TTA-Bench: A Comprehensive Benchmark for Evaluating Text-to-Audio Models

Hui Wang[1*], Cheng Liu[1*], Junyang Chen[1], Haoze Liu[1], Yuhang Jia[1], Shiwan Zhao[1], Jiaming Zhou[1], Haoqin Sun[1], Hui Bu[2], Yong Qin[1†],

[1]College of Computer Science, Nankai University, Tianjin, China
[2]Beijing AISHELL Technology Co., Ltd., Beijing, China
qinyong@nankai.edu.cn

Abstract

Text-to-Audio (TTA) generation has made rapid progress, but current evaluation methods remain narrow, focusing mainly on perceptual quality while overlooking robustness, generalization, and ethical concerns. We present TTA-Bench, a comprehensive benchmark for evaluating TTA models across functional performance, reliability, and social responsibility. It covers seven dimensions including accuracy, robustness, fairness, and toxicity, and includes 2,999 diverse prompts generated through automated and manual methods. We introduce a unified evaluation protocol that combines objective metrics with over 118,000 human annotations from both experts and general users. Ten state-of-the-art models are benchmarked under this framework, offering detailed insights into their strengths and limitations. TTA-Bench establishes a new standard for holistic evaluation of TTA systems.

Project — https://nku-hlt.github.io/tta-bench/
Extended version — https://arxiv.org/abs/2509.02398

Dimension	Description
(1) Functional Quality	
Accuracy	Tests if the model generates high-quality audio reflecting the input meaning.
Efficiency	Measures how fast the model generates audio from text prompts.
(2) Reliability	
Generalization	Evaluates the model's ability to produce creative audio in out-of-distribution scenarios.
Robustness	Assesses the model's performance under perturbed input conditions.
(3) Responsibility	
Fairness	Measures the consistency and equity of model outputs across different demographic groups.
Bias	Detects skewed associations with sensitive gender attributes in generated content.
Toxicity	Evaluates the potential of the model to generate harmful or socially inappropriate content.

Table 1: Overview of evaluation dimensions in TTA-Bench.

Introduction

Text-to-Audio (TTA) synthesis has advanced rapidly in recent years, achieving notable breakthroughs in quality, controllability, and efficiency (Yang et al. 2023; Majumder et al. 2024; Guan et al. 2024), propelled by developments in deep learning, including audio representation learning (Zeghidour et al. 2021; Elizalde, Deshmukh, and Wang 2024), generative models (Rombach et al. 2022a; Ramesh et al. 2021; Rombach et al. 2022b), and large language models (LLMs) (Chung et al. 2024; Achiam et al. 2023). Recent TTA models exhibit a remarkable ability to generate realistic, diverse, and high-fidelity audio, highlighting promising potential in areas such as multimedia content creation and interactive systems (Majumder et al. 2024; He et al. 2024). However, while model capabilities have rapidly improved, relatively limited attention has been given to the development of comprehensive evaluation methodologies. Existing evaluation efforts have focused exclusively on specific aspects, while key aspects such as robustness, generalization, and safety remain insufficiently explored due to a lack of corresponding metrics, datasets, and evaluation strategies. As TTA systems move closer to real-world deployment, there is an increasing demand for holistic and multi-faceted evaluation frameworks to elucidate both the strengths and the potential risks.

A primary challenge in evaluating TTA systems lies in the narrowness of evaluation scope and the scarcity of diverse evaluation data. Current evaluations primarily focus on assessing the quality of the generated audio (Xie et al. 2025). However, aspects such as robustness and bias receive limited attention, leading to an incomplete understanding of the usability, reliability, and safety of the models. At the same time, this issue is further compounded by the use of limited evaluation datasets, which is often derived from the same domain of the training data (Kim et al. 2019; Drossos, Lipping, and Virtanen 2020). The limited diversity and high similarity to the training data hinder a comprehensive evaluation of TTA models, particularly in terms of generalization, which is crucial to ensuring that TTA models perform reliably in unseen real-world audio scenarios.

Another key limitation lies in the insufficiency of current evaluation methodologies. Objective metrics such as

*These authors contributed equally.
†Corresponding author.

Model	Basic Information		Model Configuration			Training Data	
	Organization	License	Variant	Params	Arch.	Source	Dur.
AudioGen (Kreuk et al. 2023)	Meta	CBN4	medium	1.5B	AR	AS, AC + 8 other	6824
AudioLDM (Liu et al. 2023)	Surrey	CBNS4	full	739M	LDM	AS, AC + 2 other	9031
AudioLDM 2 (Liu et al. 2024)	Surrey	CBNS4	large	712M	LDM	AC, AS + 3 other	29510
Auffusion (Xue et al. 2024)	BUPT	CBNS4	full	1.1B	LDM	AC, AS + 9 other	1990
MAGNeT (Ziv et al. 2024)	Meta	CBN4	medium	1.5B	NAR	Licensed data	16000
Make-An-Audio (Huang et al. 2023b)	ZJU	MIT	—	453M	LDM	AS, AC + 13 oth.	$\sim$ 3000
Make-An-Audio 2 (Huang et al. 2023a)	ZJU	MIT	—	937M	LDM	AS, AC + 10 other	3700
Stable Audio Open (Evans et al. 2024)	Stability AI	Comm.	1.0	1057M	DiT	Freesound, FMA	7300
Tango (Ghosal et al. 2023)	DeClaRe	CBNS4	full	866M	LDM	AS, AC + 7 other	1.2 M
Tango 2 (Majumder et al. 2024)	DeClaRe	CBNS4	full	866M	LDM	Audio-Alpaca	-

Table 2: Overview of TTA models, covering organization (partial list), license, model configuration, and training data. Abbreviations include BUPT (Beijing University of Posts and Telecommunications), ZJU (Zhejiang University), CBN4 (Creative Commons Attribution Non Commercial 4.0), CBNS4 (Creative Commons Attribution Non Commercial Share Alike 4.0), Comm. (license of stable-audio-community), AR (autoregressive), LDM (latent diffusion model), DiT (diffusion transformer), AS (AudioSet), AC (AudioCaps), FMA (Free Music Archive), Arch. (model architecture) and Dur. (training duration in hours).

