BUILDING AUDIO SCENES WITH VOICE DESIGN AND CLONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Building audio scenes requires jointly generating multi-speaker speech, ambience, music, and sound effects. This task requires consistent and expressive speech coordinated with the surrounding audio. Speech can be generated through voice design from a scene description or through voice cloning from reference audio. However, the scarcity of high-quality annotated data and the difficulty of supporting both tasks while jointly generating multiple audio modalities further complicate audio scene generation. Therefore, we first propose VoxCaps, which cleans media data, expands it with targeted synthetic data, and uses a persona library to annotate diverse multi-level scene descriptions. Then, we introduce VoxVAE, which pairs a local anti-aliased encoder with a high-capacity decoder to provide a shared continuous representation for multiple audio modalities. Next, we propose VoxDirector, an expressive multi-speaker speech and surrounding audio generation model that supports both voice design and cloning. VoxDirector adopts reward-conditioned quality control to improve generation quality, and a Dual-Router MoE tailored to the two tasks and multiple audio modalities. Experimental results show that VoxDirector outperforms the baselines on key metrics for both voice design and cloning, while supporting the generation of complex audio scenes. Demos are available at `https://voxdirector.github.io`.

1 INTRODUCTION

Recent advances in text-to-speech (TTS) have enabled increasingly flexible generation of natural and expressive speech. Zero-shot systems transfer a voice from reference audio to new content and have expanded to long-form and multi-speaker generation (Du et al., 2025; Zhu et al., 2025a). Meanwhile, controllable TTS lets users specify attributes such as timbre, age, emotion, and speaking rate (Qwen Team, 2026a; Hu et al., 2026). However, emerging media applications, like animation, short drama, and advertising, demand richer output with multiple characters, ambience, music, and sound effects, as an audio scene. Such scenes require voices to remain consistent across turns, performances to fit each character and context, and speech to interact naturally with the surrounding audio.

Extending TTS to audio scene generation requires precise control at multiple levels. For speech, this calls for TTS that supports both voice design and voice cloning. When suitable reference audio is unavailable, voice design creates a reusable character voice based on the **Speakers** description, which specifies attributes such as timbre, age, and persona. When **reference audio** is provided, voice cloning should carry that identity into new lines in multi-speaker scenes. In addition, the **Environment** description specifies the recording space, reverberation, and persistent ambience (Vyas et al., 2023; Liu et al., 2023). The **Content** description, in turn, can specify local style, including what each speaker says, how they say it, and the accompanying local sound effects (Huang et al., 2025; Hu et al., 2026). A cascade can also implement this workflow, but it risks mismatches in timing, reverberation, and acoustic texture across dialogue, ambience, and sound effects. Audio scene generation therefore calls for a unified model that supports voice design and voice cloning while jointly generating expressive multi-speaker speech and surrounding audio.

Overall, this task presents three main challenges. **(1) Multi-level data annotation.** Public speech corpora support large-scale TTS training (Zen et al., 2019; Chen et al., 2021; Kang et al., 2024), but audio scenes require broader coverage of styles and surrounding audio (Xu et al., 2024). Moreover, audio scene data need multi-level annotations to distinguish audio modalities and separate

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

global characteristics from temporal ones. **(2) Control and quality across two tasks.** Voice design obtains speaker information from descriptions and voice cloning from reference audio, which are substantially different input modalities. Maintaining speaker consistency, instruction following, and acoustic quality across both tasks within a single model remains challenging (Qwen Team, 2026a; Leng et al., 2024). **(3) Heterogeneous audio modalities.** Speech, ambience, music, and transient sound effects have distinct temporal and spectral structures (Vyas et al., 2023), yet an audio scene places them on the same timeline. Building a shared representation and generator that preserve intelligible pronunciation, recognizable timbre, harmonic detail, and sharp transients while keeping the sequence compact for long-range scene modeling remains an open problem (Yang et al., 2023).

To address these challenges, we propose **VoxDirector**, a multi-speaker expressive TTS and surrounding audio generation model that supports both voice design and cloning. To obtain multi-level scene descriptions at this granularity, we construct **VoxCaps** from media audio. Its coverage design stage combines diverse media, along with targeted synthetic subsets to cover scarce styles. We also introduce a persona library to enrich speaker descriptions of recurring character archetypes in advertising, animation, and short drama. Then, to support generation across multiple audio modalities, we design **VoxVAE**. VoxVAE pairs a local anti-aliased encoder with a high-capacity decoder and shapes the posterior mean with latent-prediction and acoustic-readout objectives. This asymmetric design improves reconstruction while providing the flow Transformer with a compact, generation-oriented 25-Hz continuous representation. Finally, VoxDirector integrates reward-conditioned quality control into its flow Transformer, preserving style coverage from mixed-quality data while steering generation toward high-quality audio (Emmons et al., 2022). To support both tasks and diverse audio modalities, we design the Dual-Router MoE with a task and an audio router. We then use curriculum learning, moving from simpler tasks to complete scene generation (Bengio et al., 2009).

In reconstruction experiments, VoxVAE leads on key speech and general audio metrics and remains competitive on music. We then evaluate VoxDirector on zero-shot voice cloning in monologue and dialogue, and on voice design with instruction following and role-play. VoxDirector leads on key metrics across these tasks and achieves the best expressiveness scores in all of them. Finally, we construct VoxBench-Scene to evaluate complete audio scene generation, demonstrating VoxDirector's strong performance in complex tasks involving multiple audio modalities.

2 Related Work

Zero-shot TTS. Zero-shot TTS uses a reference waveform to transfer the identity of an unseen speaker to new content. VALL-E (Wang et al., 2023a) establishes codec-language-model in-context synthesis. NaturalSpeech 2 (Shen et al., 2024) uses latent diffusion, while F5-TTS (Chen et al., 2024) uses flow matching for zero-shot generation. These works primarily target speaker identity and content, with limited control over the style of each utterance. Recent methods add such local control. CosyVoice 3 (Du et al., 2025), IndexTTS2 (Zhou et al., 2025a), and VoxCPM2 (Zhou et al., 2026) support instructions for emotion, duration, or delivery under reference conditioning. In parallel, MoonCast (Ju et al., 2025), ZipVoice-Dialog (Zhu et al., 2025a), FireRedTTS-2 (Xie et al., 2025b), and VibeVoice (Peng et al., 2025) explore dialogue speech. However, researchers usually study local control and dialogue separately, and neither includes sound effects as part of local control. VoxDirector supports local control and sound effect generation in voice cloning.

Controllable TTS. Controllable TTS generates speech while allowing explicit control over speaker characteristics or delivery through labels, acoustic variables, learned embeddings, or natural-language descriptions. Early systems expose predefined or learned factors. FastSpeech 2 (Ren et al., 2021) controls duration, pitch, and energy, while Global Style Tokens (Wang et al., 2018) learn reusable style embeddings without explicit labels. Prompt-based methods then move control into natural language. PromptTTS (Guo et al., 2022) separates style from content, and PromptTTS 2 (Leng et al., 2024) models voice variation left unspecified by text descriptions. Parler-TTS (Lyth & King, 2024) scales descriptive supervision, while VoxInstruct (Zhou et al., 2024) studies multilingual instruction following. More recent systems extend description-based control to voice creation, including Qwen3-TTS (Qwen Team, 2026a), VoxCPM2 (Zhou et al., 2026), MOSS-VoiceGenerator (Huang et al., 2026), and VoiceSculptor (Hu et al., 2026). However, existing methods remain focused on single-speaker speech control. VoxDirector separately controls ambience, the global voice styles of multiple speakers, and local delivery or sound effects.

Unified audio generation and representation. Unified audio generation aims to model speech, sound, and music with shared representations or generators under text and audio conditions. Neural codecs first make long waveforms tractable by compressing them into shorter sequences. EnCodec (Defossez et al., 2022) and DAC (Kumar et al., 2023b) emphasize high-fidelity reconstruction. SpeechTokenizer (Zhang et al., 2024) adapts discrete representations for language modeling by incorporating semantic structure. Built on compact representations, AudioGen (Kreuk et al., 2023) establishes text-conditioned general-audio generation. Later systems broaden the task scope. UniAudio (Yang et al., 2023) formulates multiple audio modalities as tasks in one model; Audiobox (Vyas et al., 2023) adds transcript-, description-, and example-based control; Fugatto (Valle et al., 2025) follows free-form instructions for synthesis and transformation. These systems organize conditions by clips or tasks, while VoxDirector models them with explicit multi-level conditions. Moreover, VoxVAE provides a shared continuous representation for multiple audio modalities.

3 Data Pipeline: VoxCaps

As shown in Figure 1, we introduce VoxCaps, a data pipeline to clean and annotate audio scenes.

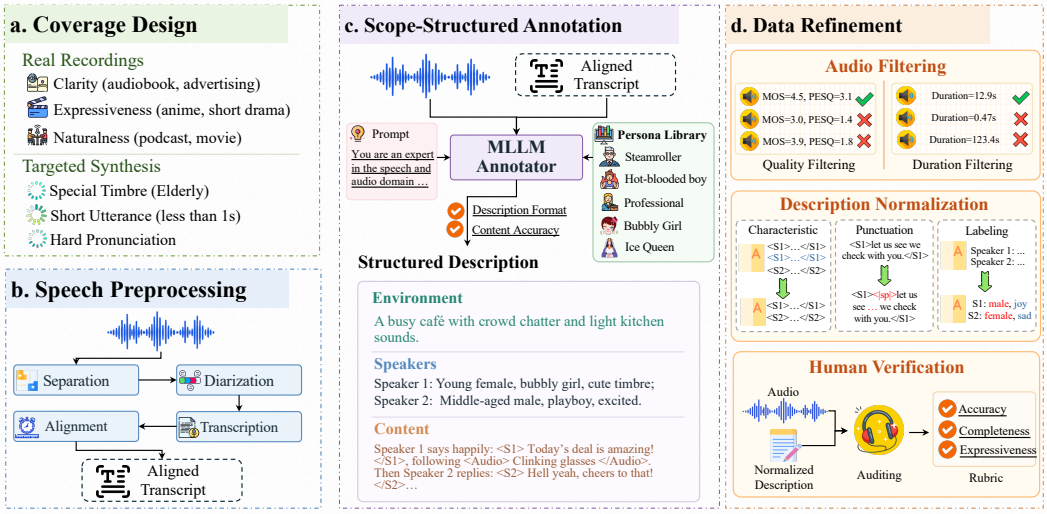

Figure 1: The VoxCaps pipeline consists of coverage design, speech preprocessing, scope-structured annotation, and data refinement.

3.1 Coverage Design

Media coverage. We collect Chinese and English recordings from different kinds of media, including both clear speech and scenes where speech coexists with ambience, music, and sound effects. Audiobooks and advertising provide clear speech, while animation, short drama, and podcasts provide character performances. Moreover, films supply natural dialogue and surrounding audio.

Targeted synthesis. Beyond media coverage, many voice styles are scarce in real recordings but necessary in applications, like elderly voices (Chen et al., 2025b). Very short turns such as names and 'Hello', along with pronunciation-sensitive texts involving polyphonic characters and mixed-language phrases, are also rare and hard to annotate accurately. We therefore use CosyVoice 3 (Du et al., 2025) as a teacher with good content accuracy to generate about 300K additional training examples. For pronunciation-sensitive text, pypinyin (Huang et al., 2023) supplies reading hints so the teacher produces the intended pronunciation. More details can be found in Appendix A.1.

3.2 Speech Preprocessing

Multi-speaker scene descriptions need reliable speech spans, speaker turns, and pauses. We therefore preprocess speech in the original recordings before annotation. First, we use Ultimate Vocal

Remover (Anjok07 & aufr33, 2020) for **separation**, extracting a vocal track for better ASR and alignment. The original mixture remains the input for scene annotation and the waveform target. Then we conduct **diarization** with a pipeline based on 3D-Speaker (Chen et al., 2025a), VAD, and CAM++ speaker embeddings (Wang et al., 2023b) to split long recordings at speaker turns into clips of up to 120 seconds. For **transcription**, Fun-ASR-Nano (An et al., 2025) transcribes each segment of the vocal track. Finally, we adopt Qwen3-ForcedAligner (Shi et al., 2026) for **alignment**, which aligns the transcript with the vocal track. We record word boundaries and pauses for punctuation correction during data refinement. Further settings appear in Appendix A.2.

3.3 Scope-Structured Annotation

Scene ambience can last through a recording. Speaker voice traits recur across turns, while emotion and sound effects may change from line to line. VoxCaps therefore keeps these scopes separate in each scene description so that stable traits and local changes have distinct control targets. We define **Environment** to describe the recording space, reverberation, and persistent sounds, while **Speakers** lists the people who actually speak and describes their gender, age, timbre, persona, articulation, accent, and habitual delivery. Finally, **Content** follows speech turns, local styles, and sound effects in order. It uses speaker turn tags such as `<S{id}>` and `<Audio>` tags for sound effects.

We use Qwen3.5-Omni (Qwen Team, 2026b) to fill the three fields from the original mixture audio and transcript. In animation, drama, and advertising, we found that direct annotation often falls back on generic adjectives and fails to capture a character persona. However, users may instead want to specify a concrete character, such as Steamroller. Therefore, we build a **persona library** that organizes candidate voice traits and performance cues by media type. The annotator keeps only cues supported by the audio, placing stable traits in Speakers and momentary delivery in Content. Figure 1 illustrates the persona library alongside an example of a scene description. For more details on the description format and the persona library, please refer to Appendix A.3.

