ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-6874-1/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICASSP49660.2025.10890307

# MusicEval: A Generative Music Dataset with Expert Ratings for Automatic Text-to-Music Evaluation

Cheng Liu[1,†], Hui Wang[1,†], Jinghua Zhao[1], Shiwan Zhao[1], Hui Bu[2], Xin Xu[2]
Jiaming Zhou[1], Haoqin Sun[1], and Yong Qin[1,*]
[1] TMCC, College of Computer Science, Nankai University, Tianjin, China, [2] Beijing AISHELL Technology Co., Ltd.
Email: liucheng_hlt@mail.nankai.edu.cn, wanghui_hlt@mail.nankai.edu.cn

***Abstract*—The technology for generating music from textual descriptions has seen rapid advancements. However, evaluating text-to-music (TTM) systems remains a significant challenge, primarily due to the difficulty of balancing performance and cost with existing objective and subjective evaluation methods. In this paper, we propose an automatic assessment task for TTM models to align with human perception. To address the TTM evaluation challenges posed by the professional requirements of music evaluation and the complexity of the relationship between text and music, we collect MusicEval, the first generative music assessment dataset. This dataset contains 2,748 music clips generated by 31 advanced and widely used models in response to 384 text prompts, along with 13,740 ratings from 14 music experts. Furthermore, we design a CLAP-based assessment model built on this dataset, and our experimental results validate the feasibility of the proposed task, providing a valuable reference for future development in TTM evaluation. The dataset is available at https://www.aishelltech.com/AISHELL_7A.**

***Index Terms*—mean opinion score, text-to-music generation, automatic quality assessment**

## I. Introduction

In recent years, music-generative models have achieved significant advances and shown considerable potential for applications in areas such as gaming and education. Among these, text-to-music (TTM) systems [1]–[3], which generate music from natural language prompts, offer superior expression, personalization, diversity, and user-friendliness, making them more advantageous than traditional music generation systems. Many studies have utilized techniques such as diffusion models [4] and language models [5] to address TTM task, achieving notable results [2], [3]. However, despite the extensive focus on generative techniques, research on evaluation methods for TTM systems remains relatively underexplored. Assessing the quality of generated music is both essential and challenging as it shapes the future development of these models. Such evaluations not only play a key role in refining current-generation techniques but also stimulate innovation in model architectures and training strategies, ensuring that advancements align with the intended musical outcomes.

A universally accepted evaluation paradigm for assessing the quality of AI-generated music has not yet been established. Traditionally, methods for evaluating the quality of generative music involve a combination of objective metrics and subjective assessments. Objective metrics, such as Fréchet Audio Distance (FAD) [6] and Inception Score (IS) [7], offer quick and convenient evaluations, yet they often show a weak correlation with human-perceived quality. In contrast, subjective evaluations directly reflect human judgment, but are time-consuming, labour-intensive, and lack reproducibility, making it difficult to compare results across different works [8], [9]. Therefore, there is an urgent need for an efficient and reliable method to accurately assess TTM systems.

In light of this background, we propose an automatic music quality assessment task for TTM models that aligns with human perception. Although several studies have explored automatic prediction of Mean Opinion Scores (MOS) for enhanced speech [10]–[12], synthesized speech [13]–[15], and voice singing [16], which yield promising results, the automatic evaluation of generative music aligned with human perception remains largely unexplored. Compared to the evaluation of speech quality and singing quality, the assessment of TTM systems presents unique challenges. On the one hand, assessing the quality of music involves various complex factors, such as melody, harmony, and rhythm, making it a demanding task and often requiring the opinions of professionals with specialized knowledge in music. On the other hand, compared to tasks like TTS, the relationship between the generated music and the input text descriptions in TTM systems is inherently more intricate and less direct, necessitating a deeper level of consideration and analysis. This dual challenge of evaluating both artistic quality and semantic alignment adds considerable complexity to the overall assessment process.

To address these challenges, we develop an evaluation mechanism for the TTM system based on professional music knowledge. This mechanism includes two dimensions: overall musical impression and alignment with the text prompt, which respectively emphasizes the importance of both the quality of the generated music and its consistency with the given text prompt. Based on this mechanism, we compile a generative music evaluation dataset with contributions from music experts, referred to as **MusicEval**. The MusicEval dataset contains music generated by 31 prevalent and advanced TTM models in response to 384 text prompts, with each piece of generated music scored by five experts in a back-to-back assessment. To the best of our knowledge, MusicEval is the first dataset specifically designed for the automatic evaluation

† These authors contributed equally to this work.
* Corresponding author. This work has been supported by the National Key R&D Program of China (Grant No.2022ZD0116307) and NSF China (Grant No.62271270).

