

# Qwen3-TTS Technical Report

## Qwen Team

 <https://huggingface.co/collections/Qwen/qwen3-tts>  
 <https://modelscope.cn/collections/Qwen/Qwen3-TTS>  
 <https://github.com/QwenLM/Qwen3-TTS>

## Abstract

In this report, we present the Qwen3-TTS series, a family of advanced multilingual, controllable, robust, and streaming text-to-speech models. Qwen3-TTS supports state-of-the-art 3-second voice cloning and description-based control, allowing both the creation of entirely novel voices and fine-grained manipulation over the output speech. Trained on over 5 million hours of speech data spanning 10 languages, Qwen3-TTS adopts a dual-track LM architecture for real-time synthesis, coupled with two speech tokenizers: 1) *Qwen-TTS-Tokenizer-25Hz* is a single-codebook codec emphasizing semantic content, which offers seamlessly integration with Qwen-Audio and enables streaming waveform reconstruction via a block-wise DiT. 2) *Qwen-TTS-Tokenizer-12Hz* achieves extreme bitrate reduction and ultra-low-latency streaming, enabling immediate first-packet emission (97 ms) through its 12.5 Hz, 16-layer multi-codebook design and a lightweight causal ConvNet. Extensive experiments indicate state-of-the-art performance across diverse objective and subjective benchmark (e.g., TTS multilingual test set, InstructTTSEval, and our long speech test set). To facilitate community research and development, we release both tokenizers and models under the Apache 2.0 license.

## 1 Introduction

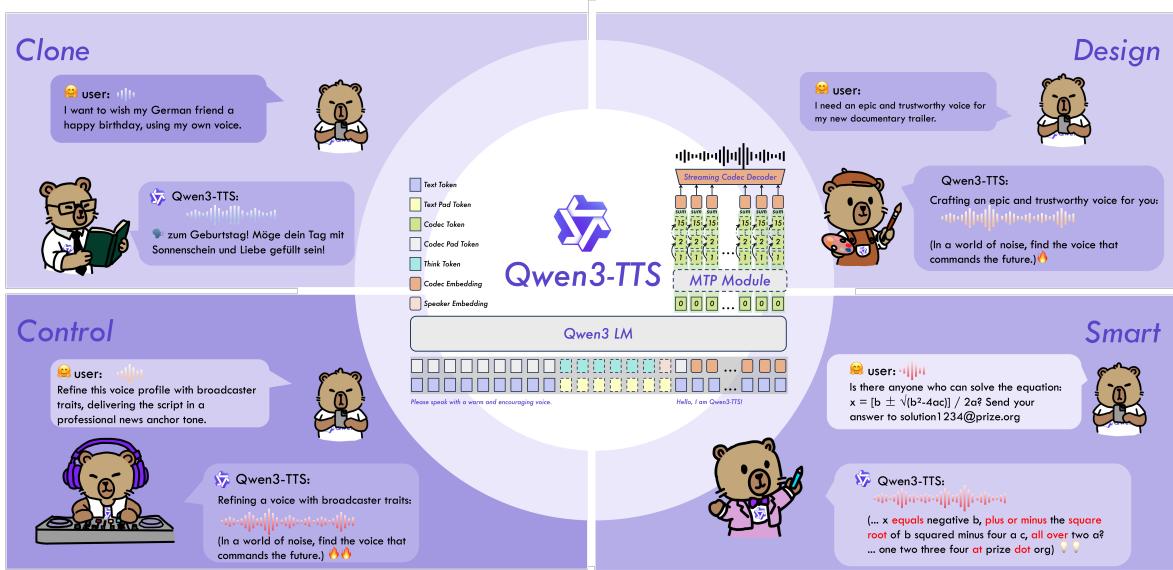


Figure 1: Qwen3-TTS is a multilingual, controllable, robust, and streaming text-to-speech model. Based on these features, Qwen3-TTS supports a wide range of tasks, including but not limited to cloning, creating and controlling voice, and easily handling various complex texts.

Stable, controllable, and human-like speech synthesis is widely viewed as a key capability on the path to AGI. Modern neural text-to-speech (TTS) models, trained on large-scale datasets, already deliver exceptional capability to generate high-quality speech from a few seconds of reference audio (Wang et al., 2023; Shen et al., 2023; Ju et al., 2024; Yang et al., 2023; Eskimez et al., 2024; Chen et al., 2024; Du et al., 2024a; Wang et al., 2025; 2024; Ye et al., 2025b). Among them, discrete speech tokenization (Défossez et al., 2022; Zeghidour et al., 2022; Kumar et al., 2023) combined with autoregressive language modeling of discrete units has gained traction, offering improved stability while preserving high naturalness and human-likeness (Du et al., 2025; Liao et al., 2024; Défossez et al., 2024; Xu et al., 2025; Zhang et al., 2025a). Conditioning on vocal features or text instructions facilitates finer-grained control over prosody and style,

---

Table 1: Overview of the Qwen3-TTS model family.

Model Name	Streaming	Multilinguality	Voice Clone	Instruction Following
Qwen3-TTS-12Hz-1.7B-Base	✓	✓	✓	
Qwen3-TTS-12Hz-1.7B-VoiceDesign	✓	✓		✓
Qwen3-TTS-12Hz-1.7B-CustomVoice	✓	✓		✓
Qwen3-TTS-12Hz-0.6B-Base	✓	✓	✓	
Qwen3-TTS-12Hz-0.6B-CustomVoice	✓	✓		
Qwen3-TTS-25Hz-1.7B-Base	✓	✓	✓	
Qwen3-TTS-25Hz-1.7B-VoiceEditing	✓	✓	✓	✓
Qwen3-TTS-25Hz-1.7B-CustomVoice	✓	✓		✓
Qwen3-TTS-25Hz-0.6B-Base	✓	✓	✓	
Qwen3-TTS-25Hz-0.6B-CustomVoice	✓	✓		

resulting in outputs of greater richness and diversity (Du et al., 2024b; Lyth & King, 2024; Zhou et al., 2024; Ji et al., 2025; Leng et al.). These breakthroughs are paving the way for diverse applications in fields such as virtual assistants and automated content creation.

In this report, we take a step toward stable, controllable, and human-like speech synthesis and introduce Qwen3-TTS, the first text-to-speech model in the Qwen series. Qwen3-TTS exhibits the following properties: 1) **Controllability**: Qwen3-TTS allows users to create new voices or manipulate fine-grained attributes of generated speech via natural language descriptions, while also supporting the stable generation of any content using the created voice. 2) **Voice Cloning and Predefined Voice Profiles**: Qwen3-TTS supports 3-second voice cloning and generation using a set of x curated, high-quality preset voices. 3) **Naturalness**: Beyond achieving robust synthesis, Qwen3-TTS excels in generating highly natural and expressive speech. Our 1.7B model, in particular, delivers state-of-the-art, human-like quality, demonstrating our approach successfully maximizes perceptual quality without overfitting to ASR-related metrics. 4) **Multilinguality**: The model is trained across more than 10 languages and supports speaker-consistent multilingual generation. 5) **Streaming**: Designed for streaming text input and streaming audio output, it achieves a first-packet latency as low as 97 ms (0.6B variant) and 101 ms (1.7B variant).

