# Enhancing Instrument-Level Controllability in Text-to-Music Generation for Professional Songwriting Applications

Dongseok Heo
2023-24137
Interdisciplinary Program in AI
ty8900@snu.ac.kr

Hanwool Sul
2023-37098
Interdisciplinary Program in AI
kangsul@snu.ac.kr

Sangmin Kim
2020-11769
Dept. of Computer Science and Engineering
ksme456@snu.ac.kr

Jungmin Seo
2020-12385
Dept. of Electrical and Computer Engineering
jmseo1204@snu.ac.kr

Sooyoung Ryu
2020-11769
Dept. of Fashion and Textiles
ryuswimming@snu.ac.kr

## Abstract

*Recent advancements in deep learning have transformed the field of audio generation, enabling the creation of highly realistic and contextually rich soundscapes. While AI-driven audio systems have demonstrated impressive capabilities in applications such as text-to-speech and music composition, challenges remain in bridging the gap between cutting-edge research and practical adoption in professional workflows. Current music generation approaches excel in creating complex compositions but lack fine-grained control mechanisms, limiting their usability in professional settings. Unlike image generation, which benefits from tools like inpainting and layer editing, music generation still struggles to isolate or manipulate individual instrumental components.*

*To address these limitations, this study introduces a novel framework for music generation that emphasizes adaptability and control. The proposed approach generates diverse instrumental tracks from input melodies by leveraging advanced conditioning methods, such as latent input-based and attention-based mechanisms. Additionally, the integration of Low-Rank Adaptation (LoRA) enables efficient and modular fine-tuning, allowing for flexibility and precision in creating instrument-specific tracks while maintaining coherence in multi-layered compositions.*

*By aligning with existing industry practices, this work bridges the gap between state-of-the-art generative models and practical applications in music and audio production. The proposed framework lays the foundation for developing systems that seamlessly integrate into creative workflows, empowering professionals to fully exploit the potential of AI-driven audio generation.*

## 1. Introduction

Recent advancements in deep learning have revolutionized the field of audio generation, enabling the creation of highly realistic and contextually rich soundscapes. From text-to-speech (TTS) synthesis to music generation, these technologies have unlocked new possibilities for creative and practical applications. Notable models such as Tacotron[18] for speech generation and MusicLM[5] for music composition highlight the growing potential of AI-driven audio systems. However, significant challenges remain in bridging the gap between state-of-the-art research and practical adoption in professional workflows.

The performance of audio generation models is heavily influenced by their audio representations and architectures. Commonly used representations include raw waveforms, which preserve high fidelity at the cost of computational complexity, and mel-spectrograms, which compress frequency information based on human auditory perception. Mel-spectrograms have gained popularity for their efficiency and adaptability in applications such as TTS and music synthesis. On the architectural front, models like au-

toencoders, Generative Adversarial Networks (GANs), normalizing flows, transformers, and diffusion models have played pivotal roles in advancing the field. Among these, transformers excel at modeling long-term dependencies, while diffusion models leverage iterative noise reduction to generate high-quality audio. These architectures have enabled breakthroughs in generating waveforms directly or synthesizing intermediate representations for various audio applications.

The field of music generation has also witnessed remarkable progress in recent years, closely tied to advancements in latent diffusion models. Commercial services such as Udio[3] and Suno[2] have demonstrated impressive capabilities, even achieving recognition in composition competitions. Despite these achievements, significant challenges remain for professionals seeking to incorporate such technologies into their workflows.

Although these advances are promising, significant limitations persist. Models like Jukebox[7] and AudioLDM2[14] excel at creating complex compositions but lack tools for fine-grained control. Unlike image generation, where methods like inpainting and layer editing allow targeted modifications, current music generation models are unable to isolate or manipulate specific elements, such as individual instruments, limiting their usability for professionals.

To address these challenges, this study proposes a novel framework to enhance control and adaptability in music generation. Specifically, the framework focuses on generating diverse instrumental tracks from input melodies, leveraging advanced conditioning methods such as latent input-based and attention-based mechanisms. Additionally, the integration of Low-Rank Adaptation (LoRA)[10] facilitates efficient and modular fine-tuning, enabling flexibility and precision in professional workflows.