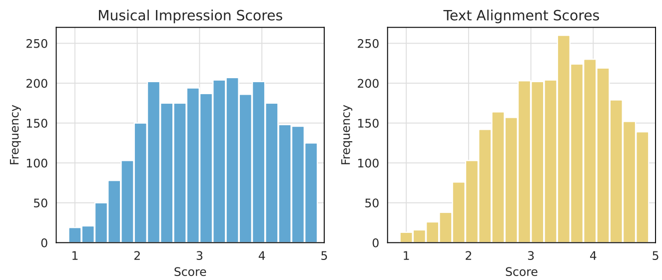



Fig. 1: The distribution of musical impression scores and textual alignment scores in the MusicEval dataset.

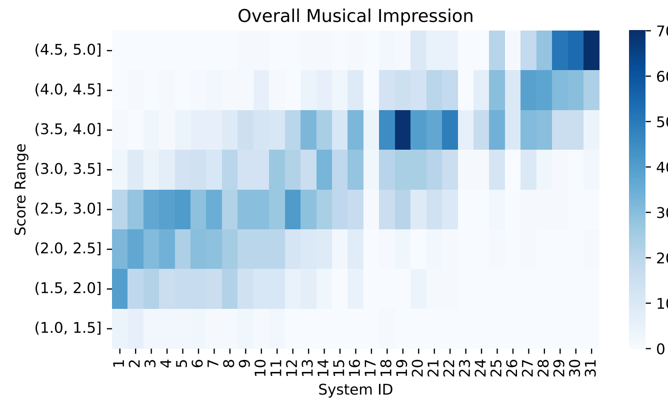



Fig. 2: The distribution chart of the musical impression scores for each system in the MusicEval dataset.

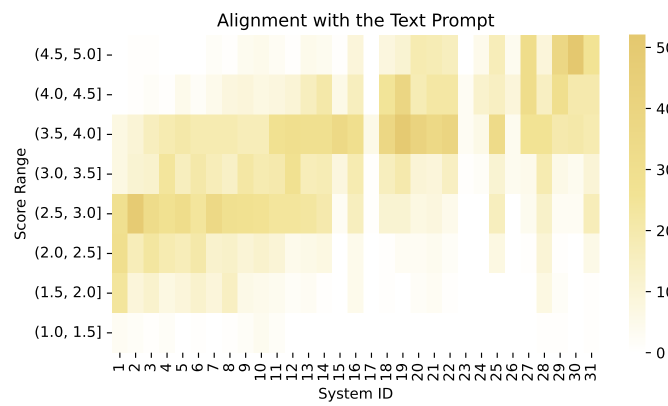



Fig. 3: The distribution of the textual alignment scores for each system in the MusicEval dataset.

of TTM systems, which offers a diverse, professional, and comprehensive data foundation for this task. Furthermore, we develop an automatic scoring system for the generated music using CLAP [17], which assesses both the overall quality of the music and its alignment with the corresponding text prompts.

Our contributions are as follows:

1) We introduce the task of automatic evaluation for TTM systems and propose a comprehensive evaluation framework specifically tailored for TTM tasks.
2) With the participation of music experts, we collect the first generative music assessment dataset, named **MusicEval**, which contains music from diverse and advanced systems in response to a combination of dataset-extracted and manually crafted prompts.
3) We develop a CLAP-based model to evaluate both the quality of generated music and its alignment with textual prompts. The experimental results demonstrate the feasibility of this task and establish a valuable baseline for future research.

## II. The Dataset Overview

### A. Basic Information

The MusicEval dataset is a generative music evaluation dataset with a total duration of 16.62 hours, comprising 2,748 music clips in mono audio and 13,740 ratings from 14 music experts. These clips are generated by 21 different systems (spanning 31 models) in response to 384 prompts. To ensure consistency, we use ffmpeg to resample all generated music to a 16 kHz mono format. Each music clip is evaluated by five musical experts from conservatories, who score the clips based on two criteria: overall musical impression and alignment with the text prompts. The selection of the system and the method for designing text prompts are discussed in detail in Section III, while the specific details of the collection and processing of scoring data are introduced in Section IV.

### B. Data Distribution

Figure 1 presents the distribution of overall musical impression scores and textual alignment scores, with the x-axis representing score ranges from 1 to 5 and the y-axis indicating frequency. Both distributions exhibit similar characteristics, following a roughly normal distribution pattern, with the highest frequency occurring in the mid-range scores and relatively fewer instances of extremely high or low scores.