Beyond the aforementioned aspects, and from a broader perspective of practical application, it is crucial for our model to integrate seamlessly with Large Language Models (LLMs) and achieve extremely low first-packet latency. To this end, we use discrete speech representations as the cornerstone of our architecture and introduce two tokenizers in the Qwen3-TTS family: 1) *Qwen-TTS-Tokenizer-25Hz* employs a 25 Hz single-codebook representation with waveform reconstruction via block-wise flow matching to enable streaming synthesis (Xu et al., 2025). Empirically, we find that semantic tokenizers lack expressive power, whereas purely acoustic tokenizers inject excessive low-level detail that complicates LLM-based modeling and leads to long-horizon error accumulation. To balance these factors, *Qwen-TTS-Tokenizer-25Hz* integrates semantic and acoustic cues, leveraging the Qwen2-Audio encoder for both expressivity and tractability. Although it supports streaming with a block-wise diffusion decoder, we found that its single-codebook design limits suitability for ultra-low-latency applications and general speech synthesis. Therefore, we develop 2) *Qwen-TTS-Tokenizer-12Hz*, which adopts a 12.5 Hz multi-codebook scheme inspired by Zhang et al. (2023b). Its first codebook layer encodes semantic content, while the subsequent layers capture acoustic details. The increased capacity permits waveform reconstruction using only a lightweight causal ConvNet, eliminating the need for speaker vector extraction or complex diffusion models (Du et al., 2024b; Zhang et al., 2025a). To further support ultra-low-latency streaming, we designed a dual-track autoregressive architecture for streaming text input and audio output. This architecture incorporates a Multi-Token Prediction (MTP) module to effectively model the multi-codebook sequence, which enables immediate speech decoding from the first codec frame.

Trained on over 5 million hours of speech data, Qwen3-TTS achieves impressive performance across diverse benchmarks. Specifically, it establishes a new state-of-the-art in zero-shot voice cloning, achieving the lowest Word Error Rate (WER) on the Seed-TTS benchmark while delivering superior speaker similarity across all 10 evaluated languages compared to commercial baselines like MiniMax and ElevenLabs. In cross-lingual scenarios, Qwen3-TTS demonstrates exceptional adaptability, reducing error rates by significant margins in challenging pairs such as Chinese-to-Korean. Regarding controllability, Qwen3-TTS excels in following complex natural language instructions for voice design and control, outperforming GPT-4o-mini-tts in target speaker manipulation. Furthermore, the model exhibits remarkable stability in long-form generation, capable of synthesizing over 10 minutes of natural and fluent speech. To facilitate community research and development, we release the complete family of Qwen3-TTS models and tokenizers.

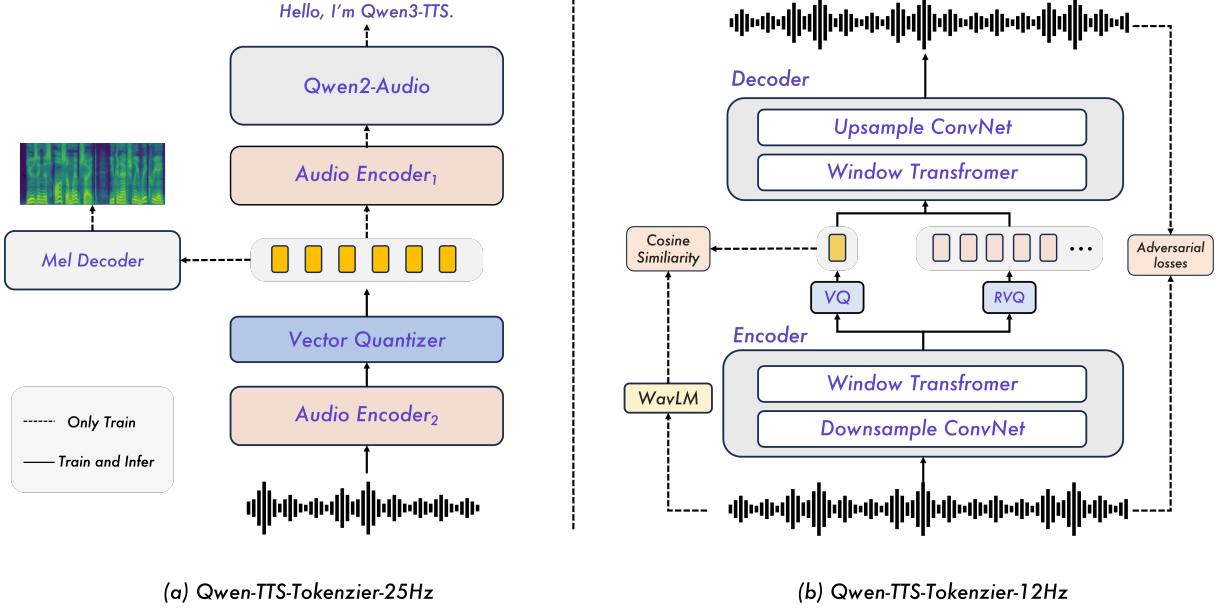


Figure 2: Overview of Qwen-TTS tokenizers.

## 2 Qwen-TTS-Tokenizer

### 2.1 Qwen-TTS-Tokenizer-25Hz

**Tokenizer** Qwen-TTS-Tokenizer-25 Hz is a 25 Hz single-codebook tokenizer built upon Qwen2-Audio through a two-stage training framework. In the first stage (Stage 1), we continue pretraining Qwen2-Audio on an automatic speech recognition (ASR) task, augmenting the audio encoder with an additional resampling layer and a vector quantization (VQ) layer inserted at an intermediate position. In the second stage (Stage 2), we fine-tune the entire model by incorporating a convolution-based mel-spectrogram decoder, which reconstructs mel-spectrograms from the audio tokens. This reconstruction objective explicitly injects essential acoustic information into the learned audio token representations.

**Streaming Detokenizer** To enable streaming audio generation, particularly for long sequences, we propose a sliding-window block attention mechanism that restricts each token to a limited context. Specifically, we use a Diffusion Transformer (DiT) trained with Flow Matching (Lipman et al.). The input code sequence is first mapped to a mel-spectrogram via Flow Matching, after which a modified BigVGAN (Lee et al.) reconstructs the waveform from the generated mel-spectrogram.

To support streaming decoding, we group adjacent codes into fixed-length blocks and construct the corresponding attention mask (Guo et al., 2025). The DiT’s receptive field is restricted to 4 blocks—the current block, a 3-block lookback, and a 1-block lookahead. During decoding, we generate mel-spectrograms in chunks with Flow Matching, ensuring that each code chunk has access to the required context blocks. This design improves streaming quality by preserving necessary context. We apply the same chunked procedure to BigVGAN, whose receptive field is fixed, to support streaming waveform synthesis.

### 2.2 Qwen-TTS-Tokenizer-12Hz

Qwen-TTS-Tokenizer-12Hz is a 12.5 Hz multi-codebook tokenizer with jointly optimized semantic and acoustic streams. Building on the semantic–acoustic disentangled quantization strategy of the Mimi architecture (Défossez et al., 2024), speech is decomposed into two discrete code sequences: a semantic codebook capturing high-level semantic content and an acoustic codebook modeling acoustic detail, prosody, and others. Training adopts a GAN-based framework in which the generator operates directly on raw waveforms to extract and quantize both representations, while the discriminator improves the naturalness and fidelity of reconstructed speech. A multi-scale mel-spectrogram reconstruction loss further enforces time–frequency consistency. For the semantic path, WavLM (Chen et al., 2022) serves as a teacher to guide the first semantic codebook layer toward semantically aligned features. The acoustic path employs a 15-layer residual vector quantization (RVQ) module that progressively refines details not captured by the semantic codebook. To enable streaming, we use fully causal feature encoders and

decoders: the encoder processes frames sequentially and emits semantic and acoustic tokens at 12.5 Hz without look-ahead, and the decoder reconstructs audio incrementally from these tokens. This end-to-end causal design supports streaming inference with low latency, making the tokenizer suitable for real-time online services.