By aligning with existing industry practices, this work aims to bridge the gap between cutting-edge generative models and practical applications in music and audio production. This research not only introduces innovative methodologies, but also lays the groundwork for developing systems that seamlessly integrate into creative workflows, empowering professionals to fully harness the potential of AI in audio generation. The results of our experiments can be accessed at the followed:
Style transfer results: https://url.kr/lct3mp
Autoregressive music generatioin: https://url.kr/zlljjh

## 2. Related works

### 2.1. Music Generation Models

Deep generative models have significantly advanced music generation, enabling both symbolic and waveform-based synthesis. Among the most widely used architectures are Variational Autoencoders (VAEs), Generative Adversarial Networks (GANs), Transformers, and Diffusion Models, each with distinct methodological advantages. VAEs provide a probabilistic framework that enables structured latent spaces, facilitating the generation of coherent musical patterns. GANs leverage adversarial training to generate realistic outputs by optimizing a generator and discriminator in competition. Transformers excel at capturing long-term dependencies in sequential data, making them well-suited for modeling complex temporal relationships in music. Diffusion models, on the other hand, iteratively refine noise into high-quality audio through forward and reverse diffusion processes, demonstrating robustness in generating diverse and high-fidelity outputs.

Despite these advancements, existing music generation systems face critical limitations in fine-grained control over individual elements within a composition. Current approaches excel at producing complete pieces but lack the ability to isolate or manipulate specific components, such as individual instrument tracks or harmonic layers. This limitation reduces their practical utility for professional musicians and sound designers, who require precise control for creative workflows. Our work builds on these existing methods by introducing a conditional framework that allows targeted generation of instrumental sounds based on input melodies. This approach addresses the need for controllable and adaptable music generation.

### 2.2. Latent Diffusion Models

Diffusion models have recently gained prominence as a robust generative approach for audio synthesis, including music generation. Inspired by non-equilibrium thermodynamics, diffusion models consist of two processes: a forward diffusion process, which gradually adds noise to the data until it becomes a standard Gaussian distribution, and a reverse process, which reconstructs the data iteratively by learning to denoise it. This method allows for the generation of high-quality and diverse outputs, making it particularly suited for complex tasks like waveform and symbolic music synthesis.

Latent diffusion models (LDMs) further optimize this approach by operating within a compressed latent space rather than directly on high-dimensional data such as raw waveforms. By leveraging encoder-decoder architectures, LDMs reduce computational overhead while maintaining high-fidelity outputs. In music generation, this enables efficient modeling of intricate patterns and long-term dependencies, such as harmonic progressions and instrument arrangements. Methods like Jukebox[7] and AudioLDM2[14] demonstrate the capability of diffusion-based techniques to generate high-quality music conditioned on specific inputs, such as melodies or textual descriptions.

The adoption of latent diffusion models addresses some of the key challenges in music generation, including scalability and quality. However, fine-grained control over individual elements within compositions remains limited. In this study, our aim is to build upon these advancements by integrating conditioning mechanisms and modular adaptation techniques, enabling precise control over instrumental sounds and enhancing the practical applicability of diffusion models in music production workflows.

### 2.3. Low Rank Adaptation

Low-Rank Adaptation (LoRA)[10] has emerged as an efficient fine-tuning technique for large-scale models, significantly reducing the number of trainable parameters by introducing trainable low-rank matrices into each layer of a pretrained model. Originally proposed for large language models , LoRA has been effectively applied to image generation tasks, particularly in fine-tuning diffusion models for text-to-image synthesis. By freezing the original model weights and optimizing only the low-rank matrices, LoRA enables rapid adaptation to new tasks with minimal computational resources.

In the context of image generation, LoRA has been utilized to personalize models for specific visual concepts or styles without the need for extensive retraining. For instance, integrating multiple LoRA modules allows for the composition of various visual elements, enhancing the model's ability to generate complex images that align with user prompts. Additionally, methods like LoRA Fusion apply attention mechanisms to effectively merge multiple LoRA modules, capturing user intent more accurately in generated images. These applications demonstrate LoRA's versatility in efficiently adapting large models to diverse image generation tasks.

