimental results as follows. Section 4.1 describes the experimental setup. Section 4.2 evaluates the multi-granularity decoding performance of the proposed SL-Audio-Tokenizer, in comparison with previous tokenizers and audio codecs. Sections 4.3-4.5 demonstrate the versatility of SL-Audio across a range of audio-language tasks, including text-to-audio generation (Section 4.3), audio-to-text captioning (Section 4.4), and text-queried audio separation (Section 4.5).

4.1 SETUP

SL-Audio is initialized from Llama2-chat-13B (Touvron et al., 2023). We adapt LoRA (Hu et al., 2022) for parameter-efficient optimization in both the pre-training and instruction-tuning stage. Specifically, LoRA is applied to the LLM, and jointly optimized with lightweight audio-related modules including the layer-specific input and output projection layers. SL-Audio-Tokenizer is trained on the same dataset used for the text-queried audio separation task and remains frozen throughout the training process of SL-Audio. Detailed descriptions of the experimental setup are provided in Appendix A.4.

4.2 PERFORMANCE OF THE SL-AUDIO-TOKENIZER

Table 1 compares the proposed SL-Audio-Tokenizer with representative neural audio tokenizers on AudioCaps test set, including neural audio codecs (Encodec (Défossez et al., 2022) and DAC (Kumar et al., 2023)), semantic-oriented codecs (SemanticCodec (Liu et al., 2024a) and WavTokenizer (Ji et al., 2024)), and tokenizers used in text-to-audio generation models (AudioLDM2-Tokenizer (Liu et al., 2024b)). We evaluate reconstruction quality using signal-level fidelity metrics (MEL and STFT distance) and perceptual metrics (ViSQOL (Hines et al., 2015) and LPAPS (Manor & Michaeli, 2024)), and assess semantic preservation using CLAP-A and AFSim (Yin et al., 2025), which measure the similarity between original and reconstructed audio in the CLAP and Audio Flamingo embedding spaces. For efficient tokenizer evaluation, we sample 300 clips from the AudioCaps test set.

We can see from Table 1 that acoustic codecs, such as DAC and EnCodec, primarily focus on waveform reconstruction and typically rely on long token sequences to achieve high-fidelity decoding. They often struggles to preserve high-level semantics required for audio-language modeling, leading to low semantic scores. SemantiCodec improves semantic representation efficiency by integrating semantic encoder AudioMAE (Huang et al., 2022) and leveraging diffusion-based decoding. It compresses audio into fewer tokens while maintaining strong reconstruction quality and improved semantic scores. Nevertheless, its token length still remains relatively high, which limits scalability for autoregressive audio language models. AudioLDM2-Tokenizer compresses audio into a highly abstract semantic latent space by pooling AudioMAE features into only 8 continuous tokens, together with T5 textual embeddings. This extreme compression significantly simplifies the learning problem for generative models and leads to strong semantic alignment, but sacrifices fine-grained acoustic details required for faithful waveform reconstruction.

SL-Audio-Tokenizer bridges the gap between semantic abstraction and acoustic fidelity by providing a unified multi-layer representation that supports both semantic alignment and fine-grained recon-

Table 2: Text-to-audio generation results on AudioCaps testset. **Bold**: best, underline: second-best.

Model	FAD↓	FD↓	KL↓	IS↑	CLAP-T↑
AudioGen	1.69	13.79	1.52	11.77	0.47
AudioLDM2-FULL	4.83	28.02	1.48	8.92	0.34
Make-an-Audio 2	**1.07**	**12.52**	1.34	11.52	0.40
Stable Audio Open	3.08	34.61	2.25	11.71	0.30
AudioX	1.91	19.54	1.51	11.72	0.38
Ours					
SL-Audio	3.35	17.30	1.25	12.39	0.50
SL-Audio-gen	2.82	16.71	**1.24**	**14.55**	**0.53**

Table 3: Audio caption results on AudioCaps testset. **Bold**: best, underline: second-best.

