2025 IEEE INTERNATIONAL WORKSHOP ON MACHINE LEARNING FOR SIGNAL PROCESSING, AUG. 31– SEP. 3, 2025, ISTANBUL, TURKEY

FROM AESTHETICS TO HUMAN PREFERENCES: COMPARATIVE PERSPECTIVES OF EVALUATING TEXT-TO-MUSIC SYSTEMS

Huan Zhang[⋆] *Jinhua Liang*[⋆] *Huy Phan*[†] *Wenwu Wang*[‡] *Emmanouil Benetos*[⋆]

[⋆] Centre for Digial Music, Queen Mary University of London , United Kingdom
[†] Reality Labs, Meta, France
[‡] Centre for Vision, Speech and Signal Processing (CVSSP), University of Surrey, United Kingdom

ABSTRACT

Evaluating generative models remains a fundamental challenge, particularly when the goal is to reflect human preferences. In this paper, we use music generation as a case study to investigate the gap between automatic evaluation metrics and human preferences. We conduct comparative experiments across five state-of-the-art music generation approaches, assessing both perceptual quality and distributional similarity to human-composed music. Specifically, we evaluate synthesis music from various perceptual dimensions and examine reference-based metrics such as Mauve Audio Divergence (MAD) and Kernel Audio Distance (KAD). Our findings reveal significant inconsistencies across the different metrics, highlighting the limitation of the current evaluation practice. To support further research, we release a benchmark dataset comprising samples from multiple models. This study provides a broader perspective on the alignment of human preference in generative modeling, advocating for more human-centered evaluation strategies across domains.

***Index Terms*—** Text-to-Music Generation, Human Preference Alignment, Evaluation Metrics, Music Quality Assessment, Generative Audio Models

1. INTRODUCTION

With the rapid advances in generative models, a fundamental question arises: *How well do these models truly perform, and how can we evaluate them more thoroughly and systematically?* While early efforts often relied on proxy objectives or handcrafted metrics [1, 2, 3], the growing complexity of generative systems demands more human-centered and rigorous evaluation protocols.

Across fields such as language [4, 5], vision [6], and speech [7], aligning model outputs with human preference has emerged as a promising direction for both training and evaluation. In the domain of audio and music, a series of models have been proposed to enhance the generation quality by incorporating human feedback, such as MusicRL [8], Tango2 [9], BATON [10], and DRAGON [11]. These systems typically introduce a reward model trained on human-annotated preference datasets, guiding the generator toward outputs perceived as more desirable. Preference data can be collected in various formats, including individual scores or pairwise rankings [11, 12], often reflecting subjective criteria such as music fidelity and content enjoyment. Despite these advances, understanding how well these human-centered methods generalize — especially in perceptually complex domains like music — remains an open question [13].

The work does not relate to Huy Phan's work at Meta.

In this paper, we use music generation as a case study to provide a broader view of human-centric evaluation challenges. Specifically, we focus on the consistency and bias of different evaluation methods, benchmarking five recent text-to-music models using both human ratings and automatic reference-based scores.

This paper is structured around **three main hypotheses**, each investigated through dedicated experiments:

- **First**, we ask whether *existing human-annotated preference datasets* align with aesthetic predictors trained independently on the human aesthetic annotations.
- **Second**, we benchmark *state-of-the-art text-to-music models* in different perceptual dimensions to understand how performance varies under different evaluation perspectives.
- **Third**, we examine *distribution alignment* among the generated outputs by the models as well as human-composed datasets using reference-based scores like Mauve Audio Divergence (MAD) [14] and Kernel Audio Distance (KAD) [15].

To our knowledge, this is the first systematic study of human preference alignment in music generation. In addition to the experimental findings, we release a benchmark dataset containing generated music samples and evaluation results to prompt transparent, reproducible, and human-centered evaluation of music generative models. We hope this research will inspire further research on evaluating and optimizing generative models to better reflect human aesthetic judgments.

2. RELATED WORKS

In this work, we propose a categorization of performance evaluations for music generation systems based on three dimensions: objectives, human involvement, and modality, as summarized in Table 1. Since our focus is on aligning music generation with human preferences, this section is organized into objective and subjective evaluations, depending on whether human annotators are involved.

