11-785 Team 23
WAME: Waveform Music Editor
A MusicLLM Model for Music Editing

Huankang Chen | Qixiao Zhu | Yuding Wang
Carnegie Mellon University

Abstract

The project, WAME.AI, introduces an advanced AI-driven music editor that enhances music editing capabilities through conversational AI interactions. By leveraging a fine-tuned MusicGen model [1] and a multitrack dataset from the "Mixing Secrets" Free Multitrack Download Library, this project uniquely utilizes LLMs [2] to generate data-driven prompts and employs automatic tag matching for more effective model training. These innovations facilitate deeper musical comprehension and dynamic editing based on complex user prompts. Preliminary results are promising, offering a solid foundation for future advancements in AI-enabled creative music editing.

Code implementations: https://github.com/YudingWang/GenMusic

1 Introduction

Music transcends the simple act of sound production, embodying a form of art that integrates expression and carries the cultural identities of the human race. From the resonance of ancient drums to the synthetic tones of modern synthesizers, musical instruments have always been pivotal in turning abstract thoughts into auditory reality. The last century marked significant advancements with the introduction of electronic musical instruments like the theremin in the 1920s and the rise of digital synthesizers such as the Yamaha DX7 in the 1950s, which catalyzed a new era in electronic music production [3].

Despite these advancements, the process of musical composition remains intricate, requiring deep knowledge of music theory and considerable time. The advent of digital technology brought early attempts to algorithmically understand and recreate musical behaviors as early as the 1950s [4]. Progressing from these rudimentary systems, the recent application of deep learning has revolutionized music generation, with models like MelodyRNN [5], WaveNet [6], and more recently, MusicLDM [7] and MusicGen [1], setting new benchmarks in the field. Furthermore, the emergence of Large Language Models (LLMs) such as ChatGPT [2] and Llama [8] has transformed human-computer interaction, enabling users to manipulate digital environments with conversational language, similar to how DALL·E enables users to create and modify images through textual prompts [9].

However, while these developments have stabilized the generation of music using AI, they predominantly produce outputs that can be unpredictable and often not fully controllable, limiting their utility to more novelty applications rather than robust tools for serious music creation. Most existing models are adept at generating music but lack mechanisms to effectively edit or refine generated music in response to nuanced user feedback, thereby retaining minimal information from one iteration to the next.

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

In response to these limitations, we propose WAME.AI, a novel system that leverages both music generation models and the contextual understanding capabilities of LLMs. Our model is designed to interpret language input and generate music that not only aligns with the user's initial prompts but also adaptively refines itself in response to subsequent user input. By integrating a dynamic editing mechanism, WAME.AI aims to transform AI-generated music from a mere novelty into a valuable tool for both amateur and professional music production. This approach is largely unexplored in existing music LLM models, and through WAME.AI, we intend to pioneer this promising new avenue in AI-enabled creative music editing.

2 Literature Review

The music generation field has seen significant evolution over recent decades, moving from simple probability models to the sophisticated use of deep learning techniques. Insights from surveys like "A Comprehensive Survey on Deep Music Generation" [10] and "A Survey on Deep Learning for Symbolic Music Generation" [11] provide a broad understanding of the various tasks involved in music creation at different levels.

Previously, computer-based music creation was largely rule-based, attempting to replicate traditional musical theory. However, the advent of deep learning has revolutionized this process, enabling computers to analyze vast amounts of music and generate original compositions. Deep learning models, such as RNNs [12, 13, 14], LSTMs [15], and Transformers [16], have propelled the field forward. For instance, the Transformer model has been effectively used for music infilling tasks, as seen in "Melody Infilling With User-Provided Structural Context." [17] Furthermore, the integration of Variational Autoencoders (VAEs) and Generative Adversarial Networks (GANs) for music generation, inspired by dance videos in "Dance-conditioned Artistic Music Generation by Creative-GAN," [18] illustrates the creative synergy between data of multiple modalities.