Model	METEOR↑	SPICE↑	SPIDEr↑
SALMONN	14.39	16.43	32.88
AudioFlamingo	10.56	11.24	17.51
AudioFlamingo2	10.65	12.01	21.19
AudioFlamingo3	**14.55**	18.87	**40.74**
Qwen2.5-omni	12.47	13.77	26.83
Ours			
SL-Audio	12.44	17.15	34.66
SL-Audio-cap	13.12	**19.06**	37.75

struction, meeting diverse task requirements. Using only the coarsest layer (f_0), SL-Audio-Tokenizer achieves strong semantic scores with a single token, demonstrating that f_0 effectively captures global semantic information. Notably, the relatively worse signal-level metrics do not indicate degraded audio quality, as illustrated by the audio examples on our demo page [1]. Since decoding from f_0 does not constrain fine-grained time-frequency realizations, the generated audio may differ substantially at the signal level while preserving the same semantic events (e.g., different realizations of the same sound category). By incorporating the mid-level layer, acoustic reconstruction metrics improve substantially while preserving high semantic alignment. This demonstrates that f_1 effectively captures coarse acoustic structure complementary to the global semantic layer f_0. With all three layers enabled, SL-Audio-Tokenizer achieves the best overall reconstruction performance. The finest layer f_2 preserves detailed time-frequency information critical for high-fidelity decoding and precise audio manipulation.

Overall, the SL-Audio-Tokenizer consistently outperforms existing tokenizers across multiple evaluation dimensions, validating its effectiveness as a unified and scalable audio representation for downstream audio-language modeling.

4.3 Text-to-Audio Generation

Table 2 reports text-to-audio generation results on the AudioCaps test set. We compare SL-Audio with representative text-to-audio generation models, most of which are specialized expert systems, including AudioGen (Kreuk et al., 2022), AudioLDM2 (Liu et al., 2024b), Make-an-Audio 2 (Huang et al., 2023), Stable Audio (Evans et al., 2025) and AudioX (Tian et al., 2025). We note that text-to-audio generation is inherently a one-to-many problem, where multiple valid acoustic realizations can correspond to the same textual description. In such settings, overly fine-grained representations are not always optimal, as they may unnecessarily constrain the diversity and flexibility of generation (see Appendix A.5). Therefore, we directly take the output at the coarsest semantic layer f_0 as the final generated result, without further refining it into finer-grained layers. The results show that the pretrained SL-Audio already achieves clear advantages in both perceptual audio quality and text-audio alignment over prior methods. This setting highlights the effectiveness of the high-level semantic layer f_0, which captures abstract sound events and concepts in a form that aligns well with LLM. When further fine-tuned with task-specific instruction tuning for text-to-audio generation, SL-Audio-gen shows the potential for additional performance gains.

4.4 Audio-to-Text Captioning

We compare SL-Audio with representative audio LLMs, including SALMONN (Tang et al., 2024), Qwen2.5-omni (Xu et al., 2025), and AudioFlamingo series (Kong et al., 2024; Ghosh et al., 2025; Goel et al., 2025). Following prior work, we evaluate performance using METEOR (Banerjee & Lavie, 2005), SPICE (Anderson et al., 2016) and SPIDEr (Liu et al., 2017). Table 3 presents comparison results on the AudioCaps test set. We could see from the results that the pre-trained SL-Audio already exhibits reasonable captioning performance, and caption-specific fine-tuning further improves performance. Notably, compared with understanding-oriented systems such as the AudioFlamingo3, which rely on substantially larger and more sophisticated reasoning-oriented training data, SL-Audio is trained with relatively smaller-scale data within a unified framework for understanding and generation, yet still achieves competitive captioning performance, demonstrating

[1] Audio samples are available at https://sl-audio.github.io

Table 4: Text-queried separation results on AudioCaps and FSD50K test sets under the extraction setting.