3.4 Data Refinement

Collected audio and automatically generated scene descriptions can still carry errors into training. Therefore, we use Torchaudio-SQUIM (Kumar et al., 2023a) to estimate speech quality, then remove clips that fail quality or duration checks. Next, we normalize the scene descriptions by repairing misplaced or unpaired tags. Then, according to the pause duration, we can determine where to insert commas, periods, and `<|sp|>` or remove wrong punctuation. We extract gender, age, emotion, and persona labels from each description, so the model can learn stronger control over each attribute. This data pipeline yields 29.22M training examples with a mean duration of 7.4 seconds. To further improve data quality, we add human review. Auditors assess each description for accuracy and completeness. For expressiveness, they apply Best-Worst Scaling (Kiritchenko & Mohammad, 2017) to four matched candidates and select the most and least expressive samples. One judgment yields five implied pairwise preferences, much more efficient and accurate than a two-sample comparison. Their judgments form a human preference partial order used to select a high-expressiveness subset. Finally, we get an 8.2M expressive SFT set. Further details appear in Appendix A.4.

4 Method: VoxDirector

We first outline the method shown in Figure 2 and then introduce each core component.

4.1 Overview

Let c denote a task-dependent structured description. As shown in Figure 2(a), VoxDirector maps the inputs of voice design and cloning to a latent audio trajectory. Voice design uses the full c, including Environment, Speakers, and Content, whereas voice cloning uses only the Content field and relies on reference audio x_{ref} for speaker identity. Because these conditions interact across the whole audio sequence, we use a non-causal flow-based Transformer, which can also avoid the repetition and skipping errors reported for autoregressive TTS (Shen et al., 2024). At inference, VoxVAE encodes x_{ref} into fixed reference latents for voice cloning. For both tasks, the VoxVAE decoder converts the latents generated by the flow-based Transformer into the output waveform $\hat{x}$.

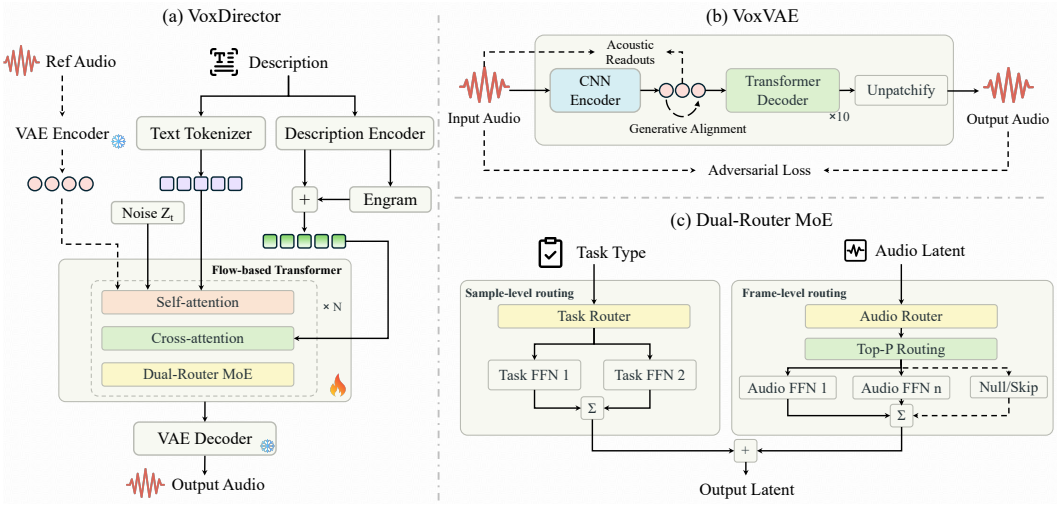

Figure 2: Architecture of VoxDirector, sharing a latent flow model across voice design and cloning. Dual-Router MoE uses a task router at the sample level and an audio router at the frame level.

4.2 VoxVAE

Long audio scenes require a short latent sequence, but speech, music, ambience, and sound effects still depend on fine local detail. As shown in Figure 2(b), VoxVAE maps 48-kHz mono audio to shared 25-Hz, 96-dimensional continuous latents with a local anti-aliased encoder, while placing most waveform-synthesis capacity in the decoder. A fixed low-pass filter precedes each of the encoder's five downsampling stages, while a small learned shortcut retains high-frequency envelope cues. The encoder predicts the mean and log-variance of a diagonal Gaussian posterior from about 0.95 seconds of context, leaving longer-range scene structure to VoxDirector's flow Transformer. During VoxVAE training, the decoder receives posterior samples and uses ten local-attention blocks to reconstruct each 40-ms latent interval as six 320-sample waveform patches (Parker et al., 2026).

Moreover, accurate waveform reconstruction does not ensure that the posterior mean is easy for the downstream flow model to predict (Skorokhodov et al., 2025). We therefore train a lightweight auxiliary flow-matching predictor on posterior-mean sequences to provide a direct signal about their learnability by a flow model. A causal predictor estimates future latent patches from preceding context, encouraging locally predictable changes. Acoustic readouts supervise frame energy, multi-scale chroma, and bandwise energy allocation so that these local cues remain represented alongside the prediction objectives. Specifically, we downweight the gradients from these auxiliary objectives to the encoder, keeping waveform reconstruction and KL regularization as the main constraints on the latent space. The full configuration is given in Appendix B.

4.3 Flow-based Transformer

Transcription and description. Spoken words must remain accurate and align locally with the audio, while descriptions in the Speakers and Environment fields may span the whole scene. Since the two conditions act at different temporal scales, VoxDirector encodes transcription and description separately in DiT to maintain lexical accuracy and scene-level acoustic control. For **transcription** y, the CosyVoice 2.0 (Du et al., 2024b) tokenizer supplies tokens aligned with speaker-turn embeddings. A lightweight Transformer then encodes them, and its states are aligned with the latent timeline and combined with noised latents. Meanwhile, Qwen3.0-Instruct-8B (Yang et al., 2025) encodes the **description** c for cross-attention. We add aligned label embeddings to distinguish description levels and Engram (Cheng et al., 2026) for fixed phrases, as detailed in Appendix C.3.

Reward-conditioned quality control. To retain this large dataset and its style coverage without treating variable recording quality as the target, we condition generation on a quality description and a quality level. We discretize STOI, SI-SDR, PESQ, and MOS from VoxCaps into levels de-

scribing speech clarity, noise level, signal naturalness, and listening quality, respectively. These levels form a quality description placed before Content. We also derive an overall quality level $Q \in \{\text{low}, \text{normal}, \text{high}, \text{unknown}\}$ from the four estimates and add its embedding to the global AdaLN-Zero condition. At inference, both the quality description and level request high quality. Following outcome-conditioned learning (Kumar et al., 2019; Emmons et al., 2022), we treat the quality estimates as conditions rather than rollout rewards, so moderate-quality samples still contribute diverse style coverage without a separate reward model. Appendix C.2 gives more details.

4.4 DUAL-ROUTER MOE

VoxDirector generates multi-speaker speech with surrounding audio in one network. Speech frames must preserve words, speaker identity, and prosody, while ambience is often sustained, and local sound effects require sharply timed changes. Music introduces pitch and harmonic structure that transcription alone does not describe. Using the same dense feed-forward parameters for all these patterns makes heterogeneous audio compete for one set of weights.

As shown in Figure 2(c), we introduce Dual-Router MoE to separate task-level conditioning from frame-level acoustic modeling. The task remains fixed throughout a sample, but speech turns and surrounding sounds change along the latent sequence, motivating routing at both sample and frame levels. Specifically, every second Transformer feed-forward layer is replaced by a sparse layer, with the intervening dense layers retained as a shared path.

Expert layout and routing. Each sparse layer contains two task experts, one selected by the task router for all frames of a sample, alongside acoustic experts and null candidates routed independently at each frame. The audio router uses the hidden state after self- and cross-attention, which incorporates transcription, description, and reference audio, together with the flow-time and quality condition. Its acoustic experts provide capacity for diverse speakers, changes in vocal delivery, and local sound effects. Stable ambience and near-silent frames need less additional processing, so dynamic Top-P routing selects a variable number of acoustic and null candidates rather than assigning every frame a fixed Top-K. Null candidates retain their routing probability but produce no feed-forward output, reducing the acoustic experts' contribution when selected together. If all selected candidates are null, the acoustic branch is skipped while the task expert remains active. A learned flow-time-dependent budget adjusts the Top-P threshold and null preference. To keep discrete routing trainable and expert use balanced, we use annealed Gumbel selection (Jang et al., 2016) and a load-correction bias (Wang et al., 2024a). We further impose capacity limits to avoid expert overload and a router z-loss (Zoph et al., 2022) to stabilize training, while an all-null penalty discourages the router from bypassing acoustic experts on every frame. Appendix C.4 gives the routing details.

4.5 CURRICULUM LEARNING

To make voice cloning and design, along with multiple audio modalities, easier to learn, we first train voice cloning and dialogue generation to establish acoustic quality, pronunciation accuracy, and prosody, then introduce description-conditioned speech and full audio scenes to add controllability. Introducing descriptions before text alignment and speaker continuity are stable would make both alignment and conditional learning harder (Bengio et al., 2009). **1) Cloning foundation.** We start with single-speaker voice cloning, then add concatenated and natural dialogue with up to four speakers to learn speaker turns and continuity. We drop reference audio for half of the samples to prevent speaker control from relying entirely on the reference. **2) Dense description adaptation.** We next use clean, attribute-rich speech from VoxBox (Wang et al., 2025) to teach the model to follow templated descriptions built from gender, age, and emotion labels. We also include the targeted synthetic subsets to improve coverage of rare styles and hard cases. During this phase, we sample voice design and cloning examples at 70% and 30%, respectively. **3) Full VoxCaps training.** Then, we introduce Dual-Router MoE and expand to the full VoxCaps mixture of speech, ambience, and local sound effects. The dense layers initialize the shared path, while routed experts model speaker variation and surrounding audio. We also continue to use attribute labels to preserve control over individual attributes while learning from natural-language descriptions. **4) Expressive SFT.** The full mixture broadens the range of styles, but recording expressiveness remains uneven. Therefore, we fine-tune on high-quality, expressive examples selected by human review to improve stability and expressiveness in both tasks. For the detailed schedule, please refer to Appendix C.5.

4.6 Inference Procedure

At inference, the transcription with speaker turn tags controls who says what. The description, or reference audio in voice cloning, guides the other acoustic attributes. A single guidance scale would couple these two levels of conditions, so we use decomposed classifier-free guidance (Ho & Salimans, 2022). Each condition is independently dropped with 15% probability during training. The denoiser is evaluated under null, text-and-turn, and full conditions, giving the guided velocity

$$\widetilde{v}_t = v_\varnothing + \omega_{\text{text}}(t)(v_{\text{text}} - v_\varnothing) + \omega_{\text{all}}(t)(v_{\text{full}} - v_{\text{text}}). \tag{1}$$

Here, $\omega_{\text{text}}(t)$ controls spoken text and speaker turns, while $\omega_{\text{all}}(t)$ controls the structured description, reference audio, and quality conditions. Further details are given in Appendix C.6.

5 Experiments

5.1 Experimental Setup

VoxVAE is trained for 200,000 steps on approximately 100,000 hours of speech, general audio, and music using 32 A100 GPUs. Training uses random 3.84-s crops, with waveform mixing on 25% of crops and effective-bandwidth augmentation on 1%. VoxDirector has 2B active parameters and is trained on 64 A100 GPUs. The cloning foundation uses 23M hours of single-speaker and 1.7M hours of two-speaker speech. Dense description adaptation then draws on VoxBox (Wang et al., 2025), whose public release contains 47.7M utterances (SparkAudio, 2025), for 20k steps, followed by full VoxCaps training on 29.22M examples for 10k steps and expressive SFT on 8.2M selected examples for 4k steps. At inference, the text-and-turn and full-condition branches use guidance weights of 1.5 and 3.0. Implementation and evaluation details appear in Appendices B, C, and D.

5.2 VoxVAE Reconstruction

We evaluate VoxVAE on out-of-domain subsets of VCTK (Yamagishi et al., 2019), FSD50K (Fonseca et al., 2021), and MUSDB18-HQ (Stöter et al., 2018), selecting 1,000 clips per audio domain. For speech, PESQ (Rix et al., 2001) and ViSQOL (Chinen et al., 2020) assess perceptual reconstruction quality, STOI (Taal et al., 2011) measures intelligibility, and MCD (Kubichek, 1993) measures mel-cepstral distortion. For general audio and music, ViSQOL assesses perceptual similarity, while LSD (Gray & Markel, 1976) measures log-spectral distortion. We compare VoxVAE with representative audio autoencoders and neural audio codecs across the three audio domains. Table 1 reports all evaluated models at their native representation settings.

Table 1: VoxVAE reconstruction on speech, general audio, and music. Bold and underlined values mark the best and second-best results. Dashes mark models not evaluated in that domain.

