A Unified Model for Speech, Music, and Sound Effect Generation with Text Instructions

Anonymous ACL submission

Abstract

Generative audio modeling has largely been fragmented into specialized tasks, text-to-speech (TTS), text-to-music (TTM), and text-to-audio (TTA), each operating under heterogeneous control paradigms. Unifying these modalities remains a fundamental challenge due to the intrinsic dissonance between structured semantic representations (speech/music) and unstructured acoustic textures (sound effects). In this paper, we introduce **UniSonate**, a unified flow-matching framework capable of synthesizing speech, music, and sound effects through a standardized, reference-free natural language instruction interface. To reconcile structural disparities, we propose a novel dynamic token injection mechanism that projects unstructured environmental sounds into a structured temporal latent space, enabling precise duration control within a phoneme-driven Multimodal Diffusion Transformer (MM-DiT). Coupled with a multi-stage curriculum learning strategy, this approach effectively mitigates cross-modal optimization conflicts. Extensive experiments demonstrate that UniSonate achieves state-of-the-art performance in instruction-based TTS (WER 1.47%) and TTM (SongEval Coherence 3.18), while maintaining competitive fidelity in TTA. Crucially, we observe *positive transfer*, where joint training on diverse audio data significantly enhances structural coherence and prosodic expressiveness compared to single-task baselines. Audio samples are available at https://demoanonymity.github.io/UniSonate/.

1 Introduction

The landscape of neural audio generation has long been fragmented. While specialized models for Text-to-Speech (TTS) (Chen et al., 2024; Du et al., 2024a), Text-to-Music (TTM) (Copet et al., 2023; Gong et al., 2025), and Text-to-Audio (TTA) (Liu et al., 2023) have achieved remarkable fidelity, they operate under heterogeneous control paradigms. TTS systems typically demand reference audio for timbre cloning and strict phoneme alignment; TTM models rely on lyrics or specialized tags; whereas TTA models generate unstructured textures from open-ended captions. This fragmentation creates a significant barrier to developing general-purpose audio intelligence capable of synthesizing complex auditory scenes—such as dialogue overlaid with background music and environmental effects—within a single probabilistic framework.

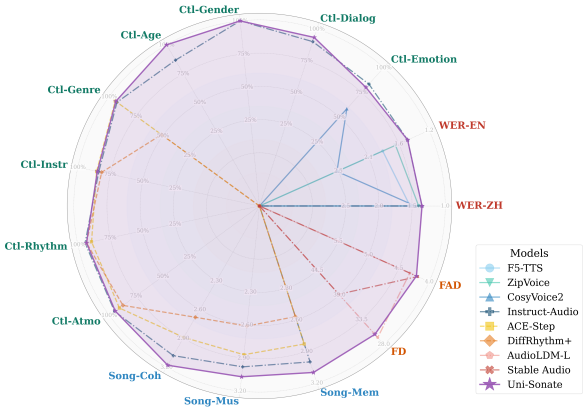

Figure 1: Holistic capability assessment across Speech (SeedTTS-WER, Control), Music (SongEval), and Sound Effects (FAD, FD). Unlike specialized baselines restricted to specific domains, UniSonate achieves *pan-modal coverage*. It demonstrates superior instruction-following and structural coherence in structured tasks (TTS/TTM) while effectively extending to unstructured TTA.

Previous attempts at unification have faced substantial limitations regarding consistency and coverage. Models like Vevo2 (Zhang et al., 2025) and CosyVoice (Du et al., 2024a) unify speech and singing but remain dependent on reference audio

∗ Corresponding author.

The name "Sonate" is derived from the musical term "Sonata", symbolizing the model's comprehensive capabilities in audio generation.

for timbre control, lacking the flexibility of natural language description. UniAudio (Yang et al., 2023) and AudioBox (Vyas et al., 2023) support multiple tasks but resort to inconsistent input formats or task-specific fine-tuning, failing to achieve a truly unified interface. To date, no single framework has simultaneously achieved (1) unified generation of speech, music, and sound effects, (2) a consistent instruction-only input format, and (3) reference-free control over fine-grained acoustic attributes. Achieving this unification presents a fundamental challenge: the intrinsic dissonance between *structured* and *unstructured* semantic representations. Speech and music require precise temporal alignment between discrete units (phonemes/notes) and acoustic realization. Conversely, sound effects (SFX) are inherently holistic and unstructured, lacking rigid temporal boundaries. Simply training a model on concatenated datasets often leads to negative transfer, where the variance of unstructured sound effects destabilizes the articulation required for high-quality speech. While InstructAudio (Qiang et al., 2025) successfully bridged speech and music via structured instruction control, the integration of unstructured environmental sounds remains an unresolved optimization conflict.

In this paper, we introduce **UniSonate**, a unified generative framework based on conditional flow matching that synthesizes speech, music, and sound effects through a standardized interface. To reconcile the structural disparities, we propose a novel *Instruction-Content Alignment* paradigm. Beyond mere format standardization, this paradigm seeks to align the semantic space of natural language instructions (e.g., "raspy male voice, sorrowful tone") with the acoustic manifold of diverse audio modalities. We decouple conditioning into two streams: Instruction for high-level attribute control, and Content for temporal structure.

To bridge the gap between discrete linguistic processing and continuous environmental audio, we introduce dynamic token injection. Theoretically, this mechanism acts as the symbolization of unstructured acoustic events, projecting holistic sound effects into a pseudo-linguistic discrete space. By injecting learnable `[SFX]` tokens, we enable the transformer to process non-verbal audio with the same discrete symbolic reasoning used for phoneme articulation. This allows the model to infer duration and progression for sound effects using shared attention mechanisms, effectively treating all audio generation as a sequence modeling problem. To harmonize these diverse modalities, UniSonate employs a dual-stream MM-DiT trained via a multi-stage curriculum learning strategy. By progressively expanding from structured speech to semi-structured music and finally to unstructured effects, we mitigate optimization conflicts and catastrophic forgetting. Our contributions are summarized as follows:

- We propose UniSonate, the first flow-matching framework to unify TTS, TTM, and TTA tasks under a consistent, reference-free natural language instruction interface, achieving deep semantic alignment between textual descriptions and acoustic features.
- We introduce Dynamic Token Injection, a mechanism that symbolically represents unstructured acoustic events, enabling precise duration control for sound effects within a phoneme-driven architecture.
- Extensive experiments demonstrate that UniSonate achieves state-of-the-art performance in instruction-based TTS (WER 1.47%) and TTM (SongEval Coherence 3.18). Crucially, we observe *positive transfer*: joint training with diverse audio data significantly enhances the structural coherence and prosodic expressiveness of generated speech compared to single-task baselines.