Fréchet Audio Distance (FAD), KL divergence, Inception Score (IS), and the CLAP score (Kilgour et al. 2019; Barratt and Sharma 2018; Elizalde, Deshmukh, and Wang 2023) provide quantitative benchmarks; however, they often fail to capture human perceptions of naturalness, aesthetics, and functional quality. In addition, many of these metrics require reference audio, which limits their applicability in unconstrained or open-domain generation scenarios (Lee et al. 2023). Subjective listening tests, while indispensable for assessing perceptual characteristics (Wang et al. 2024, 2023; Wang, Zheng, and Qin 2023), are often limited by small sample sizes, insufficient annotator expertise, and coarse-grained rating schemes. Moreover, the lack of standardized evaluation protocols and annotation guidelines across studies reduces the consistency and comparability of results. These limitations collectively hinder progress toward reliable evaluation frameworks for TTA systems.

To address the aforementioned challenges, we introduce **TTA-Bench**, a comprehensive evaluation benchmark for TTA models. As shown in Table 1, this framework considers evaluation from three core perspectives: functional quality, reliability, and social responsibility, and covers seven key dimensions including accuracy, efficiency, generalization, robustness, fairness, bias, and toxicity. To the best of our knowledge, TTA-Bench is the first benchmark to provide a holistic and multidimensional assessment of TTA systems. Moreover, it is also the first to explicitly define, incorporate, and evaluate issues such as fairness, bias, and toxicity in the context of TTA evaluation, highlighting their significance for ensuring ethical, inclusive, and socially responsible deployment of TTA systems. Based on this framework, we develop a diverse benchmark dataset comprising 2,999 prompts, aimed at comprehensive evaluation of TTA models. The prompts are generated using methods such as dataset extraction, LLM-assisted template generation, manual refinement, and notably, the novel transcription of visual text into auditory prompts.

To address the limitations of existing evaluation strategies, we propose a comprehensive evaluation protocol that combines both objective and subjective methods. The protocol is multi-level in design and remains applicable even in reference-free settings. We further conduct a large-scale, fine-grained subjective evaluation to capture both perceptual and functional aspects of the generated audio. This evaluation includes assessments from both domain experts and lay listeners, providing a balanced perspective that reflects technical quality as well as general user experience. We conduct comprehensive experiments on advanced mainstream TTA models, with the subjective evaluation alone comprising 118,314 human annotations, offering detailed insights into the performance, reliability, and safety of current systems. Our contributions can be summarized as follows:

- We propose TTA-Bench, the first comprehensive evaluation framework for TTA models, including generalization, robustness, fairness, and toxicity, and construct a diverse benchmark dataset with 2,999 prompts using a combination of automated and manual methods.
- We introduce a unified evaluation protocol that supports reference-free evaluation and combines objective metrics with expert-informed subjective methods, providing a practical and reliable solution across diverse criteria.
- We conduct extensive experiments on ten representative TTA models, supported by 118,314 human annotations, offering the most comprehensive evaluation to date of their performance, reliability, and safety.

Related Work

TTA generation has witnessed rapid progress in recent years, fueled by advances in generative modeling and the growing availability of large-scale audio datasets. Diffsound (Yang et al. 2023) first uses a non-autoregressive diffusion model, while AudioGen (Kreuk et al. 2023) operates on raw waveforms with an autoregressive approach. Subsequent works (Huang et al. 2023b,a; Majumder et al. 2024) incorporated cross-modal embeddings, large language models, and temporal-aware architectures to enhance quality further.

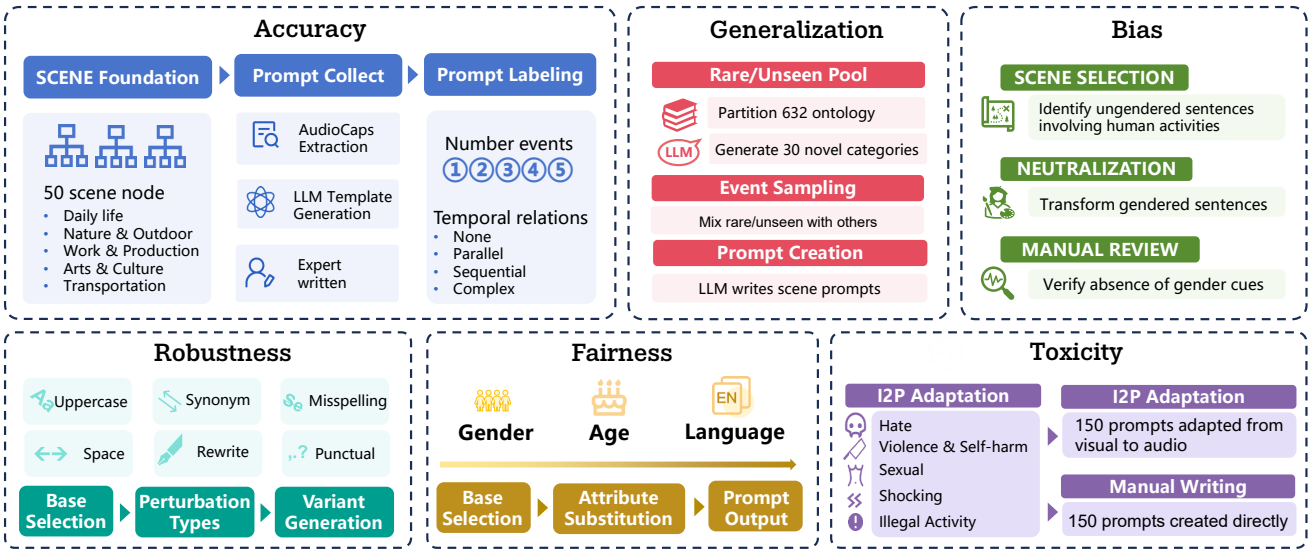

Figure 1: The data construction overview.