	Speech				General audio		Music	
Model	PESQ ↑	STOI ↑	MCD ↓	ViSQOL ↑	ViSQOL ↑	LSD ↓	ViSQOL ↑	LSD ↓
DAC	<u>4.118</u>	**0.969**	<u>1.196</u>	<u>4.159</u>	4.020	0.959	4.153	0.920
EnCodec	3.187	0.930	1.515	3.504	<u>4.114</u>	0.976	**4.298**	**0.900**
WavTokenizer Large Unify	2.142	0.843	2.939	2.379	2.860	1.097	2.697	1.061
VoxCPM2 VAE	3.999	<u>0.969</u>	1.222	4.034	—	—	—	—
MegaTTS 3 VAE	3.597	0.951	1.513	**4.235**	—	—	—	—
Stable Audio Open 1.0	—	—	—	—	4.036	**0.936**	4.136	0.924
SAME-L	—	—	—	—	3.754	1.037	4.027	0.921
ACE-Step Music-DCAE	—	—	—	—	—	—	4.176	<u>0.902</u>
VoxVAE	**4.168**	0.968	**0.964**	4.125	**4.127**	<u>0.946</u>	<u>4.262</u>	0.917

As shown in Table 1, VoxVAE achieves strong reconstruction performance across all three audio domains. On speech, it delivers the best overall reconstruction among the compared systems. On general audio, it achieves the best perceptual reconstruction and remains close to Stable Audio Open 1.0. On music, EnCodec performs best overall, while VoxVAE remains competitive with the leading systems. These results show that our design enables VoxVAE to preserve fine-grained acoustic detail in compact 25-Hz latents across all three domains. Appendix E.2 presents our ablation study results.

5.3 Zero-Shot Voice Cloning

We compare VoxDirector with representative zero-shot voice cloning systems for monologue and dialogue generation on SwanBench-Speech (Pan et al., 2026). We measure timbre and reverb consistency, sound fidelity, and content error with Chinese CER and English WER. For expressiveness, SpeechJudge (Zhang et al., 2026) assesses naturalness, while Gemini 3 Pro (Google DeepMind, 2026) scores expressive richness and hierarchy. Table 2 reports the results.

Table 2: Zero-shot voice cloning on SwanBench-Speech. Error is the average of Chinese CER and English WER. Bold and underlined values mark the best and second-best results.

Model	Timbre ↑	Reverb ↓	Fidelity ↑	Error ↓	Judge ↑	Rich. ↑	Hier. ↑
Monologue							
CosyVoice-2	0.93	2.37	3.58	0.106	2.81	2.02	2.59
CosyVoice-3	0.93	2.73	3.80	0.077	3.26	2.64	2.47
FishSpeech	0.93	2.00	**4.09**	**0.066**	**3.77**	2.37	2.90
F5TTS	0.92	2.12	2.60	0.085	2.87	2.77	2.97
GLM-TTS	<u>0.94</u>	**1.64**	3.90	0.074	3.28	1.57	2.39
IndexTTS-2	0.93	<u>1.77</u>	2.78	0.077	3.63	3.32	2.94
MegaTTS-3	0.93	2.07	3.52	<u>0.072</u>	3.22	2.40	3.01
SparkTTS	0.92	2.04	3.53	0.314	2.35	2.23	2.22
VibeVoice	0.92	2.45	3.47	0.092	<u>3.75</u>	<u>3.42</u>	<u>3.06</u>
ZipVoice	0.89	2.10	3.53	0.213	2.97	2.11	2.05
VoxDirector	**0.95**	1.83	<u>3.95</u>	0.086	<u>3.75</u>	**3.90**	**3.70**
Dialogue							
FireRedTTS-2	0.91	3.54	2.54	0.148	2.93	2.52	2.65
MoonCast	0.90	3.29	2.60	0.284	2.93	2.42	2.54
MOSS-TTSD	0.89	3.52	2.83	0.227	2.57	3.04	2.86
SoulX-Podcast	<u>0.92</u>	3.23	**3.98**	**0.101**	<u>3.89</u>	2.80	<u>3.15</u>
VibeVoice	0.89	**2.09**	2.75	0.204	3.00	<u>3.09</u>	2.83
ZipVoice-Dialog	0.90	3.49	2.48	<u>0.116</u>	3.46	2.88	2.93
VoxDirector	**0.94**	<u>3.21</u>	<u>3.73</u>	0.120	**3.92**	**3.66**	**3.85**

As shown in Table 2, we can find that VoxDirector achieves the best Timbre Consistency, Expressive Richness, and Expressive Hierarchy in both monologue and dialogue settings. Meanwhile, VoxDirector achieves the highest SpeechJudge score in dialogue and ties for second in monologue. Curriculum training and SFT help maintain stable voices and high expressiveness, while reward-conditioned quality control helps improve sound fidelity. Appendix E.3 reports additional results.

5.4 Voice Design

Table 3: Voice design on InstructTTSEval. Best scores are in **bold** and second-best are <u>underlined</u>.

Model	ZH			EN		
	APS	DSD	RP	APS	DSD	RP
Parler-TTS-large	—	—	—	60.0	45.9	31.2
VoxInstruct	47.5	52.3	42.6	54.9	57.0	39.3
VoiceSculptor	75.7	64.7	61.5	—	—	—
MiMo-Audio	75.7	74.3	61.5	80.6	77.6	59.5
Qwen3-TTS-VD	<u>85.2</u>	**81.1**	<u>65.1</u>	<u>82.9</u>	<u>82.4</u>	68.4
MOSS-VoiceGenerator	78.0	80.0	**74.0**	68.2	82.0	<u>68.7</u>
VoxCPM2	<u>85.2</u>	71.5	60.8	**84.2**	**83.2**	**71.4**
VoxDirector	**86.1**	<u>80.1</u>	64.1	**84.2**	79.2	63.6

We evaluate VoxDirector on InstructTTSEval (Huang et al., 2025) and VoxBench-Instruct, our 180-item role-play benchmark covering advertising, short drama, and general scenes. InstructTTSEval measures Chinese and English instruction-following accuracy for acoustic-parameter specification

(APS), descriptive-style directive (DSD), and role-play (RP) using Gemini 2.5 Pro. On VoxBench-Instruct, five professional listeners rate expressiveness, prosodic naturalness, audio fullness, and scene appropriateness, whose average gives the Mean MOS. We compare with representative recent instruct TTS, voice-design, and audio-language systems, with results in Tables 3 and 4.

Table 4: Voice-design on VoxBench-Instruct. Best scores are in **bold** and second-best are underlined.

Model	Mean MOS	Exp.	Nat.	Full.	App.
MiMo-Audio	3.21	3.06	3.27	3.57	2.94
MOSS-VoiceGenerator	3.60	3.42	3.62	4.06	3.31
Qwen3-TTS-VD	4.09	3.96	4.09	4.45	3.88
VoxDirector	**4.22**	**4.12**	**4.20**	**4.47**	**4.10**

According to the results, VoxDirector leads acoustic-parameter control and achieves the highest overall MOS, including the best expressiveness and scene-appropriateness scores. The expanded VoxCaps data and persona library provide diverse character and delivery examples, while the staged curriculum teaches the model to use these descriptions. More results can be found in Appendix E.4.

5.5 Complete Audio Scenes and Ablation

We construct VoxBench-Scene with 500 challenging cases covering ambience, local sound effects, music, and multi-speaker speech. The benchmark covers both voice design and voice cloning, with details in Appendix D.4. We use the `gemini-3.5-flash` judge to rate instruction accuracy, acoustic quality, and expressiveness on a 1–5 scale. Table 5 compares VoxDirector with a cascaded system and an existing audio scene generation model, alongside our ablations.

Table 5: Comparisons on VoxBench-Scene. Dasheng AudioGen does not support voice cloning. The 32B variant changes the description encoder. Best results are in **bold** and second-best are underlined.

Setting	Instruction Acc	Acoustic Quality	Expressiveness
Qwen3-TTS + TangoFlux	3.40	4.11	3.69
Dasheng AudioGen	2.71	3.52	3.64
VoxDirector	3.61	4.33	3.89
with 32B description encoder	**4.10**	**4.36**	**4.03**
w/o Dual-Router MoE	3.32	3.97	3.52
w/o Task Router	3.40	4.21	3.77
w/o Audio Router	3.36	4.03	3.59
w/o quality conditioning	3.55	3.64	3.79
w/o staged curriculum	3.31	3.44	3.48

As shown in Table 5, VoxDirector outperforms baselines across all metrics and offers broader capabilities. Removing the staged curriculum causes the largest drops in acoustic quality and expressiveness, whereas quality conditioning mainly benefits acoustic quality. Dual-Router MoE improves all three scores, while the 32B description encoder scaling mainly improves instruction accuracy.

6 Conclusion

Overall, we propose VoxDirector, an expressive multi-speaker speech and surrounding audio generation model that supports both voice design and cloning. For the data pipeline, we propose VoxCaps, which cleans media data, expands it with targeted synthetic data, and uses a persona library to annotate diverse multi-level scene descriptions. Then, we introduce VoxVAE, which pairs a local anti-aliased encoder with a high-capacity decoder to provide a shared continuous representation for multiple audio modalities. VoxVAE outperforms baselines on key reconstruction metrics across multiple audio modalities. VoxDirector also adopts reward-conditioned quality control to improve generation quality and a Dual-Router MoE tailored to the two tasks and multiple audio modalities. Across voice cloning, voice design, and complete audio scene generation tasks, VoxDirector shows strong generation quality, expressiveness, and ability to follow complex instructions.

Ethics Statement

This work studies expressive speech and audio scene generation from descriptions and reference audio. Use and redistribution of the media recordings, datasets, and pretrained models remain subject to their respective licenses and usage terms. Participants in our subjective listening evaluation provide informed consent. Realistic voice cloning and audio scenes can be misused for impersonation or misleading content. We advocate obtaining permission to use a person's voice, disclosing synthetic outputs, and respecting privacy and intellectual property.

Reproducibility Statement

Sections 3 and 4 describe the data pipeline and model. Appendices A, B, and C provide data processing and annotation procedures, VoxVAE specifications, and VoxDirector training and inference details. Evaluation sets, metrics, and baselines are described in Appendix D, with extended results and VoxVAE ablations in Appendix E. We provide an anonymous demo page for review.

AI Use Statement

Generative models support targeted synthesis, transcription, description annotation, and model-based evaluation, as described in Sections 3 and 5, Appendices A and D. An AI assistant supported manuscript polishing. The authors are responsible for the final text, references, figures, and scientific claims.

References

Keyu An, Qian Chen, Chong Deng, Zhihao Du, Changfeng Gao, et al. FunAudioLLM: Voice Understanding and Generation Foundation Models for Natural Interaction Between Humans and LLMs, 2024. URL https://arxiv.org/abs/2407.04051.

Keyu An, Yanni Chen, Zhigao Chen, et al. Fun-ASR Technical Report, 2025. URL https://arxiv.org/abs/2509.12508.

Philip Anastassiou, Jiawei Chen, Jitong Chen, et al. Seed-TTS: A Family of High-Quality Versatile Speech Generation Models, 2024. URL https://arxiv.org/abs/2406.02430.

Anjok07 and aufr33. Ultimate Vocal Remover. GitHub repository, 2020. URL https://github.com/Anjok07/ultimatevocalremovergui.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum Learning. International Conference on Machine Learning, pp. 41–48, 2009.

Boson AI. Higgs Audio V2: Redefining Expressiveness in Audio Generation. GitHub repository, 2025. URL https://github.com/boson-ai/higgs-audio/blob/main/README_V2.md.

Guoguo Chen, Shuzhou Chai, Guanbo Wang, Jiayu Du, Wei-Qiang Zhang, Chao Weng, Dan Su, Daniel Povey, Jan Trmal, Junbo Zhang, et al. GigaSpeech: An Evolving, Multi-Domain ASR Corpus with 10,000 Hours of Transcribed Audio, 2021. URL https://arxiv.org/abs/2106.06909.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, et al. WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing. IEEE Journal of Selected Topics in Signal Processing, 16(6):1505–1518, 2022.

Yafeng Chen, Siqi Zheng, Hui Wang, Luyao Cheng, Tinglong Zhu, et al. 3D-Speaker-Toolkit: An Open-Source Toolkit for Multimodal Speaker Verification and Diarization. ICASSP, pp. 1–5, 2025a.

Yang Chen, Hui Wang, Shiyao Wang, Junyang Chen, Jiabei He, Jiaming Zhou, Xi Yang, Yequan Wang, Yonghua Lin, and Yong Qin. SeniorTalk: A Chinese Conversation Dataset with Rich Annotations for Super-Aged Seniors, 2025b. URL https://arxiv.org/abs/2503.16578.

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, 2024. URL https://arxiv.org/abs/2410.06885.

Xin Cheng, Wangding Zeng, Damai Dai, Qinyu Chen, Bingxuan Wang, et al. Conditional Memory via Scalable Lookup: A New Axis of Sparsity for Large Language Models, 2026. URL https://arxiv.org/abs/2601.07372.

Michael Chinen, Felicia S. C. Lim, Jan Skoglund, Nikita Gureev, Feargus O'Gorman, and Andrew Hines. ViSQOL v3: An Open Source Production Ready Objective Speech and Audio Metric. QoMEX, pp. 1–6, 2020.

Jiayan Cui, Zhihan Yang, Naihan Li, Jiankun Tian, Xingyu Ma, Yi Zhang, Guangyu Chen, Runxuan Yang, Yuqing Cheng, Yizhi Zhou, Guochen Yu, Xiaotao Gu, and Jie Tang. GLM-TTS Technical Report, 2025. URL https://arxiv.org/abs/2512.14291.

Alexandre Defossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. High Fidelity Neural Audio Compression, 2022. URL https://arxiv.org/abs/2210.13438.