2 Related Work

2.1 Text-to-Speech

Driven by generative AI, high-fidelity TTS models based on language modeling (e.g., VALL-E (Chen et al., 2025b)) and flow matching (e.g., F5-TTS (Chen et al., 2024), ZipVoice (Zhu et al., 2025)) have emerged. While proficient in reference-based cloning, they often lack flexibility. Consequently, research has pivoted to instruction-based control. Pioneers like PromptTTS (Guo et al., 2023) and InstructTTS (Yang et al., 2024) mapped prompts to styles, while ControlSpeech (Ji et al., 2024) and CosyVoice (Du et al., 2024a) advanced style-timbre decoupling. Recently, IndexTTS2 (Zhou et al., 2025) introduced precise duration mechanisms, and LLM-native frameworks like Spark-TTS (Wang et al., 2025b) and EmoVoice (Yang et al., 2025b) leveraged chain-of-thought reasoning for fine-grained prosody control.

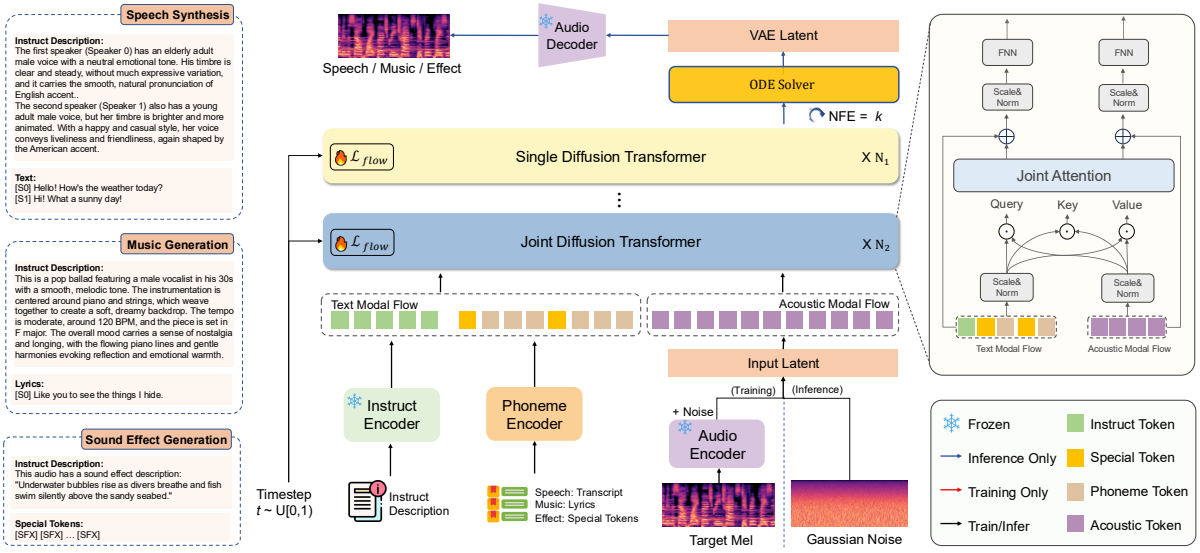

Figure 2: The overall architecture of **UniSonate**. The framework employs a dual-stream MM-DiT based on conditional flow matching. The input follows the *Instruction-Content Alignment* paradigm, unifying natural language instructions with content sequences, utilizing phonemes for speech/music and special token Injection (via learnable `[SFX]` tokens) for sound effects. These semantic conditions interact with acoustic latents (compressed by a Mel-VAE) through Joint Diffusion Transformer layers to enable unified audio generation.

2.2 Text-to-Music

Text-to-music generation has shifted from symbolic to direct audio synthesis. While early raw-waveform approaches like Jukebox (Dhariwal et al., 2020) proved inefficient, MusicGen (Copet et al., 2023) established a robust autoregressive framework using discrete tokens, recently scaled by YuE (Yuan et al., 2025) and SongGen (Liu et al., 2025) for full-length songs. Parallelly, latent diffusion models have gained traction for their controllability. AudioLDM 2 (Liu et al., 2024) unified audio generation, and MUSTANGO (Melechovsky et al., 2024) enhanced attribute control. Notably, the DiffRhythm series (Ning et al., 2025; Chen et al., 2025a) pioneered full-song synthesis with diffusion, achieving fidelity comparable to commercial systems like Suno (AI, 2024a) and Udio (AI, 2024b).

2.3 Text-to-Audio

General audio generation has evolved from discrete autoregressive models like AudioGen (Kreuk et al., 2022) to robust latent diffusion approaches exemplified by AudioLDM (Liu et al., 2023) and Make-An-Audio (Huang et al., 2023). Subsequently, research shifted toward unified foundation models like UniAudio (Yang et al., 2023) leverages LLM-based tokenization for diverse modalities, while AudioBox (Vyas et al., 2023) employs flow matching to generate speech, music, and sound within a single architecture. Recent works have further expanded cross-modal capabilities: Vevo2 (Zhang et al., 2025) bridges speech and singing via unified prosody learning, while Kling-Foley (Wang et al., 2025a) and MMAudio (Cheng et al., 2025) utilize multimodal Diffusion Transformers to achieve high-fidelity video-to-audio synchronization, demonstrating the potential of complex context modeling.

3 Method

Uni-Sonate unifies speech, music, and sound effects within a single probabilistic framework based on conditional flow matching. As shown in Figure 2, it employs a dual-stream Multimodal Diffusion Transformer (MM-DiT) that processes a standardized *Instruction-Content* input. The core innovation lies in handling structured (speech/music) and unstructured (SFX) modalities via a unified attention mechanism, enabled by our Dynamic Token Injection strategy.

3.1 Unified MM-DiT Dual-Stream Architecture

We employ a MM-DiT architecture underpinned by conditional flow matching (Lipman et al., 2022),

designed to facilitate bidirectional information flow between semantic conditions and acoustic latents. The architecture is composed of two parallel processing streams—the *Text Stream* and the *Audio Stream*, which interact via joint attention layers.

Text Modality Stream (Conditioning). To unify the heterogeneous control inputs of speech, music, and SFX, we standardize the conditioning signal into a composite sequence. For a given sample, the text stream input C_{text} is constructed by temporally concatenating the instruction embedding and the content embedding. Formally, let $E_I \in \mathbb{R}^{B\times L_I\times D}$ denote the embeddings derived from the natural language instruction (e.g., "A happy male voice," "Upbeat jazz piano," or "Footsteps on gravel"), extracted via a frozen pre-trained instruction encoder (Qwen2.5-7B). Let $E_C \in \mathbb{R}^{B\times L_C\times D}$ denote the content embeddings. The nature of E_C varies by task but remains structurally consistent to the transformer. Speech & Music: E_C corresponds to the phoneme sequence derived from the transcript or lyrics. Since SFX lacks linguistic content, E_C is composed of a sequence of learnable `[SFX]` special tokens. The length of this token sequence is dynamically adjusted to align with the target audio duration, serving as a temporal anchor for the generation process.