The emergence of the Diffusion model, particularly when combined with the Transformer architecture, marks the latest advancement, offering unprecedented creativity and output fidelity. Notable implementations include "Audiobox: Unified Audio Generation with Natural Language Prompts," [19] showcasing sophisticated audio generation from textual prompts, "MusicGen: Simple and Controllable Music Generation" [1], and "MusicLDM: Enhancing Novelty in Text-to-Music Generation Using Beat-Synchronous Mixup Strategies," [7] which demonstrate innovative text-to-music generation approaches. Additionally, "Noise2Music: Text-conditioned Music Generation with Diffusion Models" [20] leverages a pre-trained language model to interpret prompts, further enhancing the text-conditioned music generation landscape.

Furthermore, we will also draw upon models such as those described in "Symbolic Music Generation with Diffusion Models" [21], which utilize diffusion model for music-to-music generation functionalities, serving as a foundation for our implementation. Although these models often do not incorporate the ability to accept and process prompts based on Large Language Models (LLMs), their methodologies provide crucial insights into the capabilities of diffusion models in music generation contexts.

Despite these advancements, the concept of music-to-music generation guided by human prompts remains an uncharted and challenging area. This innovative approach aims to create a system that takes existing musical pieces as inputs, rather than generating music based on human's orders from scratch, and uses human prompts to guide the re-composition or modification of the music according to specific themes or instructions. Drawing inspiration from seminal works such as "Denoising Diffusion Probabilistic Models," [22] which introduces the diffusion model for generation, and "Attention Is All You Need," [16] which thoroughly elucidates the Transformer model, our project seeks to build upon these pivotal contributions. In particular, we have recently been exploring the capability of AudioSep [23] as a potential addition to our final model for stronger editorial ability, specifically when the model is prompted to adjust the volume or delete specific voices. We will also reference papers on the most relevant and recent applications in this field that align closely with our ideas, including the previously mentioned music generation models [19, 1, 20].

By adopting the principles outlined in these landmark studies and incorporating insights from these cutting-edge applications, we aim to explore the untapped potential of prompt-driven music-to-music generation. This method promises not only to open up a new avenue for interactive musical creativity

but also allows users to have a direct impact on the generation process, enabling them to customize the output to reflect their personal tastes or creative intentions.

3 Model

3.1 Baseline Model

As described in our proposal idea, our model intends to have the capability of both generating new music and editing existing music. Thus, our baseline model attempts to combine two primary parts.

The first part uses MusicGen [1], as described in the referenced paper, as the baseline model for our music generation task. MusicGen [1] is a state-of-the-art single-stage language model that generates music based on compressed discrete music representations. It introduces a novel token interleaving pattern that enables the efficient generation of high-quality mono and stereo music samples from textual descriptions or melodic prompts. This approach streamlines the music generation process by eliminating the need for multiple cascading models.

