DSA-Tokenizer: Disentangled Semantic-Acoustic Tokenization via Flow Matching-based Hierarchical Fusion

Anonymous ACL submission

Abstract

Speech tokenizers serve as the cornerstone of discrete Speech Large Language Models (Speech LLMs). Existing tokenizers either prioritize semantic encoding, fuse semantic content with acoustic style inseparably, or achieve incomplete semantic-acoustic disentanglement. To achieve better disentanglement, we propose **DSA-Tokenizer**, which explicitly disentangles speech into discrete semantic and acoustic tokens via distinct optimization constraints. Specifically, semantic tokens are supervised by ASR to capture linguistic content, while acoustic tokens focus on mel-spectrograms restoration to encode style. To eliminate rigid length constraints between the two sequences, we introduce a hierarchical **Flow-Matching** decoder that further improve speech generation quality. Furthermore, We employ a joint reconstruction-recombination training strategy to enforce this separation. DSA-Tokenizer enables high fidelity reconstruction and flexible recombination through robust disentanglement, facilitating controllable generation in speech LLMs. Our analysis highlights disentangled tokenization as a pivotal paradigm for future speech modeling. Audio samples are avaialble at https://anonymous.4open.science/w/DSA_Tokenizer_demo/. The code and model will be made publicly available after the paper has been accepted.

1 Introduction

The rapid advancement of large language models (LLMs) has catalyzed a paradigm shift in speech processing, spawning Speech LLMs (KimiTeam et al., 2025; Chen et al., 2025b; Chu et al., 2024; Xu et al., 2025) that unify speech and language processing within a single framework. Among existing architectures, **fully discrete Speech LLMs** (Zhang et al., 2025; Huang et al., 2025; Zeng et al., 2024; Nguyen et al., 2024) tokenize both input and output speech, enabling end-to-end processing in a unified discrete space with seamless LLM integration—yet their performance hinges heavily on the design of the speech tokenizer (Guo et al., 2025).

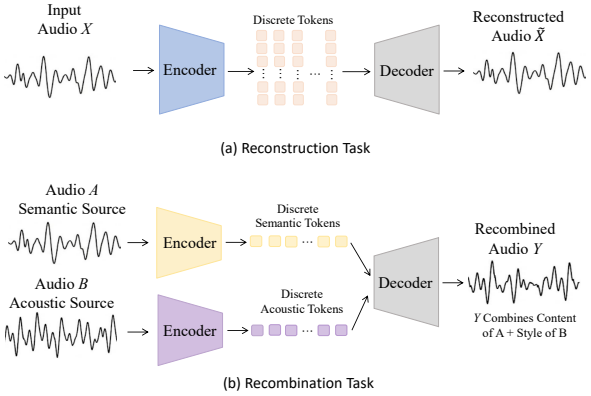

Figure 1: Illustration of (a) speech reconstruction and (b) semantic-acoustic recombination based on discrete token

Existing speech tokenizers generally fall into three categories. **Semantic tokenizers** (Lakhotia et al., 2021; Hsu et al., 2021; Du et al., 2024) prioritize linguistic information via self-supervised learning or ASR supervision. While this facilitates integration with LLMs, they often discard essential acoustic cues like timbre. Conversely, **semantic-acoustic mixed tokenizers** (Défossez et al., 2022; Ji et al., 2025) target high-fidelity reconstruction but produce entangled representations, preventing independent attribute control. Finally, **shallowly disentangled tokenizers** (Défossez et al., 2024; Zhang et al., 2023b) attempt to decouple semantic and acoustic information atop mixed architectures. However, they often suffer from incomplete disentanglement, failing to achieve a clean separation of attributes.

To rigorously evaluate the disentanglement capability of these tokenizers, rather than relying solely on reconstruction quality or ASR performance, we argue that **cross-utterance semantic-acoustic recombination** is a critical and more direct evalua-

tion task (Figure 1). Ideally, recombination task requires the tokenizer to extract pure semantics from a semantic source and exclusive acoustic style (e.g., timbre, prosody) from an acoustic source, then fuse them into a new speech utterance that retains the semantic content of the former and the acoustic style of the latter. Our experiments demonstrate that existing tokenizers exhibit substantial limitations in this task. Some models bias towards linguistic information, causing severe timbre mismatch, whereas others prioritize acoustic cues at the expense of semantic fidelity. Consequently, this issue hinders the ability of Speech LLMs in acoustic-related tasks. Furthermore, a fundamental limitation of existing tokenizers is their rigid length constraint between tokens extracted from the semantic source and the acoustic source, making it impossible to perform recombination between speech utterances of different lengths.

To address this gap, we propose **DSA-Tokenizer**. To ensure strict disentanglement, we utilize a **constrained dual-stream tokenizer** where semantic and acoustic tokens are supervised by ASR and mel-spectrograms restoration objectives, respectively. These streams are processed by a **flow-based hierarchical fusion decoder**, which allows for high-fidelity reconstruction and cross-utterance recombination free from rigid length constraints. Finally, by adopting a **joint reconstruction-recombination training strategy** that combines self-reconstruction with contextual inpainting, we enforce the robust separation of attributes necessary for controllable speech generation.

The experimental results demonstrate that our proposed DSA-Tokenizer achieves high-fidelity reconstruction and flexible recombination by effectively disentangling semantic and acoustic information without leakage. Therefore, it yields outstanding performance in acoustic-related LLM tasks compared to SAC (Chen et al., 2025c) and WavTokenizer (Ji et al., 2025). Further analysis reveals the effectiveness and robustness of the DSA-Tokenizer, highlighting the DSA-Tokenizer as a practical solution for the fully discrete speech LLMs. Our core contributions are summarized as follows:

- We introduce a novel recombination task to rigorously quantify tokenizer's semantic-acoustic disentanglement capacity, which highlight the emergence of explicit disentanglement and flexible recombination.
- We propose the **DSA-Tokenizer**, tailored for fully discrete speech LLMs, achieving strict semantic-acoustic disentanglement and seamless LLM integration.
- Our analysis reveals high reconstruction fidelity does not guarantee LLMs generation stability. Crucially, we demonstrate that strict semantic-acoustic disentanglement significantly enhances the robustness and controllability of Speech LLMs.

2 Related Work

2.1 Speech Large Language Models

Recent advances in Speech LLMs have centered on unifying speech and language processing, giving rise to two distinct technical paradigms: (1) **Thinker-talker architectures** (e.g., MinMo (Chen et al., 2025b), Qwen3-omni (Xu et al., 2025)) employ independent acoustic encoders and discrete token decoders, processing continuous speech inputs and generating discrete speech token outputs; (2) **Fully discrete architectures** (e.g., GLM-4-Voice (Zeng et al., 2024), MiMo-Audio (Zhang et al., 2025), Moshi (Défossez et al., 2024), Spirit LM (Nguyen et al., 2024)), by contrast, tokenize both input and output speech signals, which feature two core advantages: consistent cross-modal input-output formats and seamless integration with LLMs.