Model	AudioCaps					FSD50K				
	KL↓	LSD↓	PSNR↑	CLAP-T↑	AFSim↑	KL↓	LSD↓	PSNR↑	CLAP-T↑	AFSim↑
AudioSep	1.00	2.14	<u>21.17</u>	0.42	0.75	1.17	**4.31**	<u>23.66</u>	0.24	0.69
OmniSep	2.18	4.8	16.03	0.28	0.64	1.9	5.9	19.25	0.19	0.62
FlowSep	1.24	<u>2.01</u>	20.37	0.4	0.75	1.68	**4.31**	19.27	0.23	0.69
Ours										
SL-Audio-sep (f_0)	<u>0.81</u>	2.39	17.96	**0.45**	**0.79**	0.94	5.62	18.83	**0.31**	<u>0.76</u>
SL-Audio-sep (f_0, f_1)	**0.67**	2.173	20.24	<u>0.44</u>	**0.79**	<u>0.81</u>	4.99	22.03	<u>0.3</u>	**0.77**
SL-Audio-sep (f_0, f_1, f_2)	0.90	**1.95**	**22.96**	0.39	<u>0.77</u>	**0.75**	<u>4.68</u>	**26.50**	0.28	**0.77**

that the layered audio representations capture meaningful semantic information that effectively supports audio-to-text understanding.

We also observe that audio captioning benefits from access to multiple layers (see Appendix A.5). While high-level layers provide event-level semantics, additional layers offer complementary acoustic context that helps describe attributes. This differs from text-to-audio generation, where coarse layers alone are often sufficient and can be preferable for stable and diverse synthesis.

Overall, these results indicate that SL-Audio's layered audio representation provides a flexible semantic interface: captioning can leverage richer multi-layer context for semantic understanding, while generation can selectively rely on coarser layers to balance controllability and diversity.

4.5 Text-queried Audio Separation

Table 4 reports results on text-queried audio separation evaluated on the AudioCaps and FSD50K test sets. We compare SL-Audio-sep with AudioSep (Liu et al., 2024c), OmniSep (Cheng et al., 2025), FlowSep (Yuan et al., 2024). We evaluate separation performance using log spectral distance (LSD), peak signal-to-noise ratio (PSNR), and Kullback-Leibler divergence (KL) to measure spectral fidelity, signal reconstruction quality, and distributional similarity between the generated and target audio, respectively. In addition, we assess semantic correctness using the CLAP-T score and AFSim (Yin et al., 2025), which evaluate semantic alignment between the separated audio and the corresponding textual descriptions or target audio, respectively.

Overall, SL-Audio-sep achieves competitive performance across both datasets, particularly on semantic metrics. Results in Table 4 also indicate that different layer configurations reveal clear functional roles. Using only the coarsest layer f_0, SL-Audio-sep (f_0) already attains strong CLAP-T and AFSim scores, successfully separating the text-specified target sound event, but produces low signal-level fidelity due to a lack of acoustic detail. Incorporating the mid-level layer f_1 (SL-Audio-sep (f_0,f_1)) significantly improves signal metrics while preserving semantic performance. When all three layers are used, the model achieves optimal semantic alignment and detailed reconstruction.

These results demonstrate that text-queried audio separation requires both semantic discrimination and precise acoustic reconstruction. By decoupling semantic alignment from acoustic realization through layered representations and coarse-to-fine refinement, SL-Audio-sep effectively balances semantic correctness and signal fidelity, validating its suitability for complex audio separation tasks.

5 Conclusion

In this paper, we introduce SL-Audio, a layered audio-language modeling framework that enables flexible interaction between language models and audio representations at different levels. By organizing audio into a hierarchy ranging from global semantic representations to fine-grained acoustic details, SL-Audio effectively addresses the diverse requirements of various audio tasks, including audio understanding, text-to-audio generation, and text-queried audio separation, balancing semantic correctness, generation stability, and signal fidelity within a single modeling framework. Overall, SL-Audio highlights the importance of multi-layer audio representations as a shared semantic interface for LLM-based audio modeling. By decoupling high-level semantic understanding from low-level acoustic realization, this work provides a practical step toward more flexible and controllable audio-language systems, and opens new directions for general-purpose audio modeling with large language models.

AI USE STATEMENT

We used generative AI tools to assist with dataset reformatting, as well as to improve the language and readability of the manuscript. We did not use generative AI tools for any other tasks requiring disclosure under the ICLR policy. The authors reviewed the AI-assisted data processing and manuscript edits to ensure accuracy and consistency. We take full responsibility for the final content of this work, including all AI-assisted text and artifacts.