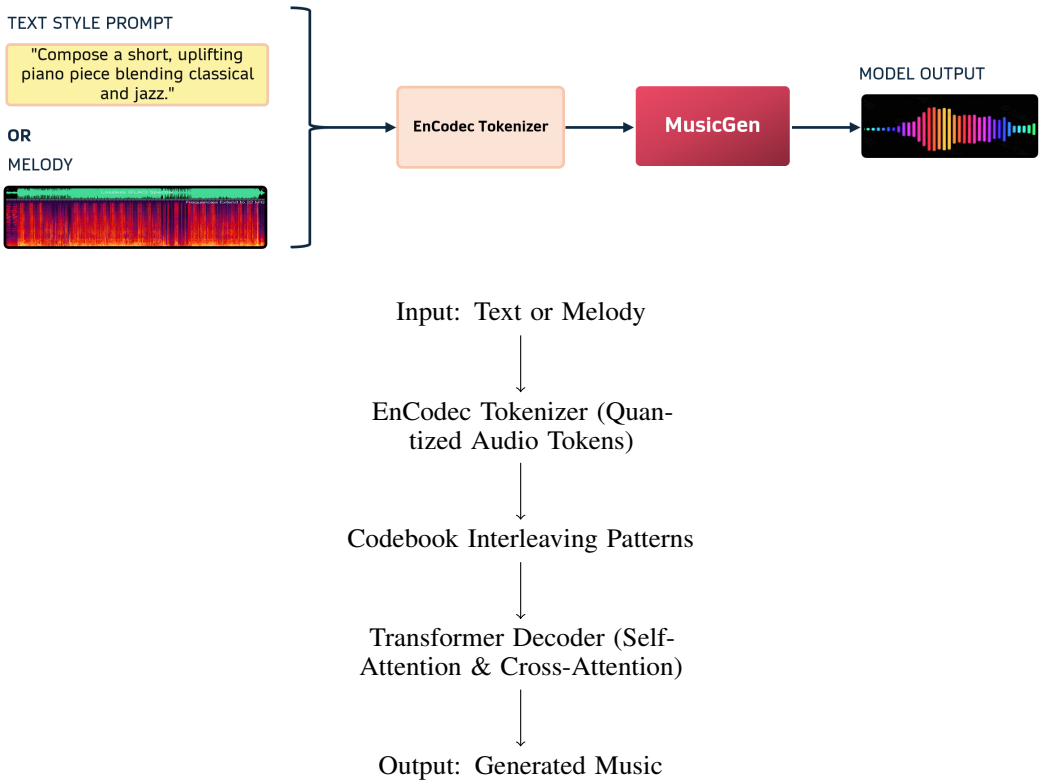

Figure 1: MusicGen Architecture Diagram [1]

The second part of the baseline model features AudioSep [23], a state-of-the-art model that takes in text user query and extracts the specified source from a mixed audio signal, to improve the editing capability of our model when prompted changing or deleting specific voices. We will likely remove the language query in our final model and stick to separating every possible voice in the mixed signal. This model separates existing music by applying a Short Term Fourier Transform (STFT) to convert the input waveform into complex spectrogram in the frequency domain and feed it into an encoder-decoder network. Loss is calculated by comparing the ground truth waveform and the generated separated waveform, both in the time domain.

Ultimately, our final product attempts to combine the two parts mentioned above into a model that contains the best of both worlds.

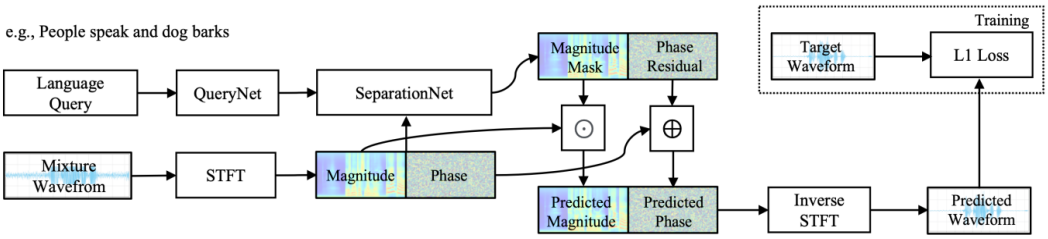

Figure 2: AudioSep Architecture Diagram [23]. "AudioSep contains two key components: a QueryNet and a SeparationNet. The QueryNet is the text encoder of CLIP [24] or CLAP [25] model. The SeparationNet is a frequency-domain ResUNet [26, 27] model." [23]

3.2 Training Objectives

Our proposed generative model [1] seeks to synthesize music by directly approximating the conditional probability distribution of musical data, X, given a context c, such as textual descriptions or melodic prompts. The training objective is formulated as the minimization of the negative log-likelihood:

$$-\log p(X|c) = -\sum_{i=1}^{N} \log p(x_i|x_{i-1}, \ldots, x_1, c), \tag{1}$$

where x_i denotes the tokens in the sequence of the music output. This objective encourages the model to learn a representation that is faithful to the structure of the music while being able to generate novel compositions when conditioned on new prompts.