The final conditioning input is $C_{\text{text}} = \text{Concat}(E_I, E_C) \in \mathbb{R}^{B\times(L_I+L_C)\times D}$.

Audio Modality Stream (Generation). The audio stream processes the noisy latent representations x_t. We utilize a pre-trained Mel-VAE to compress the 44.1kHz raw waveform into a continuous latent space with a downsampling factor of 1024, yielding a compact representation x_0. During training, x_t represents the linear interpolation between the clean latent x_0 and Gaussian noise, following the flow-matching formulation.

Joint Stream Interaction. The two streams interact through a stack of N_2 Joint Diffusion Transformer layers. In each layer, the text representations C_{text} and audio latents x_t are processed by separate self-attention blocks to model intra-modal dependencies. Subsequently, a joint attention mechanism concatenates the queries, keys, and values from both modalities, enabling the model to align semantic instructions and content tokens with acoustic textures. This allows the audio stream to attend to the instruction for global style control (e.g., timbre, genre) and the content sequence for fine-grained structural control (e.g., articulation, rhythm). Following the joint layers, the streams are decoupled. To refine the acoustic details, the audio latents pass through an additional set of N_1 Single Diffusion Transformer layers where only self-attention is applied.

Training Objective. The model is trained to estimate the vector field v_θ that transforms the noise distribution to the data distribution. The optimization objective is defined as:

$$\mathcal{L}_{\text{CFM}} = \mathbb{E}_{t,x_0,x_1,C_{\text{text}}}\left\|v_\theta(t, C_{\text{text}}, x_t) - (x_1 - x_0)\right\|^2 \tag{1}$$

where $t \in [0, 1]$ is the timestep, and $x_t = tx_1 + (1-t)x_0$. During inference, the target audio latents are reconstructed by integrating the predicted velocity field using an ODE solver (Euler method).

3.2 Unified Input Representation with Dynamic Special Tokens

A central challenge in unified audio modeling lies in reconciling the structural disparities between tasks that possess intrinsic linguistic content (speech and music) and those that do not (sound effects). To address this, we propose a standardized input paradigm, the *Instruction-Content Alignment* framework, which extends the instruction-phoneme format of InstructAudio to accommodate the unstructured nature of environmental sounds. As depicted in Figure 2, the unified text modality input is composed of two primary segments: a natural language instruction and a content sequence.

The instruction serves as the high-level semantic controller, provided as a natural language prompt. For speech synthesis (TTS), this description specifies speaker attributes such as gender, age, emotion, style, and accent. To support multi-speaker dialogue, we adopt a syntax where distinct descriptions are provided for each speaker, and their respective utterances in the content sequence are prefixed with speaker-id tokens (e.g., `[S0]`, `[S1]`). For music generation (TTM), the instruction details musical parameters including genre, instrumentation, tempo, mood, and vocal characteristics (if applicable). For sound effects (SFX), the instruction describes the acoustic event or scene (e.g., "A dog barking in a busy street," "Thunder rolling in the distance").

The content sequence provides fine-grained structural guidance. For TTS and TTM, this is straightforward: the input text or lyrics are converted into a phoneme sequence C_{text} using a Grapheme-to-Phoneme (G2P) model (Qiang et al., 2022), offering precise temporal alignment for ar-

ticulation and melody. However, SFX generation lacks textual transcripts. To integrate SFX into this phoneme-driven architecture without architectural modification, we introduce a dynamic token injection strategy. We define a learnable special token, `[SFX]`, to serve as a pseudo-phoneme unit. Crucially, the sequence length of these tokens is not arbitrary; it acts as a proxy for temporal duration, enabling the model to infer the length of the audio event.

Let T_{audio} be the target duration of the sound effect in seconds. We determine the number of special tokens, L_{sfx}, by aligning with the temporal density of speech phonemes. Specifically, we calculate a global scaling factor λ from our speech corpus, representing the average phoneme-to-duration ratio:

$$\lambda = \frac{1}{N} \sum_{i=1}^{N} \frac{\text{len}(P_i)}{\text{duration}(A_i)} \tag{2}$$

where P_i and A_i are the phoneme sequence and audio waveform of the i-th speech sample, respectively. For any given SFX query with a desired duration T_{target}, the content sequence is constructed as a repetition of the special token:

$$C_{\text{sfx}} = [\texttt{[SFX]}] \times \lfloor \lambda \cdot T_{\text{target}} \rfloor \tag{3}$$

Crucially, we employ repeated tokens rather than a single global `<duration>` embedding to create temporal anchors. These anchors provide physical "length" in the input space, allowing the MM-DiT's cross-attention to "walk" through the sequence step-by-step. This mechanism mimics the monotonic alignment of phonemes, effectively treating temporal unfolding in SFX as a sequence modeling problem. This design ensures structural integrity for long-form generation and unifies the attention mechanism across all modalities.

3.3 Multi-Stage Curriculum Learning Strategy

While the unified architecture enables joint modeling, the intrinsic complexity of the generation tasks varies significantly. Speech synthesis requires high-fidelity capture of linguistic articulation and prosody; music generation demands long-term structural coherence for melody and rhythm; sound effects involve diverse, unstructured acoustic textures. Direct joint training on all modalities from scratch often leads to optimization conflict or negative transfer, where the model struggles to converge on fine-grained speech details due to the high variance of environmental sounds. To mitigate this, we employ a multi-stage curriculum learning strategy (Algorithm 1). As shown, the training progressively expands from highly structured speech (Stage 1) to semi-structured music (Stage 2), and finally incorporates unstructured sound effects (Stage 3), ensuring robust alignment learning before generalizing to diverse acoustic scenes.