Building upon these advancements, our work aims to incorporate LoRA into music generation models, facilitating efficient fine-tuning and enabling precise control over individual musical elements. By leveraging LoRA's parameter-efficient adaptation, we seek to enhance the flexibility and scalability of music generation systems, making them more practical for professional use.

### 2.4. Audio Captioning

Recent advancements in deep learning have enabled audio captioning systems to generate coherent textual descriptions for audio inputs. Encoder-decoder architectures, particularly those leveraging transformers, have shown effectiveness in modeling the temporal and semantic aspects of audio signals to produce descriptive captions. For instance, methods like the Audio Captioning Transformer (ACT)[15] utilize transformer networks to extract global features from audio inputs, improving caption fluency and relevance. Furthermore, datasets like AudioCaps[11] and Clotho[9] provide large collections of audio-only data paired with human-written captions, facilitating model training and evaluation.

Audio captioning plays a critical role in tasks such as text-to-audio generation, where the generated captions can serve as additional annotations to augment data. By converting audio-only datasets into captioned datasets, models can access a richer set of semantic information, enhancing training efficiency and generalization for text-to-audio tasks. This approach enables the creation of more robust systems capable of generating contextually aligned sounds conditioned on textual prompts. In our work, we leverage audio captioning to enhance existing audio datasets, enabling the augmentation of textual descriptions for downstream text-to-audio generation tasks. This strategy bridges the gap between audio understanding and generation while maximizing the utility of audio-only datasets in resource-constrained environments.

## 3. Methods

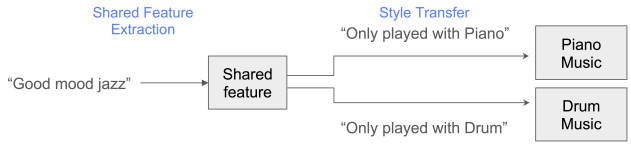


Figure 1. Shared Feature Extraction and Style Transfer process for multi-layered music generation. The shared feature, extracted from a text prompt (e.g., 'Good mood jazz'), ensures musical consistency. This feature is then transformed into instrument-specific tracks (e.g., Piano Music, Drum Music) using style transfer guided by instrument-specific prompts.

To generate multi-layered music using LoRA, we adopted a two-step approach to ensure both musical consistency and instrument-specific transformations. First, we generate a "shared feature" 2, which serves as a foundation to maintain the overall structure and coherence of the music. This shared feature can be derived from a text prompt or reference music and represents a generalized musical concept.

Next, each shared feature is transformed into single-instrument tracks using LoRA fine-tuning, combined with control prompts. This step requires precise adjustments to reflect the unique characteristics of each instrument while preserving the essence of the shared feature. To achieve this, we employed style transfer-like methods, which integrate instrumental specificity without disrupting the overall musical integrity.

Before finalizing the multi-layered architecture, we tested two distinct style transfer approaches: Language of Audio (LoA) and Reference Latent Guidance. These methods were evaluated for their ability to effectively integrate