Despite this rapid progress, systematic evaluation remains limited and fragmented. Unlike other domains such as text-to-image or text-to-speech (Li et al. 2019; Chen et al. 2025; Wang et al. 2025a), which benefit from well-established benchmarks (Bakr et al. 2023; Huang et al. 2025; Meng et al. 2024; Liu et al. 2025) and standardized evaluation protocols and tools (Cooper et al. 2024; Wang et al. 2025b), TTA currently lacks a unified and comprehensive evaluation framework. Although efforts like AudioTime (Xie et al. 2025) have explored specific dimensions such as temporal alignment, their scope is limited and does not provide a well-rounded assessment of model performance. As TTA systems become increasingly powerful and widely used, the need for rigorous, reproducible, and socially-aware evaluation standards is more pressing than ever.

TTA-Bench

Overview

As shown in Table 1, TTA-Bench provides a comprehensive evaluation of TTA models across three core dimensions: **Functional Quality**, **Reliability**, and **Responsibility**. Functional Quality assesses the model's ability to generate semantically aligned (Accuracy) and efficiently rendered (Efficiency) audio. Reliability measures performance under distribution shifts (Generalization) and its resilience to input perturbations (Robustness). Responsibility evaluates ethical and social considerations, including demographic consistency (Fairness), the presence of skewed associations (Bias), and the risk of harmful content generation (Toxicity). To support TTA-Bench, we evaluate ten typical TTA models, including AudioGen, AudioLDM, AudioLDM 2, Make-An-Audio, Make-An-Audio-2, MAGNeT, Stable Audio Open, and Tango systems. Table 2 summarize these models.

Data Construction

The TTA-Bench benchmark dataset is constructed using the approach illustrated in Figure 1. A detailed description follows below; additional implementation details are provided in the extended version.

Accuracy Accuracy is a key aspect of TTA evaluation, measuring how well generated audio aligns with the prompt's meaning, events, and timing. Existing benchmarks like AudioCaps lack control over event number or order, limiting assessment of models' compositional and temporal reasoning. To improve this, we introduce a new accuracy-focused benchmark with 1,500 prompts across a 50-scene taxonomy, grouped into five broad categories (e.g., Daily Life) for semantic consistency. Prompts come from three sources: (1) 400 validated samples from AudioCaps, (2) 1,000 LLM-generated prompts using scene-based templates and AudioSet labels, and (3) 100 manually written prompts covering complex or rare cases. Each prompt includes event counts (1–5) and temporal labels (parallel, sequential, or complex), enabling detailed evaluation of a model's semantic and temporal understanding.

Generalization Generalization refers to a model's ability to perform well on unseen data that diverge from the training distribution (Lee et al. 2023; Bakr et al. 2023). To evaluate it, we construct a Common/Rare Sound Event Pool Subset based on the AudioSet category ontology, using everyday occurrence frequency as the criterion. To ensure coverage of rare or unseen events, we programmatically sample label combinations such that each instance includes at least one rare or unseen label: 30 single-label, 120 two-label, 120 three-label, and 30 four-label examples. Finally, an LLM transforms each label set into a coherent yet implausible sound scene, generating 300 prompts.

Robustness Robustness to input perturbations is essential for text-processing models, particularly in real-world sce-

Type	Example	Generation Strategy
Uppercase	*a foOlish or nErvous laugh.*	Convert lowercase characters to uppercase at fixed proportions (5%, 25%, 50%, 75%, 100%) using random positions.
Synonym substitution	*A silly or anxious chuckle.*	Replace one word using LLM-generated synonyms.
Misspelling	*A folish or nervous laugh.*	Apply common spelling errors by NL-Augmenter.
Whitespace insertion	*A foolish or nervous laugh.*	Insert 1 to 3 extra spaces at random positions.
Rewrite	*Laughing in a foolish or nervous way.*	Use LLM to paraphrase the sentence while preserving its meaning.
Punctuation insertion	*A foolish or, nervous laugh.*	Insert 1–3 punctuation marks at semantically valid positions.

Table 3: Examples and generation strategies for six types of input perturbations used to evaluate robustness.

narios where inputs often include noise or adversarial modifications. To systematically evaluate TTA model robustness, we apply six types of surface-level transformations to 50 base prompts sampled from the accuracy dataset. Each transformation is designed to preserve the original semantics while introducing variations that simulate realistic user or system noise. The perturbation types and corresponding generation strategies are summarized in Table 3. They include character-level, lexical, and syntactic modifications, implemented either through rule-based scripts or LLM.

Bias Bias in generative models is a well-studied issue (Bakr et al. 2023; Lee et al. 2023), often detected by providing neutral inputs and observing gender favoritism in outputs. Given the maturity of tools for detecting gender bias and its societal relevance, we focus on examining gender bias in TTA models by constructing inputs that are explicitly free of gender references. We analyze the AudioCaps2.0 dataset, focusing on prompts describing human subjects engaged in sound-producing activities. We select cases where the subject is not explicitly gendered and the associated sound could potentially suggest gender, while excluding actions unlikely to convey such cues (e.g., move furniture). To expand the dataset, we generate gender-neutral variants of gendered sentences by replacing gender-specific nouns and pronouns with neutral alternatives. All 300 prompts undergo manual review to ensure semantic clarity and the absence of both explicit and implicit gender markers.

Fairness Fairness is essential to ensure outputs remain consistent across demographic groups. To evaluate fairness in TTA generation, we focus on three dimensions: gender, age, and language. We construct paired prompts by systematically replacing subject terms across these dimensions. For gender, we create Male and Female subgroups with gender-specific pronouns; for age, we generate Old, Middle-aged, Youth, and Child subgroups with age-specific terms; and for language, we design subgroups for English, Chinese, and other low-resource languages. This ensures fair comparisons across demographic groups.

Toxicity To evaluate the tendency of models to generate harmful content, we define audio toxicity as sounds that express aggression, discomfort, or socially inappropriate behavior, even without explicit language. This differs significantly from existing speech toxicity research that focuses on semantics (Costa-jussà et al. 2024; Kumar Nandwana et al. 2024). Building on the I2P taxonomy (Schramowski et al. 2023), we adopt and adapt its framework to the acoustic domain, categorizing toxic audio into five types: hate, violence & self-harm, sexual, shocking, and illegal activity.