2.1. Objective evaluation

Evaluating synthesized audio requires human annotators to rate the candidate samples according to various perceptual criteria, which is time-consuming and labour-intensive. To alleviate the annotation cost, many existing works rely on objective metrics over subjective evaluation by applying metrics without involving human listening tests [16, 17, 18, 19, 20, 21]. For example, Kong *et al.* applied Inception Score (ISc) to measure the divergence among a bunch of generated audio [1]. With regard to reference-based metrics, Fréchet

979-8-3503-2411-2/25/$31.00 ©2025 IEEE

Table 1. Summary of evaluation metrics for music generation systems.

Metric	Modality	Objective	Human Involvement
Inception Score (ISc)	audio	distribution-to-distribution	×
Fréchet Audio Distance (FAD)	audio	distribution-to-distribution	×
Per-song FAD	audio	instance-to-distribution	×
FAD-∞	audio	distribution-to-distribution	×
Log-spectral Distance (LSD)	audio	instance-to-instance	×
Kullback–Leibler (KL) Divergence	audio	distribution-to-distribution	×
Kernel Inception Distance (KID)	audio	distribution-to-distribution	×
Learned Perceptual Audio Patch Similarity (LPAPS)	audio	instance-to-instance	×
CLAP Score	audio-text	instance-wise	×
KAD (Kernel Audio Distance)	audio	distribution-to-distribution	×
MAUVE Audio Divergence (MAD)	audio	distribution-to-distribution	×
Mean Opinion Score (MOS)	audio	instance-wise	✓
Meta-AudioBox-Aesthetics	audio	instance-wise	✓
MusicEval-CLAP	audio-text	instance-wise	✓
Human Aesthetics Preference Model (DRAGON)	audio	instance-wise	✓

Audio Distance (FAD) [2], Log-spectral distance (LSD) and Kullback–Leibler (KL) divergence [22] are computed between reference and synthesized audio by feeding them into pretrained deep audio networks. Gui *et al.* further improved FAD by proposing per-song FAD, which calculates an individual FAD score for each song against the reference dataset, and FAD-∞, which estimates an unbiased FAD score by extrapolating the FAD to an infinite sample size [23]. Kernel Inception Score (KID) was used to assess the quality of generated audio by measuring the distance between ground truth distributions and synthesized audio [21]. Based on the premise that the feature within deep audio networks matches human perception better, Learned Perceptual Audio Patch Similarity (LPAPS) was used to model the distance between generated and reference audio [24, 25]. To evaluate cross-modality generation performance, the contrastive audio-language model (CLAP) score was devised to reflect the alignment of generated audio and text description [25]. KAD [15] was proposed based on Maximum Mean Discrepancy (MMD) to alleviate the reliance on Gaussian assumptions and sensitivity to sample size.

2.2. Subjective human evaluation

Human evaluation is the most straightforward to reflect the quality of synthesized audio, as well as the golden standard for music. Many works used human annotators to evaluate the generated audio and even to optimize the model for better alignment with human preference [11].

In an existing work [26] for the evaluation of text-to-speech models, neural networks are trained to predict the Mean Opinion Score (MOS). The Meta-AudioBox-Aesthetics project [27] proposes a guideline for annotating human aesthetics scores from four axes: Production Quality (PQ), Production Complexity (PC), Content Enjoyment (CE), and Content Usefulness (CU). In particular, the annotations are sourced from 158 raters who passed through a rater qualification program that maintains alignment with a golden set. After annotating 97k audio samples (500 hours of data, with one third being music), four predictors for non-intrusive and utterance-level automatic aesthetics prediction are trained. The predictor's output is verified to be closely correlated with human perception.

In the particular category of instrumental music, there exist assessment models with focus on music skill such as PianoJudges [28], PISA assessment [29] and LLaQo [30]. Such models' training are heavily reliant on the annotation of professional instrumentalists or pedagogues, and their domain of expertise is hardly transferable to the evaluation of general popular music.

The work of Huang *et al.* [14] quantified multiple desiderata and tested the robustness of objective evaluation metrics with regard to synthetic degradations. Along with their pairwise-annotated dataset MusicPrefs, their proposed MAD metric was computed on representations from a self-supervised audio embedding model that's demonstrated to align better with human preference.