Zhihao Du, Qian Chen, Shiliang Zhang, et al. CosyVoice: A Scalable Multilingual Zero-Shot Text-to-Speech Synthesizer Based on Supervised Semantic Tokens, 2024a. URL https://arxiv.org/abs/2407.05407.

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. CosyVoice 2: Scalable Streaming Speech Synthesis with Large Language Models, 2024b. URL https://arxiv.org/abs/2412.10117.

Zhihao Du, Changfeng Gao, Yuxuan Wang, Fan Yu, Tianyu Zhao, et al. CosyVoice 3: Towards In-the-Wild Speech Generation via Scaling-Up and Post-Training, 2025. URL https://arxiv.org/abs/2505.17589.

Scott Emmons, Benjamin Eysenbach, Ilya Kostrikov, and Sergey Levine. RvS: What is Essential for Offline RL via Supervised Learning?, 2022. URL https://arxiv.org/abs/2112.10751.

Zach Evans, Julian D. Parker, CJ Carr, Zack Zukowski, Josiah Taylor, and Jordi Pons. Stable Audio Open. ICASSP, 2025.

Tiago H. Falk, Chenxi Zheng, and Wai-Yip Chan. A Non-Intrusive Quality and Intelligibility Measure of Reverberant and Dereverberated Speech. IEEE Transactions on Audio, Speech, and Language Processing, 18(7):1766–1774, 2010.

Fish Audio. OpenAudio S1 Mini. Hugging Face model card, 2025. URL https://huggingface.co/fishaudio/s1-mini.

Fish Audio Team. Fish-Speech: Leveraging Large Language Models for Advanced Multilingual Text-to-Speech Synthesis, 2024. URL https://arxiv.org/abs/2411.01156.

Eduardo Fonseca, Xavier Favory, Jordi Pons, Frederic Font, and Xavier Serra. FSD50K: An Open Dataset of Human-Labeled Sound Events. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 30:829–852, 2021.

Junmin Gong, Sean Zhao, Sen Wang, Shengyuan Xu, and Joe Guo. ACE-Step: A Step towards Music Generation Foundation Model, 2025. URL https://arxiv.org/abs/2506.00045.

Yitian Gong, Botian Jiang, Yiwei Zhao, et al. MOSS-TTS Technical Report, 2026. URL https://arxiv.org/abs/2603.18090.

Google DeepMind. Gemini 3 Pro Model Card, 2026. URL https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-Pro-Model-Card.pdf.

Augustine H. Gray and John D. Markel. Distance Measures for Speech Processing. IEEE Transactions on Acoustics, Speech, and Signal Processing, 24(5):380–391, 1976.

Hao-Han Guo, Yao Hu, Kun Liu, et al. FireRedTTS: A Foundation Text-to-Speech Framework for Industry-Level Generative Speech Applications, 2024. URL https://arxiv.org/abs/2409.03283.

Zhifang Guo, Yichong Leng, Yihan Wu, Sheng Zhao, and Xu Tan. PromptTTS: Controllable text-to-speech with text descriptions. arXiv:2211.12171, 2022. URL https://arxiv.org/abs/2211.12171.

Jonathan Ho and Tim Salimans. Classifier-Free Diffusion Guidance, 2022. URL https://arxiv.org/abs/2207.12598.

Jingbin Hu, Huakang Chen, Linhan Ma, Dake Guo, Qirui Zhan, et al. VoiceSculptor: Your Voice, Designed by You, 2026. URL https://arxiv.org/abs/2601.10629.

Huang Huang, Xiaoquan Kong, et al. python-pinyin: pypinyin. Software, version 0.48.0, 2023. URL https://github.com/mozillazg/python-pinyin.

Kexin Huang, Qian Tu, Liwei Fan, et al. InstructTTSEval: Benchmarking Complex Natural-Language Instruction Following in Text-to-Speech Systems, 2025. URL https://arxiv.org/abs/2506.16381.

Kexin Huang, Liwei Fan, Botian Jiang, Yaozhou Jiang, Qian Tu, et al. MOSS-VoiceGenerator: Create Realistic Voices with Natural Language Descriptions, 2026. URL https://arxiv.org/abs/2603.28086.

Chia-Yu Hung, Navonil Majumder, Zhifeng Kong, Ambuj Mehrish, Amir Ali Bagherzadeh, Chuan Li, Rafael Valle, Bryan Catanzaro, and Soujanya Poria. TangoFlux: Super Fast and Faithful Text to Audio Generation with Flow Matching and Clap-Ranked Preference Optimization, 2024. URL https://arxiv.org/abs/2412.21037.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical Reparameterization with Gumbel-Softmax, 2016. URL https://arxiv.org/abs/1611.01144.

Won Jang, Dan Lim, Jaesam Yoon, Bongwan Kim, and Juntae Kim. UnivNet: A Neural Vocoder with Multi-Resolution Spectrogram Discriminators for High-Fidelity Waveform Generation, 2021. URL https://arxiv.org/abs/2106.07889.

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. WavTokenizer: An Efficient Acoustic Discrete Codec Tokenizer for Audio Language Modeling. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/ea1f5f0878d43ff4fb8bf64ef4a2326c-Paper-Conference.pdf.

Ziyue Jiang, Yi Ren, Ruiqi Li, Shengpeng Ji, Boyang Zhang, et al. MegaTTS 3: Sparse Alignment Enhanced Latent Diffusion Transformer for Zero-Shot Speech Synthesis, 2025. URL https://arxiv.org/abs/2502.18924.

Zeqian Ju, Dongchao Yang, Jianwei Yu, Kai Shen, Yichong Leng, et al. MoonCast: High-Quality Zero-Shot Podcast Generation, 2025. URL https://arxiv.org/abs/2503.14345.

Wei Kang, Xiaoyu Yang, Zengwei Yao, Fangjun Kuang, Yifan Yang, Liyong Guo, Long Lin, and Daniel Povey. LibriHeavy: A 50,000 Hours ASR Corpus with Punctuation, Casing and Context. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 10991–10995, 2024.

Svetlana Kiritchenko and Saif Mohammad. Best-Worst Scaling More Reliable than Rating Scales: A Case Study on Sentiment Intensity Annotation. ACL, Volume 2: Short Papers, pp. 465–470, 2017.

Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis. Advances in Neural Information Processing Systems, 33:17022–17033, 2020.

Felix Kreuk, Gabriel Synnaeve, Adam Polyak, Uriel Singer, Alexandre Défossez, Jade Copet, Devi Parikh, Yaniv Taigman, and Yossi Adi. AudioGen: Textually Guided Audio Generation. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=CYK7Rfc0zQ4.

Robert F. Kubichek. Mel-Cepstral Distance Measure for Objective Speech Quality Assessment. IEEE Pacific Rim Conference on Communications, Computers and Signal Processing, volume 1, pp. 125–128, 1993.

Anurag Kumar, Ke Tan, Zhaoheng Ni, Pranay Manocha, Xiaohui Zhang, Ethan Henderson, and Buye Xu. Torchaudio-SQUIM: Reference-less Speech Quality and Intelligibility Measures in Torchaudio. ICASSP, pp. 1–5, 2023a.

Aviral Kumar, Xue Bin Peng, and Sergey Levine. Reward-Conditioned Policies, 2019. URL https://arxiv.org/abs/1912.13465.

Rithesh Kumar, Prem Seetharaman, Alejandro Luebs, Ishaan Kumar, and Kundan Kumar. High-Fidelity Audio Compression with Improved RVQGAN, 2023b. URL https://arxiv.org/abs/2306.06546.

Yichong Leng, Zhifang Guo, Kai Shen, Zeqian Ju, Xu Tan, Yanjing Liu, Yufei Liu, Dongchao Yang, Leying Zhang, Kaitao Song, Lei He, Xiangyang Li, Sheng Zhao, Tao Qin, and Jiang Bian. PromptTTS 2: Describing and Generating Voices with Text Prompt. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/fed1ea8dcc2a13f3835cc854e8c8294c-Paper-Conference.pdf.

Shijia Liao, Yuxuan Wang, Songting Liu, et al. Fish Audio S2 Technical Report, 2026. URL https://arxiv.org/abs/2603.08823.

Haohe Liu, Yi Yuan, Xubo Liu, Xinhao Mei, Qiuqiang Kong, Qiao Tian, Yuping Wang, Wenwu Wang, Yuxuan Wang, and Mark D. Plumbley. AudioLDM 2: Learning holistic audio generation with self-supervised pretraining. arXiv:2308.05734, 2023. URL https://arxiv.org/abs/2308.05734.

Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. International Conference on Learning Representations, 2019.

Dan Lyth and Simon King. Natural Language Guidance of High-Fidelity Text-to-Speech with Synthetic Annotations, 2024. URL https://arxiv.org/abs/2402.01912.

Jiahao Mei, Heinrich Dinkel, Yadong Niu, Xingwei Sun, Gang Li, Yifan Liao, Jiahao Zhou, Junbo Zhang, Jian Luan, and Mengyue Wu. Dasheng AudioGen: A Unified Model for Generating Coherent Audio Scenes from Text, 2026. URL https://arxiv.org/abs/2605.27838.

MOSI. MOSS-TTSD. Project website, 2026. URL https://mosi.cn/models/moss-ttsd.

Changhao Pan, Rui Yang, Han Wang, Zhuan Zhou, Xuming He, Wenxiang Guo, Ziyue Jiang, Ruiqi Li, Yu Zhang, Chenyuhao Wen, Ke Lei, Xiang Yin, Jingyu Lu, Zhiyuan Zhu, and Zhou Zhao. Comprehensive Benchmarking of Long-Form Speech Generation in Diverse Scenarios, 2026. URL https://arxiv.org/abs/2605.28618.

Julian D. Parker, Zach Evans, CJ Carr, Zachary Zukowski, Josiah Taylor, Matthew Rice, and Jordi Pons. SAME: A Semantically-Aligned Music Autoencoder, 2026. URL https://arxiv.org/abs/2605.18613.

William Peebles and Saining Xie. Scalable Diffusion Models with Transformers. ICCV, pp. 4195–4205, 2023.

Zhiliang Peng, Jianwei Yu, Wenhui Wang, Yaoyao Chang, Yutao Sun, et al. VibeVoice Technical Report, 2025. URL https://arxiv.org/abs/2508.19205.

Qwen Team. Qwen3-TTS Technical Report, 2026a. URL https://arxiv.org/abs/2601.15621.

Qwen Team. Qwen3.5-Omni Technical Report, 2026b. URL https://arxiv.org/abs/2604.15804.

Chandan K. A. Reddy, Vishak Gopal, and Ross Cutler. DNSMOS: A Non-Intrusive Perceptual Objective Speech Quality Metric to Evaluate Noise Suppressors. ICASSP, pp. 6493–6497, 2021.

Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. FastSpeech 2: Fast and High-Quality End-to-End Text to Speech. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=piLPYqxtWuA.

A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra. Perceptual Evaluation of Speech Quality (PESQ): A New Method for Speech Quality Assessment of Telephone Networks and Codecs. ICASSP, volume 2, pp. 749–752, 2001.

Kai Shen, Zeqian Ju, Xu Tan, Yanjing 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. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/035a73893121b4534bb3314e831050b1-Paper-Conference.pdf.

Xian Shi, Xiong Wang, Zhifang Guo, Yongqi Wang, Pei Zhang, Xinyu Zhang, Zishan Guo, Hongkun Hao, Yu Xi, Baosong Yang, Jin Xu, Jingren Zhou, and Junyang Lin. Qwen3-ASR Technical Report, 2026. URL https://arxiv.org/abs/2601.21337.

Ivan Skorokhodov, Sharath Girish, Benran Hu, Willi Menapace, Yanyu Li, Rameen Abdal, Sergey Tulyakov, and Aliaksandr Siarohin. Improving the Diffusability of Autoencoders, 2025. URL https://arxiv.org/abs/2502.14831.

SparkAudio. VoxBox Dataset. GitHub repository, 2025. URL https://github.com/SparkAudio/VoxBox.

Fabian-Robert Stöter, Antoine Liutkus, and Nobutaka Ito. The 2018 Signal Separation Evaluation Campaign. Latent Variable Analysis and Signal Separation, pp. 293–305, 2018.

Cees H. Taal, Richard C. Hendriks, Richard Heusdens, and Jesper Jensen. An Algorithm for Intelligibility Prediction of Time–Frequency Weighted Noisy Speech. IEEE Transactions on Audio, Speech, and Language Processing, 19(7):2125–2136, 2011.

Rafael Valle, Rohan Badlani, Zhifeng Kong, Sang gil Lee, Arushi Goel, Sungwon Kim, Joao Santos, Shuqi Dai, Siddharth Gururani, Aya Aljafari, Alexander Liu, Kevin Shih, Ryan Prenger, Wei Ping, Chao-Han Huck Yang, and Bryan Catanzaro. Fugatto 1: Foundational Generative Audio Transformer Opus 1. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/0cc540ef357216738768caaa7301b615-Paper-Conference.pdf.

Apoorv Vyas, Bowen Shi, Matthew Le, Andros Tjandra, Yi-Chiao Wu, et al. Audiobox: Unified Audio Generation with Natural Language Prompts, 2023. URL https://arxiv.org/abs/2312.15821.

Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, Lei He, Sheng Zhao, and Furu Wei. Neural codec language models are zero-shot text to speech synthesizers. arXiv:2301.02111, 2023a. URL https://arxiv.org/abs/2301.02111.

Hui Wang, Siqi Zheng, Yafeng Chen, Luyao Cheng, and Qian Chen. CAM++: A Fast and Efficient Network for Speaker Verification Using Context-Aware Masking, 2023b. URL https://arxiv.org/abs/2303.00332.