For the original AudioSep [23], voices s_1 and s_2 from the training set are mixed into input x while using the loudness augmentation method proposed in [27], where the authors calculate the energy E_1 and E_2 for both voices using $E = \|s\|_2^2$ and apply a scaling factor $\alpha = \sqrt{D_1/E_2}$ to s_2 such that $x = s_1 + \alpha s_2$. They then minimize the L1 loss between the ground truth waveform s and the separated waveform $\hat{s}$:

$$l = \|s - \hat{s}\|_1 \tag{2}$$

Our model seeks to fine tune MusicGen [1] using the same training approach as AudioSep [23], in which we construct a dataset that is specific for music separation and addition. However, we use the Cross Entropy Loss function that MusicGen uses:

$$\text{CE} = -\sum_{i=1}^{N} y_i \log(p_i) \tag{3}$$

where N corresponds to all possible output values, y is the ground truth sample, and p is the predicted probability.

4 MOS - Baseline Model Evaluation

The Mean Opinion Score (MOS) is an essential metric used to evaluate subjective experiences, such as audio quality. In our project, the MOS test served as the starting point, enabling us to identify the MusicGen model's drawbacks through subjective evaluation and thus guide our efforts to enhance the model's capabilities.

4.1 Baseline Model Evaluation Process

To assess the MusicGen model's proficiency, we employed the MOS test across four different music genres: Pop, Orchestra, Rock, and Jazz. This diversity was integral in examining the model's versatility. Each genre underwent a survey with more than 10 participants who evaluated both modification and separation prompts. The core of the survey revolved around two pivotal aspects:

- Modification Quality: Participants evaluated how accurately the model generated music in line with the given prompts.
- Retain Quality: Participants assessed the extent to which the model preserved the original music's features in the modified output.

4.2 Survey Structure and Questions

Participants were presented with samples of original music and their corresponding MusicGen-generated counterparts. Each sample was modified based on prompts designed to either insert or delete musical elements. Respondents then answered two questions per prompt to express their satisfaction with the quality of the music generated and the fidelity of information retention from the original music.

For instance, one of the prompts was: "Add heavy drumbeat without changing the existing rhythm, melody, and vocals." The participants then responded to questions aimed at gauging the modification accuracy and retention of the original music's information.

4.3 Take-home Lessons from the MOS Test

The initial MOS test yielded insightful conclusions. Notably, participants generally assigned lower scores to the model's ability to retain elements of the original music. It was observed that MusicGen tended to generate new music that bore little relation to the original input, especially during tasks that involved insertion and deletion.

The scores provided by the participants are as follows:

Genre	Modification Quality	Retain Quality
Pop	2.47	2.27
Orchestra	2.80	2.49
Rock	2.56	2.47
Jazz	2.86	2.60

Table 1: Mean Opinion Scores for MusicGen generated music samples.

Based on this data, it became apparent that while MusicGen could initiate changes to music samples based on textual prompts, it struggled significantly with tasks that required a delicate balance between modification and retention of original music features. Consequently, this highlighted the necessity for a refined approach that would enable MusicGen not only to generate but also to edit music by preserving the integrity of the original input.

In response to these findings, our project's objective was set: to fine-tune the model to endow it with the capability of retaining more information from the original input music, thus ensuring a more controlled and refined music generation process.

5 Dataset Collection

5.1 Raw Dataset Source

A foundational contribution of the WAME.AI project is the creation of an extensive dataset that serves as the bedrock for training our AI models. We have sourced our raw data from the 'Mixing Secrets' Free Multitrack Download Library, courtesy of Cambridge Music Technology [28]. This dataset comprises 597 multitrack songs, each replete with separated audio files recorded for each individual microphone, corresponding to different instruments.

5.2 Dataset Challenges

One of the primary challenges with the dataset is the inconsistency in labeling. The audio files were labeled by different people, resulting in a variety of names for what may essentially be the same instrument. Moreover, it is common within the dataset to encounter multiple files pertaining to the

same instrument, which introduces additional complexity to the task of data processing and model training.

5.3 Dataset Processing

The dataset structure, reflecting the multi-track nature of the recordings, can be visualized as follows:

```
Song
  |
  |-- Mic1Instrument1
  |-- Mic2Instrument1
  |-- Mic1Instrument2
  ...
```

Our preprocessing pipeline involves scraping the entire dataset from the Cambridge Library, followed by a downsampling step from 44.1 kHz to 32 kHz to conform to the MusicGen paper's standard, and a conversion from stereo to mono to simplify the audio data to a single channel.