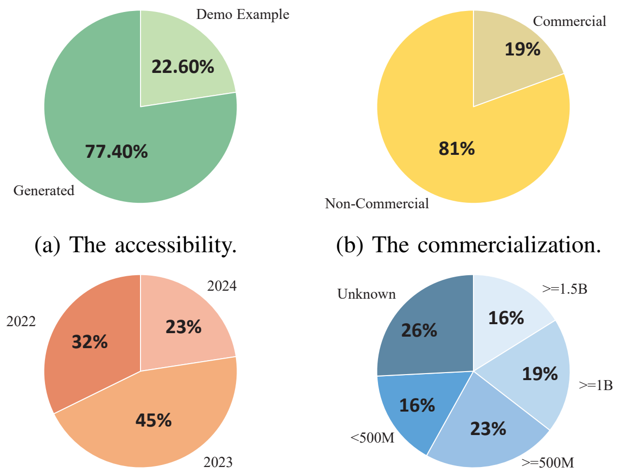
Figures 2 and 3 further illustrate the distribution of scores in the two dimensions across different systems. Within each system, the generated music demonstrates relatively consistent quality; however, notable variations in quality are observed between different systems.

## III. Generated Music Collection

### A. Systems

To ensure comprehensive coverage of various generative systems, we select 14 TTM systems [1]–[3], [18]–[22] and 7 text-to-audio (TTA) systems [23]–[29], some of them have multiple models that differ in size or training data, resulting in a total of 31 different models.

These systems exhibit significant variation across multiple dimensions. Figure 4 presents an analysis of the 31 models across four key attributes: accessibility, commercialization, development year, and model size. In terms of accessibility,




(a) The accessibility. (b) The commercialization.

(c) The year. (d) The model size.

Fig. 4: The pie chart of system information across four dimensions: accessibility, commercialization, year, and model size.

25 systems are publicly accessible, and we used them to generate music samples based on the designed prompts. For the remaining 6 systems, which are not publicly available, we rely on their demo audio as valuable supplement data. Regarding commercialization, 6 systems are commercial [30]–[35], while 25 are non-commercial, ensuring a balanced representation of advancements from both industry and academia. The temporal distribution shows that 14 systems were developed in 2023, 7 systems were developed in 2024, and 10 in 2022, contributing to a broader temporal scope within the dataset. Additionally, the dataset includes models of varying sizes, ensuring a balanced mix of large-scale and smaller models. This diversity in multiple dimensions ensures that the dataset captures a wide range of characteristics and distributions of the TTM systems.

Most TTM systems generate music by producing latent representations or discrete tokens and decoding them into a waveform in autoregressive or non-autoregressive approaches. Besides, symbolic music generation systems have also made notable progress, although their application in TTM remains limited [36] because of their inability to effectively model timbre information, as well as their heavy reliance on external components such as sound sources and rendering tools. Nevertheless, to enhance the diversity of our dataset, we include a few music samples from symbolic music generation systems [20] in our evaluation, including those based on ABC notation and MIDI formats.

### B. Text Prompts

The TTM systems use text descriptions as input, referred to as *prompts*. The MusicEval dataset includes a total of 384 prompts, comprising 80 manually crafted prompts, 20 prompts selected from the MusicCaps dataset [1], and 284 prompts extracted from system demo pages. Among these, 100 text prompts are used to generate music with open-access models, while the remaining 284 prompts correspond exclusively to music clips from demo-only systems.

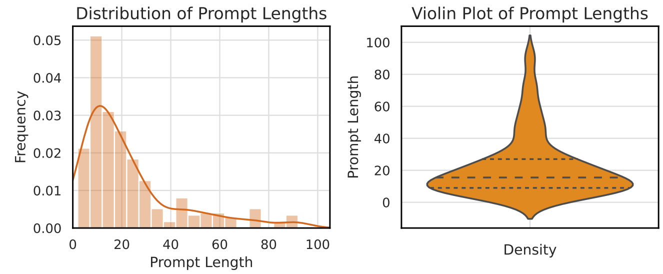



Fig. 5: The distributional analysis of prompt length.

The manually crafted prompts are carefully designed to take multiple musical aspects into account such as emotion, structure, rhythm, theme, and instrumentation. The prompts extracted from existing datasets are included because many TTM models have encountered them during training, facilitating more likely to achieve better performance and ensuring comprehensive coverage of the MusicEval's distribution.