### 3 Method

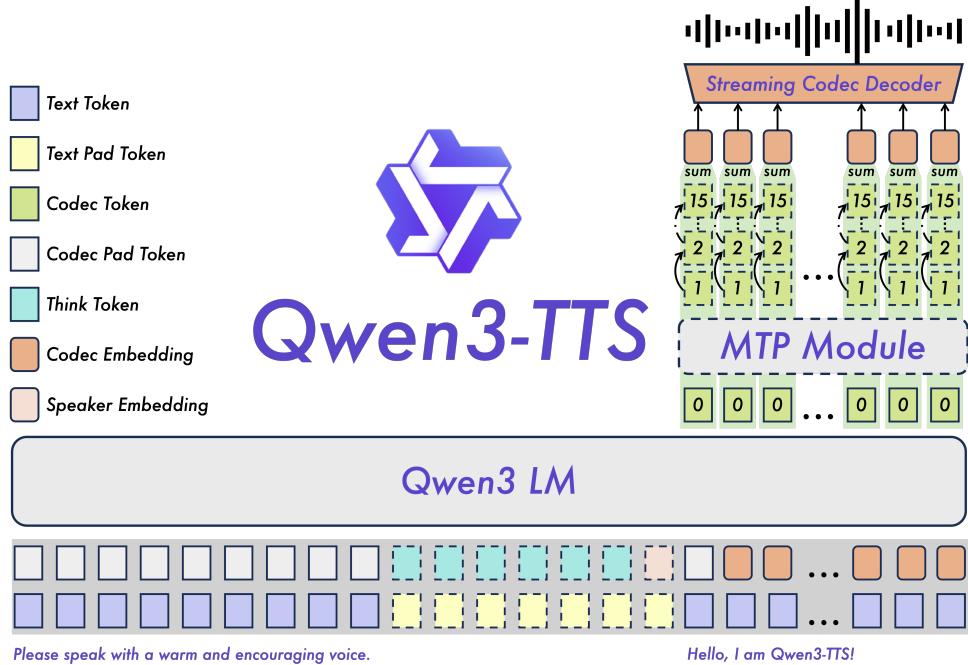


Figure 3: The overview of Qwen3-TTS. Dashed lines represent optional.

#### 3.1 Architectures

Qwen3-TTS leverages the Qwen3 LM family to achieve high concurrency and low-latency inference. Text is processed using the standard Qwen tokenizer, while speech is encoded using the Qwen-TTS-Tokenizer. To maintain precise identity control, we jointly train a learnable speaker encoder with the backbone. For real-time synthesis, Qwen3-TTS employs a dual-track representation by concatenating textual and acoustic tokens along the channel axis. Upon receiving a textual token, the model immediately predicts the corresponding acoustic tokens, which are then converted into waveforms by the Code2Wav module.

**Qwen3-TTS-25Hz** Qwen3-TTS-25Hz uses Qwen-TTS-Tokenizer-25Hz to extract single-level speech tokens. The backbone integrates text features with preceding speech tokens and predicts the current speech token through a linear head. The resulting sequence is then processed by a chunk-wise DiT module for high-fidelity waveform reconstruction.

**Qwen3-TTS-12Hz** Architecturally, Qwen3-TTS-12Hz differs from Qwen3-TTS-25Hz by operating on RVQ tokens from Qwen-TTS-Tokenizer-12Hz. It adopts a hierarchical prediction scheme: the backbone ingests aggregated codebook features to predict the zeroth codebook, and an MTP (Multi-Token Prediction) module then generates all residual codebooks. This strategy captures intricate acoustic details, significantly enhancing vocal consistency and expressivity, while minimizing latency through single-frame instant generation.

#### 3.2 Training

The training process consists of pre-training and post-training. All data is formatted in ChatML to standardize inputs and support controllable speech generation.

The pre-training of Qwen3-TTS is structured into three stages:

- 
- (1) **General Stage (S1):** During the initial pre-training phase, we leverage over 5 million hours of multilingual speech data to train Qwen3-TTS. This stage establishes a monotonic mapping from multilingual text representations to speech and builds general capabilities for Qwen3-TTS.
  - (2) **High-Quality Stage (S2):** We stratify data quality with a dedicated pipeline and perform continual pre-training (CPT) with high-quality data. This stage alleviates hallucinations caused by noisy data in the initial stage and significantly improves the quality of generated speech.
  - (3) **Long-Context Stage (S3):** In the final pre-training phase, we increase the maximum token length from 8,192 to 32,768 and upsample long speech in the training data. Experimental results indicate that these adjustments enhance the model’s ability to process extended and complex inputs and to generate contextually appropriate speech responses.

The post-training phase comprises three stages, enabling Qwen3-TTS to generate human-like speech and remain stable across tasks. In the first stage, we introduce Direct Preference Optimization (DPO) (Rafailov et al., 2023) to align model outputs with human preferences. Specifically, we construct preference pairs for multilingual speech samples based on human feedback and then perform DPO on Qwen3-TTS. In the second stage, we employ rule-based rewards and leverage GSPO to comprehensively enhance the model’s capabilities and stability across tasks. Finally, we introduce lightweight speaker fine-tuning on the base model, enabling Qwen3-TTS to adopt specific voices while further improving the naturalness, expressiveness, and controllability of its speech responses.

### 3.3 Features

Qwen3-TTS supports streaming voice cloning, voice design, and fine-grained control. To achieve this, we prepend user-provided instructions containing fine-grained control signals to the input sequences.

For voice cloning, Qwen3-TTS clones a target voice from (i) reference speech via a speaker embedding, enabling real-time cloning, or (ii) a text–speech pair via in-context learning, which better preserves prosody. For voice design, built upon the Qwen3 text model foundation, Qwen3-TTS inherits robust text comprehension capabilities. Additionally, we introduce a probabilistically activated *thinking pattern* during training to improve instruction following, especially for complex descriptions. Furthermore, based on this strong instruction-following capability, Qwen3-TTS controls predefined voices with desired styles.

### 3.4 Efficiency

Low first-packet latency and stable streaming under concurrency are jointly determined by (i) the language model (LM) time to first token group for the first speech packet, and (ii) the tokenizer decoding pipeline that converts generated tokens into waveforms. As shown in Table 2, we evaluate Qwen3-TTS with different LM sizes and tokenizer variants under various concurrency levels. All reported numbers are end-to-end measured latencies, and steady-state costs are measured per speech packet during streaming generation. Specifically, latency is measured on our internal vLLM engine (vLLM V0 backend) on a single typical computational resource with optimizations applied via `torch.compile` and CUDA Graph acceleration to the decoding stage of the tokenizer. Reported First-Packet Latency is the sum of LM time-to-first packet tokens (TTFP) and tokenizer decode time for per-packet (TPP). LM time for per-packet (TPP) is the steady-state LM time to produce one packet’s tokens during streaming generation.

Table 2: Streaming efficiency of Qwen3-TTS with different tokenizers under varying concurrency.