Algorithm 1 Multi-Stage Curriculum Learning Strategy

Datasets: $\mathcal{D}_S$ (Speech), $\mathcal{D}_M$ (Music), $\mathcal{D}_E$ (Effects)
Initialize: Model parameters θ
Hyperparameters: $E_1 = 1$ (Stage 1 Epochs), $E_2 = 2$ (Stage 2 Epochs)
$epoch \leftarrow 0$
while training not converged **do**
  $epoch \leftarrow epoch + 1$
  if $epoch \leq E_1$ **then**
    Stage 1: Speech Anchoring
    $\mathcal{D}_{curr} \leftarrow \mathcal{D}_S$
  else if $epoch \leq E_1 + E_2$ **then**
    Stage 2: Semantic Expansion
    $\mathcal{D}_{curr} \leftarrow \mathcal{D}_S \cup \mathcal{D}_M$
  else
    Stage 3: Universal Generalization
    $\mathcal{D}_{curr} \leftarrow \mathcal{D}_S \cup \mathcal{D}_M \cup \mathcal{D}_E$
  end if
  for batch $B \sim \mathcal{D}_{curr}$ **do**
    $C_{\text{text}}, x_0 \leftarrow \text{PrepareInput}(B)$
    Sample $t \sim \mathcal{U}(0,1)$, $\epsilon \sim \mathcal{N}(0, I)$
    $x_t \leftarrow t x_1 + (1-t) x_0$
    $\mathcal{L} \leftarrow \|v_\theta(t, C_{\text{text}}, x_t) - (x_1 - x_0)\|^2$
    Update θ via $\nabla_\theta \mathcal{L}$
  end for
end while

4 Experiments

We compare Uni-Sonate against domain-specific SOTA models using standard objective metrics and subjective MOS. Detailed baseline configurations and metric definitions are in Appendix A.

4.1 Datasets

We construct a large-scale unified audio corpus comprising three distinct modalities: speech, music, and sound effects. The dataset consists of 50K hours of speech and 20K hours of music collected from internet sources, consistent with InstructAudio, alongside a newly introduced collection of 1.5 million sound effect (SFX) clips. We apply a standardized internal data processing pipeline to generate unified natural language instructions across all tasks. For speech, instructions cover attributes including gender, age, emotion, style, and accent. Music instructions detail genre, instrument, rhythm, and atmosphere. For the SFX data, instruc-

Table 1: Comprehensive comparison of capabilities across all baselines. UniSonate is the only framework that supports Speech, Music, and Sound Effect generation simultaneously within a single model, while providing the most comprehensive text-based control for speech synthesis.

Model	Params	Data Scale	Generation Tasks			Control Capabilities					
			Speech	Music	SFX	Gender	Age	Emo	Style	Accent	Dialogue
TTS Models											
MaskGCT (Wang et al., 2024)	1B	100k hrs (S)	✓	✗	✗	✗	✗	✗	✗	✗	✗
E2-TTS (Eskimez et al., 2024)	333M	100k hrs (S)	✓	✗	✗	✗	✗	✗	✗	✗	✗
F5-TTS (Chen et al., 2024)	336M	100k hrs (S)	✓	✗	✗	✗	✗	✗	✗	✗	✗
ZipVoice (Zhu et al., 2025)	123M	100k hrs (S)	✓	✗	✗	✗	✗	✗	✗	✗	✗
CosyVoice1 (Du et al., 2024a)	416M	170k hrs (S)	✓	✗	✗	✗	✗	✓	✓	✓	✗
CosyVoice2 (Du et al., 2024b)	618M	167k hrs (S)	✓	✗	✗	✗	✗	✓	✓	✓	✗
TTM Models											
DiffRhythm+ (Chen et al., 2025a)	1B	120k hrs (M)	✗	✓	✗	–	–	–	–	–	–
ACE-Step (Gong et al., 2025)	3B	100k hrs (M)	✗	✓	✗	–	–	–	–	–	–
TTA Models											
AudioLDM-L (Liu et al., 2023)	739M	634k clips (E)	✗	✗	✓	–	–	–	–	–	–
Tango-FT (Ghosal et al., 2023)	866M	45k clips (E)	✗	✗	✓	–	–	–	–	–	–
EzAudio-XL (Hai et al., 2024)	875M	270k clips (E)	✗	✗	✓	–	–	–	–	–	–
Stable Audio (Evans et al., 2025)	1.0B	486k clips (E)	✗	✗	✓	–	–	–	–	–	–
GenAU-L (Haji-Ali et al., 2024)	1.2B	811k clips (E)	✗	✗	✓	–	–	–	–	–	–
Unified Models											
InstructAudio (Qiang et al., 2025)	1.3B	50k hrs (S) + 20k hrs (M)	✓	✓	✗	✓	✓	✓	✓	✓	✓
UniSonate (Ours)	1.3B	50k hrs (S) + 20k hrs (M) + 1.5M clips (E)	✓	✓	✓	✓	✓	✓	✓	✓	✓

Note: **S**=Speech, **M**=Music, **E**=Sound Effects (clips). **SFX**: Sound Effects Generation.

Table 2: Performance comparison of instruction-based TTS on control accuracy, similarity, distortion/error metrics, and subjective evaluation. UniSonate demonstrates superior signal quality and dialogue control while maintaining competitive expressiveness.

Model	Classification Control Accuracy Rate (%)↑						Similarity↑		Distortion/Error ↓				MOS↑	
	Gender	Age	Emotion	Style	Accent	Dialog	Spk	Emo	LSD	MCD	MSEP	MR	QMOS	NMOS
Ground Truth	100.00	100.00	100.00	100.00	100.00	100.00	1.00	1.00	0.00	0.00	0.00	0.00	–	–
CosyVoice2(Du et al., 2024b)	–	–	58.33	65.00	**100.00**	–	0.68	0.53	2.57	7.11	547.87	0.46	**3.90 ± 0.11**	**3.65 ± 0.22**
InstructAudio(Qiang et al., 2025)	**100.00**	**86.67**	**83.33**	**86.67**	**100.00**	90.00	0.76	**0.71**	1.88	5.71	437.58	0.33	3.73 ± 0.24	3.46 ± 0.32
UniSonate	**100.00**	**86.67**	80.00	80.00	**100.00**	**93.33**	**0.77**	0.67	**1.79**	**5.46**	**422.36**	**0.31**	3.83 ± 0.17	3.50 ± 0.18

Table 3: Comparison of Word Error Rate (WER) performance. UniSonate achieves the best recognition accuracy on both English and Chinese datasets.

Model	WER(%)↓	
	EN	ZH
Ground Truth	2.14	1.25
MaskGCT(Wang et al., 2024)	2.26	2.40
E2-TTS(Eskimez et al., 2024)	2.49	1.91
F5-TTS(Chen et al., 2024)	1.89	1.53
ZipVoice(Zhu et al., 2025)	1.70	1.40
CosyVoice1(Du et al., 2024a)	4.29	3.63
CosyVoice2(Du et al., 2024b)	2.57	1.45
InstructAudio(Qiang et al., 2025)	1.52	1.35
UniSonate (Ours)	**1.47**	**1.25**

tions describe acoustic events (e.g., "footsteps," "glass breaking") and environmental scenes. All audio samples are standardized to a 44.1kHz sampling rate, with clip durations ranging from 2 to 20 seconds. The speech data maintains balanced 1:1 ratios for Chinese-English languages and gender distribution, including 0.5% dialogue-specific data to support multi-speaker generation.