2.2 Disentangled Speech Tokenizers

Speech tokenizers largely focus on three key goals: low bitrate, high reconstruction fidelity or semantic preservation, and semantic-acoustic disentanglement. Notable recent disentanglement methods include: SpeechTokenizer uses SSL model distillation to guide the first layer of its RVQ framework for semantic encoding, with subsequent layers dedicated to acoustic detail capture; XY-Tokenizer (Gong et al., 2025) equips RVQ architecture with LLM-aligned ASR supervision and speech reconstruction to mitigate the conflict between semantic and acoustic capabilities; DualCodec employs dual-codebook RVQ to assign dedicated semantic and style partitions, enhancing first-layer semantic encoding via SSL feature prediction; SAC (Chen et al., 2025c) leverages a dedicated dual-stream structure with separate semantic and acoustic encoders to realize explicit separation.

3 Method

We propose **Disentangled Semantic-Acoustic Tokenizer** (DSA-tokenizer), a unified speech tok-

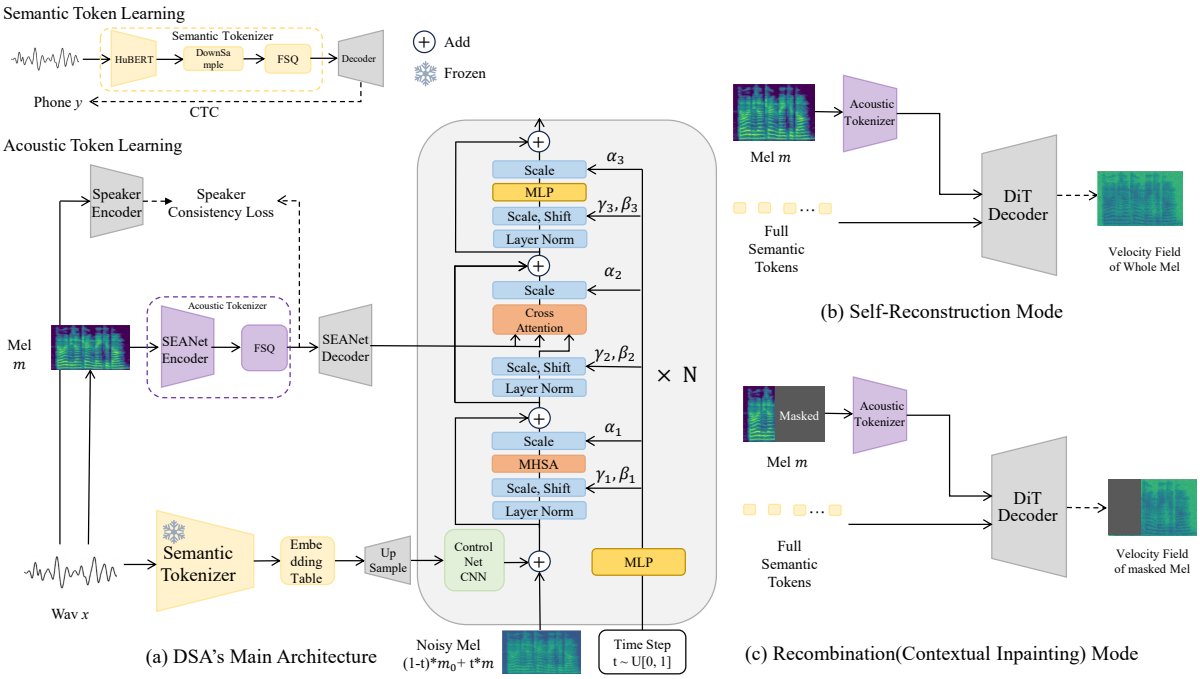

Figure 2: Overview of the proposed framework and training strategy. (a) DSA-Tokenizer framework: Input audio X is encoded into discrete semantic and acoustic tokens, which are fed into the DiT decoder for audio generation. (b) Self-Reconstruction Mode: The model learns to predict the velocity field of the full Mel-spectrogram based on the complete acoustic and semantic tokens. (c) Recombination (Contextual Inpainting) Mode: The model learns to predict the velocity field of the masked Mel-spectrogram region based on the acoustic tokens of the unmasked region and the full semantic tokens.

enization framework tailored for fully discrete Speech LLMs, as shown in Figure2(a). Its core objective is to achieve effective semantic-acoustic disentanglement, supporting both high-fidelity speech reconstruction and flexible cross-utterance content-style recombination. To this end, the framework employs two parallel discrete token streams with orthogonal goals: semantic tokens (z_s) encoding linguistic content, and acoustic tokens (z_a) capturing style attributes. These two token streams are then fused via distinct condition injection methods in a diffusion transformer for end-to-end speech generation.

3.1 Dual Token Streams with Task-Specific Constraints

The robust disentanglement capability of DSA-Tokenizer stems from the asymmetric optimization objectives enforced on each token stream. Specifically, semantic tokens are constrained to filter out all stylistic noise via ASR-based supervision, ensuring they retain only linguistic content. In contrast, acoustic tokens are optimized to enable high-quality mel-spectrogram restoration given semantic codes, with a training strategy that integrates both reconstruction and recombination modes.

3.1.1 Semantic Token Learning

To ensure semantic tokens retain only linguistic content (e.g., phonemes, tones) while excluding stylistic noise (e.g., timbre, prosody), we adopt an ASR-supervised paradigm consistent with recent discrete speech tokenization works (Chen et al., 2025a; Tao et al., 2024; DiscreteSpeech Team, 2025). Given a speech waveform x, we use a pre-trained HuBERT model as the semantic encoder, which captures robust phonetic and lexical representations via large-scale self-supervised pre-training. A Finite Scale Quantization (FSQ) (Mentzer et al., 2023) layer is appended to discretize the continuous encoder outputs into semantic tokens, with a codebook size of 1024 and a frame rate of 25Hz. Formally, semantic tokens are defined as:

$$z_s = \text{FSQ}\left(\text{HuBERT}(x)\right) \in \mathcal{Z}^{T_s \times D_s}$$

where $\mathcal{Z}$ denotes the discrete token space, T_s is the sequence length, and D_s is the FSQ codebook size. We enforce strict linguistic constraints by training the HuBERT-FSQ pipeline with Connectionist Temporal Classification (CTC) loss (Graves

et al., 2006). After training, the lightweight decoder for loss computation is discarded, and the HuBERT-FSQ module is frozen as a dedicated semantic token extractor to avoid interfering with subsequent acoustic token learning.

3.1.2 Acoustic Token Learning

Acoustic tokens are designed to capture stylistic attributes (e.g., timbre, prosody) essential for high-fidelity speech reconstruction but filtered out by semantic tokens. Specifically, mel-spectrograms m are extracted from raw speech x, downsampled via a SEANet-style (Tagliasacchi et al., 2020) encoder, and discretized through a FSQ quantizer to yield the acoustic tokens:

$$z_a = \text{FSQ}\left(\text{SEANetEncoder}(m)\right) \in \mathcal{Z}^{T_a \times D_a}$$

where T_a denotes the sequence length jointly determined by the CNN structure and mel-spectrogram length, and D_a is the FSQ codebook size. This design imposes no restrictions on the lengths of T_a and T_s, facilitating cross-utterance content-style recombination.

Unlike the pre-trained semantic token stream, the acoustic tokenizer is trained **end-to-end** with the DiT decoder (Sec. 3.2). Gradients from Flow Matching loss backpropagate through the discrete bottleneck via the straight-through estimator, optimizing the SEANet encoder. This ensures z_a encodes all spectral details required for high-fidelity reconstruction that are not captured by z_s.