ETHICS STATEMENT

This work explores layered audio-language modeling as a way to support more flexible interaction between large language models and audio representations. By enabling structured control over different semantic levels of audio, our approach may benefit researchers and developers working on audio understanding, generation, and editing tasks. Potential risks associated with audio synthesis and editing are shared with existing generative audio systems and should be mitigated through responsible use and deployment.

REFERENCES

Inclusion AI, Biao Gong, Cheng Zou, Chuanyang Zheng, Chunluan Zhou, Canxiang Yan, Chunxiang Jin, Chunjie Shen, Dandan Zheng, Fudong Wang, et al. Ming-omni: A unified multimodal model for perception and generation. *arXiv preprint arXiv:2506.09344*, 2025.

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *European conference on computer vision*, pp. 382–398. Springer, 2016.

Jisheng Bai, Haohe Liu, Mou Wang, Dongyuan Shi, Wenwu Wang, Mark D Plumbley, Woon-Seng Gan, and Jianfeng Chen. AudioSetCaps: An enriched audio-caption dataset using automated generation pipeline with large audio and language models. *IEEE Transactions on Audio, Speech and Language Processing*, 2025.

Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pp. 65–72, 2005.

Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi, et al. AudioLM: a language modeling approach to audio generation. *IEEE/ACM transactions on audio, speech, and language processing*, 31:2523–2533, 2023a.

Zalán Borsos, Matt Sharifi, Damien Vincent, Eugene Kharitonov, Neil Zeghidour, and Marco Tagliasacchi. SoundStorm: Efficient parallel audio generation. *arXiv preprint arXiv:2305.09636*, 2023b.

Sanyuan Chen, Chengyi Wang, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, et al. Neural codec language models are zero-shot text to speech synthesizers. *IEEE Transactions on Audio, Speech and Language Processing*, 33:705–718, 2025a.

Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*, 2025b.

Xize Cheng, Siqi Zheng, Minghui Fang, Ziang Zhang, Rongjie Huang, Shengpeng Ji, Jialong Zuo, Tao Jin, Zhou Zhao, et al. Omnisep: Unified omni-modality sound separation with query-mixup. In *International Conference on Learning Representations*, volume 2025, pp. 55196–55213, 2025.

Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. Qwen-Audio: Advancing universal audio understanding via unified large-scale audio-language models. *arXiv preprint arXiv:2311.07919*, 2023.

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

Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. High fidelity neural audio compression. *arXiv preprint arXiv:2210.13438*, 2022.

Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen. Clotho: An audio captioning dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 736–740. IEEE, 2020.

Zach Evans, Julian D Parker, CJ Carr, Zack Zukowski, Josiah Taylor, and Jordi Pons. Stable audio open. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025.

Eduardo Fonseca, Xavier Favory, Jordi Pons, Frederic Font, and Xavier Serra. FSD50K: an open dataset of human-labeled sound events. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:829–852, 2021.

Yuying Ge, Sijie Zhao, Jinguo Zhu, Yixiao Ge, Kun Yi, Lin Song, Chen Li, Xiaohan Ding, and Ying Shan. Seed-X: Multimodal models with unified multi-granularity comprehension and generation. *arXiv preprint arXiv:2404.14396*, 2024.

Sreyan Ghosh, Zhifeng Kong, Sonal Kumar, S Sakshi, Jaehyeon Kim, Wei Ping, Rafael Valle, Dinesh Manocha, and Bryan Catanzaro. Audio Flamingo 2: An audio-language model with long-audio understanding and expert reasoning abilities. In *Forty-second International Conference on Machine Learning*, 2025.

Arushi Goel, Sreyan Ghosh, Jaehyeon Kim, Sonal Kumar, Zhifeng Kong, Sang-gil Lee, Chao-Han Huck Yang, Ramani Duraiswami, Dinesh Manocha, Rafael Valle, and Bryan Catanzaro. Audio Flamingo 3: Advancing audio intelligence with fully open large audio language models. *arXiv preprint arXiv*, 2025.