4.2 Model Architecture

UniSonate is built upon a MM-DiT architecture comprising approximately 1.34 billion parameters. The model utilizes a flow matching feedforward dimension of 1024 and consists of 14 Joint Diffusion Transformer layers followed by 6 Single Diffusion Transformer layers, incorporating RoPE positional encoding (Su et al., 2024) for temporal awareness. For conditioning, we employ Qwen2.5-7B (Yang et al., 2025a) as the frozen instruction encoder to process natural language descriptions. The content encoder utilizes a Zipformer-based (Zhu et al., 2025) network (512 dimension) to encode phoneme sequences for speech and music, while employing

Table 4: Performance comparison of TTM on control accuracy, SongEval, and subjective evaluation. UniSonate achieves state-of-the-art results on all SongEval metrics and Musicality MOS (MMOS), demonstrating that unified training enhances musical structure.

Model	Classification Control Accuracy Rate (%)↑						SongEval↑					MOS↑	
	Genre	Instr	Gend	Age	Rhy	Atmo	Coh	Mus	Mem	Cla	Nat	QMOS	MMOS
Ground Truth	100.0	100.0	100.0	100.0	100.0	100.0	3.60	3.52	3.56	3.43	3.34	–	–
DiffRhythm+(Chen et al., 2025a)	51.33	81.67	22.22	44.44	93.33	87.22	2.68	2.61	2.57	2.48	2.37	3.04 ± 0.46	2.79 ± 0.54
ACE-Step(Gong et al., 2025)	**94.44**	**85.56**	96.11	95.00	89.44	90.56	2.89	2.87	2.83	2.77	2.71	**3.30 ± 0.28**	2.88 ± 0.20
InstructAudio(Qiang et al., 2025)	92.78	83.89	**98.89**	97.22	**94.44**	**95.00**	3.08	2.98	3.00	2.89	2.82	2.82 ± 0.26	2.91 ± 0.35
UniSonate	93.89	85.00	**98.89**	**97.78**	93.33	94.44	**3.18**	**3.07**	**3.10**	**2.99**	**2.90**	2.88 ± 0.21	**3.01 ± 0.29**

Note: Instr=Instrument, Gend=Gender, Rhy=Rhythm, Atmo=Atmosphere; Coh=Coherence, Mus=Musicality, Mem=Memorability, Cla=Clarity, Nat=Naturalness.

Table 5: Performance comparison on Sound Effects (TTA) generation benchmarks. UniSonate leverages a unified dataset of speech, music, and effects to achieve competitive fidelity.

Model	FAD↓	FD↓	KL↓	IS↑	CLAP↑
Ground Truth	0.00	0.00	0.00	–	–
AudioLDM-L (Liu et al., 2023)	4.32	29.50	1.68	8.17	0.208
Tango-FT (Ghosal et al., 2023)	2.68	15.64	**1.24**	8.78	0.291
EzAudio-XL (Hai et al., 2024)	3.64	14.98	1.29	**11.38**	**0.314**
Stable Audio (Evans et al., 2025)	4.19	39.14	2.36	10.07	0.209
GenAU-L (Haji-Ali et al., 2024)	**2.07**	**14.58**	1.36	10.43	0.300
UniSonate (Ours)	4.21	30.21	2.44	3.22	0.156

learnable special tokens for sound effects to model temporal duration. Audio is processed via a pre-trained Mel-VAE encoder that compresses 44.1kHz waveforms into continuous latent embeddings at 43 Hz, achieving a 1024× downsampling rate. Training is conducted on 32 NVIDIA Tesla A800 80GB GPUs with a batch size of 16 per GPU, utilizing the Adam optimizer (Kingma and Ba, 2014) with an initial learning rate of $1e^{-4}$.

4.3 Results and Analysis

4.3.1 Evaluation of TTS

We first evaluate the fundamental speech generation capabilities. Table 3 reports the Word Error Rate (WER) on the Seed-TTS test set. UniSonate achieves the lowest WER (1.47% on English and 1.25% on Chinese), surpassing both the dedicated TTS baselines (e.g., F5-TTS, CosyVoice2) and the previous unified model InstructAudio. This suggests that the inclusion of diverse audio data (music and sound effects) during the curriculum learning phase does not dilute speech intelligibility; rather, it appears to enhance the model's acoustic robustness.

Table 2 presents a detailed comparison of instruction-based control against the SOTA model CosyVoice2 and InstructAudio. UniSonate exhibits superior controllability and signal quality: UniSonate maintains 100% accuracy in Gender and Accent control and achieves 93.33% in Dialogue control—a capability entirely absent in CosyVoice2. Compared to InstructAudio, UniSonate improves dialogue handling. In terms of distortion metrics, UniSonate achieves the best performance with an LSD of 1.79 and MCD of 5.46. It consistently outperforms CosyVoice2, which suffers from emotion leakage due to its reliance on reference audio. While CosyVoice2 achieves a slightly higher QMOS due to reference-based guidance, UniSonate attains a comparable NMOS (3.50) using pure text instructions, significantly reducing the ambiguity inherent in one-to-many mappings.

4.3.2 Evaluation of TTM

Table 4 compares UniSonate against specialized music generation models. While specialized models like ACE-Step excel in genre classification accuracy, UniSonate demonstrates superior performance in structural and detailed attributes. Notably, UniSonate achieves state-of-the-art results on the SongEval benchmark, with the highest scores in Coherence (3.18) and Musicality (3.07). This represents a significant improvement over InstructAudio (Coherence 3.08). We hypothesize that the unified training with large-scale speech data enhances the model's ability to model long-term temporal dependencies, which transfers positively to musical structure. Subjectively, UniSonate achieves the highest Musicality MOS (3.01), validating that our unified architecture captures melodic nuances effectively without specialized music-only architectural designs.

4.3.3 Evaluation of TTA

Table 5 assesses the newly added sound effect generation capability. UniSonate achieves an FAD of 4.21 and CLAP score of 0.156, demonstrating com-

Table 6: Ablation study on Speech Synthesis (TTS). We compare the full UniSonate model against a variant trained exclusively on speech data with identical architecture. The joint training significantly improves intelligibility (WER) and signal fidelity (LSD/MCD), demonstrating that diverse audio modalities enhance speech robustness.