MusicEval [31] project collected a dataset of 2,748 music clips generated by 31 different generative methods, each of which was prompted by 384 music descriptions. Each clip was evaluated by five annotators, and their scores were averaged. Based on the data collected, a music assessment model was proposed by finetuning a CLAP model. However, as of the time of writing, neither their dataset nor model was released to the public yet.

Similarly, Bai *et al.* proposed a human aesthetics preference model for AI-generated music by collecting 1,676 human-rated clips [11]. They leverage the aesthetics model, together with 20 objective evaluation metrics, to optimize a music generative model, DRAGON. Despite their improved performance on various metrics, such as FAD, it is still debatable if the objective evaluation is aligned with human preference, as discussed in [15]. In the symbolic music domain, Nicolas *et al.* applied the Meta-AudioBox-Aesthetic predictor [27] as a reward to improve a piano MIDI model with group relative policy optimization (GRPO) [12].

However, the approach of obtaining human annotations for training or testing data of evaluation systems suffers from inherent subjectivity. For example, the personal preferences of the annotators would lead to disagreement, making it difficult to find underlying patterns for "high-quality music". A study [32] on the teacher's rating of piano performance quality has already reflected large disagreements even under strict grading guidelines.

3. HUMAN VS. HUMAN: DO `MUSICPREF` ANNOTATIONS AGREE WITH AESTHETICS SCORES?

Can different sources of human annotation reach a common ground? In this experiment, our aim is to understand the relationship between human-annotated pairwise preferences and the independently

trained aesthetics predictor, conducting a detailed comparative analysis across their perceptual dimensions, respectively.

Each entry in `MusicPref` consists of a pairwise comparison between two audio clips generated by different models, accompanied by human judgments of which audio is preferred for musicality and fidelity. Independently, we used the Meta-AudioBox-Aesthetics model [27] trained on human preferences, producing scalar scores across four aspects: Content Enjoyment (CE), Content Usefulness (CU), Production Complexity (PC), and Production Quality (PQ). For each in 2515 pairs in `MusicPref`, we computed the difference in aesthetic scores between the two audio clips and test whether this difference aligns with the direction of human preference. We evaluated the agreement between human preference direction and aesthetics score deltas using two metrics:

- **Accuracy:** The proportion of non-tie human decisions (musicality: 2049 pairs; fidelity: 1889) where the aesthetics score delta correctly predicted the preferred audio.
- **Spearman correlation:** Rank correlation between human preference direction ($+1$ for A wins, -1 for B wins, 0 for tie) and the difference in score across the dataset.

Table 2 reports the accuracy of each aesthetic metric in predicting human preferences. We can observe that Content Usefulness (CU) consistently yields the highest accuracy of 62.3% for musicality, and Production Quality (PQ) predicts higher accuracy for fidelity. However, given 50% of the random guess baseline, the results demonstrate a very poor agreement between the human annotated pairs and learnt preference scores from neural network.

Table 2. Prediction accuracy of aesthetics score deltas against human preferences.

Preference Type	CE	CU	PC	PQ
Musicality	0.604	**0.623**	0.483	0.596
Fidelity	0.572	0.590	0.432	**0.591**

Table 3 presents the Spearman correlation between aesthetic score deltas and human preferences, supporting the low correlation trend observed in the accuracy scores. CU achieves the highest correlation for musicality ($r = 0.258$) and strong performance in fidelity ($r = 0.200$), followed closely by CE and PQ. Again, the PC performs poorly and even exhibits a weak negative correlation with fidelity.

Table 3. Spearman correlation between aesthetic score deltas and human preferences.

Preference Type	CE	CU	PC	PQ
Musicality	0.237	**0.258**	-0.002	0.204
Fidelity	0.177	**0.200**	-0.087	0.181

4. TTM LEADERBOARD: WHO DOES BEST, AND UNDER WHOSE JUDGE?

In this section, we benchmark five recent text-to-music (TTM) generation models by evaluating their outputs on a standardized set of prompts, and reporting the four AudioBox-aesthetics metrics over synthezised music clips.