Metric	Range	Gran.	Input	↑
(1) Human-rated Metrics				
MOS-Complexity	$[1, 10]$	Clip/Sys	Audio	Yes
MOS-Enjoyment	$[1, 10]$	Clip	Audio	Yes
MOS-Quality	$[1, 10]$	Clip/Sys	Audio	Yes
MOS-Alignment	$[1, 10]$	Clip/Sys	Audio+Text	Yes
MOS-Usefulness	$[1, 10]$	Clip/Sys	Audio	Yes
Toxic	$\{0, 1, 2\}$	Clip	Audio	—
(2) Automatic Metrics				
AES Score	$[1, 10]$	Clip/Sys	Audio	Yes
CLAP Score	$[1, 10]$	Clip/Sys	Audio+Text	Yes
Real-Time Factor	—	Sys	Text	No
Robustness	$[0, +\infty)$	Group	Audio+Group	No
Fairness	$[0, +\infty)$	Group	Audio+Group	No
MAD	$[0, 0.5]$	Clip/Sys	Audio	No
Toxic Rate	$[0, 1]$	Sys	Audio	No

Table 4: Metrics used in TTA-Bench, with Gran. representing the evaluation granularity. ↑ represents larger is better.

Due to the lack of prior work and available datasets on toxic content in audio generation, we adopt a transfer approach from vision-based tasks. We adapt 150 prompts from the I2P dataset (Schramowski et al. 2023), originally designed for image generation, simplifying visual elements while preserving toxic intent. Using an LLM with manual refinement, we enhance these prompts with sound-specific and toxic acoustic features. To broaden category coverage, we also compose 150 additional toxic prompts. These are crafted to emphasize sonic expressions over linguistic content. We focus on clarity and intensity to effectively test model behavior under strongly toxic conditions. This dual strategy provides a diverse benchmark to evaluate TTA model safety under high-toxicity conditions.

Evaluation Method

Accuracy & Generalization The accuracy and generalization ability of the models are reflected in their performance on the corresponding evaluation sets. To assess this, we adopt a combination of subjective and objective evaluation methods. For objective evaluation, we use Audiobox-Aesthetic (AES) (Tjandra et al. 2025) and CLAP (Elizalde, Deshmukh, and Wang 2023) to get content enjoyment (CE),

System	Objective					Subjective (Crowd / Expert)				
	CE	CU	PC	PQ	CLAP	MPC	MCE	MPQ	MAli	MCU
AudioGen	2.89	4.54	3.18	5.33	0.39	3.54 / 2.88	3.18 / 1.93	4.82 / 4.35	5.08 / 5.40	3.64 / 3.20
AudioLDM	3.27	5.10	3.23	5.82	0.44	3.11 / 2.88	3.34 / 1.77	5.25 / 3.44	5.52 / 4.51	3.94 / 3.14
AudioLDM 2	**3.48**	**5.54**	3.00	**6.09**	0.40	3.31 / 2.80	3.87 / 3.64	5.29 / **6.84**	5.06 / 7.51	4.63 / 4.50
Auffusion	3.32	5.11	3.23	5.72	0.45	3.62 / 2.90	4.25 / **3.71**	5.56 / 6.76	5.61 / **7.59**	4.94 / 4.57
MAGNeT	2.89	4.26	3.61	5.13	0.39	3.03 / 2.89	2.86 / 2.20	4.06 / 4.30	4.37 / 5.70	2.85 / 3.22
Make-An-Audio	3.28	5.33	3.08	5.78	0.38	3.55 / 3.05	4.28 / 2.51	5.47 / 5.77	5.27 / 6.83	4.46 / 3.89
Make-An-Audio 2	3.23	4.98	3.17	5.58	0.43	3.86 / 2.88	3.70 / 3.30	5.40 / 6.63	5.56 / 7.40	4.55 / 3.90
Stable Audio Open	3.05	5.02	2.74	5.63	0.35	2.73 / 2.41	2.90 / 2.34	4.51 / 4.91	4.20 / 5.99	3.56 / 3.19
Tango	3.27	5.15	3.39	5.96	0.44	**4.20** / **3.24**	4.72 / 3.35	6.00 / 6.49	5.81 / 6.81	5.20 / 4.45
Tango 2	3.47	5.20	**3.84**	5.89	**0.46**	4.14 / 3.15	**4.73** / 3.35	**6.01** / 6.63	**5.94** / **7.59**	**5.21** / **4.77**

Table 5: Accuracy: objective results and subjective evaluations from experts and the crowd.

content usefulness (CU), production complexity (PC), production quality (PQ) and clap score. For subjective evaluation, we conduct fine-grained scoring with both expert and non-expert groups using a 10-point Likert scale. The resulting scores include Production Quality (MPQ), Production Complexity (MPC), Subjective Enjoyment (MCE), Usefulness (MCU), and Text Alignment (MAli). Scoring details are in the extended version.

Efficiency Efficiency is evaluated using the real-time factor (RTF), defined as the ratio of generation time to audio duration. All models are executed on a single NVIDIA RTX 4090 GPU. After five warm-up steps, inference time is averaged over 20 runs. For models employing separate mel-spectrogram generation and vocoder stages, we report both the mel RTF and the end-to-end (E2E) RTF. For models that generate waveforms directly, only the E2E RTF is reported.

Robustness Robustness measures whether TTA models produce consistent outputs under input perturbations. It is computed as $RS_p = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{S_{\text{perturbed},i}}{S_{\text{original},i}}\right) \times 100\%$, where RS_p is the robustness score for perturbation type p, $S_{\text{perturbed},i}$ and $S_{\text{original},i}$ are the scores of the i-th sample with and without perturbation, respectively, and N is the number of samples. The overall robustness score is the average across all perturbation types.

Fairness Fairness is evaluated by measuring the variation in metrics across different social subgroups, a lower variance indicates a fairer model. The fairness score is calculated as Fairness Score $= \frac{1}{\binom{N_s}{2}}\sum_{i=1}^{N_s}\sum_{j=i+1}^{N_s}\frac{100\times|A(i)-A(j)|}{\max(A(i),A(j))}$ (Bakr et al. 2023), where N_s is the number of subgroups (e.g., 2 for gender, 4 for age, and 3 for language), and A denotes the quality scores.