3.1.3 Joint Reconstruction-Recombination Training Strategy

Conventional speech tokenizers rely solely on reconstruction objectives. This approach inevitably imposes rigid length constraints between semantic and acoustic representations, hindering cross-source speech synthesis like voice conversion. Moreover, single-objective training exacerbates information leakage, leading to entangled representations where semantic and acoustic information are indistinguishable. To mitigate these issues, we propose a joint reconstruction-and-recombination training strategy: each batch is randomly assigned to one of the following two modes with equal 50% probability.

Self-Reconstruction Mode (Figure 2(b)): The full sequences of z_s (semantic tokens) and z_a (acoustic tokens) are provided as conditions. The model learns to predict the Flow Matching velocity field of the entire mel-spectrogram for high-fidelity reconstruction.

Recombination (Contextual Inpainting) Mode (Figure 2(c)): To simulate content-style recombination, we randomly sample a time-axis split point τ, then mask the mel-spectrogram segments after τ along the time axis. The DiT decoder is tasked with predicting the Flow Matching velocity field for the masked segments, conditioned only on the prefix acoustic token $z_a^{<\tau}$ and full semantic token sequence z_s. This masking strategy compels the model to infer global acoustic style from partial acoustic context while strictly following semantic guidance, which helps to decouple the two information streams.

3.2 Hierarchical Flow Matching Decoder for Semantic-Acoustic Token Fusion

Fusing semantic tokens (z_s) and acoustic tokens (z_a) poses a critical challenge: balancing the strict temporal alignment required for linguistic content with the temporal flexibility of acoustic characteristics. To address this, we propose a DiT-based (Peebles and Xie, 2023; Ho et al., 2020) Flow Matching (Lipman et al., 2023) decoder with a **hybrid injection strategy**. Specifically, we treat z_s as the structural backbone of speech and inject it as a dense temporal condition: a simplified ControlNet-like (Zhang et al., 2023a) CNN adapter processes z_s, and the output is directly added to the input noise to enforce precise temporal alignment, ensuring generated speech strictly follows semantic content. In contrast, z_a is designed to capture acoustic attributes and is injected via **cross-attention**. This allows the model to flexibly capture global acoustic cues (for recombination) and fine-grained sequential details (for reconstruction) without rigid length constraints, effectively "painting" speaker and prosodic characteristics onto the semantic skeleton.

3.2.1 Semantic ControlNet-Style Injection

To bridge the modality gap between discrete semantic tokens and continuous mel-spectrograms, we first project z_s into continuous embeddings e_s via a learnable codebook. We then upsample e_s to match the mel-spectrogram length T_{mel} using linear interpolation, yielding the aligned semantic features $\tilde{e}_s$.

Following standard Flow Matching paradigms, we derive the intermediate noisy mel-spectrogram m_t at a randomly sampled timestep $t \sim \text{U}[0, 1]$

as a linear interpolation between the initial Gaussian noise $m_0 \sim \mathcal{N}(0, I)$ and the target clean mel-spectrograms m:

$$m_t = (1 - t) \cdot m_0 + t \cdot m$$

To strictly enforce linguistic alignment, we adopt a ControlNet-inspired (Zhang et al., 2023a) injection strategy. Specifically, for the first layer, $\tilde{e}_s$ is processed by a lightweight CNN adapter and directly added to the noisy input m_t, resulting in the fused feature:

$$F_{\text{sem}} = \text{CNN}(\tilde{e}_s) + m_t$$

This F_{sem} is then fed into the DiT's multi-head self-attention layers. For subsequent blocks, m_t is replaced with the output of the previous block. By treating semantic information as the structural backbone embedded into the noisy input, we ensure the model captures long-range linguistic context prior to integrating acoustic details.

3.2.2 Acoustic Injection via Cross-Attention

Acoustic style information is integrated into the DiT backbone via cross-attention, leveraging its inherent capability for cross-sequence modeling. Specifically, acoustic tokens z_a are projected to continuous embeddings via a learnable codebook and upsampled to the mel-spectrogram length T_{mel} using a SEANet-style decoder, yielding the aligned acoustic features $\tilde{e}_a$.

Within the DiT blocks, the self-attention-enhanced semantic feature F_{sem} serves as queries, while keys and values are derived from $\tilde{e}_a$ via dedicated linear projections W^K and W^V. The cross-attention fused feature is computed as:

$$F_{\text{sem-aco}} = \text{CrossAttn}\left(F_{\text{sem}}, \tilde{e}_a W^K, \tilde{e}_a W^V\right)$$

Finally, a feed-forward network (FFN) processes $F_{\text{sem-aco}}$ and projects it to the mel-spectrogram dimension for Flow Matching velocity field prediction.

3.2.3 Training Objectives and Inference

Our framework is trained end-to-end using a hybrid objective that unifies high-fidelity generation with robust representation learning.

Flow Matching Loss The primary objective is to drive the DiT decoder to generate realistic mel-spectrograms conditioned on the dual token streams. Following the Conditional Flow Matching (CFM) (Ho and Salimans, 2022) paradigm, we optimize the decoder to approximate the conditional vector field v_t that transports the initial Gaussian noise distribution to the target mel-spectrogram distribution p_{data}. The loss is defined as the mean squared error (MSE) between the model-predicted velocity v_θ and the ground-truth conditional flow velocity $v_t = \frac{dm_t}{dt} = m - m_0$:

$$\mathcal{L}_{\text{fm}} = \mathbb{E}_{t, m_0, m}\left[\left\| v_t - v_\theta(m_t, t, \tilde{e}_s, \tilde{e}_a)\right\|^2\right]$$

where $t \sim \text{U}[0, 1]$, $m_0 \sim \mathcal{N}(0, I)$, $m \sim p_{data}$.

Speaker Consistency Loss To ensure z_a captures discriminative speaker characteristics during joint training of the acoustic tokenizer and decoder, we extract a reference speaker embedding s_{ref} from the raw speech waveform using a WavLM (Chen et al., 2022) encoder finetuned by speaker verification task[1]. We then enforce the pooled acoustic token embeddings to align with this reference embedding:

$$\mathcal{L}_{\text{spk}} = 1 - \cos\left(s_{\text{ref}}, \text{AttnPool}(e_a)\right)$$

where $\text{AttnPool}(\cdot)$ aggregates the acoustic token embeddings $\mathbf{e}_a$ into a global vector. The total training loss is a weighted sum: $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{fm}} + \lambda_{\text{spk}}\mathcal{L}_{\text{spk}}$, where $\lambda_{\text{spk}} = 1.0$.

Classifier-Free Guidance To generate high-quality speech with desired content and style, we employ Classifier-Free Guidance (CFG) (Ho and Salimans, 2022) in our framework. During training, we randomly drop the conditional inputs $\tilde{e}_s$ and $\tilde{e}_a$. During inference, the modified velocity field $\tilde{v}$ is formulated as:

$$\tilde{v} = v_\theta(m_t, t, c) + \omega \cdot (v_\theta(m_t, t, c) - v_\theta(m_t, t))$$

where ω denotes the guidance scale (set to 2), and $c = (\tilde{e}_s, \tilde{e}_a)$ is the conditional input.

4 Experiment Setup

4.1 Evaluation Tasks

4.1.1 Reconstruction and Recombination

Two evaluation scenarios are considered: (1) **Reconstruction**: using a single utterance as both semantic and acoustic source; (2) **Recombination**: using two distinct utterances, where one serves as the semantic source and the other as the acoustic source.