We analyze the distribution of prompt lengths in Figure 5, which reveals a broad range, primarily concentrated on short to mid-length prompts. The violin plot indicates high variance in prompt lengths, ensuring diversity in the generated music by accommodating both typical and atypical prompts. Moreover, the prompts in MusicEval focus on two specific genres: pop and classical music, which represent two distinct and widely recognized styles. Compared to other genres such as hip-hop or metal, most professional music evaluators have greater expertise in pop and classical music, enhancing the reliability of the evaluation results.

## IV. Mean Opinion Scores Collection

We recruit 2 professional teachers and 12 experienced students from conservatories as raters and conduct a MOS test across two dimensions: **overall musical impression (musical impression)** and **alignment with the text prompt (textual alignment)**. In total, we collect 13,740 high-quality ratings from 14 raters.

### A. Evaluation Dimension

The overall musical impression score considers factors such as the authenticity of the music, and the quality of the melody, rhythm and chords. A low score indicates that the sample lacks of musicality and is of very poor quality, or exhibits noticeable machine-generated artifacts. In contrast, a high score suggests that the sample is of excellent overall quality, characterized by a clear rhythm, a pleasant melody, and a coherent chord progression, making it difficult to distinguish whether it was composed by a human or generated by a system.

The textual alignment score assesses how well the audio sample corresponds to the given text description, reflecting the system's ability to adhere to the given prompt. A low score indicates little or no relevance between the generated music sample and the text description, while a high score suggests a strong alignment between the two.

### B. Listening Test Design

All MOS tests are conducted online by specially recruited music professionals, with data packages are distributed in multiple batches. After completing a training session, the 14 rators are instructed to perform the assessments in a quiet environment. The evaluation process requires participants to read a text description and then listen to the music generated by different systems based on that description. All the music samples can be replayed repeatedly as needed. The raters are asked to score each sample on two dimensions: overall musical impression and alignment with the text prompt, and all ratings are collected using a 5-point Likert scale. Each audio sample is scored by five different evaluators, and the average score is calculated.

### C. Quality of Ratings

Although subjective evaluation from human listeners is considered the gold standard, the ratings provided by music experts are not always reliable due to factors such as fatigue and distraction. To enhance the reliability of the scoring results, we adopt two probing methods in the evaluation. Firstly, several carefully selected human-created real music clips from the AudioSet[1] are inserted into each batch of audio data for evaluation. If a rater assigns a score lower than 3 to any of these real music samples, their scores are deemed invalid. Second, within the same batch of data, a pair of two identical audio samples are placed at different positions. If the difference in scoring results between the two samples is too large, the scores from the corresponding rater are considered invalid.

## V. Baseline

### A. Model Architecture

Inspired by the extensive success of fine-tuning pre-trained models on downstream tasks [13], [37]–[39], we adopt a similar approach. Specifically, we select the pre-trained CLAP model as the upstream audio feature extractor. The CLAP model is a pre-trained framework designed to learn joint embeddings of audio and text, enabling effective cross-modal understanding and retrieval between audio signals and textual descriptions [40]. We leverage 3-layer MLP as the downstream prediction head to predict musical impression scores and textual alignment scores, respectively. We guide the training process by the L1 loss between the predicted and true scores across two dimensions.

### B. Implementation Details

The MusicEval data set is randomly divided into a train set and a test set in an 85% and 15% ratio as a preliminary split setting. For the CLAP model, we utilize the official pre-trained checkpoint[2] which is trained on music, AudioSet, and LAION-Audio-630k. During the fine-tuning stage, we conduct the experiments using a batch size of 64 on a single NVIDIA 4090 GPU. The SGD optimizer is employed with a learning rate of 0.0005.

[1]https://research.google.com/audioset/index.html

TABLE I: Performance of our baseline model on the MusicEval test set. The prefix **U_** represents utterance-level metrics, while the prefix **S_** denotes system-level metrics.

| | U_MSE↓ | U_LCC↑ | U_SRCC↑ | U_KATU↑ |
|---|---|---|---|---|
| Musical impression | 0.647 | 0.606 | 0.633 | 0.461 |
| Textual alignment | 0.616 | 0.438 | 0.443 | 0.314 |
| | S_MSE↓ | S_LCC↑ | S_SRCC↑ | S_KATU↑ |
| Musical impression | 0.446 | 0.839 | 0.926 | 0.767 |
| Textual alignment | 0.354 | 0.757 | 0.784 | 0.617 |

### C. Test Metrics

We evaluate the performance by employing several objective indicators. We evaluate the Mean square error (MSE), the Linear Correlation Coefficient (LCC), the Spearman Rank Correlation Coefficient (SRCC) and the Kendall Tau Rank Correlation (KTAU) [13] between predictions and ground-truths.