Andrew Hines, Jan Skoglund, Anil C Kokaram, and Naomi Harte. ViSQOL: an objective speech quality model. *EURASIP Journal on Audio, Speech, and Music Processing*, 2015(1):13, 2015.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *ICLR*, 2022.

Jiawei Huang, Yi Ren, Rongjie Huang, Dongchao Yang, Zhenhui Ye, Chen Zhang, Jinglin Liu, Xiang Yin, Zejun Ma, and Zhou Zhao. Make-an-audio 2: Temporal-enhanced text-to-audio generation. *arXiv preprint arXiv:2305.18474*, 2023.

Po-Yao Huang, Hu Xu, Juncheng Li, Alexei Baevski, Michael Auli, Wojciech Galuba, Florian Metze, and Christoph Feichtenhofer. Masked autoencoders that listen. *Advances in Neural Information Processing Systems*, 35:28708–28720, 2022.

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

Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. 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)*, pp. 119–132, 2019.

Zhifeng Kong, Arushi Goel, Rohan Badlani, Wei Ping, Rafael Valle, and Bryan Catanzaro. Audio Flamingo: A novel audio language model with few-shot learning and dialogue abilities. *arXiv preprint arXiv:2402.01831*, 2024.

Felix Kreuk, Gabriel Synnaeve, Adam Polyak, Uriel Singer, Alexandre Défossez, Jade Copet, Devi Parikh, Yaniv Taigman, and Yossi Adi. Audiogen: Textually guided audio generation. *arXiv preprint arXiv:2209.15352*, 2022.

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

Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D Plumbley. Audioldm: Text-to-audio generation with latent diffusion models. *arXiv preprint arXiv:2301.12503*, 2023.

Haohe Liu, Xuenan Xu, Yi Yuan, Mengyue Wu, Wenwu Wang, and Mark D Plumbley. Semanticodec: An ultra low bitrate semantic audio codec for general sound. *IEEE Journal of Selected Topics in Signal Processing*, 2024a.

Haohe Liu, Yi Yuan, Xubo Liu, Xinhao Mei, Qiuqiang Kong, Qiao Tian, Yuping Wang, Wenwu Wang, Yuxuan Wang, and Mark D Plumbley. Audioldm 2: Learning holistic audio generation with self-supervised pretraining. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:2871–2883, 2024b.

Siqi Liu, Zhenhai Zhu, Ning Ye, Sergio Guadarrama, and Kevin Murphy. Improved image captioning via policy gradient optimization of spider. In *Proceedings of the IEEE international conference on computer vision*, pp. 873–881, 2017.

Xubo Liu, Qiuqiang Kong, Yan Zhao, Haohe Liu, Yi Yuan, Yuzhuo Liu, Rui Xia, Yuxuan Wang, Mark D Plumbley, and Wenwu Wang. Separate anything you describe. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024c.

Hao Ma, Zhiyuan Peng, Xu Li, Mingjie Shao, Xixin Wu, and Ju Liu. CLAPSep: Leveraging contrastive pre-trained model for multi-modal query-conditioned target sound extraction. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.

Hila Manor and Tomer Michaeli. Zero-shot unsupervised and text-based audio editing using ddpm inversion. *arXiv preprint arXiv:2402.10009*, 2024.

Xinhao Mei, Chutong Meng, Haohe Liu, Qiuqiang Kong, Tom Ko, Chengqi Zhao, Mark D Plumbley, Yuexian Zou, and Wenwu Wang. 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, 2024.

Xichen Pan, Satya Narayan Shukla, Aashu Singh, Zhuokai Zhao, Shlok Kumar Mishra, Jialiang Wang, Zhiyang Xu, Jiuhai Chen, Kunpeng Li, Felix Juefei-Xu, et al. Transfer between modalities with metaqueries. *arXiv preprint arXiv:2504.06256*, 2025.

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

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv:2212.04356*, 2023.

Robin Scheibler, Youna Ji, Soo-Whan Chung, Jaeuk Byun, Soyeon Choe, and Min-Seok Choi. Diffusion-based generative speech source separation. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 508–512, 2023.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun MA, and Chao Zhang. SALMONN: Towards generic hearing abilities for large language models. In *The Twelfth International Conference on Learning Representations*, 2024.