Lean Wang, Huazuo Gao, Chenggang Zhao, Xu Sun, and Damai Dai. Auxiliary-Loss-Free Load Balancing Strategy for Mixture-of-Experts, 2024a. URL https://arxiv.org/abs/2408.15664.

Xinsheng Wang, Mingqi Jiang, Ziyang Ma, Ziyu Zhang, Songxiang Liu, et al. Spark-TTS: An Efficient LLM-Based Text-to-Speech Model with Single-Stream Decoupled Speech Tokens, 2025. URL https://arxiv.org/abs/2503.01710.

Yuancheng Wang, Haoyue Zhan, Liwei Liu, et al. MaskGCT: Zero-Shot Text-to-Speech with Masked Generative Codec Transformer, 2024b. URL https://arxiv.org/abs/2409.00750.

Yuxuan Wang, Daisy Stanton, Yu Zhang, R. J. Skerry-Ryan, Eric Battenberg, Joel Shor, Ying Xiao, Ye Jia, Fei Ren, and Rif A. Saurous. Style Tokens: Unsupervised Style Modeling, Control and Transfer in End-to-End Speech Synthesis. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 5180–5189, 2018. URL https://proceedings.mlr.press/v80/wang18h.html.

Hanke Xie, Haopeng Lin, Wenxiao Cao, Dake Guo, Wenjie Tian, et al. SoulX-Podcast: Towards Realistic Long-Form Podcasts with Dialectal and Paralinguistic Diversity, 2025a. URL https://arxiv.org/abs/2510.23541.

Kun Xie, Feiyu Shen, Junjie Li, Fenglong Xie, Xu Tang, and Yao Hu. FireRedTTS-2: Towards Long Conversational Speech Generation for Podcast and Chatbot, 2025b. URL https://arxiv.org/abs/2509.02020.

Yaoxun Xu, Hangting Chen, Jianwei Yu, Qiaochu Huang, Zhiyong Wu, Shi-Xiong Zhang, Guangzhi Li, Yi Luo, and Rongzhi Gu. SECap: Speech Emotion Captioning with Large Language Model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 19323–19331, 2024.

Junichi Yamagishi, Christophe Veaux, and Kirsten MacDonald. CSTR VCTK Corpus: English Multi-Speaker Corpus for CSTR Voice Cloning Toolkit (Version 0.92), 2019.

An Yang, Anfeng Li, Baosong Yang, et al. Qwen3 Technical Report, 2025. URL https://arxiv.org/abs/2505.09388.

Dongchao Yang, Jinchuan Tian, Xu Tan, Rongjie Huang, Songxiang Liu, Xuankai Chang, Jiatong Shi, Sheng Zhao, Jiang Bian, Zhou Zhao, Xixin Wu, and Helen Meng. UniAudio: An Audio Foundation Model Toward Universal Audio Generation, 2023. URL https://arxiv.org/abs/2310.00704.

Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J. Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu. LibriTTS: A Corpus Derived from LibriSpeech for Text-to-Speech. In *Proc. Interspeech*, pp. 1526–1530, 2019.

Biao Zhang and Rico Sennrich. Root Mean Square Layer Normalization. Advances in Neural Information Processing Systems, 32, 2019.

Xin Zhang, Dong Zhang, Shimin Li, Yaqian Zhou, and Xipeng Qiu. SpeechTokenizer: Unified Speech Tokenizer for Speech Language Models. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/86d1ab582afb247ccaa84bec4a7e24f7-Paper-Conference.pdf.

Xin Zhang, Tianrui Liu, Han Li, Zhi Li, Wenbo Wang, Yunhua Zhu, Zheng Wang, et al. MiMo-Audio: Audio Language Models Are Few-Shot Learners, 2025. URL https://arxiv.org/abs/2512.23808.

Xueyao Zhang, Chaoren Wang, Huan Liao, Ziniu Li, Yuancheng Wang, Li Wang, Dongya Jia, Yuanzhe Chen, Xiulin Li, Zhuo Chen, and Zhizheng Wu. SpeechJudge: Towards Human-Level Judgment for Speech Naturalness. International Conference on Learning Representations, 2026.

Siyi Zhou, Yiquan Zhou, Yi He, Xun Zhou, Jinchao Wang, Wei Deng, and Jingchen Shu. IndexTTS2: A Breakthrough in Emotionally Expressive and Duration-Controlled Auto-Regressive Zero-Shot Text-to-Speech, 2025a. URL https://arxiv.org/abs/2506.21619.

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. ACM International Conference on Multimedia, pp. 554–563, 2024.

Yixuan Zhou, Guoyang Zeng, Xin Liu, et al. VoxCPM: Tokenizer-Free TTS for Context-Aware Speech Generation and True-to-Life Voice Cloning, 2025b. URL https://arxiv.org/abs/2509.24650.

Yixuan Zhou, Guoyang Zeng, Xin Liu, Xiang Li, Renjie Yu, Jiancheng Gui, Jiaheng Wu, Ziyang Wang, Xudong Shen, Runchuan Ye, Zhisheng Zhang, Jiuyang Zhou, Bingsong Bai, Weiyue Sun, Mengyuan Deng, Qundong Shi, Zhiyong Wu, and Zhiyuan Liu. VoxCPM2 Technical Report, 2026. URL https://arxiv.org/abs/2606.06928.

Han Zhu, Wei Kang, Liyong Guo, Zengwei Yao, Fangjun Kuang, et al. ZipVoice-Dialog: Non-Autoregressive Spoken Dialogue Generation with Flow Matching, 2025a. URL https://arxiv.org/abs/2507.09318.

Han Zhu, Wei Kang, Zengwei Yao, Liyong Guo, Fangjun Kuang, Zhaoqing Li, Weiji Zhuang, Long Lin, and Daniel Povey. ZipVoice: Fast and High-Quality Zero-Shot Text-to-Speech with Flow Matching, 2025b. URL https://arxiv.org/abs/2506.13053.

Han Zhu, Lingxuan Ye, Wei Kang, et al. OmniVoice: Towards Omnilingual Zero-Shot Text-to-Speech with Diffusion Language Models, 2026. URL https://arxiv.org/abs/2604.00688.

Barret Zoph, Irwan Bello, Sameer Kumar, Nan Du, Yanping Huang, Jeff Dean, Noam Shazeer, and William Fedus. ST-MoE: Designing Stable and Transferable Sparse Expert Models, 2022. URL https://arxiv.org/abs/2202.08906.

A VoxCaps Construction and Verification

A.1 Data Coverage and Targeted Synthesis

We organize the source data along multiple axes that characterize both speech content and its acoustic context, including media family, speaker density, recording condition, character type, location, sound field, persistent ambience, delivery pattern, and local acoustic events. Short- and long-form drama, animation, advertising, and digital-human presentation further contribute distinct speaker archetypes, speaking styles, interaction patterns, and surrounding audio conditions. Together, these dimensions guide data collection toward broad coverage of speakers, speaking scenarios, and acoustic environments.

Beyond broad source-data coverage, we construct three targeted synthetic subsets to strengthen representation in challenging or comparatively sparse regimes: elderly speaker timbres, very short utterances, and pronunciation-sensitive content. All three subsets use CosyVoice 3 (Du et al., 2025) as the TTS teacher. Table A1 summarizes their scale and coverage targets.

For pronunciation-sensitive Chinese text, we obtain reading information from pypinyin (Huang et al., 2023) and format it as pronunciation-inpainting input to the teacher. These pronunciation cues are used only during teacher-side synthesis. The difficult-pronunciation subset retains the original orthographic text as the student input, so the student does not receive auxiliary pronunciation annotations during training. This design allows the synthesized speech to provide a more reliable acoustic realization while preserving the text form encountered by the student at inference time.

Elderly-timbre augmentation is introduced to improve coverage of elderly timbres, which remain relatively limited in our speech resources (Chen et al., 2025b). The short-utterance subset emphasizes isolated words, names, short sentences, and sub-second samples in both Chinese and English, complementing the longer-form speech that dominates many general-purpose corpora. The difficult-pronunciation subset concentrates on polyphonic characters, names, brands, English insertions, and mixed-language phrases, providing targeted supervision for lexical and cross-lingual pronunciation cases.

Table A1: Targeted synthetic subsets. We use CosyVoice 3 as the TTS teacher.

Subset	Utterances	Mean duration	Coverage target
Elderly timbres	100,000	$\sim$ 10 s	Underrepresented elderly speaker timbres
Short utterances	100,000	$\sim$ 1 s	Chinese and English words, names, and short sentences, including sub-second samples
Difficult pronunciation	100,000	$\sim$ 12 s	Polyphones, names, brands, English insertions, and mixed-language phrases

A.2 Speech Preprocessing

Separation. We use Ultimate Vocal Remover (Anjok07 & aufr33, 2020) to extract a separated vocal stem for speaker and linguistic analysis. The original mixture is retained for scene annotation and as the waveform target so that background ambience, music, and sound effects are preserved rather than removed by the speech-analysis pipeline.

Diarization and segmentation. We use the 3D-Speaker (Chen et al., 2025a) pipeline, which combines voice activity detection (VAD), CAM++ (Wang et al., 2023b) speaker embeddings, and clustering to obtain candidate speaker turns. Single-speaker segments are constrained to 1–60 seconds. Multi-speaker segments can extend to 120 seconds and are constructed to include multiple speaker turns whenever the source material permits.

Transcription and alignment. Fun-ASR-Nano (An et al., 2025) first transcribes the separated vocal stem, while SenseVoice (An et al., 2024) provides complementary recognition evidence for a second-pass transcription check. Before structure annotation, punctuation is removed from the transcript so that sentence boundaries and sentence types are recovered from the speech signal rather than inherited from the initial ASR output. A script then restores sentence structure, including declarative, interrogative, and exclamatory boundaries. Finally, Qwen3-ForcedAligner (Shi et al.,

2026) aligns the transcript to the vocal stem at the word level. The resulting word boundaries and inter-word timing gaps provide temporal evidence for refining punctuation and sentence boundaries.

A.3 Description Annotation and Persona Library

The MLLM-based description annotator receives the original audio mixture, the preliminary transcript, and dataset-specific instructions, and produces a structured description following the specification below.

Environment captures scene- or recording-level properties that remain relevant across a segment, including location, room or microphone characteristics, reverberation, spatial impression, persistent music, crowds, traffic, weather, machinery, and other ambient sources.

Speakers provides an inventory of subjects who actually speak in the segment. Each speaker entry begins with perceived gender and an approximate age range, followed, when supported by the audio, by three to five stable and recurring voice characteristics. These attributes describe relatively persistent properties of the speaker rather than turn-specific delivery.

Content follows the temporal order of the recording. Tagged speech preserves the transcript, while adjacent descriptions capture local delivery attributes such as emotion, loudness, speaking rate, pauses, emphasis, hesitation, interruption, and code-switching. Local sound events are inserted at their corresponding temporal positions using <Audio> tags.

Speaker IDs are assigned by order of first appearance and are required to be contiguous. Every tagged speaker in the content field must have a corresponding entry in the speaker inventory. We further apply structural validation to check consistency between the speaker inventory and tagged transcript while preserving the original language of the transcript. Role and archetype descriptors are included only when supported by dialogue content or audible performance, including when the persona library is used.

The following example illustrates the distinction between a stable speaker characteristic and a turn-specific delivery state:

> **Environment:** {inside an old courtyard; the space is quiet, with faint <Audio>wind and water dripping from the eaves</Audio> in the distance.} **Speakers:** {Speaker 1: a young woman with a low register, restrained manner, and cool timbre.} **Content:** {With controlled anger and a probing tone, Speaker 1 says: <S1>You will return everything you took from me, piece by piece, with your own hands.</S1>}.

Here, the low register and cool timbre are treated as recurring speaker-level attributes, whereas controlled anger and a probing tone describe the local delivery of the current turn. This separation encourages the description annotator to distinguish persistent speaker characteristics from transient expressive states.

Persona library. The description annotator selects entries from the persona library when the recording provides sufficient cues about the media type or performance setting. The persona library defines candidate descriptor categories and their annotation scopes, while acoustic and dialogue evidence determines which descriptors appear in the output. Recordings without matching persona library entries follow the general annotation specification. The persona library therefore provides structured descriptor priors rather than fixed labels. It guides the selection and placement of speaker- and content-level attributes while requiring each descriptor to be supported by observable evidence. Tables A2, A3, and A4 detail the persona library for different media types.

A.4 Automatic Filtering and Human Review

We combine automatic quality control with human review to remove unreliable examples and to identify high-quality expressive speech for subsequent training.

Automatic quality filtering. We first screen speech examples using complementary quality estimates from DNSMOS (Reddy et al., 2021) and torchaudio-SQUIM (Kumar et al., 2023a). Unless otherwise specified, we retain examples with a PESQ-related estimate of at least 2.0, STOI of at

Table A2: Persona library for animation.

Aspect	Rule
Triggers	Cartoon, anime, dubbing, role-played voices, and other audible character-performance conventions.
Stable profile	Perceived gender, approximate age, supported archetype, timbre, and habitual delivery, in that order. Archetypes such as energetic lead, restrained elder, and comic supporting role require recurring performance evidence.
Local delivery	Reactions, abrupt emotion shifts, comic timing, shouting, laughter, hesitation, and temporary changes in pace, loudness, or arousal, attached to Content spans.
Acoustic evidence	Recurring pitch range, brightness, breathiness, energy, attack, pausing, restraint, or exaggeration.
Representative distinctions	Action can involve wider dynamics and sharper bursts; romance can involve breath and controlled pauses; suspense can involve restrained clarity; historical settings can involve formal diction and measured phrasing. The annotator checks these cues against the recording.