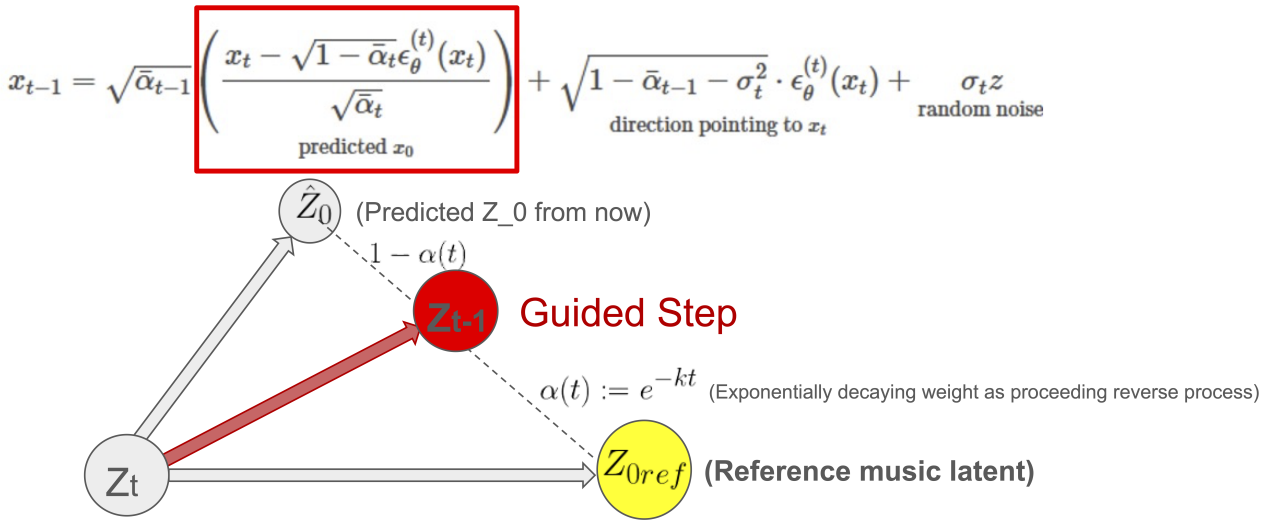


Figure 2. Illustration of Reference Latent Guidance in the DDIM denoising process. The method integrates the reference music's latent vector ($Z_{0(ref)}$) into each denoising step with an exponentially decaying weight ($\alpha(t)$). This guided step ensures that the generated latent ($Z_{t-1}$) aligns closely with the reference music's low-frequency characteristics while adapting the high-frequency details to the target instrument.

individual instrument characteristics while maintaining coherence across the multi-layered composition. By exploring these approaches, we ensured that each instrument's identity was captured without sacrificing the consistency and originality of the overall piece.

### 3.1. Dataset

we utilized an instrument-music dataset[1], with a specific focus on Guitar, Drum, and Piano recordings. The dataset contains diverse audio samples that serve as a foundation for evaluating and improving our proposed framework. Due to the presence of incorrect labeling, Violin samples were excluded from the experiments to ensure data quality and reliability. For each selected instrument, we utilized 330 samples for training and 20 samples for testing, maintaining a consistent and balanced dataset for robust evaluation.

To further enhance the dataset, we employed audio captioning to generate textual descriptions for each audio sample. Inspired by approaches such as QA-MDT[13], the captioning process incorporated three strategies:

1. Utilizing original dataset captions where available.
2. Generating captions using pre-trained caption generation models[8] to describe audio samples with greater detail and accuracy.
3. Refining captions using LLM-based[4] prompts to modify or enhance specific attributes, such as instrument type or mood descriptors, to align more closely with the target application.

The generated captions were evaluated for their semantic alignment with audio features using CLAP scores (Contrastive Language-Audio Pretraining) to ensure consistency and quality. This rigorous approach to caption generation provided detailed and contextually relevant text labels, which were used to improve the performance of LoRA-based training for instrument-level controllability. By integrating a curated selection of audio samples with high-quality textual descriptions, this dataset facilitated a deeper exploration of conditioned music generation. This enhanced dataset framework bridges textual and auditory modalities, supporting the development and evaluation of advanced generative models while maintaining a focus on data integrity and reproducibility.

### 3.2. Language of Audio(LoA)

The first style transfer approach leverages the Language of Audio (LoA) framework, directly utilizing the capabilities of the AudioLDM2[14]. This method integrates reference music and textual prompts into a unified embedding space, which serves as a guiding condition vector for the diffusion process. By combining textual and auditory representations, this approach ensures both structural coherence and stylistic alignment in the generated music.

The process begins with the conversion of text prompts (e.g., "Good mood jazz") into semantic embeddings using FLAN-T5 [6]. This model encodes the textual input into a rich contextual representation, capturing the intent and nuances of the prompt. Simultaneously, reference music is processed through CLAP(Contrastive Language Audio Pre-training), transforming its auditory features into an embedding that encapsulates its musical characteristics. These two embeddings, one representing the textual intent and the