[1] https://github.com/microsoft/UniSpeech/tree/main/downstreams/speaker_verification

The following metrics are adopted: (1) **Speech naturalness**, measured by UTMOS (Saeki et al., 2022) on a 1–5 scale, where higher scores indicate better naturalness; (2) **Content consistency**, quantified by word error rate (WER) for English and character error rate (CER) for Chinese, where lower values mean higher consistency with the text corresponding to the semantic source; (3) **Style preservation**, assessed via Speaker Similarity Score (SIM) on a -1 to 1 scale, where higher scores reflect better style retention.

4.1.2 Disentanglement Probing

This task verifies whether token streams encode only the desired attributes (semantic/style) without cross-information leakage. ASR is employed to evaluate the semantic information contained in input tokens, while speaker classification (SC) is used to assess the acoustic information. Evaluation protocols are adapted to three categories of model architectures:

(1) **Single-layer baselines** (e.g., WavTokenizer): ASR and SC are applied to the same token sequence.

(2) **Multi-layer baselines** (e.g., EnCodec, SpeechTokenizer): ASR and SC are both evaluated on Layer 0 and Layers 1–7.

(3) **DSA-Tokenizer**: ASR and SC are both conducted on semantic tokens and acoustic tokens.

A unified CNN-LSTM-MLP-based classifier is employed to process discrete token sequences, outputting recognized text for ASR tasks (scored by WER) and predicted speaker IDs for SC tasks (scored by classification accuracy).

4.1.3 LLM-based Voice Cloning

This task verifies the **compatibility** of DSA-Tokenizer with Speech LLMs and validates the **manipulability** of semantic content and acoustic style in voice cloning.

Each speech utterance is formatted into a token sequence with semantic tokens (Z_S) concatenated before acoustic tokens (Z_A). The LLM is trained on triplets $(\mathcal{S}, \mathcal{A}, \mathcal{C})$ (semantic source, acoustic source, cloned speech), taking the [SEP]-separated concatenation of $\mathcal{S}$ and $\mathcal{A}$ token sequences as input and predicting the $\mathcal{C}$ token sequence as output. During inference, the LLM generates token sequences for unseen $(\mathcal{S}', \mathcal{A}')$ pairs, which are decoded into speech waveforms via the DSA-Tokenizer decoder. For evaluation, we use UTMOS for naturalness, WER between cloned speech and semantic source text for content consistency, and SIM between cloned speech and acoustic source for style preservation.

4.2 Datasets

DSA-Tokenizer Training: We train the semantic tokenizer on 4,000 hours of open-source Chinese-English speech-text aligned data. For the acoustic tokenizer and decoder, we use the 100k-hour Chinese-English subsets of the Emilia dataset (He et al., 2024), which offers large scale and diverse speaking styles; we clean this dataset to boost performance, with detailed procedures in Appendix B. Training and architecture details of DSA-Tokenizer are provided in Appendices C.1 and A, respectively. **Waveform Evaluation**: We use the multilingual SeedTTS datasets (Anastassiou et al., 2024) (covering English and Chinese) to verify cross-language generalization. **Disentanglement Probing**: LibriSpeech (Panayotov et al., 2015) is used for ASR-based content retention evaluation, while Voxceleb1 (Nagrani et al., 2020) supports speaker classification-based style preservation evaluation. **LLM-based Voice Cloning**: We build over 350,000 training triplets $(\mathcal{S}, \mathcal{A}, \mathcal{C})$ from LibriTTS train-960, where $\mathcal{S}$ and $\mathcal{A}$ come from distinct utterances and $\mathcal{C}$ is synthesized via F5TTS (Chen et al., 2025d) as supervised targets. The test set is derived from LibriTTS test-clean, containing over 4,000 randomly matched $(\mathcal{S}', \mathcal{A}')$ pairs for inference.

4.3 Baselines

Four representative types of baseline models are selected for fair comparison: WavTokenizer (single-layer, non-decoupled), SAC and DualCodec (dual-branch, decoupled), Mimi and Encodec (multi-layer, non-decoupled), and SpeechTokenizer (multi-layer, decoupled) Cosy2 S^3 (ASR-supervised captures semantics). For LLM-based voice cloning tasks, Qwen3-0.6B (Yang et al., 2025) is adopted as the LLM backbone, with its vocabulary expanded with corresponding speech tokens.

5 Results and Discussion

5.1 Can DSA-Tokenizer Unify Reconstruction and Recombination?

DSA-Tokenizer successfully unifies high-fidelity speech reconstruction with precise cross-utterance recombination across English and Chinese. As shown in Table 1, our model achieves a superior bal-

Model	BitRate (kbps)	English			Chinese		
		UTMOS ↑	WER (%) ↓	SIM ↑	UTMOS ↑	CER (%) ↓	SIM ↑
Reconstruction Task							
WavTokenizer (75 Hz)	0.90	**3.92**	3.23	**0.84**	2.82	5.08	0.62
Mimi (8-layer)	1.10	3.30	3.46	0.74	2.37	2.93	0.73
Encodec (2-layer)	1.50	1.56	5.40	0.61	1.35	5.19	0.62
SpeechTokenizer (2-layer)	1.00	2.12	7.57	0.36	1.75	24.39	0.33
DualCodec (12.5 Hz, 6-layer)	0.925	3.78	2.58	0.76	2.95	1.62	0.80
SAC (f_s=12.5 Hz, f_a=50 Hz)	0.875	3.88	**2.03**	0.83	**2.99**	**1.53**	**0.87**
DSA-Tokenizer (f_s=25 Hz, f_a=25 Hz, $4^8, 1$)	0.70	3.41	2.09	0.72	2.60	1.78	0.78
DSA-Tokenizer (f_s=25 Hz, f_a=25 Hz, $4^7, 2$)	1.00	3.44	2.45	0.76	2.65	1.89	0.81
DSA-Tokenizer (f_s=25 Hz, f_a=25 Hz, $4^8, 2$)	1.10	3.46	2.61	0.77	2.68	1.93	0.82
DSA-Tokenizer (f_s=25 Hz, f_a=50 Hz, $4^8, 1$)	1.10	3.38	2.49	0.76	2.66	1.88	0.81
Cross-Utterance Recombination Task							
Mimi (8-layer)	1.10	2.19	107.51	0.53	1.54	100.01	0.53
Encodec (2-layer)	1.50	1.25	98.31	0.09	1.24	68.97	0.13
SpeechTokenizer (2-layer)	1.00	1.40	12.98	0.13	1.34	123.33	0.15
DualCodec (12.5 Hz, 6-layer)	0.925	2.39	17.36	0.11	1.73	16.22	0.32
SAC (f_s=12.5 Hz, f_a=50 Hz)	0.875	1.34	90.22	0.13	1.27	71.87	0.30
DSA-Tokenizer (f_s=25 Hz, f_a=25 Hz, $4^8, 1$)	0.70	3.58	6.67	0.57	2.74	3.77	0.67
DSA-Tokenizer (f_s=25 Hz, f_a=25 Hz, $4^7, 2$)	1.00	3.56	8.77	0.60	2.81	5.27	**0.70**
DSA-Tokenizer (f_s=25 Hz, f_a=25 Hz, $4^8, 2$)	1.10	3.54	8.41	**0.61**	2.81	5.02	**0.70**
DSA-Tokenizer (f_s=25 Hz, f_a=50 Hz, $4^8, 1$)	1.10	**3.67**	**5.93**	0.60	**2.92**	**3.49**	**0.70**