Table A3: Persona library for short drama and film/TV drama.

Aspect	Rule
Triggers	Short, vertical, web, film, or television drama and other staged dialogue-dominated material.
Stable profile	Perceived gender and age, a role supported by the dialogue or performance, stable timbre, and recurring manner of speech.
Local delivery	Interruption, conflict, escalation, reversal, pleading, threats, commands, hesitation, and relationship-driven vocal changes, attached to their turns.
Acoustic evidence	Pacing, diction, theatrical coloring, controlled pauses, coldness, ingratiation, and changing intensity; dialogue and vocal performance provide role evidence.
Representative distinctions	A secretary-like performance can be fast and formal, a guard-like performance terse and forceful, and a deferential court performance thin, upward-inflected, and restrained. Role descriptions refer to the performance in the recording.

Table A4: Persona library for advertising and digital-human presentation.

Aspect	Rule
Triggers	Commercial voice-over, digital-human presentation, livestream selling, recommendations, marketing, or scripted promotional narration.
Stable profile	Perceived gender and age, functionally supported persona, voice class, timbre, and recurring presentation style. Host, recommender, lecturer, and service-role labels require evidence.
Local delivery	Selling-point stress, urgency, price emphasis, calls to action, question hooks, reassurance, conversational softening, and changing excitement belong to their local spans.
Acoustic evidence	Friendliness, authority, energy, information density, warmth, cadence, pause placement, emotional range, and scriptedness. A synthetic-voice label requires audible artifacts.
Representative distinctions	Livestream hosts, recommenders, explainers, business speakers, and service roles can range from conversational storytelling to energetic selling and structured explanation.

least 0.85, SI-SDR of at least 0, and MOS of at least 2.5. For collected speech, segment duration is restricted to 1–120 seconds, with an average duration of approximately 7.4 seconds. The targeted short-utterance synthetic subset described in Section A.1 is processed separately and is not subject to the same lower duration bound.

Structural normalization and consistency checks. After description annotation, we normalize the structured descriptions and verify consistency among the transcript, speaker inventory, and local event annotations. Malformed or mismatched speaker and `<Audio>` tags are repaired for pairing and placement. Word-level alignment provides temporal evidence for punctuation restoration and

explicit pause markers. Inter-word timing gaps are used to refine commas and sentence boundaries and, when appropriate, to insert <|sp|> markers. We then extract structured attributes, including gender, age, emotion, and persona descriptors, from the normalized descriptions for category-conditioned sampling and control.

Human validity review. Samples that pass the automatic stage are further audited by human listeners. The review covers transcript correctness, speaker attribution, annotation completeness, waveform quality, intelligibility, and expressive appropriateness. Listeners additionally flag common failure modes such as source-separation artifacts, leakage from overlapping or background speakers, speaker-identity confusion in dense dialogue, unclear articulation or severe elision, unnatural accent realization, electronic artifacts, audible edits, omitted sound events, and descriptors assigned to the wrong temporal region. Samples with substantial errors are excluded before preference annotation.

Preference-based expressive selection. Human review further distinguishes expressive quality among otherwise valid samples rather than treating quality control as a binary accept–reject decision. For each group of four matched valid candidates, listeners perform Best-Worst Scaling (BWS) (Kiritchenko & Mohammad, 2017), selecting the most and least expressive realizations according to naturalness, emotional strength, prosodic variation, and contextual appropriateness. The resulting relative judgments provide preference supervision over expressive realizations and are used to construct the expressive SFT subset.

B VoxVAE Architecture and Objectives

B.1 Architecture and Representation Interface

VoxVAE uses a single checkpoint to encode and reconstruct 48-kHz mono speech, general audio, and music, providing a shared continuous representation across these audio domains. The architecture is designed around a 25-Hz latent interface, balancing temporal compression with the preservation of local acoustic structure.

Encoder. The encoder progressively transforms the waveform into a compact latent sequence through five temporal reduction stages. The product of the five encoder strides is 1,920 samples, corresponding to one latent representation every 40 ms at a 48-kHz sampling rate. The channel width increases from 64 in the early waveform-processing stages to 1,536 in the deepest encoder stage as temporal resolution decreases. Residual units use dilation rates of 1, 3, and 9 to enlarge the local temporal context without introducing additional downsampling. Before each temporal reduction, a fixed anti-aliasing low-pass operation suppresses frequency components that would otherwise alias under decimation, while a lightweight learnable shortcut is retained in parallel.

Latent interface. Two pointwise projection heads predict a 96-dimensional posterior mean and log-variance at each latent position. The resulting continuous latent sequence serves as the common representation interface between waveform modeling and the downstream generative model. Because the same latent space is used for speech, general audio, and music, downstream modeling does not require separate codec representations for different source types.

Decoder. The decoder reconstructs each 40-ms latent interval through a set of learned output tokens rather than predicting the entire waveform interval with a single projection. After projecting the latent sequence, six learned output tokens are interleaved with each latent position. Local bidirectional attention allows neighboring latent and output tokens to exchange information during waveform reconstruction. Each output token predicts a 320-sample waveform patch, and the six predicted patches are concatenated to reconstruct the 1,920 samples associated with one latent interval. This factorization provides finer-grained waveform prediction within each compressed latent step while retaining the 25-Hz latent rate exposed to the generative model. The decoder contains ten Transformer blocks operating over this interleaved latent–output-token representation.

Temporal context. The encoder has an effective waveform receptive field of approximately 0.95 seconds for each latent representation. Accounting for the local interactions introduced by the decoder, the theoretical end-to-end dependency span is approximately 3.23 seconds. The representation therefore preserves substantially more acoustic context than the 40-ms latent interval alone would suggest, while maintaining a compact temporal interface for downstream generation.

Table B1 summarizes the VoxVAE configuration.

Table B1: VoxVAE configuration.

Item	Value
Input and output	48-kHz mono waveform
Latent interface	25 Hz, 96 continuous dimensions
Encoder strides	4, 4, 4, 5, 6
Residual dilations	1, 3, 9 within each reduction stage
Encoder channel range	64 to 1,536
Decoder waveform patches	6 output tokens per latent; 320 samples per token
Encoder / bottleneck / decoder parameters	51.7M / 0.3M / 355.0M
Total inference parameters	407.0M; excludes training-only modules
Training crop	3.84 s; short recordings repeated
Training data / steps / hardware	Approximately 100,000 h / 200,000 / 32 A100 GPUs
KL coefficient	0.02

B.2 Waveform-Side Training

Waveform-side training combines multi-resolution reconstruction, variational regularization, and adversarial supervision to preserve both large-scale acoustic structure and fine waveform detail.

Reconstruction objectives. We supervise reconstruction in complementary spectral and energy domains:

$$\mathcal{L}_{\mathrm{rec}} = \lambda_{\mathrm{stft}}\mathcal{L}_{\mathrm{stft}} + \lambda_{\mathrm{mel}}\mathcal{L}_{\mathrm{mel}} + \lambda_{\mathrm{eng}}\mathcal{L}_{\mathrm{eng}}.$$

The complex STFT objective compares reconstructed and reference spectra at multiple time–frequency resolutions, providing supervision for both spectral magnitude and phase structure. Multi-band Mel objectives further constrain perceptually relevant spectral envelopes across different frequency regions and analysis window sizes. The energy objective compares log energy on the same 40-ms temporal grid as the latent representation, encouraging the reconstructed waveform to preserve coarse temporal dynamics.

Latent regularization. The encoder parameterizes a diagonal Gaussian posterior, which is regularized toward a unit Gaussian through a KL-divergence term. During training, the decoder operates on reparameterized posterior samples rather than directly on the posterior mean. This regularization encourages a smooth continuous latent space while retaining stochasticity during reconstruction training.

Adversarial waveform supervision. Reconstruction losses alone primarily constrain averaged spectral structure, so we additionally employ complementary discriminator families to improve local waveform fidelity. We use MPD (Kong et al., 2020) to capture periodic waveform structure, MRD (Jang et al., 2021) to evaluate signals at multiple time–frequency resolutions, and MBCSD (Kumar et al., 2023b) to discriminate complex-STFT representations across frequency bands up to 24 kHz. Together, these discriminators provide complementary supervision over periodicity, time–frequency detail, and high-frequency reconstruction.

The complete waveform-side objective is

$$\mathcal{L}_{\mathrm{wav}} = \mathcal{L}_{\mathrm{rec}} + \lambda_{\mathrm{KL}}\mathcal{L}_{\mathrm{KL}} + \lambda_{\mathrm{adv}} \sum_{D\in\mathcal{D}} w_D \mathcal{L}^D_{\mathrm{adv}} + \lambda_{\mathrm{fm}} \sum_{D\in\mathcal{D}} \mathcal{L}^D_{\mathrm{fm}},$$

where $\mathcal{D}$ denotes the set of discriminator families and w_D controls the contribution of each family to adversarial training. Feature matching compares intermediate discriminator representations between real and reconstructed audio, providing a more stable perceptual training signal in addition to the adversarial objective. The coefficients λ_{stft}, λ_{mel}, and λ_{eng} balance the reconstruction terms, while λ_{KL}, λ_{adv}, and λ_{fm} control latent regularization, adversarial supervision, and feature matching, respectively.

B.3 Generative Predictors and Acoustic Readouts

Beyond waveform reconstruction, we introduce auxiliary predictive and acoustic objectives to encourage the latent space to retain information that is useful for downstream generative modeling.

Latent prediction. Following SAME (Parker et al., 2026), an auxiliary flow predictor is trained jointly with the autoencoder to model sequences of posterior means. The predictor is optimized with its full objective, while the gradient propagated from this objective to the encoder is down-weighted. This asymmetric gradient scaling allows the auxiliary predictor to learn the latent dynamics without letting its objective dominate representation learning in the encoder. We additionally train a causal predictor to estimate future posterior-mean patches from preceding latent context:

$$\mathcal{L}_{\text{pred}} = \sum_{j} \text{SmoothL1}\left(g_\psi(\mu_{<j}), \mu_j\right).$$

Here, g_ψ predicts the posterior-mean patch μ_j from the preceding sequence $\mu_{<j}$. The causal objective encourages information in the latent sequence to evolve in a locally predictable manner while preserving a causal dependency structure. As with the auxiliary flow predictor, its contribution to the encoder update is kept weaker than the gradient used to train the predictor itself. Predictor optimization and encoder representation learning, therefore, operate with separate gradient scales.

Acoustic readouts. We further attach lightweight readout heads that predict interpretable acoustic properties directly from the latent representation. These targets include normalized frame energy, octave-specific chroma at multiple temporal scales, and normalized energy allocation across frequency bands. The chroma targets follow the harmonic supervision used in SAME and encourage the latent representation to preserve local pitch-class and harmonic structure. Frame-energy targets provide explicit supervision for local loudness dynamics, while band-energy targets characterize how acoustic energy is distributed across the spectrum. The latter also exposes differences in effective bandwidth that may not be apparent from the nominal sampling rate alone. For example, audio originally resampled from 24 kHz contains little energy above 12 kHz even when it is subsequently converted to the common 48-kHz waveform interface.

Representation tradeoff. Together, the predictive and acoustic auxiliary tasks encourage the latent sequence to expose temporally structured and acoustically meaningful information, while waveform reconstruction remains responsible for preserving fine-grained signal detail. A representation optimized only for predictability can overly smooth rapidly changing events, whereas reconstruction favors retaining such local variation. This distinction is particularly important around consonant boundaries, note attacks, transients, and brief sound effects, where the representation must change rapidly over time. Balancing these objectives therefore trades off a smooth, generatively predictable latent space against the temporal precision required for faithful waveform reconstruction.

B.4 Augmentation, Freezing, and Normalization

Waveform augmentation. Training operates on random 3.84-second waveform crops, with shorter recordings repeated to provide a full crop. This fixed-duration sampling provides a common training shape across speech, general audio, and music while exposing different local regions of longer recordings across iterations.

For 25% of crops, we apply on-the-fly waveform mixing following EnCodec (Defossez et al., 2022). Two waveforms are combined using a coefficient $\alpha \sim \mathcal{U}(0.3, 0.7)$, and the resulting mixture is used as both the encoder input and reconstruction target. Because the model reconstructs the mixture itself rather than separating its components, this augmentation exposes the autoencoder to denser and more compositionally varied acoustic content without changing the reconstruction task.

For an additional 1% of crops, we simulate reduced effective bandwidth by resampling the waveform through an intermediate sampling rate between 8 and 44.1 kHz before converting it back to the common 48-kHz interface. This augmentation encourages the representation to remain robust to source audio whose nominal 48-kHz waveform does not contain full-band high-frequency information.

Representation freezing and normalization. After 200,000 training steps, we freeze VoxVAE and use its latent space as a fixed representation interface for VoxDirector. VoxDirector is trained

on globally normalized posterior-mean latents rather than stochastic posterior samples. At generation time, predicted latents are transformed back to the original latent scale before being passed to the frozen decoder. Freezing the autoencoder at this stage prevents the representation target from drifting while the generative model is being optimized.