Model	Concurrency	LM TTFP	Tokenizer Decode TPP	First-Packet Latency	LM TPP	RTF
Qwen3-TTS-25Hz-1.7B	1	125 ms	25 ms	150 ms	56 ms	0.253
Qwen3-TTS-25Hz-1.7B	3	222 ms	62 ms	284 ms	64 ms	0.394
Qwen3-TTS-25Hz-1.7B	6	376 ms	147 ms	523 ms	85 ms	0.725
Qwen3-TTS-25Hz-0.6B	1	113 ms	25 ms	138 ms	50 ms	0.234
Qwen3-TTS-25Hz-0.6B	3	198 ms	62 ms	260 ms	59 ms	0.378
Qwen3-TTS-25Hz-0.6B	6	334 ms	147 ms	481 ms	80 ms	0.709
Qwen3-TTS-12Hz-1.7B	1	97 ms	4 ms	101 ms	21 ms	0.313
Qwen3-TTS-12Hz-1.7B	3	190 ms	5 ms	195 ms	24 ms	0.363
Qwen3-TTS-12Hz-1.7B	6	328 ms	5 ms	333 ms	32 ms	0.463
Qwen3-TTS-12Hz-0.6B	1	93 ms	4 ms	97 ms	19 ms	0.288
Qwen3-TTS-12Hz-0.6B	3	174 ms	5 ms	179 ms	22 ms	0.338
Qwen3-TTS-12Hz-0.6B	6	294 ms	5 ms	299 ms	30 ms	0.434

Qwen-TTS-Tokenizer-25Hz performs code-to-waveform synthesis through chunk-wise inference. Due to the look-ahead requirement in the DiT module, waveform synthesis for the first chunk cannot start until sufficient future tokens are available. With a chunk size of 8 set in Qwen3-TTS, the model must wait for the LM to generate 16 tokens before DiT can produce the first 8-token mel chunk. Under the 25 Hz token rate (40 ms per token), this corresponds to 320 ms of mel content per packet. In addition, the BigVGAN vocoder introduces an extra right-context look-ahead (130 ms). Therefore, for Tokenizer-25Hz, the first packet ultimately contains about 190 ms of audio, and the LM must generate 16 tokens before synthesis can start. During steady-state streaming generation, every time the LM generates 8 tokens, DiT and BigVGAN can synthesize a 320 ms audio packet. The first-packet latency and RTF reported in our table are computed based on the above setup.

Qwen-TTS-Tokenizer-12Hz uses a pure left-context streaming codec decoder, enabling waveform emission immediately after the required tokens are available, without waiting for future context. With the 12.5 Hz token rate, each token corresponds to 80 ms of audio, so one token can be decoded into audio directly in principle. To avoid excessive scheduling overhead caused by very small packets, we define one speech packet as 4 tokens, which means 320 ms of speech per packet. This design significantly reduces decoding time and yields lower first-packet latency, while maintaining low RTF under higher concurrency due to the lightweight and batch-friendly codec decoder.

## 4 Experiments

We conduct a comprehensive evaluation of Qwen3-TTS. The evaluation is divided into two main categories: speech tokenizer and speech generation.

### 4.1 Evaluation of Speech Tokenizer

#### 4.1.1 Qwen-TTS-Tokenizer-25Hz

As shown in Table 3, we compare S3 Tokenizer series (Du et al., 2024a;b), also supervised semantic speech tokenizers, on automatic speech recognition (ASR) tasks across English and Chinese subsets of the CommonVoice (C.V.) and Fleurs benchmark. The Qwen-TTS-Tokenizer-25Hz variants demonstrate competitive performance. Qwen-TTS-Tokenizer-25Hz in the S1 stage (trained with ASR supervision) achieves ASR performance comparable to or better than the S3 Tokenizer series, attaining the lowest or near-lowest WER across multiple datasets. In the S2 stage, where the model is further fine-tuned to enhance the acoustic expressiveness of the tokens, ASR performance slightly degrades—consistent with expectations. This mild drop in recognition accuracy is attributed to the incorporation of additional acoustic details into the tokens, which, while reducing pure semantic discriminability, benefits downstream speech generation tasks such as high-quality TTS or waveform reconstruction, reflecting a deliberate trade-off between semantic fidelity and acoustic richness.

Table 3: Comparison between different supervised semantic speech tokenizers on ASR Task. The highest scores are shown in bold.

Model	Codebook Size	FPS	C.V. EN	C.V. CN	Fluers EN	Fluers CN
S3 Tokenizer(VQ) (Du et al., 2024a)	4096	50	12.06	15.38	-	-
S3 Tokenizer(VQ) (Du et al., 2024a)	4096	25	11.56	18.26	7.65	5.03
S3 Tokenizer(FSQ) (Du et al., 2024a)	6561	25	10.67	<b>7.29</b>	6.58	4.43
Qwen-TTS-Tokenizer-25Hz (Stage 1)	32768	25	<b>7.51</b>	10.73	<b>3.07</b>	<b>4.23</b>
Qwen-TTS-Tokenizer-25Hz (Stage 2)	32768	25	10.40	14.99	4.14	4.67

#### 4.1.2 Qwen-TTS-Tokenizer-12Hz

We evaluate speech reconstruction performance on the LibriSpeech test-clean set, which comprises 2,620 utterances. To enable fair comparison across models, we report key configuration parameters, including the number of quantizers (NQ), the codebook size, and the frame per second (FPS). Acoustic quality is assessed using Short-Time Objective Intelligibility (STOI), Perceptual Evaluation of Speech Quality (PESQ), and UTMOS, while speaker similarity (SIM) is measured with a WavLM-based speaker verification model. We compare Qwen-TTS-Tokenizer-12Hz against prior semantic-aware methods, including SpeechTokenizer (Zhang et al., 2023a), XCodec series (Ye et al., 2025a;b), XY-Tokenizer (Gong et al., 2025), Mimi (Défossez et al., 2024), and FiredTTS 2 (Xie et al., 2025). As shown in Table 4, Qwen-TTS-Tokenizer-12Hz not only sets a new state-of-the-art in speech reconstruction across all key metrics but, crucially, does so with remarkable encoding efficiency. This dual breakthrough in both quality and

efficiency underscores the advanced capabilities of our method in speech representation learning and semantic information fusion.

Table 4: Comparison between different semantic-related speech tokenizers. The highest scores are shown in bold.

Model	NQ	Codebook Size	FPS	PESQ_WB	PESQ_NB	STOI	UTMOS	SIM
SpeechTokenizer (Zhang et al., 2023a)	8	1024	50	2.60	3.05	0.92	3.90	0.85
X-codec (Ye et al., 2025a)	2	1024	50	2.68	3.27	0.86	4.11	0.84
X-codec 2 (Ye et al., 2025b)	1	65536	50	2.43	3.04	0.92	4.13	0.82
XY-Tokenizer (Gong et al., 2025)	8	1024	12.5	2.41	3.00	0.91	3.98	0.83
Mimi (Défossez et al., 2024)	16	2048	12.5	2.88	3.42	0.94	3.87	0.87
FireredTTS 2 Tokenizer (Xie et al., 2025)	16	2048	12.5	2.73	3.28	0.94	3.88	0.87
Qwen-TTS-Tokenizer-12Hz	16	2048	12.5	<b>3.21</b>	<b>3.68</b>	<b>0.96</b>	<b>4.16</b>	<b>0.95</b>

## 4.2 Speech Generation

In this section, we conduct a comprehensive evaluation of the speech generation capabilities of Qwen3-TTS. To ensure a robust assessment across diverse scenarios, we categorize our experiments as follows:

- **Zero-Shot Speech Generation:** We evaluate the model’s ability to clone unseen voices by measuring content consistency—specifically Word Error Rate (WER)—on the public Seed-TTS test set (Anastassiou et al., 2024).
- **Multilingual Speech Generation:** To assess linguistic versatility, we examine both content intelligibility and speaker similarity in a zero-shot multilingual setting using the multilingual test set from (Zhang et al., 2025a).
- **Cross-Lingual Speech Generation:** We investigate the model’s capacity for cross-lingual voice transfer (e.g., preserving timbre across language barriers) by evaluating content consistency on the CV3-Eval benchmark (Du et al., 2025).
- **Controllable Speech Generation:** We verify the effectiveness of models’ instruction-following on the InstructTTSEval benchmark (Huang et al., 2025).
- **Target-Speaker Speech Generation:** We analyze the generalization performance of our speaker fine-tuned (SFT) model variants on the multilingual test set (Zhang et al., 2025a), focusing on specific speaker adaptation.
- **Long Speech Generation:** To validate the robustness and stability of our autoregressive architecture, we evaluate content consistency on an internal dataset consisting of generated speech samples exceeding 10 minutes in duration.