4.1. Models

- `JASCO`: A large-scale model optimized for musical generation with strong coherence and aesthetic fluency, tuned for structured control including melody, chord and drum track [33]. Its output is 10 seconds.
- `Stable-Audio-Open`: A diffusion-based model developed by Stability AI, trained on free-licensed music [34]. The output is 47 seconds.
- `MusicGen`: A Transformer-based model from Meta AI for TTM generation, trained on licensed music [35]. We use the `Large` model in our experiments and their output is 20 seconds.
- `YuE`: A two-stage auto-regressive system for teh generation of full-length songs [36]. In practice, its output is around 50 seconds depending on the length of the input lyrics.
- `DiffRhythm`: A recent diffusion-based song generation framework that excels in inference speed [37]. We generate outputs of 95 seconds from this model.

4.2. Dataset

We chose LP-MusicCaps [38] as the prompt to generate a fixed set of music by various models, and the prompts were derived from the original tag from the MusicCaps dataset. We looked at the entire training and testing set combined, a total of 5,521 prompts. We also included the audio ground truth of the MusicCaps (human-made recording, 10s each) in the aesthetics evaluation.

With the growing emphasis on fine-grained control in music generation, several of the TTM models evaluated in this study require additional conditioning inputs.

For models such as `DiffRhythm` and `YuE`, lyrics are required as input modality. To support this, we generated a set of 50 lyrics using `GPT-4o` spanning diverse themes and structured them according to common popular song forms that each model is capable of handling. The `JASCO` model, on the other hand, requires both drum tracks and chord progressions as additional input. We fetched a collection of 50 drum tracks from *The Drum Tamer*[1], covering a wide range of genres, tempi, and drum kit styles. For chord progression inputs, we generated a set of 100 symbolic chord sequences using `GPT-4o` that encompass a variety of harmonic patterns and musical styles.

4.3. Main Results

We report the average (*mean*) and variability (*standard deviation*) of the aesthetics scores along four dimensions for each model, as summarized in Table 4. Notably, `JASCO` consistently achieves the highest scores across most dimensions, particularly in CU (7.81 ± 0.42) and PQ (7.80 ± 0.50), suggesting its strength in generating coherent and stylistically rich compositions that are perceived as both creative and high-quality. `DiffRhythm`, although a newer diffusion-based model, shows strong performance in PC (6.39 ± 0.93), outperforming all baselines in this category, and remains competitive in CE and PQ. This indicates its potential in capturing fine-grained rhythmic structure and production fidelity, aided by rhythm-aware conditioning.

In contrast, `Stable-Audio-Open` underperforms in CE and PC, with relatively low scores (4.55 ± 0.56 and 2.83 ± 0.58, respectively), although it maintains a high PQ score (7.01 ± 0.37), suggesting that the outputs retain aesthetic appeal despite weaker

[1] `https://www.adoptabeat.co.uk`

Table 4. Mean ± standard deviation of aesthetics scores (CE, CU, PC, PQ) for each model. `(no low)*` is the MusicCaps GT subset that removes the recordings with "low quality" tag.

Model	Content Enjoyment	Content Usefulness	Production Complexity	Production Quality
`JASCO`	**7.33 ± 0.60**	**7.81 ± 0.42**	5.26 ± 0.87	**7.80 ± 0.50**
`Stable-Audio-Open`	4.55 ± 0.56	6.82 ± 0.43	2.83 ± 0.58	7.01 ± 0.37
`MusicGen-Large`	6.77 ± 1.07	7.57 ± 0.66	5.03 ± 0.79	7.46 ± 0.66
`DiffRhythm`	6.81 ± 1.19	7.18 ± 0.86	**6.39 ± 0.93**	7.51 ± 0.85
`YUE`	7.08 ± 0.78	7.60 ± 0.37	5.01 ± 0.93	7.77 ± 0.94
`MusicCaps GT`	6.21 ± 1.41	6.69 ± 1.32	5.41 ± 1.46	6.93 ± 1.15
`MusicCaps GT (no low)*`	6.25 ± 1.39	6.75 ± 1.27	5.44 ± 1.46	6.98 ± 1.12

structural or conceptual expression. `MusicGen-Large` offers a more balanced profile, trailing behind `JASCO` in CU and PQ, but outperforming `Stable-Audio-Open` across all metrics.

The ground truth reference, `MusicCaps GT`, serves as a human baseline, but is surpassed by both `JASCO` and `MusicGen-Large` in several dimensions. Given that a non-trivial amount of low quality recordings exists in the MusicCaps dataset (under the "low quality" tag or captions), we have also computed the aethetics score for the subset without the "low quality" tag, totaling 4384 recordings. However, the `MusicCaps GT (no low)` subset does not achieve a score that is significantly higher than the full set.