Table 1: Performance comparison of different tokenizers on speech reconstruction and cross-utterance recombination tasks for English and Chinese speech. For DSA-Tokenizer, the notation (f_s, f_a, N^C, L) denotes the **semantic token rate** f_s, **acoustic token rate** f_a, acoustic FSQ codebook with N levels per channel and C channels, and L stacked FSQ layers. Baseline models with similar bitrate are selected for fair comparison, annotated with their codebook layers or token rates (semantic f_s and acoustic f_a where applicable).

ance compared to all baselines. In standard reconstruction, it delivers competitive UTMOS scores and high speaker similarity while maintaining low WER and CER. Crucially, in the challenging cross-utterance recombination task, DSA-Tokenizer significantly outperforms competitors, demonstrating exceptional semantic-acoustic disentanglement. It secures a substantial lead in audio quality and style preservation while achieving drastically lower error rates for content control. While Mimi maintains decent SIM (0.53) by merely extracting acoustic features without effective fusion, our model surpasses it while ensuring content accuracy. Furthermore, where conventional baselines (e.g., Mimi, Encodec, SAC) and SpeechTokenizer fail to retain semantic integrity—resulting in high error rates—DSA-Tokenizer yields the lowest WER/CER, successfully reconciling high-fidelity reconstruction with rigorous semantic control.

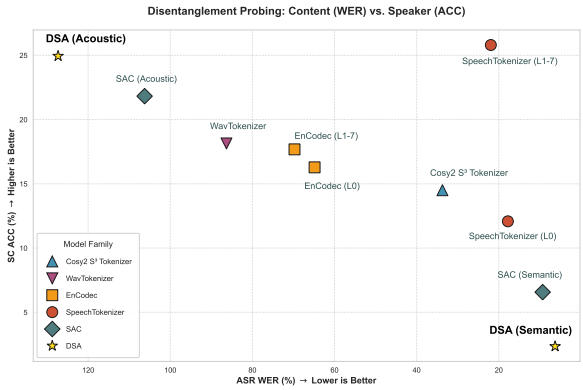

Figure 3: Disentanglement probing evaluation results. (L0) means the first layer, (L1-7) means the second to eighth layers.

5.2 Can Discrete Tokens Capture Sufficient Information without Leakage?

Semantic tokens capture sufficient semantic information, while acoustic tokens retain adequate acoustic details, with no significant information leakage. As illustrated in Figure 3, different tokenizers exhibit distinct encoding patterns based on their objectives. First, regarding single-codebook models, the ASR-supervised CosyVoice2 S^3 Tokenizer prioritizes semantics (low WER, moderate SC ACC), whereas the reconstruction-optimized WavTokenizer captures mixed information (high WER, slightly higher SC ACC). This confirms that ASR supervision steers encoding toward semantics, while reconstruction objectives preserve acoustic details. Second, multi-layer tokenizers demon-

Tokenizer	UTMOS ↑	WER (%) ↓	SIM ↑
WavTokenizer	3.75	89.44	0.28
SAC	3.30	24.21	0.37
DSA-Tokenizer	**3.90**	**23.95**	**0.41**

Table 2: LLM-based voice cloning performance.

	Reconstruction			Recombination		
Model	UTMOS	WER	SIM	UTMOS	WER	SIM
DSA-Tokenizer	3.38	2.49	0.76	3.67	5.93	0.60
w/o $\mathcal{L}_{\text{spk}}$	3.42	2.23	0.56	3.73	4.95	0.36
w/o Recombination	3.43	2.36	0.72	2.67	107.68	0.58

Table 3: Ablation study of speaker loss and recombination mode on reconstruction and recombination tasks.

strate varying disentanglement capabilities. EnCodec, lacking explicit constraints, shows uniform performance across layers, suggesting semantic and acoustic entangled. In contrast, SpeechTokenizer achieves shallow disentanglement by distilling semantics into layer 0, which yields significantly lower WER and SC ACC compared to layers 1–7. Third, dual-token baselines such as SAC achieve partial separation. While SAC's semantic tokens prioritize content (low WER) and its acoustic tokens capture style (high SC ACC), the model suffers from incomplete disentanglement and information leakage between two streams.Finally, our DSA-Tokenizer realizes strict disentanglement through a dedicated dual-stream design, with semantic tokens achieving ultra-low WER (6.28%) and minimal SC ACC (2.35%).

5.3 Can disentanglement boost LLM's performance of acoustic-related task?

Effective semantic-acoustic disentanglement is the key to success in acoustic-related task. As shown in Table 2, despite achieving high fidelity in reconstruction, the non-disentangled WavTokenizer performs the worst across all metrics and even leads to unstable LLM inference (e.g., endless generation). This is because its lack of disentanglement hinders the separate modeling of semantics and acoustics. This observation highlights that **high reconstruction fidelity does not equate to generation stability.** While both SAC and DSA-Tokenizer adopt disentanglement designs, our model's superiority stems from its finer-grained separation mechanism: this enables LLMs to independently model the two information streams, ultimately achieving an optimal balance of speech quality, semantic fidelity and timbre similarity.

5.4 Are Speaker Loss and Recombination Mode Essential?

Speaker loss plays a critical role in preserving speaker timbre. Recombination Mode is vital for ensuring semantic consistency Table 3 presents the ablation results verifying the contribution of the speaker loss ($\mathcal{L}_{\text{spk}}$) and the recombination training mode. We observe that removing $\mathcal{L}_{\text{spk}}$ causes a sharp reduction in SIM scores across tasks, confirming its critical role in preserving speaker timbre. Furthermore, excluding the recombination mode results in slight SIM degradation during reconstruction but leads to a severe performance collapse in the recombination task (evidenced by drastically lower UTMOS and extremely high WER). This demonstrates that the recombination mode is essential for ensuring semantic consistency and robust acoustic modeling. In conclusion, both components are indispensable, and their synergistic effect ensures that the DSA-Tokenizer achieves superior, balanced performance across all tasks.

6 Conclusion

This paper presents DSA-Tokenizer, a dual-stream tokenizer that disentangles speech into discrete semantic and acoustic tokens. Semantic tokens capture linguistic content via ASR supervision, while acoustic tokens encode style features through waveform restoration. Combined with a joint reconstruction-recombination training strategy and hierarchical Flow Matching decoder, DSA-Tokenizer outperforms SOTA baselines in recombination, disentanglement probing and LLM-based voice conversion, while maintaining high-fidelity speech reconstruction. Our work validates that explicit semantic-acoustic disentanglement is a viable paradigm for advancing controllable speech generation in Speech LLMs.

Limitations

Despite the promising capabilities of the DSA-Tokenizer in disentanglement and reconstruction, several limitations remain. First, the inference latency poses a challenge for real-time applications. Since our decoder comprises a deep stack of 22 DiT blocks, the iterative Flow Matching sampling process is computationally intensive compared to GAN-based counterparts. Future work will investigate acceleration techniques to reduce the number of sampling steps. Second, our current study focuses exclusively on speech signals.

While the model demonstrates robust performance in this domain, its generalization to other audio modalities—such as music or environmental sounds—remains unexplored. We plan to extend the DSA-Tokenizer to general audio modeling in future iterations.