Training Configuration	WER-EN↓	WER-ZH↓	Sim-Spk↑	Sim-Emo↑	LSD↓	MCD↓	MSEP↓	MR↓
UniSonate (TTS-Only Data)	2.24	1.40	0.63	0.51	2.63	8.70	574.67	0.426
UniSonate (Joint Data)	**1.47**	**1.25**	**0.77**	**0.67**	**1.79**	**5.46**	**422.36**	**0.31**

Table 7: Ablation study on Music Generation (TTM). Comparing the full unified model against a music-only variant. Joint training yields improvements across all SongEval metrics, indicating that large-scale structured speech data helps the model learn better musical coherence.

Training Configuration	SongEval↑				
	Coh	Mus	Mem	Cla	Nat
UniSonate (TTM-Only Data)	3.11	3.00	3.04	2.92	2.84
UniSonate (Joint Data)	**3.18**	**3.07**	**3.10**	**2.99**	**2.90**

petitive performance comparable to widely used baselines such as AudioLDM-L (FAD 4.32) and Stable Audio (FAD 4.19). While there is a performance gap compared to the specialized SOTA model GenAU-L, we consider this trade-off acceptable given UniSonate's unique position as a unified multi-task model. Unlike specialized TTA models that focus exclusively on a single modality, UniSonate accommodates speech, music, and sound effects within one framework. Crucially, the results confirm that UniSonate successfully learns to generate non-linguistic acoustic events using our proposed dynamic token injection strategy. This validates that the multi-stage curriculum learning effectively integrates unstructured sound effects into a phoneme-driven architecture without causing catastrophic forgetting of speech or music capabilities.

Across all three domains, UniSonate demonstrates that a single unified model can achieve performance superior to domain-specific specialists in structured tasks (TTS and TTM) while maintaining competitive fidelity in unstructured tasks (TTA). The improvements over InstructAudio in both TTS and TTM metrics indicate that scaling up data diversity through sound effects and employing curriculum learning leads to positive transfer across varying acoustic modalities, proving the viability of a truly unified audio generation model.

4.3.4 Effectiveness of Joint Training (Ablation Study)

To rigorously validate the superiority of unified modeling over single-task approaches, we conducted a controlled ablation study. We retrained the exact same UniSonate architecture under two restricted data configurations: one using exclusively speech data (TTS-Only) and another using exclusively music data (TTM-Only), while keeping all hyperparameters and model size constant. As shown in Table 6, the joint-trained UniSonate (Speech+Music+SFX) significantly outperforms its TTS-only counterpart. The English WER drops from 2.24% to 1.47%, and spectral fidelity metrics (LSD, MCD) show marked improvements. This confirms that exposing the model to the rich acoustic diversity of music and sound effects enhances its generalization capabilities, allowing the shared encoder to learn more robust acoustic features that benefit speech reconstruction. Table 7 reveals a similar trend in music generation. The unified model surpasses the TTM-only variant across all SongEval metrics. We attribute this to the inclusion of 50K hours of highly structured speech data. The strict alignment requirements of speech training likely force the model to learn better temporal attention mechanisms, which *positively transfers* to music generation, resulting in improved structural coherence (Coh) and rhythm stability.

5 Conclusions

We presented Uni-Sonate, a unified flow-matching framework that synthesizes speech, music, and sound effects under a single architecture. By introducing Dynamic Token Injection and a multi-stage curriculum learning strategy, we successfully harmonized structured and unstructured audio modalities. Our results demonstrate not only state-of-the-art performance in instruction-based TTS and TTM but also, crucially, that unified training induces positive transfer, enhancing the generation quality of individual tasks. Uni-Sonate paves the way for general-purpose audio intelligence capable of complex auditory scene synthesis.

6 Limitations

While UniSonate demonstrates the potential of a unified framework for speech, music, and sound effect generation, several limitations remain to be addressed in future work. As shown in Table 5, although UniSonate achieves competitive performance in sound effect generation, there is still a noticeable gap in Fréchet Audio Distance (FAD) compared to specialized SOTA models like GenAU-L (4.21 vs. 2.07). This suggests that while the unified representation is effective, the model may struggle to capture the extreme diversity of unstructured acoustic environments as effectively as models dedicated solely to that modality. Currently, our training and evaluation focus primarily on audio clips ranging from 2 to 20 seconds. While the model excels at short-context coherence, generating consistent long-form content (e.g., full songs exceeding 3 minutes or extended audiobooks) remains challenging. The attention mechanism's memory constraints and the lack of a hierarchical structure for long-term planning limit the model's ability to maintain musical structure or narrative consistency over extended durations. Relying solely on natural language instructions introduces inherent one-to-many mapping ambiguity. Unlike reference-based methods that provide explicit acoustic cues, text descriptions (e.g., "a sad song") can correspond to vastly different acoustic realizations. This sometimes results in generations that, while faithful to the text, may not align with the user's specific unstated preferences, leading to slight variances in perceived naturalness compared to reference-conditioned systems. As a 1.3B parameter diffusion model requiring multiple denoising steps, UniSonate is computationally intensive during inference compared to lightweight, non-autoregressive TTS systems. This currently limits its applicability in real-time scenarios requiring low-latency synthesis.

7 Ethical Considerations

The development of high-fidelity unified audio generation models brings significant capabilities but also necessitates careful consideration of potential risks and ethical implications. UniSonate's ability to generate realistic speech and dialogue via text instructions poses a risk of misuse for creating misleading content, disinformation, or "deepfakes." Although our model relies on descriptive prompts (e.g., "young male") rather than direct voice cloning from reference audio—which theoretically reduces the risk of impersonating specific individuals without their consent—the high quality of the output could still be exploited to deceive listeners. Our model is trained on large-scale datasets collected from the internet. Consequently, it may inherit biases present in the training data, such as gender stereotypes associated with certain professions in speech, or Western-centric biases in musical genres. There is a risk that the model may default to these biases when instructions are underspecified. We are committed to further analyzing these biases and developing methods to ensure more equitable representation. The music generation capability raises concerns regarding copyright and artistic style mimicry. While the model generates original compositions based on text, the training process utilizes existing musical works. We emphasize that this tool is intended to assist creators rather than replace human artists. Future releases will strictly adhere to copyright laws, and we are exploring mechanisms such as dataset filtering and output watermarking to respect intellectual property rights. o mitigate these risks, we plan to release the model weights under a license that prohibits malicious use. Furthermore, we advocate for the development and integration of synthetic audio detection tools (watermarking) to help users distinguish between human-produced and AI-generated audio content.

References

Suno AI. 2024a. Suno: Ai music generation platform. https://suno.com.

Udio AI. 2024b. Udio: Ai music creation platform. https://udio.com.

Philip Anastassiou, Jiawei Chen, Jitong Chen, Yuanzhe Chen, Zhuo Chen, Ziyi Chen, Jian Cong, Lelai Deng, Chuang Ding, Lu Gao, et al. 2024. Seed-tts: A family of high-quality versatile speech generation models. *arXiv preprint arXiv:2406.02430*.