It is worth noting that some models, such as `JASCO`, `DiffRhythm`, and `Yue`, benefit from additional conditioning inputs (e.g., chord progressions, drum tracks, or lyrics), which may provide them with an advantage in generating more structured or expressive outputs. While this reflects their intended usecases, it also introduces variability in input richness, which should be considered when interpreting the comparisons.

4.4. Do our aesthetic preferences lean towards particular kinds of music?

To investigate whether the learned aesthetics metrics exhibit bias with respect to specific music types, we cluster descriptive tags (from the original tags of LP-MusicCaps) using sentence-level embeddings from a pre-trained language model (MiniLM) [2] and KMeans clustering. Each tag is assigned to one of 15 semantic clusters, and we compute the most frequent tag within each cluster as its representative label. These labels are used on the x-axis in Figure 1 to improve interpretability.

Among the four aesthetics dimensions, we only show the **Production Complexity (PC)** demonstrates the most prominent separation between models and cluster types, and other dimensions such as CE or CU show comparatively little inter-model or inter-cluster variability.

Figure 1 compares PC scores across clusters for all models. Interestingly, despite architectural and training differences, most models exhibit *similar scoring patterns across clusters*. This suggests that our learned aesthetics scoring function is systematically influenced by certain types of music content—clusters associated with rhythmic or electronic elements (e.g., "punchy kick", "synth") tend to receive higher PC scores, while clusters with simpler or lo-fi instrumentation (e.g., "low quality", "mono") receive lower scores.

Moreover, the ground-truth human-composed reference dataset (`musiccaps_gt`) displays higher variance in PC scores across clusters. This could reflect the broader range of styles and production techniques in real-world music compared to the more homogeneous outputs of generative models. These findings highlight the importance of considering semantic content when interpreting evaluation scores and suggest that aesthetic predictors may encode latent content preferences shaped by the training data or annotator biases.

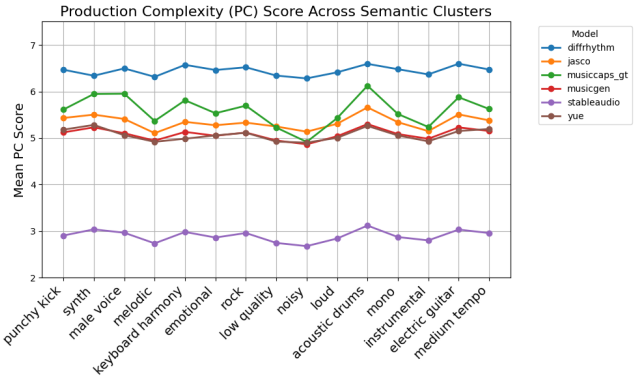

Fig. 1. Production Complexity (PC) score across semantic clusters, labeled by most frequent tag. While all models follow a similar score pattern across clusters, the human reference shows greater variance.

5. DIVERSITY OR DRIFT? DATASET-LEVEL ALIGNMENT USING REFERENCE-BASED SCORES

While the previous experiments centered on aesthetics scores and pairwise human preference alignment, this section investigates the distributional alignment among each generated music sets and human-composed set. We examine reference-based evaluation metrics on the generated dataset computed in Section 4, offering a complementary perspective to subjective assessments. Specifically, we report results using two metrics: MAD [14] and KAD [15].

MAD [14] combines the KL divergence-based notion to measure the distributional divergence between generated and reference audio embeddings. A lower MAD score indicates a closer alignment to the real music distribution. KAD [15], on the other hand, quantifies the average kernelized distance between the generated and reference distributions in a learned perceptual space. We employ the `panns-wavegram-logmel` [39] for reference embeddings, using its given implementation.