5.4 Instrument Extraction

Our approach to addressing the labeling challenge began with extracting all potential file names from the dataset. We then tasked ChatGPT with generating a corresponding list of instrument names encompassing all file names found in the dataset. This process yielded a dictionary mapping each instrument to all file names that may resemble that instrument. This dictionary was meticulously optimized by our team to ensure precise and useful instrument identification.

5.5 Rich Description Prompt Generation

With the assistance of ChatGPT, we generated two dictionaries aimed at editing operations: one for insertion prompts and another for deletion prompts. Each dictionary correlates instruments with a variety of descriptive prompts, laying the groundwork for complex and intuitive music editing tasks.

5.6 Excluded & Full Audio File Pairing

For each track, we determined the involved instruments by consulting the extensive dictionary created previously. This facilitated the creation of song-specific dictionaries mapping instruments to audio file paths. For each instrument identified, we combined the remaining instrument tracks to produce a version excluding the selected instrument. Additionally, we merged all audio files to generate a complete version of the song with all instruments included. From these paired versions, we extracted a consistent 15-second audio segment to serve as the foundation for the model to process editing commands based on the generated prompts.

5.7 File Structure

The final structure of the data for each song is outlined as follows:

```
Song
  |
  |-- Instrument1
  |     |-- excluded.wav
  |     |-- full.wav
  |     |-- sepPrompt.json
  |     |-- insertPrompt.json
  |
  |-- Instrument2
  ...
```

In conclusion, the dataset collection phase is a pivotal contribution to the WAME.AI project, enhancing the AI's capability to discern and manipulate multifaceted musical compositions accurately. This

section captures all aspects of the dataset's sourcing, preparation, and optimization that are integral to developing a sophisticated AI model for music editing.

6 Fine-Tuning Process of MusicGen Models

To enhance the MusicGen model's performance and address the issues identified during the initial MOS evaluation, we embarked on a fine-tuning process. This process was meticulously crafted to refine both the Small and Large versions of the MusicGen model, which possess 300 million and 3.3 billion parameters, respectively.

6.1 Model Architecture

Both versions of the MusicGen model are built upon the DoRA (Weight-Decomposed, Rank-Adapted) architecture. This structure is specifically designed to optimize the efficiency of the model while maintaining its capacity for high-quality music generation. The DoRA architecture allows for rapid adaptation to new data, making it particularly suitable for fine-tuning tasks.

6.2 Fine-Tuning Objectives

The primary objective of the fine-tuning process was to improve the models' ability to retain original music characteristics while performing modifications as instructed by prompts. Given the substantial parameter count, particularly in the Large model, fine-tuning also aimed at enhancing the models' capability to comprehend and execute complex musical edits that require a nuanced understanding of music theory and composition.

6.3 Fine-Tuning Strategy

The fine-tuning strategy was twofold:

- For the Small model, with its lower parameter count, the goal was to achieve rapid adaptation without overfitting to the new data. This required a careful balance of learning rate, epochs, and batch sizes to ensure effective learning while preserving the model's generalization abilities.
- The Large model, with its extensive parameter set, was geared towards deep learning, allowing the model to internalize more complex patterns and relationships within the music data. The challenge here was to manage the computational demands and prevent the model from becoming too specialized to the fine-tuning dataset.

6.4 Fine-Tuning Parameters and Execution

The fine-tuning process for the Small and Large models involved several iterations of training with adjusted hyperparameters. For the Small model, the training was executed with a higher batch size over fewer epochs to promote broader learning. Conversely, for the Large model, a larger number of epochs with a smaller batch size was employed to encourage deeper understanding, albeit with a significant increase in computational resources and time.

6.5 Outcomes of the Fine-Tuning Process

The fine-tuning process resulted in both the Small and Large MusicGen models achieving greater proficiency in music modification tasks. Specifically, the models demonstrated improved capability in generating music that adheres more closely to user prompts while maintaining the essence of the original input music. This enhancement is pivotal in transitioning MusicGen from a purely generative model to an editing-capable model, fostering more creative and controlled music production.