Huakang Chen, Yuepeng Jiang, Guobin Ma, Chunbo Hao, Shuai Wang, Jixun Yao, Ziqian Ning, Meng Meng, Jian Luan, and Lei Xie. 2025a. Diffrhythm+: Controllable and flexible full-length song generation with preference optimization. *arXiv preprint arXiv:2507.12890*.

Sanyuan Chen, Chengyi Wang, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, Lei He, Sheng Zhao, and Furu Wei. 2025b. Neural codec language models are

zero-shot text to speech synthesizers. *IEEE Transactions on Audio, Speech and Language Processing*, 33:705–718.

Yushen Chen, Zhikang Niu, Ziyang Ma, Keqi Deng, Chunhui Wang, Jian Zhao, Kai Yu, and Xie Chen. 2024. F5-tts: A fairytaler that fakes fluent and faithful speech with flow matching. *arXiv preprint arXiv:2410.06885*.

Ho Kei Cheng, Masato Ishii, Akio Hayakawa, Takashi Shibuya, Alexander Schwing, and Yuki Mitsufuji. 2025. Mmaudio: Taming multimodal joint training for high-quality video-to-audio synthesis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 28901–28911.

Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. 2023. Simple and controllable music generation. *Advances in Neural Information Processing Systems*, 36:47704–47720.

Prafulla Dhariwal, Heewoo Jun, Christine Payne, Jong Wook Kim, Alec Radford, and Ilya Sutskever. 2020. Jukebox: A generative model for music. *arXiv preprint arXiv:2005.00341*.

Zhihao Du, Qian Chen, Shiliang Zhang, Kai Hu, Heng Lu, Yexin Yang, Hangrui Hu, Siqi Zheng, Yue Gu, Ziyang Ma, et al. 2024a. Cosyvoice: A scalable multilingual zero-shot text-to-speech synthesizer based on supervised semantic tokens. *arXiv preprint arXiv:2407.05407*.

Zhihao Du, Yuxuan Wang, Qian Chen, Xian Shi, Xiang Lv, Tianyu Zhao, Zhifu Gao, Yexin Yang, Changfeng Gao, Hui Wang, et al. 2024b. Cosyvoice 2: Scalable streaming speech synthesis with large language models. *arXiv preprint arXiv:2412.10117*.

Sefik Emre Eskimez, Xiaofei Wang, Manthan Thakker, Canrun Li, Chung-Hsien Tsai, Zhen Xiao, Hemin Yang, Zirun Zhu, Min Tang, Xu Tan, et al. 2024. E2 tts: Embarrassingly easy fully non-autoregressive zero-shot tts. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pages 682–689. IEEE.

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

Deepanway Ghosal, Navonil Majumder, Ambuj Mehrish, and Soujanya Poria. 2023. Text-to-audio generation using instruction guided latent diffusion model. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 3590–3598.

Junmin Gong, Sean Zhao, Sen Wang, Shengyuan Xu, and Joe Guo. 2025. ACE-Step: A step towards music generation foundation model. *arXiv preprint arXiv:2506.00045*.

Zhifang Guo, Yichong Leng, Yihan Wu, Sheng Zhao, and Xu Tan. 2023. Prompttts: Controllable text-to-speech with text descriptions. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Jiarui Hai, Yong Xu, Hao Zhang, Chenxing Li, Helin Wang, Mounya Elhilali, and Dong Yu. 2024. Ezaudio: Enhancing text-to-audio generation with efficient diffusion transformer. *arXiv preprint arXiv:2409.10819*.

Moayed Haji-Ali, Willi Menapace, Aliaksandr Siarohin, Guha Balakrishnan, and Vicente Ordonez. 2024. Taming data and transformers for audio generation. *arXiv preprint arXiv:2406.19388*.

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

Shengpeng Ji, Jialong Zuo, Wen Wang, Minghui Fang, Siqi Zheng, Qian Chen, Ziyue Jiang, Hai Huang, Zehan Wang, Xize Cheng, et al. 2024. Controlspeech: Towards simultaneous zero-shot speaker cloning and zero-shot language style control with decoupled codec. *arXiv preprint arXiv:2406.01205*.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D Plumbley. 2020. Panns: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2880–2894.

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

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. 2022. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*.

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

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

Zihan Liu, Shuangrui Ding, Zhixiong Zhang, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Dahua Lin, and Jiaqi Wang. 2025. Songgen: A single stage auto-regressive transformer for text-to-song generation. *arXiv preprint arXiv:2502.13128*.

Jan Melechovsky, Zixun Guo, Deepanway Ghosal, Navonil Majumder, Dorien Herremans, and Soujanya Poria. 2024. Mustango: Toward controllable text-to-music generation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8293–8316.

Ziqian Ning, Huakang Chen, Yuepeng Jiang, Chunbo Hao, Guobin Ma, Shuai Wang, Jixun Yao, and Lei Xie. 2025. DiffRhythm: Blazingly fast and embarrassingly simple end-to-end full-length song generation with latent diffusion. *arXiv preprint arXiv:2503.01183*.

Chunyu Qiang, Peng Yang, Hao Che, Jinba Xiao, Xiaorui Wang, and Zhongyuan Wang. 2022. Back-translation-style data augmentation for mandarin chinese polyphone disambiguation. In *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 1915–1919. IEEE.

Chunyu Qiang, Kang Yin, Xiaopeng Wang, Yuzhe Liang, Jiahui Zhao, Ruibo Fu, Tianrui Wang, Cheng Gong, Chen Zhang, Longbiao Wang, et al. 2025. Instructaudio: Unified speech and music generation with natural language instruction. *arXiv preprint arXiv:2511.18487*.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. 2024. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063.

Apoorv Vyas, Bowen Shi, Matthew Le, Andros Tjandra, Yi-Chiao Wu, Baishan Guo, Jiemin Zhang, Xinyue Zhang, Robert Adkins, William Ngan, et al. 2023. Audiobox: Unified audio generation with natural language prompts. *arXiv preprint arXiv:2312.15821*.

Jun Wang, Xijuan Zeng, Chunyu Qiang, Ruilong Chen, Shiyao Wang, Le Wang, Wangjing Zhou, Pengfei Cai, Jiahui Zhao, Nan Li, et al. 2025a. Klingfoley: Multimodal diffusion transformer for high-quality video-to-audio generation. *arXiv preprint arXiv:2506.19774*.

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

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

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

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, and et al. Bo Zheng. 2025a. Qwen2.5:technical report.