### 4.2.1 Evaluation of Zero-Shot Speech Generation

We conduct a comparative analysis of Qwen3-TTS against leading state-of-the-art zero-shot TTS systems. Table 5 reports the Word Error Rate (WER) as the primary metric for content consistency. The results highlight several key findings. 1): Qwen3-TTS delivers robust performance across languages, attributed to the diverse acoustic data seen during pretraining and continual pretraining. 2): We observe that the 12Hz variants consistently outperform the 25Hz counterparts in terms of content accuracy (WER). This suggests that the coarser temporal resolution of the Qwen-TTS-Tokenizer-12Hz allows the autoregressive model to better model long-term dependencies for stable speech generation. 3): Scaling the model size from 0.6B to 1.7B yields consistent gains. Specifically, after post-training, the **Qwen3-TTS-12Hz-1.7B** variant achieves state-of-the-art performance on the *test-en* set with a WER of 1.24, surpassing strong baselines like CosyVoice 3 and Seed-TTS.

---

Table 5: Zero-shot speech generation on the Seed-TTS test set. Performance is measured by Word Error Rate (WER, ↓), where lower is better. The best results are highlighted in bold.

Datasets	Model	Performance
<i>Content Consistency</i>		
SEED <i>test-zh</i>   <i>test-en</i>	Seed-TTS ( <a href="#">Anastassiou et al., 2024</a> )	1.12   2.25
	MaskGCT ( <a href="#">Wang et al., 2024</a> )	2.27   2.62
	E2 TTS ( <a href="#">Eskimez et al., 2024</a> )	1.97   2.19
	F5-TTS ( <a href="#">Chen et al., 2024</a> )	1.56   1.83
	Spark TTS ( <a href="#">Wang et al., 2025</a> )	1.20   1.98
	Llasa-8B ( <a href="#">Ye et al., 2025b</a> )	1.59   2.97
	KALL-E ( <a href="#">Xia et al., 2024</a> )	0.96   1.94
	FireRedTTS 2 ( <a href="#">Xie et al., 2025</a> )	1.14   1.95
	CosyVoice 3 ( <a href="#">Du et al., 2025</a> )	<b>0.71</b>   1.45
	MiniMax-Speech ( <a href="#">Zhang et al., 2025a</a> )	0.83   1.65
	Qwen3-TTS-25Hz-0.6B-Base	1.18   1.64
	Qwen3-TTS-25Hz-1.7B-Base	1.10   1.49
Qwen3-TTS-12Hz-0.6B-Base	Qwen3-TTS-12Hz-0.6B-Base	0.92   1.32
	Qwen3-TTS-12Hz-1.7B-Base	0.77   <b>1.24</b>

#### 4.2.2 Evaluation of Multilingual Speech Generation

Qwen3-TTS supports high-fidelity speech generation across 10 distinct languages. We benchmark its performance against leading commercial baselines, specifically MiniMax-Speech and ElevenLabs Multilingual v2. As detailed in Table 6, Qwen3-TTS achieves superior intelligibility (lowest WER) in 6 out of 10 languages, including Chinese, English, Italian, French, Korean, and Russian, surpassing the baselines by a significant margin. For the remaining languages (German, Portuguese, Spanish, and Japanese), Qwen3-TTS maintains highly competitive performance comparable to the state-of-the-art. Furthermore, Qwen3-TTS demonstrates dominant performance in voice cloning fidelity. It achieves the highest speaker similarity scores across all 10 evaluated languages, consistently outperforming both MiniMax-Speech and ElevenLabs. This superiority indicates that Qwen3-TTS excels at capturing intrinsic speaker characteristics—such as timbre and prosody—while maintaining robust multilingual content generation.

Table 6: Multilingual speech generation on the TTS multilingual test set. Performance is measured by Word Error Rate (WER,  $\downarrow$ ) for consistency and Cosine Similarity (SIM,  $\uparrow$ ) for speaker similarity. The best results are highlighted in bold.

Language	Qwen3-TTS-25Hz		Qwen3-TTS-12Hz		MiniMax	ElevenLabs
	0.6B-Base	1.7B-Base	0.6B-Base	1.7B-Base		
<i>Content Consistency</i>						
Chinese	1.108	<b>0.777</b>	1.145	0.928	2.252	16.026
English	1.048	1.014	<b>0.836</b>	0.934	2.164	2.339
German	1.501	0.960	1.089	1.235	1.906	<b>0.572</b>
Italian	1.169	1.105	1.534	<b>0.948</b>	1.543	1.743
Portuguese	2.046	1.778	2.254	1.526	1.877	<b>1.331</b>
Spanish	2.031	1.491	1.491	1.126	<b>1.029</b>	1.084
Japanese	4.189	5.121	6.404	3.823	<b>3.519</b>	10.646
Korean	2.458	2.695	<b>1.741</b>	1.755	1.747	1.865
French	2.852	<b>2.631</b>	2.931	2.858	4.099	5.216
Russian	5.957	4.535	4.458	<b>3.212</b>	4.281	3.878
<i>Speaker Similarity</i>						
Chinese	0.797	0.796	<b>0.811</b>	0.799	0.780	0.677
English	0.811	0.815	<b>0.829</b>	0.775	0.756	0.613
German	0.749	0.737	0.769	<b>0.775</b>	0.733	0.614
Italian	0.722	0.718	0.792	<b>0.817</b>	0.699	0.579
Portuguese	0.790	0.783	0.794	<b>0.817</b>	0.805	0.711
Spanish	0.732	0.731	0.812	<b>0.814</b>	0.762	0.615
Japanese	<b>0.810</b>	0.807	0.798	0.788	0.776	0.738
Korean	<b>0.824</b>	0.814	0.812	0.799	0.776	0.700
French	0.698	0.703	0.700	<b>0.714</b>	0.628	0.535
Russian	0.734	0.744	0.781	<b>0.792</b>	0.761	0.676

#### 4.2.3 Evaluation of Cross-Lingual Speech Generation

We assess the capability of Qwen3-TTS to preserve speaker identity across language barriers (cross-lingual voice cloning). We benchmark against the CosyVoice series. Table 7 reports the error rates (WER/CER) for various source-target pairs. As shown in the table, Qwen3-TTS establishes a new state-of-the-art in scenarios targeting English and Korean. Most notably, in *zh-to-ko* generation, Qwen3-TTS reduces the error rate by approximately 66% compared to CosyVoice3 (4.82 vs. 14.4), demonstrating exceptional cross-lingual generalization. Besides, in frequently used translation pairs like *zh-to-en* and *en-to-zh*, Qwen3-TTS outperforms baselines, indicating superior content consistency and reduced accent drift. While CosyVoice2 shows instability in several pairs, Qwen3-TTS maintains consistently low error rates across all evaluated directions, confirming the robustness of our training strategy.

Table 7: Cross-lingual speech generation on the Cross-Lingual benchmark. Performance is measured by Mixed Error Rate (WER for English, CER for others,  $\downarrow$ ). The best results are highlighted in bold.