Figure 2 visualizes the asymmetric pairwise MAUVE Audio Divergence (MAD) scores across the models' outputs. Notably, `MusicGen-Large` exhibits the lowest MAD score with respect to `MusicCaps GT` (2.45), followed by `JASCO` (3.89) when evaluated against the other models. These results suggest that

[2] `https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2`

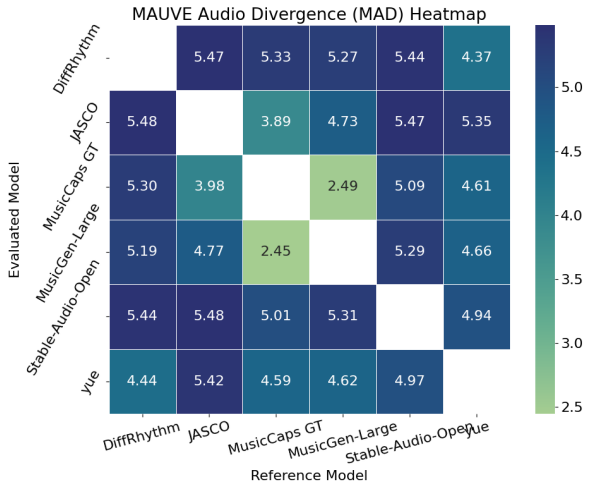

Fig. 2. Asymmetric heatmap of pairwise MAD scores across the evaluated models and the ground-truth MusicCaps dataset. Lower scores indicate closer distributional alignment.

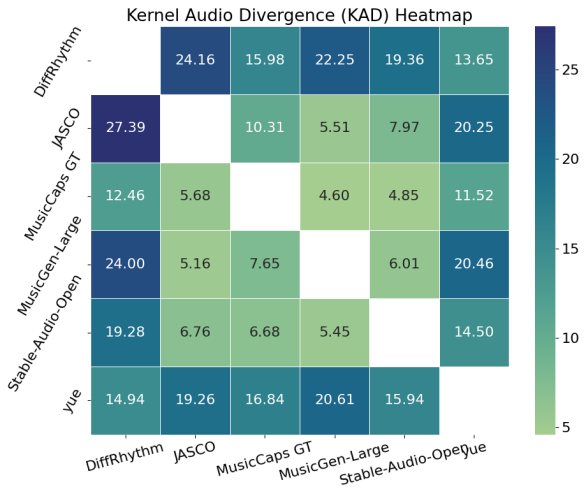

Fig. 3. Pairwise KAD scores between the evaluated models and the MusicCaps ground-truth dataset. Lower scores indicate closer proximity in the learned perceptual space.

`MusicGen-Large` is best aligned with the reference distribution, while models such as `DiffRhythm`, `Stable-Audio-Open`, and `YuE` exhibit consistently higher MAD scores, indicating larger distributional shifts.

The Kernel Audio Distance (KAD) scores in Figure 3 further corroborate these findings. Again, `MusicGen-Large` and `JASCO` achieve the lowest distances relative to the `MusicCaps GT` reference (7.65 and 5.51, respectively), underscoring their strong alignment with perceptual aspects of real music. In contrast, `DiffRhythm`, `Stable-Audio-Open`, and especially `YuE` show substantially higher distances from all other sets, suggesting that their output distributions deviate more significantly. Comparatively, the `DiffRhythm` and `YuE` are relatively close in terms of both MAD and KAD values, which is explainable given that both models have a focus on song generation.

6. CONCLUSION AND FUTURE WORK

In this work, we presents a cross-referenced discussion of text-to-music generation evaluations. Our results reveal substantial inconsistencies between different evaluation perspectives, highlighting the challenges of fully capturing human judgment through automated proxies. By benchmarking five recent models and releasing the generated dataset, we aim to support more transparent, reproducible, and human-centric evaluation practices in music generation research. Future work includes exploring richer annotation schemes and developing evaluation methods that better reflect detailed music preferences that varies across culture background, music training, listening habits.

7. REFERENCES

[1] Qiuqiang Kong, Yong Xu, Turab Iqbal, Yin Cao, Wenwu Wang, and Mark D. Plumbley, "Acoustic scene generation with conditional SampleRNN," in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 925–929.

[2] Felix Kreuk, Gabriel Synnaeve, Adam Polyak, Uriel Singer, Alexandre Défossez, Jade Copet, Devi Parikh, Yaniv Taigman, and Yossi Adi, "AudioGen: Textually guided audio generation," in *The Eleventh International Conference on Learning Representations*, 2023.

[3] Huan Zhang, Shreyan Chowdhury, Carlos Eduardo Cancino-Chacón, Jinhua Liang, Simon Dixon, and Gerhard Widmer, "DExter: Learning and Controlling Performance Expression with Diffusion Models," *Applied Sciences*, vol. 14, no. 15, 2024.