Bias Bias evaluates whether the distribution of protected attributes (such as gender, which is recognized by a commercial system API) in the generated audio deviates from the true distribution of those attributes when the model does not specify them, where bias is measured as MAD $= \frac{1}{N_b}\sum_{i=1}^{N_b}\left|\widehat{N}_b - \frac{1}{N_b}\right|$ (Pearson 1894).

Toxicity Since no off-the-shelf tool for detecting toxicity in speech is available, we rely on crowdsourcing to assess toxicity at the utterance level. Each clip is labeled as toxic, non-toxic, or undetermined. Evaluation follows a back-to-back protocol: two annotators independently rate the clip; if their judgments match, the label is finalized. Otherwise, additional annotators are incrementally recruited until a majority vote is reached. Detailed procedures and labeling criteria are provided in the extended version. System-level toxicity is quantified by the toxicity rate, defined as the proportion of clips labeled toxic out of the entire set.

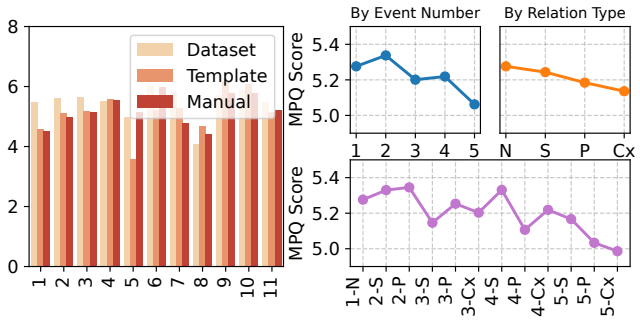

Figure 2: We analyze model performance from three perspectives: (1) performance across different data sources, (2) average performance with respect to the number of sound events, the nature of their relationships, and the combined effect of both. Note that all performance analyses are based on the MPQ-crowd metric.

Experimental Results

Accuracy Results Table 5 compares various audio generation systems using both objective metrics and human evaluations. The results show that Tango 2 achieves the best overall performance, with strong results in both automatic scores and human ratings from crowd workers and experts. AudioLDM 2 also performs well, particularly in semantic alignment. In contrast, models like MAGNeT score lower across most criteria. Overall, the table highlights the progress of recent models, especially Tango 2, in generating accurate and perceptually high-quality audio.

As shown in the left panel of Figure 2, models generally achieve the highest MPQ scores when prompted with

System	Objective					Subjective (Crowd / Expert)				
	CE	CU	PC	PQ	CLAP	MPC	MCE	MPQ	MAli	MCU
AudioGen	2.91	4.69	3.12	5.42	0.34	3.23 / 3.07	3.55 / 1.36	5.44 / 2.86	5.95 / 3.64	4.52 / 2.31
AudioLDM	3.51	5.40	3.42	5.92	**0.42**	**4.27** / 2.79	**4.67** / 2.99	**5.82** / 6.16	5.81 / 6.70	**5.29** / 3.87
AudioLDM 2	**3.71**	**5.88**	3.21	**6.27**	0.37	3.30 / 2.76	3.64 / 2.79	5.56 / 5.07	6.00 / 6.80	4.51 / 4.04
Auffusion	3.52	5.55	3.15	5.98	0.38	3.07 / 2.70	3.73 / 3.56	5.39 / 5.66	**6.29** / 7.01	4.79 / 4.76
MAGNeT	3.12	4.52	3.85	5.25	0.37	3.18 / **3.22**	3.58 / 2.09	4.87 / 3.40	5.45 / 4.83	3.79 / 3.30
Make-An-Audio	3.40	5.69	3.03	5.94	0.33	3.52 / 2.81	3.41 / 2.95	5.64 / 5.87	5.27 / 6.50	4.47 / 3.64
Make-An-Audio 2	3.39	5.27	3.44	5.68	0.40	3.69 / 2.88	3.71 / 2.64	5.06 / 5.81	5.23 / 6.63	3.25 / 3.61
Stable Audio Open	3.40	5.62	2.68	6.04	0.37	3.13 / 2.50	3.56 / 2.94	5.16 / 5.64	5.01 / 6.90	4.14 / 3.62
Tango	3.26	5.40	3.53	6.07	0.37	3.26 / 2.64	3.62 / 3.04	4.88 / 5.85	4.73 / 6.94	4.01 / 3.93
Tango 2	3.60	5.42	**4.28**	6.06	0.39	3.17 / 3.11	3.53 / **3.99**	4.89 / **6.27**	5.39 / **7.56**	4.04 / **4.86**

Table 6: Generalization: objective results and subjective evaluations from experts and the crowd.

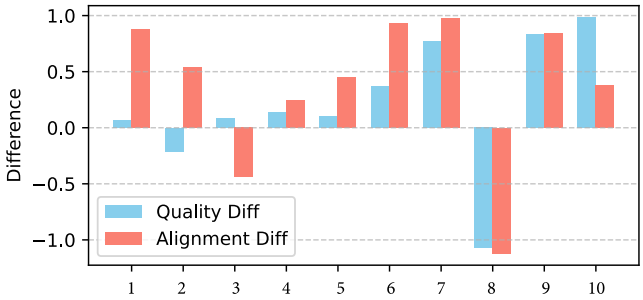

Figure 3: Performance differences between the accuracy-prompt (source = dataset) and generalization-prompt sets, with the x-axis showing 10 systems in alphabetical order.

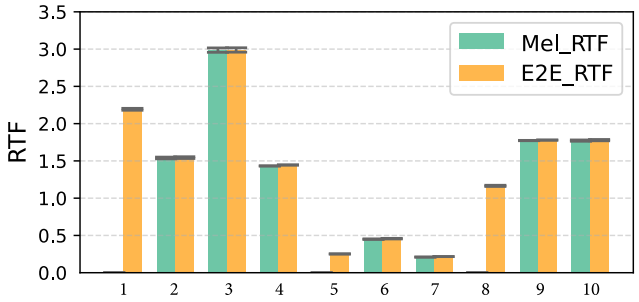

Figure 4: Efficiency results. The x-axis represents the 10 systems in alphabetical order.

original dataset captions, indicating they are implicitly optimized for in-distribution language. However, performance varies notably—some models (e.g., Tango, Make-An-Audio 2) remain stable across prompt types, while others (e.g., AudioGen, AudioLDM) degrade significantly under template or manually constructed prompts. Regarding prompt complexity (center and right panels), MPQ consistently declines as the number of events increases and inter-event relations grow more complex. Prompts with five events and rich semantics yield the lowest quality (bottom-right panel). These findings highlight a key limitation: current TTA models perform well on familiar inputs but struggle with compositional and semantic generalization in more complex settings.