Dongchao Yang, Songxiang Liu, Rongjie Huang, Chao Weng, and Helen Meng. 2024. Instructtts: Modelling expressive tts in discrete latent space with natural language style prompt. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:2913–2925.

Dongchao Yang, Jinchuan Tian, Xu Tan, Rongjie Huang, Songxiang Liu, Xuankai Chang, Jiatong Shi, Sheng Zhao, Jiang Bian, Xixin Wu, et al. 2023. Uniaudio: An audio foundation model toward universal audio generation. *arXiv preprint arXiv:2310.00704*.

Guanrou Yang, Chen Yang, Qian Chen, Ziyang Ma, Wenxi Chen, Wen Wang, Tianrui Wang, Yifan Yang, Zhikang Niu, Wenrui Liu, et al. 2025b. Emovoice: Llm-based emotional text-to-speech model with freestyle text prompting. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 10748–10757.

Jixun Yao, Guobin Ma, Huixin Xue, Huakang Chen, Chunbo Hao, Yuepeng Jiang, Haohe Liu, Ruibin Yuan, Jin Xu, Wei Xue, et al. 2025. Songeval: A benchmark dataset for song aesthetics evaluation. *arXiv preprint arXiv:2505.10793*.

Ruibin Yuan, Hanfeng Lin, Shuyue Guo, Ge Zhang, Jiahao Pan, Yongyi Zang, Haohe Liu, Yiming Liang, Wenye Ma, Xingjian Du, et al. 2025. Yue: Scaling open foundation models for long-form music generation. *arXiv preprint arXiv:2503.08638*.

Xueyao Zhang, Junan Zhang, Yuancheng Wang, Chaoren Wang, Yuanzhe Chen, Dongya Jia, Zhuo Chen, and Zhizheng Wu. 2025. Vevo2: Bridging controllable speech and singing voice generation via unified prosody learning. *arXiv preprint arXiv:2508.16332*.

Siyi Zhou, Yiquan Zhou, Yi He, Xun Zhou, Jinchao Wang, Wei Deng, and Jingchen Shu. 2025. Indextts2: A breakthrough in emotionally expressive and duration-controlled auto-regressive zero-shot text-to-speech. *arXiv preprint arXiv:2506.21619*.

Han Zhu, Wei Kang, Zengwei Yao, Liyong Guo, Fangjun Kuang, Zhaoqing Li, Weiji Zhuang, Long Lin, and Daniel Povey. 2025. Zipvoice: Fast and high-quality zero-shot text-to-speech with flow matching. *arXiv preprint arXiv:2506.13053*.

A Compared Methods and Evaluation Metrics

To evaluate UniSonate's unified generation capabilities across speech, music, and sound effects, we compare against state-of-the-art (SOTA) specialized models in each domain as well as the previous unified model, InstructAudio(Qiang et al., 2025).

A.1 Baselines

For TTS, we benchmark fundamental generation quality against MaskGCT(Wang et al., 2024), E2-TTS(Eskimez et al., 2024), F5-TTS(Chen et al., 2024), ZipVoice(Zhu et al., 2025), CosyVoice1(Du et al., 2024a), and CosyVoice2(Du et al., 2024b) (Table 1 & 3). We specifically compare instruction-based control performance against CosyVoice2 and InstructAudio (Table 2). Consistent with previous settings, since UniSonate is purely instruction-controlled, we use neutral text descriptions with randomized speakers for Seed-TTS WER evaluation. For CosyVoice2, which requires reference audio for timbre, we provide matching reference samples and map instructions to its supported control tags. For Music (TTM), we compare with DiffRhythm+(Chen et al., 2025a), ACE-Step(Gong et al., 2025), and InstructAudio (Table 4). As DiffRhythm+ lacks support for short-duration synthesis, we generate longer sequences and truncate them for fair comparison. For Sound Effects (TTA), we benchmark against specialized latent diffusion models including AudioLDM-L(Liu et al., 2023), Tango-FT(Ghosal et al., 2023), EzAudio-XL(Hai et al., 2024), Stable Audio(Evans et al., 2025), and GenAU-L(Haji-Ali et al., 2024) (Table 5).

A.2 Evaluation Metrics

We employ a comprehensive suite of objective and subjective metrics tailored to each modality. Speech Metrics: We evaluate intelligibility using Word Error Rate (WER) on the Seed-TTS(Anastassiou et al., 2024) test set. Acoustic fidelity and similarity are measured via Speaker Similarity*, Emotion Similarity†, Log-Spectral Distance (LSD), Mel-Cepstral Distortion (MCD), Mean Squared Error of Pitch (MSEP), and Voiced/Unvoiced Mismatch Rate (MR). Music Metrics: We utilize the SongEval(Yao et al., 2025) benchmark to assess musical attributes including coherence, musicality, and memorability. Sound Effect Metrics: We adopt standard TTA metrics on the AudioCaps test set. Fréchet Audio Distance (FAD) for audio quality, Fréchet Distance (FD) based on PANNs(Kong et al., 2020), Inception Score (IS) for generation diversity, and CLAP Score(Wu et al., 2023) for text-audio alignment. Control & Subjective Metrics: We assess control capability via Classification Control Accuracy through human listening tests, where annotators verify if generated samples match specific attributes (e.g., Age, Genre, Atmosphere). Subjective quality is evaluated using Quality Mean Opinion Score (QMOS), Naturalness MOS (NMOS), and Musicality MOS (MMOS).

A.3 Subjective Evaluation Details

To rigorously assess the perceptual quality of the synthesized audio, we conducted subjective listening tests following the standard Mean Opinion Score (MOS) protocol.

We recruited 20 volunteer listeners with normal hearing. For speech evaluation, all participants were native speakers of Chinese. To ensure consistent acoustic conditions, participants were provided with high-quality monitoring headphones and instructed to perform the evaluation in a quiet, sound-isolated environment.

Participants rated samples on a 5-point Likert scale (1 = Bad, 5 = Excellent, with 0.5 increments). The evaluation focused on three distinct dimensions corresponding to the unified tasks: Naturalness MOS (NMOS):Evaluated specifically for speech (TTS), focusing on prosody, intonation, and human-like articulation. Musicality MOS (MMOS):Evaluated for music (TTM), focusing on melodic coherence, rhythmic stability, and harmony. Quality MOS (QMOS):Evaluated across all modalities (including SFX), focusing on

A.4 Test Sets

For speech, we use the complete Seed-TTS test set for WER and a manually annotated set of 500 instruction-phoneme pairs for control evaluation. For music, we construct a 500-sample test set with descriptions covering genre, instrument, and atmosphere. For sound effects, evaluations are conducted on the standard AudioCaps test split.

*https://github.com/resemble-ai/Resemblyzer

†https://huggingface.co/emotion2vec