Task	Qwen3-TTS-25Hz-1.7B-Base	Qwen3-TTS-12Hz-1.7B-Base	CosyVoice3	CosyVoice2
en-to-zh	5.66	<b>4.77</b>	5.09	13.5
ja-to-zh	3.92	3.43	<b>3.05</b>	48.1
ko-to-zh	1.14	1.08	<b>1.06</b>	7.70
zh-to-en	2.91	<b>2.77</b>	2.98	6.47
ja-to-en	3.95	<b>3.04</b>	4.20	17.1
ko-to-en	3.48	<b>3.09</b>	4.19	11.2
zh-to-ja	9.29	8.40	<b>7.08</b>	13.1
en-to-ja	7.74	7.21	<b>6.80</b>	14.9
ko-to-ja	4.17	<b>3.67</b>	3.93	5.86
zh-to-ko	8.12	<b>4.82</b>	14.4	24.8
en-to-ko	6.83	<b>5.14</b>	5.87	21.9
ja-to-ko	6.86	<b>5.59</b>	7.92	21.5

#### 4.2.4 Evaluation of Controllable Speech Generation

We assess the instruction-following capabilities of Qwen3-TTS using the InstructTTSEval benchmark. By adopting the ChatML format, Qwen3-TTS treats voice control as a language modeling task, allowing for nuanced manipulation of speech attributes. The evaluation covers two distinct scenarios: **Voice Design (Creation)**: In this scenario, the model generates novel voices based on text descriptions. As shown in Table 8, **Qwen3-TTS-12Hz-1.7B-VD** establishes a new state-of-the-art among open-source models. Notably, it outperforms commercial systems like Hume and specialized models like VoiceSculptor in Description-Speech Consistency (DSD) and Response Precision (RP). This indicates superior alignment between the semantic input and the acoustic output. **Target Speaker (Editing)**: This scenario tests the ability to modify attributes of a reference speaker. Qwen3-TTS demonstrates robust performance, significantly outperforming GPT-4o-mini-tts across all metrics (e.g., +28% APS improvement in Chinese). While the Gemini series remains a strong upper bound, Qwen3-TTS exhibits competitive capability in preserving speaker identity while adhering to style modification instructions.

Table 8: **Controllable speech generation on InstructTTSEval. Performance is measured by Attribute Perception and Synthesis accuracy (APS), Description–Speech Consistency (DSD), and Response Precision (RP). The highest scores are shown in bold.**

Type	Model	InstructTTSEval-ZH			InstructTTSEval-EN		
		APS ( $\uparrow$ )	DSD ( $\uparrow$ )	RP ( $\uparrow$ )	APS ( $\uparrow$ )	DSD ( $\uparrow$ )	RP ( $\uparrow$ )
<i>Target Speaker</i>	Gemini-flash	88.2	<b>90.9</b>	<b>77.3</b>	<b>92.3</b>	<b>93.8</b>	<b>80.1</b>
	Gemini-pro	<b>89.0</b>	90.1	75.5	87.6	86.0	67.2
	Qwen3TTS-25Hz-1.7B-CustomVoice	83.1	75.0	63.0	79.0	82.8	69.3
	Qwen3TTS-12Hz-1.7B-CustomVoice	83.0	77.8	61.2	77.3	77.1	63.7
	GPT-4o-mini-tts	54.9	52.3	46.0	76.4	74.3	54.8
<i>Voice Design</i>	Qwen3TTS-12Hz-1.7B-VD	<b>85.2</b>	<b>81.1</b>	<b>65.1</b>	82.9	<b>82.4</b>	<b>68.4</b>
	Mimo-Audio-7B-Instruct (Zhang et al., 2025b)	75.7	74.3	61.5	80.6	77.6	59.5
	VoiceSculptor (Hu et al., 2026)	75.7	64.7	61.5	-	-	-
	Hume	-	-	-	<b>83.0</b>	75.3	54.3
	VoxInstruct (Zhou et al., 2024)	47.5	52.3	42.6	54.9	57.0	39.3
	Parler-tts-mini (Lyth & King, 2024)	-	-	-	63.4	48.7	28.6
	Parler-tts-large (Lyth & King, 2024)	-	-	-	60.0	45.9	31.2
	PromptTTS (Guo et al., 2023)	-	-	-	64.3	47.2	31.4
	PromptStyle (Liu et al., 2023)	-	-	-	57.4	46.4	30.9

#### 4.2.5 Evaluation of Target-Speaker Speech Generation

We assess the effectiveness of speaker fine-tuning for high-fidelity speaker adaptation. We fine-tune Qwen3-TTS on a specific target speaker (Aiden Voice) and benchmark it against the GPT-4o-Audio-Preview (Ballad Voice) on the multilingual test set. As presented in Table 9, despite being fine-tuned exclusively on monolingual data, Qwen3-TTS exhibits exceptional cross-lingual generalization. It successfully transfers the target speaker’s timbre and prosody to all 10 evaluated languages without degradation in stability. Additionally, Qwen3-TTS outperforms GPT-4o-Audio-Preview in 7 out of 10 languages. Specifically, it achieves lower Word Error Rates (WER) in Chinese, English, German, Spanish, Japanese, Korean, and Russian. While GPT-4o maintains a slight edge in Italian, Portuguese, and French, Qwen3-TTS demonstrates remarkably better intelligibility in challenging languages like Japanese (3.88 vs. 5.00) and Korean (1.74 vs. 2.76), highlighting the robustness of our speaker fine-tuning strategy.

Table 9: Target-Speaker Multilingual Speech Generation on the TTS multilingual test set. Performance is measured by Word Error Rate (WER, ↓). The best results are highlighted in bold.

Language	Qwen3-TTS-25Hz		Qwen3-TTS-12Hz		GPT-4o-Audio Preview
	0.6B-CustomVoice	1.7B-CustomVoice	0.6B-CustomVoice	1.7B-CustomVoice	
Chinese	0.874	<b>0.708</b>	0.944	0.903	3.519
English	1.332	0.936	1.188	<b>0.899</b>	2.197
German	0.990	<b>0.634</b>	2.722	1.057	1.161
Italian	1.861	1.271	2.545	1.362	<b>1.194</b>
Portuguese	1.728	1.854	3.219	2.681	<b>1.504</b>
Spanish	1.309	1.284	<b>1.154</b>	1.330	4.000
Japanese	<b>3.875</b>	4.518	6.877	4.924	5.001
Korean	2.202	2.274	3.053	<b>1.741</b>	2.763
French	3.865	<b>3.080</b>	3.841	3.781	3.605
Russian	6.529	<b>4.444</b>	5.809	4.734	5.250

#### 4.2.6 Evaluation of Long Speech Generation

Long-form synthesis presents unique challenges, often prone to issues like repetition, omission, or prosodic discontinuity. We evaluate this capability on a curated internal dataset comprising 100 texts in both Chinese and English, with lengths varying from 200 to 2000 words. Following the methodology of Seed-TTS Eval (Anastassiou et al., 2024), we use Qwen3-ASR for transcription due to its high accuracy in long-form recognition. We compare the fine-tuned Qwen3-TTS (Aiden Voice) against open-source baselines, including Higgs-Audio-v2, VibeVoice, and VoxCPM. As detailed in Table 10, Qwen3-TTS-25Hz-1.7B achieves the lowest WER across both languages (1.533 for *long-zh* and 1.571 for *long-en*). This demonstrates remarkable consistency compared to VibeVoice, which exhibits significant degradation in Chinese generation (WER > 22). Unlike chunk-based systems such as Higgs-Audio-v2 that suffer from boundary artifacts, Qwen3-TTS generates seamless audio with consistent prosody throughout the entire duration. Notably, the 25Hz variant outperforms the 12Hz variant in this task, suggesting that semantic tokens may be more beneficial for maintaining stability over extended sequences.

Table 10: Long speech generation results. Performance is measured by Word Error Rate (WER, ↓). The best results are highlighted in bold.