Inference configuration. Auxiliary predictors, acoustic readout heads, and adversarial discriminators are used only to shape the representation during VAE training and are removed after the autoencoder is frozen. They therefore introduce no additional computation into the deployed VoxVAE encoder–decoder path. Accordingly, the reported VoxVAE inference parameter count includes only the encoder, variational bottleneck, and decoder.

C VoxDirector Implementation Details

C.1 Conditioning Paths and Reference Masks

Conditioning modes. Voice design conditions the generation on the full scene description. Voice cloning additionally uses the target text and its Content description, while reference audio provides speaker identity and recording characteristics. The two tasks share the same latent generation framework and differ in whether part of the latent sequence is fixed by the reference audio.

Let $z^\star$ denote the normalized VoxVAE posterior-mean sequence containing the target audio and, for voice cloning, the reference audio. We use a binary mask $m \in \{0,1\}^T$ to indicate frames fixed by the reference, where $m_i = 1$ denotes a reference frame and $m_i = 0$ denotes a frame to be generated. Voice design sets $m = 0$, whereas voice cloning exposes the masked reference context $r = m \odot z^\star$.

Masked flow matching. Given Gaussian noise ϵ and flow time t, interpolation is applied only to the generation region:

$$\widetilde{z}_t = (1-m) \odot [(1-t)\epsilon + t z^\star]\,. \tag{2}$$

Reference positions are therefore excluded from the noising path and are instead supplied explicitly through the reference context r. A frame-role embedding further distinguishes generated positions from fixed reference positions. The reference latents are provided through a separate context path, allowing the model to condition on them without treating them as generation targets.

Let y denote the lexical and speaker-turn conditions, c the description condition, and Q the quality level. The masked flow-matching objective is

$$\mathcal{L}_{\text{flow}} = \mathbb{E}\left[\frac{\|(1-m)\odot[v_\theta(\widetilde{z}_t, t, y, c, r, m, Q) - (z^\star - \epsilon)]\|_F^2}{\max\left(1, \sum_{i=1}^{T}(1-m_i)\right)}\right]. \tag{3}$$

The loss accumulates squared velocity errors only over generated positions and latent dimensions, then normalizes by the number of generated frames. This normalization keeps the objective comparable across examples with different amounts of reference context.

Lexical and description conditioning. The lexical branch consists of the CosyVoice 2.0 (Du et al., 2024b) tokenizer followed by a lightweight Transformer. Its text hidden states are interpolated onto the audio timeline and combined with the noised latent sequence. Speaker-turn embeddings are derived from the structured speaker tags and provide the corresponding turn structure.

For the description branch, Qwen3.0-Instruct-8B (Yang et al., 2025) provides contextual description states, with a 32B variant evaluated separately. Label embeddings explicitly distinguish spoken text, local sound effects, Environment descriptions, and other descriptive prose. These embeddings expose content type directly to cross-attention instead of requiring the model to infer the distinction solely from lexical wording.

DiT conditioning. Each DiT block combines latent self-attention with cross-attention to the description states and a dense or sparse feed-forward branch. Consequently, description information is available throughout the denoising network rather than being injected only at its input. AdaLN-Zero (Peebles & Xie, 2023) provides flow-time conditioning, while RMSNorm (Zhang & Sennrich,

2019) normalizes the hidden-state stream. The quality-level embedding is combined with the time embedding in the global conditioning stream.

C.2 Quality Metadata and Dropout

Quality metadata. When all quality estimates are available, we insert the following quality description before the Content field:

> Quality: speech clarity {STOI level}; noise level {SI-SDR level}; signal naturalness {PESQ level}; listening quality {MOS level}.

The quality description is placed before the Content field so that it is available as global contextual information rather than being associated with a specific local event. The same quality estimates are also discretized into one of four levels: low, normal, high, or unknown. At inference time, both the quality description and the discrete quality level are set to high.

Quality conditioning and dropout. The quality information enters the model through two complementary conditioning paths. The textual quality description is encoded together with the scene description and provided through cross-attention, while the discrete quality-level embedding is injected into the global conditioning stream together with the time embedding. During training, the discrete quality level is randomly replaced with unknown. This dropout prevents the model from relying exclusively on the discrete quality token while retaining the textual quality description as an additional conditioning signal. Both quality channels are learned jointly with the remaining scene and generation conditions under the flow-matching objective.

C.3 Engram Adaptation

We adapt Engram (Cheng et al., 2026) to provide local lexical memory for the bidirectional description encoder. For a token sequence $(c_1, \ldots, c_L)$, we use centered lookup windows rather than causal windows so that memory retrieval is consistent with bidirectional encoding. We use lookup orders $\mathcal{N} = \{2, 3\}$, corresponding to local windows at two granularities.

Memory lookup. For each position i, let $w_i^{(n)}$ denote the centered window of order $n \in \mathcal{N}$. With K head-specific memory tables, the entries retrieved across lookup orders and heads are concatenated as

$$e_i = \mathrm{concat}_{n\in\mathcal{N},\, k=1}^{K} \mathrm{Engram}_{n,k}\left(\mathrm{hash}_k(w_i^{(n)})\right).$$

This multi-order lookup exposes the encoder to complementary local lexical patterns before they are fused back into the hidden representation.

Gated memory injection. After projecting the description state to u_i, the retrieved memory is incorporated through a gated residual update:

$$\widetilde{u}_i = u_i + \sigma\left(\frac{\mathrm{RMSNorm}(u_i)^\top \mathrm{RMSNorm}(W_K e_i)}{\sqrt{d}} + b\right) W_V e_i.$$

The gate measures the compatibility between the current description state and the retrieved memory, allowing the lookup contribution to be modulated before the residual update. A negative initial bias b keeps the memory contribution small at the beginning of training.

We use two parallel gate branches that share the lookup tables and value projection but use distinct key projections. Their outputs are averaged before being added to the description representation. Together, the centered lookup windows and gated residual injection adapt Engram's memory mechanism to the bidirectional description encoder while preserving a controlled contribution from retrieved lexical patterns.

C.4 DUAL-ROUTER MOE

Layer structure and dual routing. Each sparse layer contains R routed acoustic experts, two task experts, and U null candidates, and alternates with a dense feed-forward layer. A task router selects the expert associated with the generation mode $\tau \in \{\text{design}, \text{clone}\}$, which is shared across all frames of the sample. In parallel, an audio router operates independently on each frame after self- and cross-attention.

For hidden state $h_{\ell,m}$ and time embedding e_t, the audio-router logits are

$$r_{\ell,m} = h_{\ell,m} + W_t e_t, \qquad \ell_{\ell,m} = W_g r_{\ell,m} + b^{\text{null}}(t), \qquad a_{\ell,m} = \ell_{\ell,m} + b.$$

Here, ℓ denotes the base logits used for mixture weighting, while a includes an additional load-balancing bias used only for expert selection. The bias b applies only to routed acoustic experts.

Auxiliary-loss-free load balancing. Let f_i denote the recent fraction of non-null assignments routed to acoustic expert i. Its load bias is updated as

$$b_i \leftarrow \text{clip}\left(b_i + \eta\left(\frac{1}{R} - f_i\right), -B, B\right).$$

The update is applied only to layers with routed assignments, balancing expert utilization without modifying the base logits used for mixture weights (Wang et al., 2024a).

Budgeted Top-P routing. A learned scalar budget

$$b_t = \sigma(W_b e_t)$$

jointly controls the routing mass, null preference, and expert capacity.

Specifically,

$$p(t) = p_{\min} + (p_{\max} - p_{\min}) b_t,$$

$$b_{\text{null}}(t) = b^{\text{null}}_{\max} + \left(b^{\text{null}}_{\min} - b^{\text{null}}_{\max}\right) b_t, \qquad c(t) = c_{\min} + (c_{\max} - c_{\min}) b_t.$$

A larger budget therefore increases the cumulative routing mass and available capacity while reducing the preference for null candidates.

During training, Gumbel noise

$$g_i = -\log(-\log u_i), \qquad u_i \sim \mathcal{U}(0,1),$$

is added to the selection logits. Candidates are ranked by

$$\text{softmax}\left(\frac{a+g}{\tau_g(n)}\right),$$

and the shortest prefix whose cumulative mass reaches $p(t)$ is selected. The temperature $\tau_g(n)$ is annealed toward a positive floor during training. Mixture weights use the same Gumbel samples with the base logits ℓ and are renormalized over the selected set. At inference, Gumbel noise is removed; deterministic selection uses $\text{softmax}(a)$, while active-set mixture weights are proportional to $\exp(\ell_i)$.

Null routing and expert capacity. Null candidates retain their selected probability mass but return zero output, thereby reducing the contribution of routed acoustic experts when mixed with non-null candidates. If all selected candidates are null, the routed acoustic output is zero, while the task-expert and residual paths remain active. The sparse feed-forward output combines the selected task-expert output with the weighted acoustic-expert output before the DiT residual connection.

For M non-null frame–expert assignments before truncation, the per-expert capacity is

$$\text{cap} = \left\lceil \frac{c(t)M}{R} \right\rceil.$$

When capacity is exceeded, assignments with the largest mixture weights are retained. Dispatch and truncation are skipped when all assignments are null. Expert dropout further reduces dependence on a fixed combination of routed experts.

C.5 Router Objectives and Training Curriculum

The auxiliary router objective is

$$\mathcal{L}_{\text{MoE}} = \lambda_z \mathcal{L}_z + \lambda_{\text{null}} \mathcal{L}_{\text{null}}, \qquad \mathcal{L}_z = \mathbb{E}_{\ell,m}\left[\left(\log \sum_j \exp(\ell_{\ell,m,j})\right)^2\right].$$

The null-collapse term penalizes the mean base probability assigned to null candidates. The full loss is $\mathcal{L}_{\text{flow}} + \omega_{\text{MoE}}(n)\mathcal{L}_{\text{MoE}}$, with the router-regularization coefficient decreasing linearly to a small positive floor.

The four supervised stages expand the task gradually (Bengio et al., 2009). The foundation starts with single-speaker cloning, then adds concatenated and natural dialogue with one to four speakers. Reference audio is dropped with probability 50%. Dense adaptation introduces descriptive speech, sampled as 70% design and 30% cloning examples. Full VoxCaps training adds surrounding audio and sparse experts, with 70% reference dropout. Final SFT uses the high-quality expressive subset selected by automatic checks and human review. Table C1 summarizes the training stages and schedules.

Table C1: VoxDirector training stages.

Stage	Training data	Schedule
Cloning foundation	23M hours of single-speaker audio; 1.7M hours of two-speaker audio	50% reference dropout
Dense description adaptation	VoxBox (Wang et al., 2025) speech	20,000 steps; 70% design, 30% cloning
Full VoxCaps training	29.22M scene-rich examples	10,000 steps; 70% reference dropout
Expressive SFT	8.2M selected examples	4,000 steps

VoxDirector training uses 64 A100 GPUs. AdamW (Loshchilov & Hutter, 2019) has $(\beta_1, \beta_2) = (0.9, 0.999)$, a peak learning rate of 7.5×10^{-5}, and a minimum of 4.0×10^{-5}.

C.6 Sampling

Guidance schedule. We use a base guidance scale of 1.5 for text-and-turn conditioning and 3.0 for full conditioning. Both guidance terms share the following time-dependent multiplier:

$$\gamma(t) = a + b(1-t)^p, \qquad (a, b, p) = (0.6, 0.6, 1.0).$$

Accordingly, the time-dependent guidance weights in Equation 1 are given by $\omega_k(t) = \gamma(t)\bar{\omega}_k$.

Integration schedule. We use a non-uniform Euler grid obtained by mapping uniformly spaced $u \in [0, 1]$ to

$$t(u) = 1 - \cos\left(\frac{\pi u}{2}\right),$$

which allocates more integration steps near the noise endpoint (Chen et al., 2024).

At each integration step, classifier-free guidance evaluates three predictions: unconditional, text-and-turn conditioned, and fully conditioned. These predictions are combined using the time-dependent guidance weights defined above. The text-and-turn branch retains only lexical and speaker-turn conditions, while the unconditional branch removes both lexical and acoustic conditioning.

D Evaluation Protocols

We evaluate the system across acoustic reconstruction, zero-shot voice cloning, voice design, and complete audio scene generation. The following protocols specify the evaluation sets, metrics, and comparison systems used in each setting.

D.1 Acoustic Reconstruction

Evaluation set. We sample 1,000 evaluation clips from each of three domains: VCTK for speech, FSD50K for general audio, and the MUSDB18-HQ test set for music. Any clips overlapping with the training data are excluded from evaluation.

Metrics. Speech reconstruction is evaluated with PESQ (Rix et al., 2001), STOI (Taal et al., 2011), MCD (Kubichek, 1993), and ViSQOL (Chinen et al., 2020). General audio and music are evaluated with ViSQOL and LSD (Gray & Markel, 1976). All reconstruction metrics compare the decoded waveform against the corresponding input recording.

Baselines and representation rate. We group reconstruction baselines by representation type. The neural codecs DAC (Kumar et al., 2023b), 48-kHz EnCodec (Defossez et al., 2022), and WavTokenizer Large Unify (Ji et al., 2025) are evaluated across all three domains. For speech, we additionally compare with the speech autoencoders VoxCPM2 AudioVAE V2 (Zhou et al., 2026) and MegaTTS 3 WaveVAE (Jiang et al., 2025). For general audio and music, we include the autoencoders from Stable Audio Open 1.0 (Evans et al., 2025) and SAME-L (Parker et al., 2026), with ACE-Step Music-DCAE (Gong et al., 2025) added for music. Each model is evaluated using its native representation. Table E1 reports nominal representation rates, using 16 bits per scalar for continuous latents and codebook-index bits for discrete latents, without accounting for entropy coding or side information.