In conclusion, the fine-tuning process was a critical phase in our project, enabling the MusicGen models to better meet the challenges of music editing. By leveraging the DoRA architecture and carefully calibrated training strategies, we have significantly advanced the state-of-the-art in AI-driven music generation and editing.

7 Evaluation and Results

As we progressed with our project, we evaluated the performance of the original MusicGen model and the fine-tuned models. Through this comprehensive analysis, we identified key areas for enhancement and conducted a series of evaluations to determine the efficacy of our fine-tuning process.

7.1 Addressing Inherited Challenges

We identified three major challenges within the original MusicGen model and addressed them through fine-tuning:

1. **Randomness and Rhythm Consistency**: We introduced additional regularization techniques to impose structure in the model's randomness, thereby improving rhythm consistency.
2. **Enhanced Music Understanding**: Leveraging the DoRA structure, we trained the model to better interpret complex musical styles, enabling it to generate more genre-specific and stylistically accurate compositions.
3. **Broadening Functional Capabilities**: We expanded the model's capabilities to include better interpretation of musical elements and instructions, particularly for handling nuanced feedback and negative prompts.

7.2 Evaluation Metrics

We employed a combination of objective metrics to assess the final music generation model's performance compared to the original MusicGen model.

Objective Metric - FAD[29]:

- **Fréchet Audio Distance (FAD)**: Measures the closeness of the distribution of generated audio to that of real audio.

Innovative Metric - PAM[30]:

- **Prompting Audio-Language Models (PAM)**: A novel metric that assesses audio quality without a reference, leveraging contrasting quality descriptors and audio-text paired data.

Metric Implementation:

$$\text{FAD} = ||\boldsymbol{\mu}_{\text{generated}} - \boldsymbol{\mu}_{\text{real}}||^2 + \text{Tr}\left(\boldsymbol{\Sigma}_{\text{generated}} + \boldsymbol{\Sigma}_{\text{real}} - 2\left(\boldsymbol{\Sigma}_{\text{generated}}\boldsymbol{\Sigma}_{\text{real}}\right)^{\frac{1}{2}}\right), \tag{4}$$

$$\text{PAM} = \frac{\exp(\text{similarity\_score}_{\text{high\_quality}})}{\exp(\text{similarity\_score}_{\text{high\_quality}}) + \exp(\text{similarity\_score}_{\text{low\_quality}})}. \tag{5}$$

7.3 Discussion of Achieved Enhancements

To evaluate the fine-tuned models, a hold-out dataset featuring diverse genres of music was introduced.

For assessing the retention of information from the ground truth, we employed the Frechet Audio Distance (FAD). Notable enhancements were observed in both the Small and Large fine-tuned models when compared to the baseline MusicGen models. This improvement suggests that fine-tuning has enhanced the similarity between the generated music and the ground truth, thereby proving the improved retention capabilities of the models.

To evaluate the quality of the generated music, we applied the Perceptual Audio Metric (PAM) to compare it with the ground truth. The results indicated that the music generated by the fine-tuned Small model exhibited a slight improvement, whereas the PAM score of music generated by the fine-tuned Large MusicGen models decreased relative to the ground truth. We hypothesize that this decline is due to constraints related to the quality of the training dataset, as well as the relatively small size of the dataset used for fine-tuning the Large model. These limitations are expected to be

addressed in the next phase of the project, which will incorporate SIMBOLIC-type music input and allocate more substantial time and computing resources for training on a larger dataset.

The following comparative form illustrates the differences across various iterations of the MusicGen models:

Model	**FAD** $\downarrow$	**PAM Score** $\uparrow$
Original MusicGen (Small)	3.645	0.776
Fine-tuned MusicGen (Small)	3.643	0.797
Original MusicGen (Large)	3.638	0.797
Fine-tuned MusicGen (Large)	3.606	0.784
Ground Truth	-	0.834

Table 2: Comparative performance of Original and Fine-tuned MusicGen models. **FAD** is calculated by comparing generated music with ground truth; inherently, there is no ground truth FAD score. Lower FAD score indicates closer output to the input (indicated by $\downarrow$). **PAM** is calculated individually from the music generated by each model, hence original music forms the **Ground Truth** (0.834). Higher PAM score signifies less noisy music. (indicated by $\uparrow$).