[4] DeepSeek-AI, "DeepSeek-R1: Incentivizing reasoning capability in llms via reinforcement learning," *arXiv:2501.12948*, 2025.

[5] Qwen Team, "Qwen2.5 technical report," *arXiv:2412.15115*, 2025.

[6] Huaisheng Zhu, Teng Xiao, and Vasant G Honavar, "DSPO: Direct score preference optimization for diffusion model alignment," in *The Thirteenth International Conference on Learning Representations*, 2025.

[7] Seed Team, "Seed-TTS: A family of high-quality versatile speech generation models," *arXiv:2406.02430*, 2024.

[8] Geoffrey Cideron, Sertan Girgin, Mauro Verzetti, Damien Vincent, Matej Kastelic, Zalán Borsos, Brian McWilliams, Victor Ungureanu, Olivier Bachem, Olivier Pietquin, Matthieu Geist, Léonard Hussenot, Neil Zeghidour, and Andrea Agostinelli, "Musicrl: Aligning music generation to human preferences," 2024.

[9] 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*, vol. 1, Association for Computing Machinery, 2024.

[10] Huan Liao, Haonan Han, Kai Yang, Tianjiao Du, Rui Yang, Qinmei Xu, Zunnan Xu, Jingquan Liu, Jiasheng Lu, and Xiu Li, "BATON: Aligning text-to-audio model using human preference feedback," in *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, 8 2024, pp. 4542–4550.

[11] Yatong Bai, Jonah Casebeer, Somayeh Sojoudi, and Nicholas J. Bryan, "DRAGON: Distributional rewards optimize diffusion generative models," *arXiv:2504.15217*, 2025.

[12] Nicolas Jonason, Luca Casini, and Bob L. T. Sturm, "SMART: Tuning a symbolic music generation system with an audio domain aesthetic reward," *arXiv:2504.16839*, 2025.

[13] Ahmet Solak, Florian Gr"otschla, Luca A Lanzend"orfer, and Roger Wattenhofer, "Benchmarking music generation models and metrics via human preference studies," in *Audio Imagination: NeurIPS 2024 Workshop AI-Driven Speech, Music, and Sound Generation*, 2024.

[14] Yichen Huang, Zachary Novack, Koichi Saito, Jiatong Shi, Shinji Watanabe, Yuki Mitsufuji, John Thickstun, and Chris Donahue, "Aligning text-to-music evaluation with human preferences," *arXiv:2503.16669*, 2025.

[15] Yoonjin Chung, Pilsun Eu, Junwon Lee, Keunwoo Choi, Juhan Nam, and Ben Sangbae Chon, "KAD: No more FAD! An effective and efficient evaluation metric for audio generation," *arXiv:2502.15602*, 2025.

[16] Jinhua Liang, Huan Zhang, Haohe Liu, and Yin Cao, "WavCraft: Audio editing and generation with large language models," in *ICLR 2024 Workshop on Large Language Model (LLM) Agents*, 2024.

[17] Xubo Liu, Zhongkai Zhu, Haohe Liu, Yi Yuan, Meng Cui, Qiushi Huang, Jinhua Liang, Yin Cao, Qiuqiang Kong, Mark D Plumbley, et al., "WavJourney: Compositional audio creation with large language models," *IEEE Transactions on Audio, Speech and Language Processing*, 2025.

[18] Tanisha Hisariya, Huan Zhang, and Jinhua Liang, "Bridging paintings and music – exploring emotion based music generation through paintings," *arXiv:2409.07827*, 2024.

[19] Wen Qing Lim, Jinhua Liang, and Huan Zhang, "Hierarchical symbolic pop music generation with graph neural networks," *arXiv:2409.08155*, 2024.

[20] Huan Zhang, Akira Maezawa, and Simon Dixon, "Renderbox: Expressive performance rendering with text control," *arXiv:2502.07711*, 2025.

[21] Yi Yuan, Haohe Liu, Jinhua Liang, Xubo Liu, Mark D. Plumbley, and Wenwu Wang, "Leveraging pre-trained audioldm for sound generation: A benchmark study," in *2023 31st European Signal Processing Conference (EUSIPCO)*, 2023, pp. 765–769.