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

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, et al. Qwen2. 5-Omni technical report. *arXiv preprint arXiv:2503.20215*, 2025.

Dongchao Yang, Jinchuan Tian, Xu Tan, Rongjie Huang, Songxiang Liu, Haohan Guo, Xuankai Chang, Jiatong Shi, Sheng Zhao, Jiang Bian, Zhou Zhao, Xixin Wu, and Helen Meng. Uniaudio: towards universal audio generation with large language models. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Dongchao Yang, Songxiang Liu, Haohan Guo, Jiankun Zhao, Yuanyuan Wang, Helin Wang, Zeqian Ju, Xubo Liu, Xueyuan Chen, Xu Tan, et al. Almtokenizer: A low-bitrate and semantic-rich audio codec tokenizer for audio language modeling. In *International Conference on Machine Learning*, pp. 70850–70872. PMLR, 2025.

Xinlei Yin, Xiulian Peng, Xue Jiang, Zhiwei Xiong, and Yan Lu. Text-queried audio source separation via hierarchical modeling. *IEEE Transactions on Audio, Speech and Language Processing*, 33: 4928–4940, 2025. doi: 10.1109/TASLPRO.2025.3633078.

Yi Yuan, Xubo Liu, Haohe Liu, Mark D Plumbley, and Wenwu Wang. Flowsep: Language-queried sound separation with rectified flow matching. In *IEEE International Conference on Acoustic, Speech and Signal Procssing (ICASSP)*, 2024.

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

A Appendix

A Technical appendices and supplementary material

A.1 Multi-Task Sequence Construction in SL-Audio

In this section, we describe how task-specific sequences are constructed for different audio tasks in SL-Audio.

Audio-to-Text Understanding Audio understanding aims to generate a natural language description for a given audio input. All audio representations at different scales are provided as conditioning inputs, and the model predicts the textual caption t_{cap}. Given an audio clip represented at three scales (f_0, f_1, f_2), the sequence is constructed as:

s_cap = [<BOS> [INST] This is an audio-to-text caption task with 3 audio scales from coarse to fine. Please describe the audio. Audio at scale 0 <audio_scale0_start> f_0 <audio_scale0_end> Audio at scale 1 <audio_scale1_start> f_1 <audio_scale1_end> Audio at scale 2 <audio_scale2_start> f_2 <audio_scale2_end> [/INST] Here is the caption. t_{cap} <EOS>]

Text-to-Audio Generation Text-to-audio generation aims to synthesize an audio signal that is semantically consistent with a given textual caption. In SL-Audio, audio is represented at three scales from coarse to fine. Given a caption t_{cap}, the model is required to generate audio representations at all three scales (f_0, f_1, f_2), the three layered audio representations are generated progressively in the *Next-Layer-Prediction* manner, as introduced in the Sec. 3.2.1. The sequence is constructed as follows:

s_gen = [<BOS> [INST] This is a text-to-audio generation task with 3 audio scales from coarse to fine. Please generate audio at all scales according to the following caption: t_{cap} [/INST] Generated audio at scale 0 <audio_scale0_start> f_0 <audio_scale0_end> Generated audio at scale 1 <audio_scale1_start> f_1 <audio_scale1_end> Generated audio at scale 2 <audio_scale2_start> f_2 <audio_scale2_end> <EOS>]

Text-queried Audio Separation Text-queried audio separation aims to modify a mixture audio according to a textual instruction, such as sound extraction or removal. The model is required to generate edited audio representations at all scales autoregressively in a coarse-to-fine manner, conditioned on both the instruction and the mixture audio. Given an editing instruction t_{edit}, a mixture audio represented at three scales $(f_0^{mix}, f_1^{mix}, f_2^{mix})$ and the ground truth target signal $(f_0^{gt}, f_1^{gt}, f_2^{gt})$, the sequence is constructed as:

s_sep = [<BOS> [INST] This is an audio separation task with 3 audio scales from coarse to fine. Please edit the audio according to the following instruction: t_{edit}. Mixture audio at scale 0 <audio_scale0_start> f_0^{mix} <audio_scale0_end> Mixture audio at scale 1 <audio_scale1_start> f_1^{mix} <audio_scale1_end> Mixture audio at scale 2 <audio_scale2_start> f_2^{mix} <audio_scale2_end> [/INST] Here is the edited audio. Edited audio at scale 0 <audio_scale0_start> f_0^{gt} <audio_scale0_end> Edited audio at scale 1 <audio_scale1_start> f_1^{gt} <audio_scale1_end> Edited audio at scale 2 <audio_scale2_start> f_2^{gt} <audio_scale2_end> <EOS>]

A.2 Pre-training and Instruction Tuning Datasets

As summarized in Table 5, we construct the multimodal pre-training corpus by combining several large-scale audio-text datasets. Notably, AudioSetCaps, VGGSound, and YouTube-4M provide large-scale audio-text pairs with relatively long and descriptive captions (around 30 words on average), we preprocess these captions to simplify and emphasize primary sound events by GPT-4[2]).

For text-queried audio separation, we construct a large-scale supervised dataset by synthetically mixing audio clips from multiple captioned and labeled audio datasets, including AudioCaps, ESC-50, Clotho v2, FSD50K, and WavCaps. Source audio clips are normalized to a fixed duration and sampling rate, and mixtures are created by randomly selecting pairs of audio clips under the constraint that the selected clips do not share overlapping audio event classes. Each mixture is paired with a target audio and a natural-language instruction, forming triplets of (mixture audio, target audio, instruction). The instructions include both extraction and removal formulations, enabling the model to learn bidirectional semantic control over audio events. Overall, this procedure results in approximately 1.4M mixture-target pairs, corresponding to over 4K hours of mixed audio.

A.3 Implementation Details of SL-Audio-Tokenizer

The proposed hierarchical tokenizer is built upon a latent diffusion framework conditioned on multi-scale CLAP representations. Specifically, we adopt HTSAT-tiny CLAP encoders to extract hierarchical audio features at three semantic scales (1/16/256) from different Swin Transformer blocks. The resulting multi-scale CLAP embeddings, with original dimensions of 512, 768, and 384, are projected into a unified 768-dimensional space through linear layers. The Diffusion Transformer (DiT) is initialized from Stable Audio Open (Evans et al., 2025). Detailed configurations are provided in Table 6.

[2]https://chatgpt.com/

Table 5: Overview of datasets used for multimodal pre-training and instruction tuning.

Stage	Dataset	Samples	Duration
Pre-training			
Audio-Text Pre-training	AudioCaps	38K	∼100 h
	WavCaps	369K	∼1K h
	AudioSetCaps	1.68M	∼4.7K h
	VGGSound	182K	∼506 h
	YouTube-4M	3.96M	∼11K h
Instruction Tuning			
Audio-to-Text Captioning	Same as pre-training datasets		
Text-to-Audio Generation	Same as pre-training datasets		
Text-queried Audio Separation	Mixture of AudioCaps, ESC-50, Clotho v2, FSD50K, WavCaps	∼1.4M	∼4K h

The tokenizer is trained on fixed-length 10-second audio segments. During training, we use the AdamW optimizer with a learning rate of 5×10^{-5} and weight decay of 10^{-3}. Exponential moving average (EMA) is further applied for training stabilization. Training is conducted on 4 NVIDIA A100 GPUs with a batch size of 160 per GPU.

Table 6: Configuration of the proposed hierarchical tokenizer.

Component	Configuration
Audio Encoder	
Encoder Backbone	CLAP (HTSAT-tiny)
Semantic Scales	3 scales (1 / 16 / 256)
Original Feature Dimensions	512 / 768 / 384
Projected Feature Dimension	768
VAE Module	
Input Length	10 s
Latent Dimension	64
Downsampling Ratio	2048
Diffusion Transformer	
Backbone	DiT
Transformer Depth	24
Hidden Dimension	1536
Attention Heads	24
Conditioning Dimension	768

A.4 Training Details of SL-Audio

LoRA setting To enable parameter-efficient optimization, we adopt LoRA for fine-tuning the LLM, applying it to the attention and feed-forward projection modules. LoRA rank is set to 32. In addition to LoRA parameters, we update the token embedding layer, language modeling head, and layer normalization modules, while keeping the remaining backbone parameters frozen. We employ a cosine learning rate scheduler with a peak learning rate of 4e-4 throughout the entire training process. All models are trained on 8 NVIDIA A100 GPUs.