7.4 Metric Justification

The selected metrics enabled us to quantitatively measure the generated music from various technical perspectives. The improvements in the Fine-tuned models were evident from the lower FAD scores, indicating a closer resemblance to real audio distributions. The PAM scores also reflected an improvement in the perceived quality of audio generation, underscoring the effectiveness of our fine-tuning efforts.

In conclusion, the fine-tuning process has successfully mitigated the initial challenges identified in the MusicGen model, resulting in a refined system capable of producing high-quality music that aligns closely with users' creative intentions.

8 Future Directions

Building upon the foundation established by the WAME.AI project, we have charted out several avenues for future exploration. These next steps are designed to address the remaining challenges and to further the capabilities of AI in music editing and generation.

8.1 Technological Advancements

Exploration of cutting-edge models and architectures will remain at the forefront of our research. We aim to experiment with the HuggingFace Transformers fine-tuning API, which promises to bring improvements in model performance and accessibility.

8.2 Symbolic Music Research

Given the lower computational requirements and the potential for higher-quality outputs, a focus on symbolic music will allow for more granular control over the music generation process, providing a pathway to breakthroughs in AI music comprehension and creation.

8.3 Data and Instrument Separation

Enhancements in data quality and labeling will be pursued to refine the model's understanding of complex musical compositions. Research will be directed towards improved instrument separation techniques to support more nuanced music editing tasks.

8.4 Integration with Large Language Models

The fusion of our music generation model with LLMs will open up possibilities for more sophisticated conversational AI interactions, leading to a music editing experience that is both intuitive and powerful.

By navigating these future directions, we anticipate significant contributions to the domains of AI, music generation, and creative expression. The exciting journey of the WAME.AI project continues as an ongoing endeavor in the world of AI research.

9 Conclusion

The journey of fine-tuning the MusicGen models has been both challenging and enlightening. We embarked on this project with the vision to push the boundaries of AI in music generation, aiming to create a system that not only produces novel musical compositions but also maintains the essence of user-input prompts. Through rigorous evaluation and methodical fine-tuning, we have made significant strides towards achieving this goal.

Our efforts have addressed the initial shortcomings of the baseline MusicGen model, particularly in terms of randomness, rhythm consistency, and the depth of music understanding. The fine-tuned models now demonstrate a marked improvement in generating compositions that are harmonically and stylistically aligned with the user's intentions, as evidenced by the objective metrics—FAD, KL Divergence, Chroma Cosine Similarity, and PAM.

As we look to the future, the prospects for further advancements are promising. With plans to explore new data sources, integrate with real-time music editing tools, and employ unsupervised learning techniques, the path ahead is ripe with potential. The collaboration with musicians and artists will continue to be an invaluable aspect of our work, ensuring that the developed tools are not only technologically advanced but also musically inspiring and intuitive to use.

In conclusion, the project has laid a solid foundation for the next generation of AI-powered music generation tools. We believe that the advancements made will not only serve the music industry but also contribute to the broader field of creative AI, encouraging others to explore the fusion of technology and artistry in new and innovative ways.

References

[1] Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. Simple and controllable music generation. *arXiv:2306.05284*, 2023.

[2] OpenAI. ChatGPT: Openai's language model. `https://www.openai.com/research/chatgpt`, 2024. Feb 18.

[3] Trevor J. Pinch and Frank Trocco. *Analog Days: The Invention and Impact of the Moog Synthesizer*. Harvard University Press, 2002.

[4] Lejaren Arthur Hiller and Leonard M Isaacson. *Experimental Music: Composition with an electronic computer*. McGraw-Hill Book Co, 1959.

[5] Elliot Waite. Generating long-term structure in songs and stories. `https://magenta.tensorflow.org/2016/07/15/lookback-rnn-attention-rnn`, 2016.

[6] Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. *arXiv:1609.03499*, 2016.

[7] 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. *arXiv:2308.01546*, 2023.

[8] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. LLaMA: Open and efficient foundation language models. *arXiv:2302.13971*, 2023.

[9] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. *arXiv:2102.12092*, 2021.