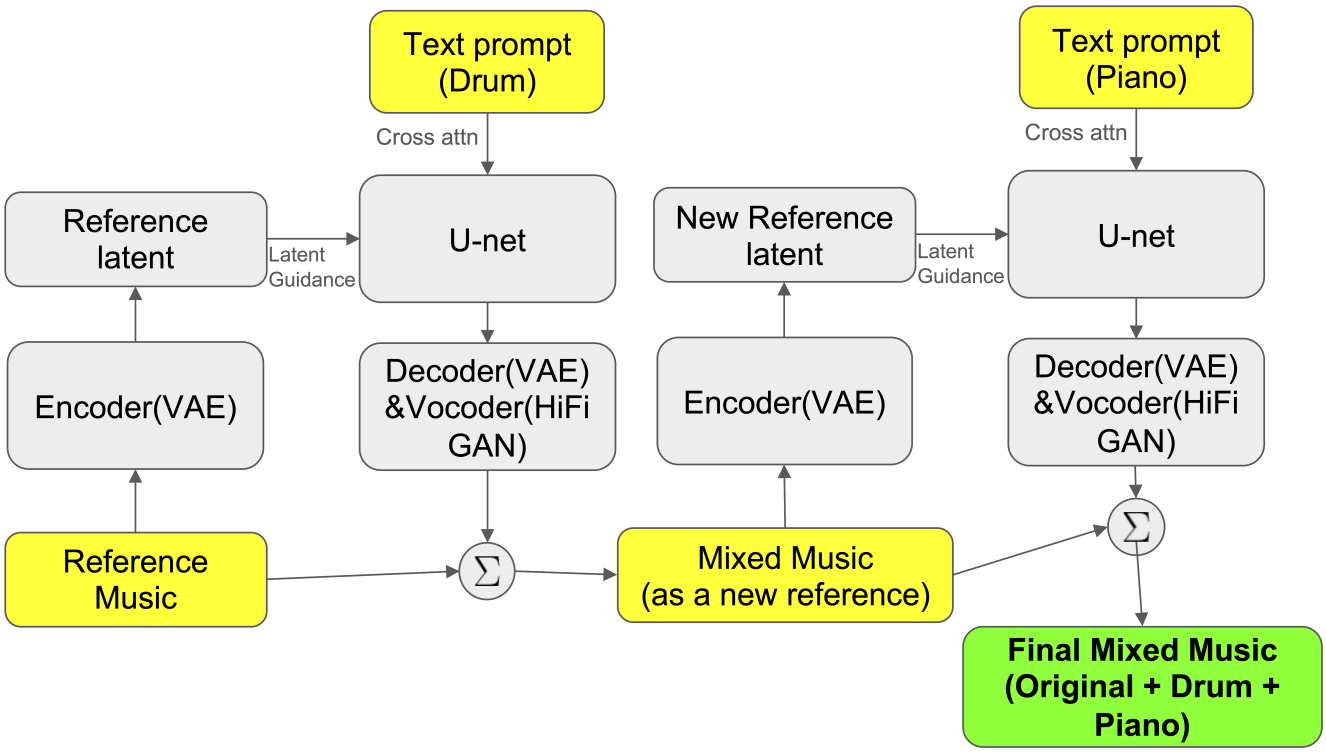


Figure 3. Autoregressive Music Generation process. The reference music is encoded into a latent space using a VAE encoder and guided by latent vectors in a step-wise manner. The latent is combined with text prompts (e.g., 'Drum' or 'Piano') using cross-attention within the U-net to generate instrument-specific tracks. These tracks are decoded and vocoded using HiFi-GAN [12], then combined into a mixed composition. The mixed composition is iteratively used as a new reference to add additional layers, resulting in the final output: a cohesive piece that integrates the original music with drum and piano loops.

other reflecting the auditory features of the reference music, are then aligned into a shared representation space.

To achieve this alignment, the embeddings are passed through an intermediate projection model, which harmonizes the representations into a unified vector space. Following this, GPT-2 [16] is employed to compose the aligned embeddings into a cohesive Language of Audio (LoA) vector. This LoA vector serves as the Unet condition vector within the AudioLDM2 diffusion model, guiding the generation process to reflect both the structural integrity of the shared feature and the stylistic attributes of the reference music.