D.2 Zero-Shot Voice Cloning

Evaluation set. SwanBench-Speech (Pan et al., 2026) evaluates zero-shot voice cloning for both monologue and dialogue under reference-audio conditioning.

Metrics. Timbre Consistency is computed as the average pairwise cosine similarity among sliding-window WavLM-TDCNN (Chen et al., 2022) embeddings. For dialogue, the metric is first computed independently for each speaker and then aggregated. Reverb Consistency is measured as the standard deviation of window-level SRMR (Falk et al., 2010) values. Sound Fidelity is measured using the reference-free SQUIM-PESQ (Kumar et al., 2023a) estimate. Content Error is the equal-weight average of Chinese CER and English WER after both are expressed as fractions. SpeechJudge (Zhang et al., 2026) is used to assess speech naturalness. Gemini 3 Pro scores Expressive Richness and Expressive Hierarchy.

Baselines. The language-model-based monologue systems are CosyVoice-2 (Du et al., 2024b), CosyVoice-3 (Du et al., 2025), FishSpeech (Fish Audio Team, 2024), GLM-TTS (Cui et al., 2025), IndexTTS-2 (Zhou et al., 2025a), and SparkTTS (Wang et al., 2025). Flow- or diffusion-based monologue systems include F5TTS (Chen et al., 2024), MegaTTS-3 (Jiang et al., 2025), and ZipVoice (Zhu et al., 2025b); we also include the long-form model VibeVoice (Peng et al., 2025). For dialogue, we compare with the multi-speaker or podcast-oriented systems FireRedTTS-2 (Xie et al., 2025b), MoonCast (Ju et al., 2025), MOSS-TTSD (MOSI, 2026), SoulX-Podcast (Xie et al., 2025a), VibeVoice (Peng et al., 2025), and ZipVoice-Dialog (Zhu et al., 2025a).

D.3 VOICE DESIGN

Instruction-following evaluation. InstructTTSEval (Huang et al., 2025) evaluates Acoustic-Parameter Specification, Descriptive-Style Directive, and Role-Play accuracy using its Gemini 2.5 Pro judging protocol.

Voice-design baselines. Parler-TTS-large (Lyth & King, 2024) and VoxInstruct (Zhou et al., 2024) provide description- or instruction-conditioned TTS. VoiceSculptor (Hu et al., 2026), Qwen3-TTS-VD (Qwen Team, 2026a), MOSS-VoiceGenerator (Huang et al., 2026), and VoxCPM2 (Zhou et al., 2026) support voice creation or design from descriptions. MiMo-Audio (Zhang et al., 2025) is included as a general audio-language-model baseline. The baseline results reported in Table 3 are taken from the VoxCPM2 report (Zhou et al., 2026).

Role-play evaluation. VoxBench-Instruct contains 180 scene descriptions, comprising 60 advertising scenes, 60 short-drama scenes, and 60 general scenes. Each item is rated by five professional listeners on a 1–5 scale for Overall Expressiveness, Prosodic Naturalness, Audio Fullness, and Scene Appropriateness. Mean MOS is computed as the average of the four dimension-level scores.

Role-play baselines. MOSS-VoiceGenerator (Huang et al., 2026) and Qwen3-TTS-12Hz-1.7B-VD (Qwen Team, 2026a) are voice-design systems, while MiMo-Audio-7B-Instruct (Zhang et al., 2025) is a general audio-language model.

D.4 COMPLETE AUDIO SCENE GENERATION

Evaluation set. VoxBench-Scene contains 500 everyday audio scenes covering Chinese and English, with ambience, local sound effects, music, and speech. Of these, 200 involve two to four speakers, 300 include local sound effects, and 100 contain only non-speech audio and ambience. These coverage categories are not mutually exclusive.

Conditioning settings. The benchmark evaluates both voice design and voice cloning in complete audio scenes. For voice cloning, 200 cases provide reference audio for speaker identity and Content descriptions for local control. The remaining 300 cases use scene descriptions without reference audio, covering voice design and non-speech audio generation. Voice design uses the full Environment, Speakers, and Content descriptions.

Metrics. The `gemini-3.5-flash` judge rates Instruction Accuracy, Acoustic Quality, and Overall Expressiveness on a 1–5 scale. For 100 audio clips rated by human listeners, Spearman rank correlations between the judge scores and mean human ratings are 0.722, 0.678, and 0.641 for Instruction Accuracy, Acoustic Quality, and Overall Expressiveness, respectively.

Baselines. Our cascaded baseline combines Qwen3-TTS (Qwen Team, 2026a) for speech with TangoFlux (Hung et al., 2024) for non-speech audio. Speech is generated turn by turn, using the 1.7B VoiceDesign variant with speaker and delivery descriptions for voice design, and the 1.7B Base variant with speaker-specific reference audio for voice cloning. The same speaker description or reference audio is reused across turns. TangoFlux generates ambience, music, and local sound effects from their descriptions and requested durations. We arrange speech turns and local sound effects according to Content, place persistent background tracks across the scene, and mix the tracks with fixed level-balancing rules. Scenes without speech use only TangoFlux. We also include Dasheng AudioGen (Mei et al., 2026), which jointly generates speech, music, sound effects, and ambience from structured descriptions. As it does not support reference-audio voice cloning, its evaluation uses only description-only cases.

E EXTENDED RESULTS

E.1 RECONSTRUCTION AND GENERATION RESULTS

Table 1 reports three-domain reconstruction, and Table 2 reports zero-shot voice cloning in monologue and dialogue. This appendix lists native latent configurations, VoxVAE ablations, Seed-TTS-Eval voice cloning results, and category-level scene results.

Table E1: Native latent configurations. Continuous nominal rates assume 16 bits per scalar; discrete rates count codebook indices only.

Model	Type	Sample rate (kHz)	Frame rate (Hz)	Latent/frame	Nominal kbps
ACE-Step Music-DCAE	Continuous	44.1	10.77	128 scalars	22.05
VoxCPM2 AudioVAE V2	Continuous	$16 \rightarrow 48$	25	64 scalars	25.60
Stable Audio Open 1.0	Continuous	44.1	21.53	64 scalars	22.05
SAME-L	Continuous	44.1	10.77	256 scalars	44.10
MegaTTS 3 WaveVAE	Continuous	24	25	32 scalars	12.80
DAC	RVQ	44.1	86.13	$9 \times$ 10-bit	7.75
EnCodec	RVQ	48	150	$16 \times$ 10-bit	24.00
WavTokenizer Large Unify	VQ	24	40	$1 \times$ 12-bit	0.48
VoxVAE	Continuous	48	25	96 scalars	38.40

Among the 25-Hz continuous models in Table E1, VoxVAE uses the widest per-frame latent (96 scalars) while keeping a short sequence for long scenes. Its nominal 38.40-kbps rate reflects this allocation of capacity to local acoustic detail in a shared 48-kHz representation, rather than a measured compressed-audio bitrate.

E.2 VoxVAE Ablations

Table E2 compares VoxVAE reconstruction across speech, general audio, and music using the same test sets as Table 1. The standard VoxVAE uses 96-dimensional latents at 25 Hz; the latent-size variant uses 64 dimensions.

The full VoxVAE leads all four speech metrics, with the largest PESQ and MCD degradation when acoustic readouts are removed. Removing AR loss increases LSD on general audio and music but improves ViSQOL, indicating a trade-off between spectral fidelity and perceptual similarity. Removing flow loss or acoustic readouts has smaller, mixed effects on general audio, while the full model retains the lowest music LSD.

Table E2: VoxVAE reconstruction ablations on speech, general audio, and music. The standard VoxVAE uses 96-dimensional latents.

	Speech				General audio		Music	
Setting	PESQ ↑	STOI ↑	MCD ↓	ViSQOL ↑	ViSQOL ↑	LSD ↓	ViSQOL ↑	LSD ↓
VoxVAE	4.168	0.968	0.964	4.125	4.127	0.946	4.262	0.917
w/o AR Loss	3.918	0.956	1.220	3.987	4.317	0.961	4.404	0.933
w/o Flow Loss	4.045	0.964	1.045	3.952	4.119	0.945	4.223	0.921
w/o Acoustic Readout	3.768	0.956	1.332	3.891	4.147	0.944	4.147	0.922
Latent size = 64	3.954	0.961	1.142	3.889	4.116	0.953	4.120	0.949

E.3 Seed-TTS-Eval Voice Cloning

VoxDirector ranks first in speaker similarity on Test-EN and Test-ZH-Hard and second on Test-ZH. Its content errors remain competitive with several TTS-focused baselines, despite its broader multi-task and expressive objectives.

Compared systems. The open-source baselines are F5-TTS (Chen et al., 2024), MaskGCT (Wang et al., 2024b), CosyVoice (Du et al., 2024a), CosyVoice 2 (Du et al., 2024b), CosyVoice 3 (0.5B) (Du et al., 2025), Spark-TTS (Wang et al., 2025), FireRedTTS (Guo et al., 2024), FireRedTTS-2 (Xie et al., 2025b), OpenAudio-s1-mini (Fish Audio, 2025), and IndexTTS2 (Zhou et al., 2025a). They also include VibeVoice (Peng et al., 2025), HiggsAudio-v2 (Boson AI, 2025), ZipVoice (Zhu et al., 2025b), MOSS-TTS (Gong et al., 2026), Qwen3-TTS (Qwen Team, 2026a), Fish Audio S2 (Liao et al., 2026), OmniVoice (Zhu et al., 2026), VoxCPM (Zhou et al., 2025b), and VoxCPM2 (Zhou et al., 2026).

Table E3: Zero-shot voice cloning on Seed-TTS-Eval (Anastassiou et al., 2024) against open-source baselines. Baseline results are from Zhou et al. (2026). SIM, WER, and CER are percentages. Bold and underlined values mark the best and second-best results among the systems shown; dashes denote unreported values.

	Test-EN		Test-ZH		Test-ZH-Hard	
Model	SIM ↑	WER ↓	SIM ↑	CER ↓	SIM ↑	CER ↓
F5-TTS	67.0	2.00	76.0	1.53	71.3	8.67
MaskGCT	71.7	2.62	77.4	2.27	–	–
CosyVoice	60.9	4.29	72.3	3.63	70.9	11.75
CosyVoice 2	65.9	3.09	75.7	1.38	72.4	6.83
CosyVoice 3 (0.5B)	71.8	2.02	78.0	1.16	<u>75.8</u>	<u>6.08</u>
Spark-TTS	57.3	3.14	66.0	1.54	–	–
FireRedTTS	46.0	3.82	63.5	1.51	62.1	17.45
FireRedTTS-2	66.5	1.95	73.6	1.14	–	–
OpenAudio-s1-mini	55.0	1.94	68.5	1.18	64.3	23.37
IndexTTS2	70.6	2.23	76.5	1.03	75.5	7.12
VibeVoice	68.9	3.04	74.4	1.16	–	–
HiggsAudio-v2	67.7	2.44	74.0	1.50	65.6	55.07
ZipVoice	66.8	1.64	75.1	1.40	–	–
MOSS-TTS	73.4	1.85	78.8	1.20	–	–
Qwen3-TTS	71.7	<u>1.23</u>	77.0	1.22	74.8	6.76
Fish Audio S2	–	**0.99**	–	**0.54**	–	**5.99**
OmniVoice	74.1	1.60	77.7	<u>0.84</u>	–	–
VoxCPM	72.9	1.85	77.2	0.93	73.0	8.87
VoxCPM2	<u>75.3</u>	1.84	**79.5**	0.97	75.3	8.13
VoxDirector	**77.2**	2.34	<u>78.9</u>	1.21	**77.6**	7.12

E.4 VoxBench-Instruct Category Results

Tables E4, E5, and E6 report the category-level results. In advertising scenes, VoxDirector leads every rated dimension in Table E4, with the largest margin over the next-best model in scene appropriateness (3.73 versus 3.38). Its higher audio-fullness and appropriateness scores indicate fuller voices and better delivery for the specified scene.

Table E4: VoxBench-Instruct advertising. Best results are in **bold**.

Model	Mean	Exp.	Nat.	Full.	App.
MOSS-VoiceGenerator	3.13	2.89	3.10	3.75	2.76
MiMo-Audio	3.13	2.96	3.17	3.49	2.89
Qwen3-TTS-VD	3.71	3.62	3.73	4.12	3.38
VoxDirector	**3.88**	**3.76**	**3.82**	**4.21**	**3.73**

Table E5: VoxBench-Instruct short drama. Best results are in **bold**.

Model	Mean	Exp.	Nat.	Full.	App.
MiMo-Audio	3.23	3.06	3.24	3.72	2.89
MOSS-VoiceGenerator	4.06	3.94	4.11	4.33	3.87
Qwen3-TTS-VD	4.35	4.20	4.35	**4.63**	4.21
VoxDirector	**4.45**	**4.36**	**4.47**	4.60	**4.36**

Table E6: VoxBench-Instruct general scenes. Best results are in **bold**.

Model	Mean	Exp.	Nat.	Full.	App.
MiMo-Audio	3.28	3.17	3.39	3.49	3.05
MOSS-VoiceGenerator	3.62	3.42	3.65	4.10	3.29
Qwen3-TTS-VD	4.22	4.05	4.18	4.59	4.04
VoxDirector	**4.34**	**4.24**	**4.30**	**4.60**	**4.21**