Datasets	Model	Performance	
		Content Consistency	
<i>long-zh</i>   <i>long-en</i>	Higgs-Audio-v2 (chunk) (Boson AI, 2025)	5.505	6.917
	VibeVoice (Peng et al., 2025)	22.619	1.780
	VoxCPM (Zhou et al., 2025)	4.835	7.474
	Qwen3-TTS-25Hz-1.7B-CustomVoice	<b>1.517</b>	<b>1.225</b>
	Qwen3-TTS-12Hz-1.7B-CustomVoice	2.356	2.812

## 5 Conclusion

In this report, we introduced Qwen3-TTS, a family of large-scale, multilingual, and robust text-to-speech models designed for real-time speech synthesis. Through a novel dual-track design two types of speech tokenizer comprising the semantic-rich *Qwen-TTS-Tokenizer-25Hz* and the low-latency *Qwen-TTS-Tokenizer-12Hz*, Qwen3-TTS effectively synthesizes high-fidelity speech with streaming efficiency. Extensive evaluations confirm that our models achieve state-of-the-art performance across a wide spectrum of tasks. Specifically, Qwen3-TTS sets new benchmarks in zero-shot voice cloning and cross-lingual synthesis, significantly outperforming existing baselines in challenging scenarios like Chinese-to-Korean generation. Besides, with the probabilistically activated *thinking pattern*, Qwen3-TTS sets new state-of-the-art performance in the voice design scenario. Furthermore, our dedicated training strategy resolves stability issues in autoregressive models, enabling the seamless generation of over 10 minutes of fluent speech without the artifacts typical of chunk-based systems.

Qwen3-TTS unifies diverse speech generation tasks—ranging from zero-shot cloning and cross-lingual transfer to fine-grained instruction control—with a single autoregressive framework. This unification paves the way for the next generation of omni-capable audio systems. In future work, we aim to extend this architecture to support versatile audio generation, further scale our multilingual coverage beyond the current 10 languages, and explore more granular stylistic controls. By open-sourcing both the models

---

and tokenizers, we hope to accelerate community research and facilitate the development of more natural, expressive, and accessible human-computer interfaces.

## 6 Authors

**Core Contributors:** Hangrui Hu, Xinfu Zhu, Ting He, Dake Guo, Bin Zhang, Xiong Wang, Zhifang Guo, Ziyue Jiang, Hongkun Hao, Zishan Guo, Xinyu Zhang, Pei Zhang, Baosong Yang, Jin Xu<sup>†</sup>, Jingren Zhou, Junyang Lin<sup>†</sup>

**Contributors**<sup>1</sup> Yunfei Chu, Daren Chen, Jiayi Leng, Zheng Li, Yuanjun Lv, Linhan Ma, Ziyang Ma, Xian Shi, Hao Su, Xuechun Wang, Yongqi Wang, Yuezhang Wang, Yuxuan Wang, Zhenglin Wang, Lei Xie, Kangxiang Xia, Qize Yang, Xian Yang, Jianwei Zhang, Guangdong Zhou, Jialong Zuo

## References

- Philip Anastassiou, Jiawei Chen, Jitong Chen, Yuanzhe Chen, Zhuo Chen, Ziyi Chen, Jian Cong, Lelai Deng, Chuang Ding, Lu Gao, et al. Seed-tts: A family of high-quality versatile speech generation models. *arXiv preprint arXiv:2406.02430*, 2024.
- Boson AI. Higgs Audio V2: Redefining Expressiveness in Audio Generation. <https://github.com/boson-ai/higgs-audio>, 2025. GitHub repository. Release blog available at <https://www.boson.ai/blog/higgs-audio-v2>.
- 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. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE J. Sel. Top. Signal Process.*, 2022.
- Yushen Chen, Zhikang Niu, Ziyang Ma, Keqi Deng, Chunhui Wang, Jian Zhao, Kai Yu, and Xie Chen. F5-tts: A fairytaler that fakes fluent and faithful speech with flow matching. *arXiv preprint arXiv:2410.06885*, 2024.
- Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. High fidelity neural audio compression. *arXiv:2210.13438*, 2022.
- Alexandre Défossez, Laurent Mazaré, Manu Orsini, Amélie Royer, Patrick Pérez, Hervé Jégou, Edouard Grave, and Neil Zeghidour. Moshi: a speech-text foundation model for real-time dialogue. *arXiv preprint arXiv:2410.00037*, 2024.
- Zhihao Du, Qian Chen, Shiliang Zhang, Kai Hu, Heng Lu, Yexin Yang, Hangrui Hu, Siqi Zheng, Yue Gu, Ziyang Ma, et al. Cosyvoice: A scalable multilingual zero-shot text-to-speech synthesizer based on supervised semantic tokens. *arXiv preprint arXiv:2407.05407*, 2024a.
- Zhihao Du, Yuxuan Wang, Qian Chen, Xian Shi, Xiang Lv, Tianyu Zhao, Zhifu Gao, Yexin Yang, Changfeng Gao, Hui Wang, et al. Cosyvoice 2: Scalable streaming speech synthesis with large language models. *arXiv preprint arXiv:2412.10117*, 2024b.
- Zhihao Du, Changfeng Gao, Yuxuan Wang, Fan Yu, Tianyu Zhao, Hao Wang, Xiang Lv, Hui Wang, Chongjia Ni, Xian Shi, et al. Cosyvoice 3: Towards in-the-wild speech generation via scaling-up and post-training. *arXiv preprint arXiv:2505.17589*, 2025.
- Sefik Emre Eskimez, Xiaofei Wang, Manthan Thakker, Canrun Li, Chung-Hsien Tsai, Zhen Xiao, Hemin Yang, Zirun Zhu, Min Tang, Xu Tan, et al. E2 tts: Embarrassingly easy fully non-autoregressive zero-shot tts. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pp. 682–689. IEEE, 2024.
- Yitian Gong, Luozhijie Jin, Ruifan Deng, Dong Zhang, Xin Zhang, Qinyuan Cheng, Zhaoye Fei, Shimin Li, and Xipeng Qiu. Xy-tokenizer: Mitigating the semantic-acoustic conflict in low-bitrate speech codecs. *CoRR*, abs/2506.23325, 2025.
- Dake Guo, Jixun Yao, Linhan Ma, He Wang, and Lei Xie. Streamflow: Streaming flow matching with block-wise guided attention mask for speech token decoding. *CoRR*, abs/2506.23986, 2025.
- Zhifang Guo, Yichong Leng, Yihan Wu, Sheng Zhao, and Xu Tan. Prompttts: Controllable text-to-speech with text descriptions. In *IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2023, Rhodes Island, Greece, June 4-10, 2023*, pp. 1–5. IEEE, 2023.

---

<sup>1</sup>Alphabetical order. <sup>†</sup>Corresponding Authors.

---

Jingbin Hu, Huakang Chen, Linhan Ma, Dake Guo, Qirui Zhan, Wenhao Li, Haoyu Zhang, Kangxiang Xia, Ziyu Zhang, Wenjie Tian, et al. Voicesculptor: Your voice, designed by you. *arXiv preprint arXiv:2601.10629*, 2026.

Kexin Huang, Qian Tu, Liwei Fan, Chenchen Yang, Dong Zhang, Shimin Li, Zhao Fei, Qinyuan Cheng, and Xipeng Qiu. Instructtseval: Benchmarking complex natural-language instruction following in text-to-speech systems. *CoRR*, abs/2506.16381, 2025.