References

Keyu An, Qian Chen, Chong Deng, Zhihao Du, Changfeng Gao, Zhifu Gao, Yue Gu, Ting He, Hangrui Hu, Kai Hu, and 1 others. 2024. Funaudiollm: Voice understanding and generation foundation models for natural interaction between humans and llms. *arXiv preprint arXiv:2407.04051*.

Philip Anastassiou, Jiawei Chen, Jitong Chen, Yuanzhe Chen, Zhuo Chen, Ziyi Chen, Jian Cong, Lelai Deng, Chuang Ding, Lu Gao, Mingqing Gong, Peisong Huang, Qingqing Huang, Zhiying Huang, Yuanyuan Huo, Dongya Jia, Chumin Li, Feiya Li, Hui Li, and 27 others. 2024. Seed-tts: A family of high-quality versatile speech generation models. *Preprint*, arXiv:2406.02430.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the twelfth language resources and evaluation conference*, pages 4218–4222.

Hervé Bredin and Antoine Laurent. 2021. End-to-end speaker segmentation for overlap-aware resegmentation. In *Proc. Interspeech 2021*, Brno, Czech Republic.

Guoguo Chen, Shuzhou Chai, Guanbo Wang, Jiayu Du, Wei-Qiang Zhang, Chao Weng, Dan Su, Daniel Povey, Jan Trmal, Junbo Zhang, and 1 others. 2021. Gigaspeech: An evolving, multi-domain asr corpus with 10,000 hours of transcribed audio. *arXiv preprint arXiv:2106.06909*.

Kai Chen, Yunhao Gou, Runhui Huang, Zhili Liu, Daxin Tan, Jing Xu, Chunwei Wang, Yi Zhu, Yihan Zeng, Kuo Yang, and 1 others. 2025a. Emova: Empowering language models to see, hear and speak with vivid emotions. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5455–5466.

Qian Chen, Yafeng Chen, Yanni Chen, Mengzhe Chen, Yingda Chen, Chong Deng, Zhihao Du, Ruize Gao, Changfeng Gao, Zhifu Gao, Yabin Li, Xiang Lv, Jiaqing Liu, Haoneng Luo, Bin Ma, Chongjia Ni, Xian Shi, Jialong Tang, Hui Wang, and 17 others. 2025b. Minmo: A multimodal large language model for seamless voice interaction. *Preprint*, arXiv:2501.06282.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. 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.

Wenxi Chen, Xinsheng Wang, Ruiqi Yan, Yushen Chen, Zhikang Niu, Ziyang Ma, Xiquan Li, Yuzhe Liang, Hanlin Wen, Shunshun Yin, Ming Tao, and Xie Chen. 2025c. Sac: Neural speech codec with semantic-acoustic dual-stream quantization. *Preprint*, arXiv:2510.16841.

Yushen Chen, Zhikang Niu, Ziyang Ma, Keqi Deng, Chunhui Wang, Jian Zhao, Kai Yu, and Xie Chen. 2025d. F5-tts: A fairytaler that fakes fluent and faithful speech with flow matching. *Preprint*, arXiv:2410.06885.

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

Yu-An Chung, Yu Zhang, Wei Han, Chung-Cheng Chiu, James Qin, Ruoming Pang, and Yonghui Wu. 2021. W2v-bert: Combining contrastive learning and masked language modeling for self-supervised speech pre-training. *Preprint*, arXiv:2108.06209.

DiscreteSpeech Team. 2025. Dstk: Discrete speech tokenization toolkit. `https://huggingface.co/DiscreteSpeech/DSTK`.

Jiayu Du, Xingyu Na, Xuechen Liu, and Hui Bu. 2018. Aishell-2: Transforming mandarin asr research into industrial scale. *arXiv preprint arXiv:1808.10583*.

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

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

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

Xinjie Fan, Qifei Wang, Junjie Ke, Feng Yang, Boqing Gong, and Mingyuan Zhou. 2021. Adversarially adaptive normalization for single domain generalization. *Preprint*, arXiv:2106.01899.

Daniel Galvez, Greg Diamos, Juan Ciro, Juan Felipe Cerón, Keith Achorn, Anjali Gopi, David Kanter, Maximilian Lam, Mark Mazumder, and Vijay Janapa Reddi. 2021. The people's speech: A large-scale diverse english speech recognition dataset for commercial usage. *arXiv preprint arXiv:2111.09344*.

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. *Preprint*, arXiv:2506.23325.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, pages 369–376.

Yiwei Guo, Zhihan Li, Hankun Wang, Bohan Li, Chongtian Shao, Hanglei Zhang, Chenpeng Du, Xie Chen, Shujie Liu, and Kai Yu. 2025. Recent advances in discrete speech tokens: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–20.

Haorui He, Zengqiang Shang, Chaoren Wang, Xuyuan Li, Yicheng Gu, Hua Hua, Liwei Liu, Chen Yang, Jiaqi Li, Peiyang Shi, Yuancheng Wang, Kai Chen, Pengyuan Zhang, and Zhizheng Wu. 2024. Emilia: An extensive, multilingual, and diverse speech dataset for large-scale speech generation. *Preprint*, arXiv:2407.05361.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851.

Jonathan Ho and Tim Salimans. 2022. Classifier-free diffusion guidance. *Preprint*, arXiv:2207.12598.

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.

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

Shengpeng Ji, Ziyue Jiang, Wen Wang, Yifu Chen, Minghui Fang, Jialong Zuo, Qian Yang, Xize Cheng, Zehan Wang, Ruiqi Li, Ziang Zhang, Xiaoda Yang, Rongjie Huang, Yidi Jiang, Qian Chen, Siqi Zheng, and Zhou Zhao. 2025. Wavtokenizer: an efficient acoustic discrete codec tokenizer for audio language modeling. In *The Thirteenth International Conference on Learning Representations*.

Wei Kang, Xiaoyu Yang, Zengwei Yao, Fangjun Kuang, Yifan Yang, Liyong Guo, Long Lin, and Daniel Povey. 2024. Libriheavy: A 50,000 hours asr corpus with punctuation casing and context. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 10991–10995. IEEE.

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

Kushal Lakhotia, Evgeny Kharitonov, Wei-Ning Hsu, Yossi Adi, Adam Polyak, Benjamin Bolte, Tu-Anh Nguyen, Jade Copet, Alexei Baevski, Adelrahman Mohamed, and Emmanuel Dupoux. 2021. Generative spoken language modeling from raw audio. *Preprint*, arXiv:2102.01192.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. 2023. Flow matching for generative modeling. *Preprint*, arXiv:2210.02747.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. *Preprint*, arXiv:1711.05101.

Fabian Mentzer, David Minnen, Eirikur Agustsson, and Michael Tschannen. 2023. Finite scalar quantization: Vq-vae made simple. *Preprint*, arXiv:2309.15505.

Arsha Nagrani, Joon Son Chung, Weidi Xie, and Andrew Zisserman. 2020. Voxceleb: Large-scale speaker verification in the wild. *Computer Speech & Language*, 60:101027.