[10] Shulei Ji, Jing Luo, and Xinyu Yang. A comprehensive survey on deep music generation: Multi-level representations, algorithms, evaluations, and future directions. *arXiv:2011.06801*, 2011.

[11] Shulei Ji, Jing Luo, and Xinyu Yang. A survey on deep learning for symbolic music generation: Representations, algorithms, evaluations, and challenges, 2023.

[12] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning representations by back-propagating errors. *Nature volume 323, pages533–536 (1986)*, 1986.

[13] J.L. Elman. Finding structure in time. *Cognitive Science Volume 14, Issue 2, April–June 1990, Pages 179-211*, 1990.

[14] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation (1997) 9 (8): 1735–1780*, 1997.

[15] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks. *arXiv:1409.3215*, 2014.

[16] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017.

[17] Chih-Pin Tan, Alvin W.Y. Su, and Yi-Hsuan Yang. Melody infilling with user-provided structural context. *arXiv:2210.02829*, 2022.

[18] Jiang HUANG, Xianglin HUANG, Lifang YANG, and Zhulin TAO. Dance-conditioned artistic music generation by creative-gan. *2023EAP1059*, 2023.

[19] Apoorv Vyas, Bowen Shi, Matthew Le, Andros Tjandra, Yi-Chiao Wu, Baishan Guo, Jiemin Zhang, Xinyue Zhang, Robert Adkins, William Ngan, Jeff Wang, Ivan Cruz, Bapi Akula, Akinniyi Akinyemi, Brian Ellis, Rashel Moritz, Yael Yungster, Alice Rakotoarison, Liang Tan, Chris Summers, Carleigh Wood, Joshua Lane, Mary Williamson, and Wei-Ning Hsu. Audiobox: Unified audio generation with natural language prompts. *arXiv:2312.15821*, 2023.

[20] Qingqing Huang, Daniel S. Park, Tao Wang, Timo I. Denk, Andy Ly, Nanxin Chen, Zhengdong Zhang, Zhishuai Zhang, Jiahui Yu, Christian Frank, Jesse Engel, Quoc V. Le, William Chan, Zhifeng Chen, and Wei Han. Noise2music: Text-conditioned music generation with diffusion models. *arXiv:2302.03917*, 2023.

[21] Gautam Mittal, Jesse Engel, Curtis Hawthorne, and Ian Simon. Symbolic music generation with diffusion models. *arXiv:2103.16091*, 2021.

[22] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *arXiv:2006.11239*, 2020.

[23] Xubo Liu, Qiuqiang Kong, Yan Zhao, Haohe Liu, Yi Yuan, Yuzhuo Liu, Rui Xia, Yuxuan Wang, Mark D. Plumbley, and Wenwu Wang. Separate anything you describe. *arXiv:2308.05037*, 2023.

[24] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763, 2021.

[25] Y. Wu, K. Chen, T. Zhang, Y. Hui, T. Berg-Kirkpatrick, and S. 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)*, pages 1–5, 2023.

[26] Q. Kong, Y. Cao, H. Liu, K. Choi, and Y. Wang. Decoupling magnitude and phase estimation with deep resunet for music source separation. In *ISMIR*, 2021.

[27] Q. Kong, K. Chen, H. Liu, X. Du, T. Berg-Kirkpatrick, S. Dubnov, and M. D. Plumbley. Universal source separation with weakly labelled data. arXiv:2305.07447, 2023.

[28] Mike Senior. Cambridge music technology. `https://cambridge-mt.com/ms/mtk/`, 2024. Accessed: March 20, 2024.

[29] Kevin Kilgour, Mauricio Zuluaga, Dominik Roblek, and Matthew Sharifi. Frechet audio distance: A reference-free metric for evaluating music generation. *arXiv:1812.08466*, 2018.

[30] Soham Deshmukh, Dareen Alharthi, Benjamin Elizalde, Hannes Gamper, Mahmoud Al Ismail, Rita Singh, Bhiksha Raj, and Huaming Wang. Pam: Prompting audio-language models for audio quality assessment. *arXiv:2402.00282*, 2024.