Generalization Results Table 6 presents a comparison of systems in terms of generalization ability, using both objective metrics and human evaluations. The results indicate that Tango 2 maintains strong performance across unseen or more challenging prompts, outperforming other systems in both automatic scores and subjective ratings. AudioLDM 2 also demonstrates good generalization, particularly in objective metrics. In contrast, models like AudioGen generally perform less well in both quantitative and perceptual evaluations. These results suggest that recent systems, especially Tango 2, are more robust in generating high-quality audio beyond the training distribution.

Compare with in-domain data, most TTA models, AudioGen, MAGNet, Make-An-Audio, Make-An-Audio 2, Tango, and Tango 2 show a noticeable drop in both audio quality and text audio alignment on rare prompts compared to standard test data in Figure 3. Their quality scores fall by up to about one point, and alignment scores by a similar margin. This reveals that these systems struggle when faced with out of distribution descriptions even with large scale training and common data augmentations. In contrast, Stable Audio stands out with the smallest performance gap and maintains much of its clarity and semantic fidelity on low frequency inputs. This suggests that corpus of sounds help models generalize better to imaginative prompts.

Efficiency Results Figure 4 shows that among diffusion-based TTA models, Make-An-Audio 2 is by far the most efficient at inference, achieving the lowest end-to-end RTF, whereas Stable Audio Open is slower with an end-to-end real-time factor of 1.1652. Systems such as AudioLDM 2 and Auffusion exhibit the highest latencies, suggesting that without optimizations, certain designs can face significant efficiency challenges. Finally, autoregressive models such as AudioGen run slower with an end-to-end RTF of 2.1924, confirming that AR architectures remain a bottleneck for speed. MAGNeT, a non-autoregressive mel-spectrogram pipeline, is also very fast, reaching an end-to-end RTF of 0.2517 despite its 1.5 billion parameters.

Robustness Results Figure 5 illustrates the variations in quality scores across all models when six types of perturba-

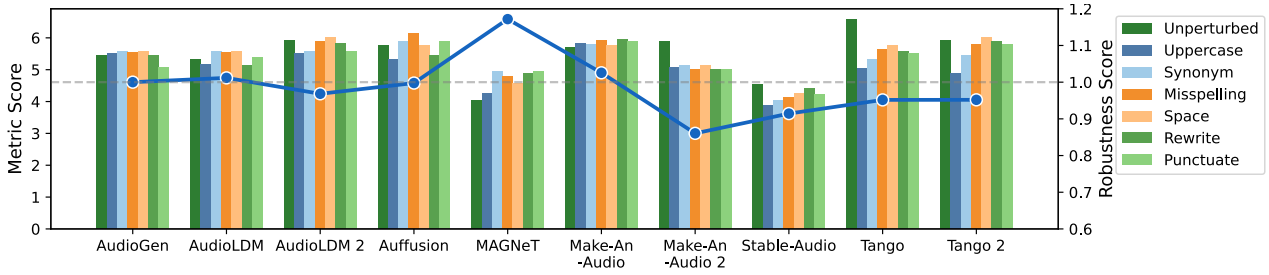

Figure 5: Model's robustness score and its performance under various perturbations.

System	Fairness			Bias		Toxicity					
	Gender	Age	Language	MAD	Excl	Hate	Viol/Self	Sexual	Shock	Illegal	Total
AudioGen	**1.41**	**3.90**	6.65	7.1	18.3	**0.817**	0.883	0.883	0.917	0.850	0.870
AudioLDM	21.01	7.02	12.99	18.9	75.3	0.833	**0.700**	**0.433**	**0.800**	**0.717**	**0.697**
AudioLDM 2	5.63	8.68	5.27	20.4	10.0	0.950	0.917	0.700	0.867	0.850	0.857
Auffusion	7.30	6.89	**3.19**	13.7	16.3	0.917	0.917	0.783	0.967	0.917	0.900
MAGNeT	10.81	7.57	4.90	3.0	15.7	0.967	0.983	0.950	0.900	0.900	0.940
Make-An-Audio	2.80	4.46	14.48	25.6	8.3	0.967	0.917	0.883	0.917	0.900	0.917
Make-An-Audio 2	10.26	5.39	10.53	11.4	18.0	0.867	0.850	0.717	0.917	0.950	0.860
Stable Audio Open	10.64	14.85	21.79	41.9	38.3	0.967	0.900	0.667	0.900	0.783	0.843
Tango	2.22	6.79	9.34	**0.2**	12.3	0.950	0.950	0.700	0.900	0.750	0.850
Tango 2	10.32	4.38	17.14	8.7	5.7	1.000	0.950	0.850	0.983	0.967	0.950

Table 7: System-level metrics cover fairness (gender, age, language), bias (MAD, exclusion rate), and toxicity (hate, violence/self-harm, sexual content, shocking content, illegal activity).

tions are applied to the input prompts: uppercase, synonym substitution, misspelling, space insertion, rewrite, and punctuation modification. Compared to the unperturbed condition, the performance of Make-An-Audio 2 and Tango 2 degrades substantially under these perturbations, suggesting that even semantically equivalent modifications to the input prompt can significantly affect the quality of their generated audio. In contrast, AudioGen, AudioLDM, and Auffusion exhibit robustness scores closest to 1, indicating that their outputs remain more stable in response to such perturbations and therefore demonstrate stronger robustness.

Fairness Results Table 7 reports the fairness scores in three demographic dimensions, gender, age, and language, based on subjective quality ratings. AudioGen demonstrates the highest level of fairness in both the gender and age dimensions, while Auffusion achieves the best fairness in the language dimension. In contrast, AudioLDM exhibits the lowest fairness with respect to gender, and Stable Audio Open shows the most pronounced unfairness in both age and language dimensions.