Tu Anh Nguyen, Benjamin Muller, Bokai Yu, Marta R. Costa-jussa, Maha Elbayad, Sravya Popuri, Christophe Ropers, Paul-Ambroise Duquenne, Robin Algayres, Ruslan Mavlyutov, Itai Gat, Mary Williamson, Gabriel Synnaeve, Juan Pino, Benoit Sagot, and Emmanuel Dupoux. 2024. Spirit lm: Interleaved spoken and written language model. *Preprint*, arXiv:2402.05755.

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.

William Peebles and Saining Xie. 2023. Scalable diffusion models with transformers. *Preprint*, arXiv:2212.09748.

Takaaki Saeki, Detai Xin, Wataru Nakata, Tomoki Koriyama, Shinnosuke Takamichi, and Hiroshi Saruwatari. 2022. Utmos: Utokyo-sarulab system for voicemos challenge 2022. *Preprint*, arXiv:2204.02152.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. 2023. Roformer: Enhanced transformer with rotary position embedding. *Preprint*, arXiv:2104.09864.

Marco Tagliasacchi, Yunpeng Li, Karolis Misiunas, and Dominik Roblek. 2020. Seanet: A multi-modal speech enhancement network. *Preprint*, arXiv:2009.02095.

Dehua Tao, Daxin Tan, Yu Ting Yeung, Xiao Chen, and Tan Lee. 2024. Toneunit: A speech discretization approach for tonal language speech synthesis. *arXiv preprint arXiv:2406.08989*.

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

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

Zehui Yang, Yifan Chen, Lei Luo, Runyan Yang, Lingxuan Ye, Gaofeng Cheng, Ji Xu, Yaohui Jin, Qingqing Zhang, Pengyuan Zhang, and 1 others. 2022. Open source magicdata-ramc: A rich annotated mandarin conversational (ramc) speech dataset. *arXiv preprint arXiv:2203.16844*.

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

Binbin Zhang, Hang Lv, Pengcheng Guo, Qijie Shao, Chao Yang, Lei Xie, Xin Xu, Hui Bu, Xiaoyu Chen, Chenchen Zeng, and 1 others. 2022. Wenetspeech: A 10000+ hours multi-domain mandarin corpus for speech recognition. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6182–6186. IEEE.

Dong Zhang, Gang Wang, Jinlong Xue, Kai Fang, Liang Zhao, Rui Ma, Shuhuai Ren, Shuo Liu, Tao Guo, Weiji Zhuang, and 1 others. 2025. Mimo-audio: Audio language models are few-shot learners. *arXiv preprint arXiv:2512.23808*.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023a. Adding conditional control to text-to-image diffusion models. *Preprint*, arXiv:2302.05543.

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

A Model Architecture

A.1 SEANetEncoder

For the SEANetEncoder, we employ a four-layer architecture with dimensions of [512, 1024, 1024, 1024]. The downsampling ratios are set to [2, 1, 1] and [2, 2, 1] for acoustic token rates of 50Hz and 25Hz, respectively.

A.2 SEANetDecoder

For the SEANetDecoder, we utilize a four-layer architecture with dimensions of [1024, 1024, 1024, 1024]. The upsampling ratios are set to [1, 1, 2] and [1, 2, 2] for acoustic token rates of 50Hz and 25Hz, respectively.

A.3 DiT blocks

For the DSA-Tokenizer, we utilize a stack of 22 DiT blocks with a hidden dimension of 1024. We incorporate AdaNorm (Fan et al., 2021) for both multi-head self-attention and multi-head cross-attention mechanisms. The FFN consists of two linear layers with an inner dimension set to 4096. And RoPE(Su et al., 2023) is set as the position embedding for both multi-head self-attention and multi-head cross-attention.

A.3.1 ControlNet-Style CNN Adapter

For the ControlNet-Style CNN Adapter, we employ a lightweight stack of three 1D convolutional layers to extract contextual information while projecting the feature dimension from 512 to 1024, as the embedding dim of semantic token is 512. Each layer utilizes a kernel size of 3 and a padding of 1 to preserve the temporal sequence length.

A.4 Attention pooling layer

Let $e_a \in \mathbb{R}^{T\times D}$ denote the input acoustic sequence, where T is the sequence length and $D = 1024$ is the feature dimension. To aggregate the variable-length sequence into a fixed-size embedding, we employ a Masked Attentive Statistics Pooling (ASP) layer followed by a projection adapter.

First, the input sequence is transposed to $\mathbf{X} \in \mathbb{R}^{D\times T}$. A channel-wise attention mechanism computes the importance scores $\mathbf{S} \in \mathbb{R}^{D\times T}$ for each time step and feature dimension:

$$\mathbf{S} = \mathbf{W}_2 \left(\tanh\left(\mathbf{W}_1\mathbf{X} + \mathbf{b}_1\right)\right) + \mathbf{b}_2,$$

where $\mathbf{W}_1 \in \mathbb{R}^{D_{attn}\times D}$, $\mathbf{W}_2 \in \mathbb{R}^{D\times D_{attn}}$ are convolution weights with kernel size 1, and D_{attn}

is the bottleneck dimension, $\mathbf{b_1} \in \mathbb{R}^D$ and $\mathbf{b}_2 \in \mathbb{R}^D$ are bias weights.

To handle variable sequence lengths, we apply a temporal mask $M \in \{0, -\infty\}^T$ based on the valid length of the sequence. The normalized attention weights $\boldsymbol{\alpha} \in \mathbb{R}^{D \times T}$ are obtained via a masked softmax operation along the temporal dimension:

$$\alpha_{d,t} = \frac{\exp(S_{d,t} + M_t)}{\sum_{\tau=1}^{T} \exp(S_{d,\tau} + M_\tau)},$$

where $d \in [1, D]$ and $t \in [1, T]$.

Using these weights, we calculate the weighted mean vector $\boldsymbol{\mu} \in \mathbb{R}^D$ and the weighted standard deviation vector $\boldsymbol{\sigma} \in \mathbb{R}^D$:

$$\boldsymbol{\mu}_d = \sum_{t=1}^{T} \alpha_{d,t} \cdot X_{d,t},$$

$$\boldsymbol{\sigma}_d = \sqrt{\sum_{t=1}^{T} \alpha_{d,t} \cdot (X_{d,t} - \boldsymbol{\mu}_d)^2 + \epsilon}$$

where ϵ is a small constant for numerical stability.

Finally, the statistics are concatenated and projected to the target dimension $D_{out} = 256$:

$$\mathbf{h} = \mathbf{W}_{proj}[\boldsymbol{\mu}; \boldsymbol{\sigma}] + \mathbf{b}_{proj},$$

where $[\cdot;\cdot]$ denotes concatenation resulting in a $2D$-dimensional vector (2048), and $\mathbf{W}_{proj} \in \mathbb{R}^{D_{out} \times 2D}$, $\mathbf{b}_{proj} \in \mathbb{R}^{2D}$. The resulting $\mathbf{h}$ serves as the pooled acoustic token embeddings

B Dataset Cleaning

We find that some samples of Emilia Dataset (He et al., 2024) contains more than two speakers. As we have Recombination (Contextual Inpainting) Mode, samples contains more than two speakers will hurt the performance of our model, therefore, we utilize speaker-diarization model [2] (Bredin and Laurent, 2021) to remove those dirty samples.