### D. Experimental Results

As shown in Table I, we evaluate our model on the MusicEval test set across two dimensions: Musical impression and Textual alignment. For the first evaluation dimension at utterance-level and system-level, the MSE is relatively low, while the LCC, SRCC, and KTAU are all high, indicating a strong correlation with human evaluations. In contrast, the second result shows slightly higher MSE and lower correlation metrics, suggesting that there is more difficulty in predicting textual alignment compared to the musical impression. In general, the results demonstrate that the scores predicted by the fine-tuned prediction model are strongly correlated with those of human experts, confirming the feasibility of using deep neural networks to automatically assess the quality of the generated music.

## VI. Conclusion

In this paper, we present MusicEval, the first expert-scored dataset for generative music evaluation, which includes a wide range of music generated by diverse TTM systems and MOS ratings on a Likert scale from music experts. Furthermore, we propose a baseline model utilizing CLAP to predict both overall musical impressions and alignment with textual descriptions of generated music. Our results demonstrate the effectiveness of this automatic evaluation method. In future work, we aim to enrich the dataset to encompass more additional genres and explore more advanced architectures for the automatic evaluation of TTM.

[2]https://huggingface.co/lukewys/laion_clap/blob/main/music_audioset_epoch_15_esc_90.14.pt

## REFERENCES


[1] Andrea Agostinelli, Timo I Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, et al., "Musiclm: Generating music from text," *arXiv preprint arXiv:2301.11325*, 2023.

[2] Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Defossez, "Simple and controllable music generation," in *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds. 2023, vol. 36, pp. 47704–47720, Curran Associates, Inc.

[3] Ke Chen, Yusong Wu, Haohe Liu, Marianna Nezhurina, Taylor Berg-Kirkpatrick, and Shlomo Dubnov, "Musicldm: Enhancing novelty in text-to-music generation using beat-synchronous mixup strategies," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 1206–1210.

[4] Jonathan Ho, Ajay Jain, and Pieter Abbeel, "Denoising diffusion probabilistic models," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, Eds. 2020, vol. 33, pp. 6840–6851, Curran Associates, Inc.

[5] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. 2017, vol. 30, Curran Associates, Inc.

[6] Kevin Kilgour, Mauricio Zuluaga, Dominik Roblek, and Matthew Sharifi, "Fréchet audio distance: A metric for evaluating music enhancement algorithms," 2019.

[7] Shane Barratt and Rishi Sharma, "A note on the inception score," *arXiv preprint arXiv:1801.01973*, 2018.

[8] Zeyu Xiong, Weitao Wang, Jing Yu, Yue Lin, and Ziyan Wang, "A comprehensive survey for evaluation methodologies of ai-generated music," 2023.

[9] Ashvala Vinay and Alexander Lerch, "Evaluating generative audio systems and their metrics," *arXiv preprint arXiv:2209.00130*, 2022.

[10] Chandan KA Reddy, Vishak Gopal, and Ross Cutler, "Dnsmos: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6493–6497.

[11] Gabriel Mittag, Babak Naderi, Assmaa Chehadi, and Sebastian Möller, "Nisqa: A deep cnn-self-attention model for multidimensional speech quality prediction with crowdsourced datasets," in *Interspeech 2021*, 2021, pp. 2127–2131.

[12] Lianwu Chen, Xinlei Ren, Xu Zhang, Xiguang Zheng, Chen Zhang, Liang Guo, and Bing Yu, "Impairment representation learning for speech quality assessment.," in *INTERSPEECH*, 2022, pp. 3323–3327.

[13] Erica Cooper, Wen-Chin Huang, Tomoki Toda, and Junichi Yamagishi, "Generalization ability of mos prediction networks," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 8442–8446.

[14] Hui Wang, Shiwan Zhao, Xiguang Zheng, and Yong Qin, "Ramp: Retrieval-augmented mos prediction via confidence-based dynamic weighting," in *INTERSPEECH 2023*, 2023, pp. 1095–1099.

[15] Hui Wang, Shiwan Zhao, Jiaming Zhou, Xiguang Zheng, Haoqin Sun, Xuechen Wang, and Yong Qin, "Uncertainty-aware mean opinion score prediction," in *Interspeech 2024*, 2024, pp. 1215–1219.