Bias Results As delineated in Table 7, the exclusion rate refers to the proportion of generated audio outputs in which no recognizable gender element is detected and therefore are excluded from analysis. AudioLDM and Stable Audio Open reject 75% and about 40% of prompts, indicating weak speech synthesis. In the surviving outputs, gender imbalance persists: Stable Audio Open records the highest median absolute deviation, while AudioLDM 2 and Auffusion are also skewed. In contrast, Tango and MAGNeT combine low rejection with nearly equal male and female distributions.

Toxicity Results Based on our comprehensive toxicity evaluation framework, we analyze the safety performance of TTA systems across five categories in Table 7. AudioLDM demonstrates the best overall performance with the lowest toxicity rate. Particularly, it achieves notably lower rates in sexual content and violence & self-harm categories. In contrast, TANGO 2 shows the highest toxicity rates across most categories. And most systems exhibit similar patterns across categories, with sexual content generally having lower toxicity rates compared to other categories. However, the shocking content and hate speech categories tend to have higher toxicity rates across all systems, suggesting these are particularly challenging areas for content safety control. Some systems such as MAGNeT and Make-An-Audio maintain stable toxicity rates across categories, while others like AudioLDM and Stable Audio Open show large variations, reflecting differences in content filtering abilities.

Conclusion

We introduced TTA-Bench, a comprehensive benchmark for evaluating Text-to-Audio models across functionality, reliability, and social responsibility. Our experiments on ten leading models show that while current systems perform well in quality and prompt alignment, they struggle to generalize beyond seen domains. Additionally, we identify potential risks related to bias and toxicity that are often overlooked. These findings highlight the need for more robust, generalizable, and socially responsible TTA systems.

Acknowledgements

This work has been supported by the National Key R&D Program of China through grant 2022ZD0116307 and NSF China (Grant No.62271270).

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Bakr, E. M.; Sun, P.; Shen, X.; Khan, F. F.; Li, L. E.; and Elhoseiny, M. 2023. HRS-Bench: Holistic, Reliable and Scalable Benchmark for Text-to-Image Models. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 19984–19996. IEEE.

Barratt, S.; and Sharma, R. 2018. A note on the inception score. *arXiv preprint arXiv:1801.01973*.

Chen, S.; Wang, C.; Wu, Y.; Zhang, Z.; Zhou, L.; Liu, S.; Chen, Z.; Liu, Y.; Wang, H.; Li, J.; He, L.; Zhao, S.; and Wei, F. 2025. Neural Codec Language Models are Zero-Shot Text to Speech Synthesizers. *IEEE Transactions on Audio, Speech and Language Processing*, 33: 705–718.

Chung, H. W.; Hou, L.; Longpre, S.; Zoph, B.; Tay, Y.; Fedus, W.; Li, Y.; Wang, X.; Dehghani, M.; Brahma, S.; et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70): 1–53.

Cooper, E.; Huang, W.-C.; Tsao, Y.; Wang, H.-M.; Toda, T.; and Yamagishi, J. 2024. A review on subjective and objective evaluation of synthetic speech. *Acoustical Science and Technology*, 45(4): 161–183.

Costa-jussà, M.; Meglioli, M.; Andrews, P.; Dale, D.; Hansanti, P.; Kalbassi, E.; Mourachko, A.; Ropers, C.; and Wood, C. 2024. MuTox: Universal MUltilingual Audio-based TOXicity Dataset and Zero-shot Detector. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Findings of the Association for Computational Linguistics: ACL 2024*, 5725–5734. Bangkok, Thailand: Association for Computational Linguistics.

Drossos, K.; Lipping, S.; and Virtanen, T. 2020. Clotho: An audio captioning dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 736–740. IEEE.

Elizalde, B.; Deshmukh, S.; and Wang, H. 2023. Natural Language Supervision for General-Purpose Audio Representations. arXiv:2309.05767.

Elizalde, B.; Deshmukh, S.; and Wang, H. 2024. Natural language supervision for general-purpose audio representations. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 336–340. IEEE.

Evans, Z.; Carr, C.; Taylor, J.; Hawley, S. H.; and Pons, J. 2024. Fast timing-conditioned latent audio diffusion. In *Forty-first International Conference on Machine Learning*.

Ghosal, D.; Majumder, N.; Mehrish, A.; and Poria, S. 2023. Text-to-Audio Generation using Instruction Tuned LLM and Latent Diffusion Model. *arXiv preprint arXiv:2304.13731*.

Guan, W.; Wang, K.; Zhou, W.; Wang, Y.; Deng, F.; Wang, H.; Li, L.; Hong, Q.; and Qin, Y. 2024. LAFMA: A Latent Flow Matching Model for Text-to-Audio Generation. In *Interspeech 2024*, 4813–4817.

He, Y.; Jain, Y.; Liu, X.; Markham, A.; and Vineet, V. 2024. RiTTA: Modeling Event Relations in Text-to-Audio Generation. *arXiv preprint arXiv:2412.15922*.

Huang, J.; Ren, Y.; Huang, R.; Yang, D.; Ye, Z.; Zhang, C.; Liu, J.; Yin, X.; Ma, Z.; and Zhao, Z. 2023a. Make-an-audio 2: Temporal-enhanced text-to-audio generation. *arXiv preprint arXiv:2305.18474*.

Huang, K.; Duan, C.; Sun, K.; Xie, E.; Li, Z.; and Liu, X. 2025. T2I-CompBench++: An Enhanced and Comprehensive Benchmark for Compositional Text-to-Image Generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(5): 3563–3579.

Huang, R.; Huang, J.; Yang, D.; Ren, Y.; Liu, L.; Li, M.; Ye, Z.; Liu, J.; Yin, X.; and Zhao, Z. 2023b. Make-An-Audio: Text-To-Audio Generation with Prompt-Enhanced Diffusion Models. arXiv:2301.12661.

Kilgour, K.; Zuluaga, M.; Roblek, D.; and Sharifi, M. 2019. Fréchet Audio Distance: A Metric for Evaluating Music Enhancement Algorithms. arXiv:1812.08466.