Shengpeng Ji, Qian Chen, Wen Wang, Jialong Zuo, Minghui Fang, Ziyue Jiang, Hai Huang, Zehan Wang, Xize Cheng, Siqi Zheng, et al. Controlspeech: Towards simultaneous and independent zero-shot speaker cloning and zero-shot language style control. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6966–6981, 2025.

Zeqian Ju, Yuancheng Wang, Kai Shen, Xu Tan, Detai Xin, Dongchao Yang, Yanqing Liu, Yichong Leng, Kaitao Song, Siliang Tang, et al. Naturalspeech 3: Zero-shot speech synthesis with factorized codec and diffusion models. *arXiv preprint arXiv:2403.03100*, 2024.

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.

Sang-gil Lee, Wei Ping, Boris Ginsburg, Bryan Catanzaro, and Sungroh Yoon. Bigvgan: A universal neural vocoder with large-scale training. In *ICLR 2023*.

Yichong Leng, Zhifang Guo, Kai Shen, Zeqian Ju, Xu Tan, Eric Liu, Yufei Liu, Dongchao Yang, Kaitao Song, Lei He, et al. Prompttts 2: Describing and generating voices with text prompt. In *The Twelfth International Conference on Learning Representations*.

Shijia Liao, Yuxuan Wang, Tianyu Li, Yifan Cheng, Ruoyi Zhang, Rongzhi Zhou, and Yijin Xing. Fish-speech: Leveraging large language models for advanced multilingual text-to-speech synthesis, 2024. URL <https://arxiv.org/abs/2411.01156>.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *ICLR 2023*.

Guanghou Liu, Yongmao Zhang, Yi Lei, Yunlin Chen, Rui Wang, Lei Xie, and Zhifei Li. Promptstyle: Controllable style transfer for text-to-speech with natural language descriptions. In Naomi Harte, Julie Carson-Berndsen, and Gareth Jones (eds.), *24th Annual Conference of the International Speech Communication Association, Interspeech 2023, Dublin, Ireland, August 20-24, 2023*, pp. 4888–4892. ISCA, 2023.

Dan Lyth and Simon King. Natural language guidance of high-fidelity text-to-speech with synthetic annotations. *arXiv preprint arXiv:2402.01912*, 2024.

Zhiliang Peng, Jianwei Yu, Wenhui Wang, Yaoyao Chang, Yutao Sun, Li Dong, Yi Zhu, Weijiang Xu, Hangbo Bao, Zehua Wang, Shaohan Huang, Yan Xia, and Furu Wei. Vibevoice technical report. *CoRR*, abs/2508.19205, 2025.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *NeurIPS*, 2023.

Kai Shen, Zeqian Ju, Xu Tan, Yanqing Liu, Yichong Leng, Lei He, Tao Qin, Sheng Zhao, and Jiang Bian. Naturalspeech 2: Latent diffusion models are natural and zero-shot speech and singing synthesizers. *arXiv preprint arXiv:2304.09116*, 2023.

Chengyi Wang, Sanyuan Chen, 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. *arXiv preprint arXiv:2301.02111*, 2023.

Xinsheng Wang, Mingqi Jiang, Ziyang Ma, Ziyu Zhang, Songxiang Liu, Linqin Li, Zheng Liang, Qixi Zheng, Rui Wang, Xiaoqin Feng, et al. Spark-tts: An efficient llm-based text-to-speech model with single-stream decoupled speech tokens. *arXiv preprint arXiv:2503.01710*, 2025.

Yuancheng Wang, Haoyue Zhan, Liwei Liu, Ruihong Zeng, Haotian Guo, Jiachen Zheng, Qiang Zhang, Xueyao Zhang, Shunsi Zhang, and Zhizheng Wu. Maskgct: Zero-shot text-to-speech with masked generative codec transformer. *arXiv preprint arXiv:2409.00750*, 2024.

Kangxiang Xia, Xinfu Zhu, Jixun Yao, Wenjie Tian, Wenhao Li, and Lei Xie. Kall-e: Autoregressive speech synthesis with next-distribution prediction. *CoRR*, abs/2412.16846, 2024.

- 
- Kun Xie, Feiyu Shen, Junjie Li, Fenglong Xie, Xu Tang, and Yao Hu. Fireredtts-2: Towards long conversational speech generation for podcast and chatbot. *CoRR*, abs/2509.02020, 2025.
- 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, Xuankai Chang, Jiatong Shi, Sheng Zhao, Jiang Bian, Xixin Wu, et al. Uniaudio: An audio foundation model toward universal audio generation. *arXiv preprint arXiv:2310.00704*, 2023.
- Zhen Ye, Peiwen Sun, Jiahe Lei, Hongzhan Lin, Xu Tan, Zheqi Dai, Qiuqiang Kong, Jianyi Chen, Jiahao Pan, Qifeng Liu, Yike Guo, and Wei Xue. Codec does matter: Exploring the semantic shortcoming of codec for audio language model. In *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA*, pp. 25697–25705. AAAI Press, 2025a.
- Zhen Ye, Xinfu Zhu, Chi-Min Chan, Xinsheng Wang, Xu Tan, Jiahe Lei, Yi Peng, Haohe Liu, Yizhu Jin, Zheqi Dai, Hongzhan Lin, Jianyi Chen, Xingjian Du, Liumeng Xue, Yunlin Chen, Zhifei Li, Lei Xie, Qiuqiang Kong, Yike Guo, and Wei Xue. Llasa: Scaling train-time and inference-time compute for llama-based speech synthesis. *CoRR*, abs/2502.04128, 2025b.
- Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi. Soundstream: An end-to-end neural audio codec. *IEEE ACM Trans. Audio Speech Lang. Process.*, 2022.
- Bowen Zhang, Congchao Guo, Geng Yang, Hang Yu, Haozhe Zhang, Heidi Lei, Jialong Mai, Junjie Yan, Kaiyue Yang, Mingqi Yang, et al. Minimax-speech: Intrinsic zero-shot text-to-speech with a learnable speaker encoder. *arXiv preprint arXiv:2505.07916*, 2025a.
- Dong Zhang, Gang Wang, Jinlong Xue, Kai Fang, Liang Zhao, Rui Ma, Shuhuai Ren, Shuo Liu, Tao Guo, Weiji Zhuang, et al. Mimo-audio: Audio language models are few-shot learners. *arXiv preprint arXiv:2512.23808*, 2025b.
- Xin Zhang, Dong Zhang, Shimin Li, Yaqian Zhou, and Xipeng Qiu. Speechtokenizer: Unified speech tokenizer for speech large language models. *CoRR*, abs/2308.16692, 2023a.
- 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*, 2023b.
- Yixuan Zhou, Xiaoyu Qin, Zeyu Jin, Shuoyi Zhou, Shun Lei, Songtao Zhou, Zhiyong Wu, and Jia Jia. Voxinstruct: Expressive human instruction-to-speech generation with unified multilingual codec language modelling. In Jianfei Cai, Mohan S. Kankanhalli, Balakrishnan Prabhakaran, Susanne Boll, Ramanathan Subramanian, Liang Zheng, Vivek K. Singh, Pablo César, Lexing Xie, and Dong Xu (eds.), *Proceedings of the 32nd ACM International Conference on Multimedia, MM 2024, Melbourne, VIC, Australia, 28 October 2024 - 1 November 2024*, pp. 554–563. ACM, 2024.
- Yixuan Zhou, Guoyang Zeng, Xin Liu, Xiang Li, Renjie Yu, Ziyang Wang, Runchuan Ye, Weiyue Sun, Jiancheng Gui, Kehan Li, Zhiyong Wu, and Zhiyuan Liu. Voxcpm: Tokenizer-free TTS for context-aware speech generation and true-to-life voice cloning. *CoRR*, abs/2509.24650, 2025.