C Training Details

C.1 Training Details of DSA-Tokenizer

All DSA-Tokenizer are trained witicdynamic batching with 30k frames of speech per batch. We train the tokenizer for 400k steps using AdamW (Loshchilov and Hutter, 2019) with a learning rate of 7.5×10^{-5} and 32k warmup steps. The dataset used for the training of the semantic tokenizer is sampled from LibriSpeech (Panayotov et al., 2015), GigaSpeech (Chen et al., 2021), Libri-Heavy (Kang et al., 2024), CommonVoice (Ardila et al., 2020), AISHELL-2 (Du et al., 2018), WenetSpeech (Zhang et al., 2022), MagicData-RAMC (Yang et al., 2022), and People's Speech (Galvez et al., 2021). The DSA-Tokenizer has a parameter size of approximately 430M. We trained the models on the Ascend platform for 15 days.

C.2 Training Details of Disentanglement Probing

In the disentanglement probing experiments, tokens extracted from the tokenizers are first mapped to 128-dimensional embeddings, followed by processing via a two-layer CNN and a 256-dimensional bidirectional LSTM. For the scenario where multi-layer tokens are used as input, the token embeddings of all layers are first summed before being processed by the CNN and LSTM.

For the ASR experiment, English character sequences (including space symbols) are used as training targets, and the sequential output of the LSTM is optimized against these sequences using CTC loss. In the speaker classification experiment, the LSTM output is aggregated by a pooling layer and then mapped to speaker IDs through an MLP layer, with optimization performed using cross-entropy loss.

For dataset configurations: the LibriSpeech train-960 subset is used for training, dev-clean for validation, and test-clean for testing in the ASR experiment. In the speaker classification task, the VoxCeleb1 train, dev, and test subsets serve as the training, validation, and testing sets, respectively. All models are trained for 30 epochs, with the model achieving the lowest validation WER (for ASR) or the highest speaker classification accuracy (for speaker classification) selected for testing and metric calculation. The LibriSpeech training set contains nearly 1000 hours of speech data, while VoxCeleb1 includes the same 1251 speakers in both its training and testing sets.

C.3 Training Details of LLM-based Voice Cloning

In LLM-based voice cloning tasks, Qwen3-0.6B serves as the LLM backbone. Speech symbols that depend on the tokenizer's codebook size and number of layers are first integrated into the LLM's vocabulary, with adjustments made to the LLM's

[2] https://huggingface.co/pyannote/speaker-diarization

embedding table and prediction head. We train the LLM using the supervised fine-tuning (SFT) paradigm, where only the loss of the response part is computed. The model is trained for 5 epochs with a learning rate of 1e-5 and a warmup ratio of 0.1.

D Baseline Details

Four representative types of baseline models are selected for fair comparison: WavTokenizer (single-layer, non-decoupled), SAC and DualCodec (dual-branch, decoupled), Mimi and Encodec (multi-layer, non-decoupled), and SpeechTokenizer (multi-layer, decoupled) Cosy2 S^3 (ASR-supervised captures semantics).

WavTokenizer is built on the VQ-GAN framework and utilizes a single VQ layer with a codebook size of 4096. Operating at a frame rate of 75 Hz, it achieves a bitrate of 0.9 kbps.

SAC is also built on the VQ-GAN framework. It uses a pretrained semantic tokenizer, while utilizing the DAC framework for the acoustic stream with a total frame rate of 62.5 Hz. The semantic codebook size is 16,384 and the acoustic codebook size is 16,384, achieving a bitrate of 0.875 kbps.

DualCodec is a semantic-enhanced tokenizer that directly encodes semantic-rich SSL features (Chung et al., 2021) into its first layer. In our experimental setting, for a fair comparison, we select the model with six codebook layers that operates at the frame rate of 12.5Hz. The first codebook contains 16,384 entries, while the remaining five each contain 4,096 entries, achieving a bitrate of 0.925 kbps.

Mimi utilizes features from WavLM for semantic distillation. It employs eight codebooks, each of size 2,048, at a 12.5 Hz frame rate, resulting in a bitrate of 1.1 kbps.

Encodec is an RVQ-based neural audio codec operating at a frame rate of 75 Hz. In our experimental setting, for a fair comparison, we use only the first two codebook layers, yielding a bitrate of 1.5 kbps.

SpeechTokenizer shares a similar architecture with Mini but enhances the semantic richness of the first-layer tokens via semantic distillation using HuBERT. In our experimental setting, for a fair comparison, we use only the first two codebook layers, yielding a bitrate of 1 kbps.

Model	ASR WER (%) ↓	SC ACC (%) ↑
Cosy2 S^3 Tokenizer	33.72	14.50
WavTokenizer	86.31	18.14
EnCodec		
Layer 0	64.85	16.28
Layer 1-7	69.74	17.68
SpeechTokenizer		
Layer 0	17.78	12.08
Layer 1-7	21.93	25.80
SAC		
Semantic token	9.31	6.57
Acoustic token	106.24	21.82
DSA-Tokenizer		
Semantic token	6.28	2.35
Acoustic token	127.31	24.94

Table 4: Disentanglement probing evaluation results

CosyVoice2 S^3 Tokenizer inserts an FSQ module into the encoder of the SenseVoice-Large ASR model (An et al., 2024), which discretizes intermediate continuous representations into discrete tokens with textual supervision. This FSQ module consists of a single-layer FSQ codebook with 8 channels and 3 levels per channel, resulting in a total of 6561 code entries. With a token rate of 25 Hz, this tokenizer achieves a bitrate of approximately 0.317 kbps.

E Disentanglement Probing Result

The detailed result of disentanglement probing is listed in Table 4.

F Evaluation tools

The UTMOS score is computed using the pretrained model available at [3]. Speaker similarity (SIM) scores are calculated with a speaker encoder based on WavLM, fine-tuned for the speaker verification task1. For English speech, the word error rate (WER) is evaluated with Whisper-large-V3[4], and for Chinese speech, the character error rate (CER) is evaluated with Paraformer[5].

G Broader Impact

This work introduces a novel dual-stream audio tokenizer capable of disentangling semantic and

[3] https://huggingface.co/spaces/sarulab-speech/UTMOS-demo/tree/main

[4] https://huggingface.co/openai/whisper-large-v3

[5] https://huggingface.co/funasr/paraformer-zh

acoustic information. Our research holds significant potential for positive societal impact, particularly in the development of Speech LLMs. The disentangled representation offers fine-grained control for speech. However, the capability to clone voices with high similarity raises concerns regarding misuse for deepfakes, unauthorized voice impersonation, and misinformation.To prevent abuse, we recommend developing robust deepfake detection tools, speech watermarking, and clear reporting mechanisms.

H License Discussion

The speaker-diarization model(Bredin and Laurent, 2021) and the model[3] used to calculate the UTMOS score are released under the MIT license. The speaker verification fine-tuned WavLM1 is released under the CC BY-SA 3.0 license. Whisper-large-V3[4] is released under the Apache 2.0 license. And the model[5] calculating CER is released under model-license license.

The Emilia dataset, LibriSpeech, MagicData-RAMC and LibriTTS are under the license of cc-by-4.0. GigaSpeech, Libri-Heavy, AISHELL-2, WenetSpeech are released under the Apache 2.0 License. CommonVoice is released under the MPL-2.0 license. MagicData-RAMC and VoxCeleb1 are under the license CC BY-NC-ND 4.0. People's Speech is under the license of CC-BY-SA and CC-BY 4.0.

I AI Usage

We used LLMs for grammatical checking and polishing in Sections 1, section 5 and the Appendix.