Loss weighting During pre-training, we weight the text token prediction loss by 1.0 and the audio regression loss by 5.0. For the audio regression terms, we apply layer-specific weights for layers f_0, f_1, and f_2, respectively, to compensate for the different magnitudes of regression losses across layers and keep them at a comparable scale during training.

A.5 HIERARCHICAL VS. FLATTEN ACROSS DIFFERENT TASKS

To further understand the role of hierarchical representations, we analyze the performance of different representation granularities across tasks. For generation and captioning, we randomly sample 500 cases from the AudioCaps test set to construct a lightweight evaluation subset for this ablation study. We compare the proposed hierarchical tokenization paradigm with single-scale flattened tokenization.

Table 7: Hierarchical vs. Flatten across different tasks.

Model	Generation					Caption			Editing				
	FAD ↓	FD ↓	KL ↓	IS ↑	CLAP-T ↑	METEOR ↑	SPICE ↑	SPIDEr ↑	KL ↓	LSD ↓	PSNR ↑	CLAP-T ↑	AFSim ↑
Flatten-Audio (f_0)	4.08	25.59	1.06	9.04	0.51	11.90	17.04	32.95	-	-	-	-	-
Flatten-Audio (f_2)	19.72	103.12	3.95	1.66	0.04	13.10	17.50	33.98	1.41	2.03	22.17	0.34	0.71
SL-Audio ($f_0 \rightarrow f_1 \rightarrow f_2$)	7.93	68.49	2.88	3.92	0.14	13.60	20.36	41.55	0.90	1.95	22.96	0.39	0.77
SL-Audio (best)	4.06	24.21	1.11	10.14	0.53	13.60	20.36	41.55	0.90	1.95	22.96	0.39	0.77

As shown in Table 7, single-scale representations (Flatten-Audio) are insufficient to support diverse audio-language tasks. Coarse representations (Flatten-Audio (f_0)) provide strong semantic alignment but lack fine-grained details, making them less suitable for tasks such as audio editing that require precise reconstruction of acoustic details. In contrast, fine-grained representations (Flatten-Audio (f_2)) are more expressive but harder to model, and may degrade performance in open-ended generation. This observation is particularly relevant for text-to-audio generation, We observe that when taking fine-grained representations f_2 for generation (e.g., Flatten-Audio (f_2) or SL-Audio ($f_0 \rightarrow f_1 \rightarrow f_2$) which progressively refines representations through the next-layer prediction paradigm), the generated audio tends to contain more acoustic artifacts or noisy components. This suggests that accurately predicting fine-grained acoustic details remains challenging in open-ended generation, and overly constraining low-level realizations may introduce instability. These results highlight an important advantage of the proposed next-layer prediction paradigm: it enables flexible control over generation granularity, allowing the model to prioritize semantic diversity and stability by primarily leveraging coarse representations in open-ended generation tasks.

For audio captioning, semantic understanding is the primary requirement. We find that coarse representations (Flatten-Audio (f_0)) are sufficient to achieve strong performance, while incorporating finer layers further improves results (SL-Audio (best)).

Tasks such as audio editing require precise control over acoustic details, where finer-grained representations (e.g., f_2) become necessary. Moreover, the hierarchical formulation (SL-Audio ($f_0 \rightarrow f_1 \rightarrow f_2$)) outperforms using f_2 alone (Flatten-Audio (f_2)), indicating that progressively generating multi-level representations is more effective than relying solely on fine-grained tokens. Overall, these results demonstrate that no single representation granularity is optimal across all tasks. Instead, hierarchical representations enable task-dependent abstraction, allowing the model to flexibly select the most suitable level of detail for each task.