[22] Dongchao Yang, Jianwei Yu, Helin Wang, Wen Wang, Chao Weng, Yuexian Zou, and Dong Yu, "Diffsound: Discrete diffusion model for text-to-sound generation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 1720–1733, 2023.

[23] Azalea Gui, Hannes Gamper, Sebastian Braun, and Dimitra Emmanouilidou, "Adapting frechet audio distance for generative music evaluation," in *2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024.

[24] Hila Manor and Tomer Michaeli, "Zero-shot unsupervised and text-based audio editing using DDPM inversion," in *ICML 2024 Workshop on Structured Probabilistic Inference & Generative Modeling*, 2024.

[25] Jinhua Liang, Yi Yuan, Dongya Jia, Xiaobin Zhuang, Zhengxi Liu, Yuanzhe Chen, Zhuo Chen, Yuping Wang, and Yuxuan Wang, "AudioMorphix: Training-free audio editing with diffusion probabilistic models," *arxiv*, 2024.

[26] Alessandro Ragano, Emmanouil Benetos, Michael Chinen, Helard Becerra Martinez, Chandan K A Reddy, Jan Skoglund, and Andrew Hines, "A comparison of deep learning MOS predictors for speech synthesis quality," in *2023 34th Irish Signals and Systems Conference (ISSC)*, 2023, pp. 1–6.

[27] Andros Tjandra, Yi-Chiao Wu, Baishan Guo, John Hoffman, Brian Ellis, Apoorv Vyas, Bowen Shi, Sanyuan Chen, Matt Le, Nick Zacharov, Carleigh Wood, Ann Lee, and Wei-Ning Hsu, "Meta audiobox aesthetics: Unified automatic quality assessment for speech, music, and sound," *arXiv:2502.05139*, 2025.

[28] Huan Zhang, Jinhua Liang, and Simon Dixon, "From Audio Encoders to Piano Judges: Benchmarking Performance Understanding for Solo Piano," in *Proceeding of the 25t International Society on Music Information Retrieval (ISMIR)*, 2024.

[29] Paritosh Parmar, Jaiden Reddy, and Brendan Morris, "Piano Skills Assessment," in *IEEE 23th International Workshop on Multimedia Signal Processing (MMSP)*, 2021.

[30] Huan Zhang, Vincent Cheung, Hayato Nishioka, Simon Dixon, and Shinichi Furuya, "LLaQo: Towards a Query-Based Coach in Expressive Music Performance Assessment," in *In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025.

[31] Cheng Liu, Hui Wang, Jinghua Zhao, Shiwan Zhao, Hui Bu, Xin Xu, Jiaming Zhou, Haoqin Sun, and Yong Qin, "MusicEval: A generative music dataset with expert ratings for automatic text-to-music evaluation," *arXiv:2501.10811*, 2025.

[32] Huan Zhang, Vincent Cheung, Hayato Nishioka, Simon Dixon, and Shinichi Furuya, "How does the teacher rate? Observations from the NeuroPiano dataset," in *Late-Breaking and Demo Session of the International Society on Music Information Retrieval (ISMIR)*, 2024.

[33] Or Tal, Alon Ziv, Itai Gat, Felix Kreuk, and Yossi Adi, "Joint Audio and Symbolic Conditioning for Temporally Controlled Text-to-Music Generation," in *Proceeding of the 25t International Society on Music Information Retrieval (ISMIR)*, 2024.

[34] Zach Evans, Julian D. Parker, CJ Carr, Zack Zukowski, Josiah Taylor, and Jordi Pons, "Stable Audio Open," *Arxiv preprint arXiv:2407.14358*, jul 2024.

[35] Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez, "Simple and controllable music generation," in *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[36] HKUST and MAP, "Yue: Scaling open foundation models for long-form music generation," *arXiv:2503.08638*, 2025.

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

[38] SeungHeon Doh, Keunwoo Choi, Jongpil Lee, and Juhan Nam, "LP-MusicCaps: LLM-based pseudo music captioning," in *Proceeding of the 24th International Society on Music Information Retrieval (ISMIR)*, 2023.

[39] Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D. Plumbley, "PANNs: Large-scale pretrained audio neural networks for audio pattern recognition," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 28, pp. 2880–2894, Nov. 2020.