[16] Yuxun Tang, Jiatong Shi, Yuning Wu, and Qin Jin, "Singmos: An extensive open-source singing voice dataset for mos prediction," 2024.

[17] Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov, "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)*, 2023, pp. 1–5.

[18] Qingqing Huang, Daniel S Park, Tao Wang, Timo I Denk, Andy Ly, Nanxin Chen, Zhengdong Zhang, Zhishuai Zhang, Jiahui Yu, Christian Frank, et al., "Noise2music: Text-conditioned music generation with diffusion models," *arXiv preprint arXiv:2302.03917*, 2023.

[19] Pengfei Zhu, Chao Pang, Yekun Chai, Lei Li, Shuohuan Wang, Yu Sun, Hao Tian, and Hua Wu, "Ernie-music: Text-to-waveform music generation with diffusion models," *arXiv preprint arXiv:2302.04456*, 2023.

[20] Peiling Lu, Xin Xu, Chenfei Kang, Botao Yu, Chengyi Xing, Xu Tan, and Jiang Bian, "Musecoco: Generating symbolic music from text," *arXiv preprint arXiv:2306.00110*, 2023.

[21] Alon Ziv, Itai Gat, Gael Le Lan, Tal Remez, Felix Kreuk, Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi, "Masked audio generation using a single non-autoregressive transformer," *arXiv preprint arXiv:2401.04577*, 2024.

[22] Flavio Schneider, Ojasv Kamal, Zhijing Jin, and Bernhard Schölkopf, "Moûsai: Text-to-music generation with long-context latent diffusion," 2023.

[23] Haohe Liu, Zehua Chen, Yiitan Yuan, Xinhao Mei, Xubo Liu, Danilo P. Mandic, Wenwu Wang, and MarkD . Plumbley, "Audioldm: Text-to-audio generation with latent diffusion models," in *International Conference on Machine Learning*, 2023.

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

[25] Deepanway Ghosal, Navonil Majumder, Ambuj Mehrish, and Soujanya Poria, "Text-to-audio generation using instruction guided latent diffusion model," in *Proceedings of the 31st ACM International Conference on Multimedia*, New York, NY, USA, 2023, MM '23, p. 3590–3598, Association for Computing Machinery.

[26] Navonil Majumder, Chia-Yu Hung, Deepanway Ghosal, Wei-Ning Hsu, Rada Mihalcea, and Soujanya Poria, "Tango 2: Aligning diffusion-based text-to-audio generations through direct preference optimization," *arXiv preprint arXiv:2404.09956*, 2024.

[27] Rongjie Huang, Jiawei Huang, Dongchao Yang, Yi Ren, Luping Liu, Mingze Li, Zhenhui Ye, Jinglin Liu, Xiang Yin, and Zhou Zhao, "Make-an-audio: Text-to-audio generation with prompt-enhanced diffusion models," in *International Conference on Machine Learning*. PMLR, 2023, pp. 13916–13932.

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

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

[30] Suno, "Suno official website," 2024.

[31] Udio, "Udio official website," 2024.

[32] SkyMusic, "Skymusic official website," 2024.

[33] qqmusic, "qqmusic official website," 2024.

[34] mubert, "mubert official website," 2024.

[35] riffusion, "riffusion official website," 2024.

[36] Shulei Ji, Xinyu Yang, and Jing Luo, "A survey on deep learning for symbolic music generation: Representations, algorithms, evaluations, and challenges," *ACM Computing Surveys*, vol. 56, no. 1, pp. 1–39, 2023.

[37] Wei-Cheng Tseng, Chien yu Huang, Wei-Tsung Kao, Yist Y. Lin, and Hung yi Lee, "Utilizing self-supervised representations for mos prediction," 2021.

[38] Haoqin Sun, Shiwan Zhao, Xuechen Wang, Wenjia Zeng, Yong Chen, and Yong Qin, "Fine-grained disentangled representation learning for multimodal emotion recognition," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 11051–11055.

[39] Haoqin Sun, Shiwan Zhao, Xiangyu Kong, Xuechen Wang, Hui Wang, Jiaming Zhou, and Yong Qin, "Iterative prototype refinement for ambiguous speech emotion recognition," in *Interspeech 2024*, 2024, pp. 3200–3204.

[40] Yusong Wu*, Ke Chen*, Tianyu Zhang*, Yuchen Hui*, Taylor Berg-Kirkpatrick, and Shlomo Dubnov, "Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, 2023.