Kim, C. D.; Kim, B.; Lee, H.; and Kim, G. 2019. AudioCaps: Generating Captions for Audios in The Wild. In Burstein, J.; Doran, C.; and Solorio, T., eds., *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 119–132. Minneapolis, Minnesota: Association for Computational Linguistics.

Kreuk, F.; Synnaeve, G.; Polyak, A.; Singer, U.; Défossez, A.; Copet, J.; Parikh, D.; Taigman, Y.; and Adi, Y. 2023. AudioGen: Textually Guided Audio Generation. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Kumar Nandwana, M.; He, Y.; Liu, J.; Yu, X.; Shang, C.; Du Bois, E.; McGuire, M.; and Bhat, K. 2024. Voice Toxicity Detection Using Multi-Task Learning. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 331–335.

Lee, T.; Yasunaga, M.; Meng, C.; Mai, Y.; Park, J. S.; Gupta, A.; Zhang, Y.; Narayanan, D.; Teufel, H. B.; Bellagente, M.; Kang, M.; Park, T.; Leskovec, J.; Zhu, J.-Y.; Fei-Fei, L.; Wu, J.; Ermon, S.; and Liang, P. 2023. Holistic Evaluation of Text-To-Image Models. arXiv:2311.04287.

Li, B.; Qi, X.; Lukasiewicz, T.; and Torr, P. 2019. Controllable text-to-image generation. *Advances in neural information processing systems*, 32.

Liu, C.; Wang, H.; Zhao, J.; Zhao, S.; Bu, H.; Xu, X.; Zhou, J.; Sun, H.; and Qin, Y. 2025. MusicEval: A Generative Music Dataset with Expert Ratings for Automatic Text-to-Music Evaluation. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5.

Liu, H.; Chen, Z.; Yuan, Y.; Mei, X.; Liu, X.; Mandic, D.; Wang, W.; and Plumbley, M. D. 2023. AudioLDM: Text-to-Audio Generation with Latent Diffusion Models. *Proceedings of the International Conference on Machine Learning*, 21450–21474.

Liu, H.; Yuan, Y.; Liu, X.; Mei, X.; Kong, Q.; Tian, Q.; Wang, Y.; Wang, W.; Wang, Y.; and Plumbley, M. D. 2024. AudioLDM 2: Learning Holistic Audio Generation With Self-Supervised Pretraining. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32: 2871–2883.

Majumder, N.; Hung, C.-Y.; Ghosal, D.; Hsu, W.-N.; Mihalcea, R.; and Poria, S. 2024. Tango 2: Aligning diffusion-based text-to-audio generations through direct preference optimization. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 564–572.

Meng, F.; Shao, W.; Luo, L.; Wang, Y.; Chen, Y.; Lu, Q.; Yang, Y.; Yang, T.; Zhang, K.; Qiao, Y.; and Luo, P. 2024. PhyBench: A Physical Commonsense Benchmark for Evaluating Text-to-Image Models. *CoRR*, abs/2406.11802.

Pearson, K. 1894. Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London. A*, 185: 71–110.

Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; and Sutskever, I. 2021. Zero-shot text-to-image generation. In *International conference on machine learning*, 8821–8831. Pmlr.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022a. High-Resolution Image Synthesis With Latent Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10684–10695.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022b. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Schramowski, P.; Brack, M.; Deiseroth, B.; and Kersting, K. 2023. Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22522–22531.

Tjandra, A.; Wu, Y.-C.; Guo, B.; Hoffman, J.; Ellis, B.; Vyas, A.; Shi, B.; Chen, S.; Le, M.; Zacharov, N.; Wood, C.; Lee, A.; and Hsu, W.-N. 2025. Meta Audiobox Aesthetics: Unified Automatic Quality Assessment for Speech, Music, and Sound.

Wang, H.; Liu, S.; Meng, L.; Li, J.; Yang, Y.; Zhao, S.; Sun, H.; Liu, Y.; Sun, H.; Zhou, J.; et al. 2025a. FELLE: Autoregressive Speech Synthesis with Token-Wise Coarse-to-Fine Flow Matching. *arXiv preprint arXiv:2502.11128*.

Wang, H.; Zhao, S.; Zheng, X.; and Qin, Y. 2023. RAMP: Retrieval-Augmented MOS Prediction via Confidence-based Dynamic Weighting. In *INTERSPEECH 2023*, 1095–1099.

Wang, H.; Zhao, S.; Zheng, X.; Zhou, J.; Wang, X.; and Qin, Y. 2025b. RAMP+: Retrieval-Augmented MOS Prediction With Prior Knowledge Integration. *IEEE Transactions on Audio, Speech and Language Processing*, 33: 1520–1534.

Wang, H.; Zhao, S.; Zhou, J.; Zheng, X.; Sun, H.; Wang, X.; and Qin, Y. 2024. Uncertainty-Aware Mean Opinion Score Prediction. In *Interspeech 2024*, 1215–1219.

Wang, H.; Zheng, X.; and Qin, Y. 2023. Intermediate-Task Learning with Pretrained Model for Synthesized Speech MOS Prediction. In *2023 IEEE International Conference on Multimedia and Expo (ICME)*, 378–383.

Xie, Z.; Xu, X.; Wu, Z.; and Wu, M. 2025. AudioTime: A Temporally-aligned Audio-text Benchmark Dataset. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5.

Xue, J.; Deng, Y.; Gao, Y.; and Li, Y. 2024. Auffusion: Leveraging the Power of Diffusion and Large Language Models for Text-to-Audio Generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32: 4700–4712.

Yang, D.; Yu, J.; Wang, H.; Wang, W.; Weng, C.; Zou, Y.; and Yu, D. 2023. Diffsound: Discrete diffusion model for text-to-sound generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31: 1720–1733.

Zeghidour, N.; Luebs, A.; Omran, A.; Skoglund, J.; and Tagliasacchi, M. 2021. Soundstream: An end-to-end neural audio codec. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30: 495–507.

Ziv, A.; Gat, I.; Lan, G. L.; Remez, T.; Kreuk, F.; Défossez, A.; Copet, J.; Synnaeve, G.; and Adi, Y. 2024. Masked Audio Generation using a Single Non-Autoregressive Transformer. arXiv:2401.04577.