This approach ensures that the generated music aligns closely with the desired instrument characteristics while maintaining the essence of the original input. By integrating textual and auditory data into a unified framework, the LoA-based method provides a robust and flexible foundation for achieving precise and consistent style transfer in multi-layered music generation.

### 3.3. Reference Latent Guidance

The second style transfer approach, Reference Latent Guidance, integrates latent features of reference music into the generation process to achieve precise style transfer while preserving the core auditory structure of the original sample. This method, inspired by Classifier-Free Guidance, modifies the diffusion process to incorporate the reference music's characteristics at every step.

The process begins by encoding the reference music into the latent space using a VAE encoder. The encoded latent vector captures the semantic and auditory features of the reference music in a compact representation. During the DDIM [17] denoising process, this latent vector is injected at each step to guide the generation process toward the reference music.

To balance the influence of the reference music on the generated output, an exponentially decaying coefficient is applied to the latent vector during the denoising steps. This ensures that the low-frequency domain of the reference music, which defines its structure and tonal foundation, is strongly preserved. Meanwhile, the high-frequency domain, which contributes to the finer details, is adapted to

emphasize the unique characteristics of the target instrument.

By dynamically adjusting the weight of the reference latent vector, this approach achieves a balance between preserving the original music and transforming it to align with the desired instrument. As a result, the generated music retains the core essence of the reference music while successfully incorporating the specific instrumental style.

The Reference Latent Guidance method offers a direct and effective way to perform style transfer, particularly in scenarios where maintaining the structural integrity of the original music is essential. Its flexibility and precision make it a robust alternative to embedding-based approaches, enabling high-quality and instrument-specific transformations.

### 3.4. Autoregressive Music Generation

Building on the style transfer methods, we implemented a novel comprehensive autoregressive music generation framework to create multi-layered compositions. This approach ensures that each layer of the generated music is integrated consistently while preserving the structural and stylistic integrity of the original reference.

For example, the process begins by transforming the reference music into a drum track using the Reference Latent Guidance method. This step ensures that the generated drum track aligns with the tonal and rhythmic structure of the reference music while introducing the desired percussive elements.

Next, the original reference music and the generated drum track are combined into a mixed composition. This mixed composition serves as the new reference, maintaining the overall consistency while allowing for the iterative addition of new layers.

The process is repeated iteratively, with the mixed composition being transformed into additional instrument tracks, such as a piano loop. Each new instrument layer is added sequentially, ensuring that the previously generated layers and the original reference are cohesively integrated.

The final output consists of the original reference music enriched with multiple layers, including drum and piano loops. By adopting an autoregressive approach, the framework maintains coherence across all layers, ensuring that the generated music is both structurally consistent and stylistically aligned with the intended instrument-specific transformations. This iterative process highlights the flexibility and robustness of the proposed methods for generating complex, multi-layered musical compositions.

## 4. Experiments

### 4.1. Evaluation

Our evaluations focus on two essential aspects: originality preservation and instrumental characteristic retention. These metrics are designed to assess the fidelity of the generated music in maintaining the essence of the original composition while applying the desired instrumental style.

Originality Preservation: To verify that the generated music retains the creative essence of the original composition, we calculate the CLAP score between the original music and the generated output. This score measures the semantic similarity between the two audio representations. A higher CLAP score indicates better preservation of the original music's structure and intent during the style transfer process.

Instrumental Characteristic Retention: To ensure the generated music reflects the characteristics of the target instrument, we compare the CLAP score between the generated music and the mean embedding of single-instrument music from the dataset. This evaluation provides a quantitative measure of how well the generated music aligns with the specific traits of the intended instrument.

Comprehensive Assessment By analyzing these two aspects, we evaluate the dual objectives of our approach: First, maintaining fidelity to the original composition, ensuring the creative intent is not lost during the transformation. Second, accurately embedding the desired instrumental style, confirming the generated music exhibits the characteristics of the target instrument. This dual evaluation framework ensures that our models achieve both originality preservation and instrumental style alignment, balancing creative integrity with the adaptability required for music generation tasks.

### 4.2. Results

Figure 4 presents the results of our evaluation on originality preservation. Our findings demonstrate that the LOA method preserves the originality of the music more effectively compared to the Guidance method. This indicates that LOA is better at retaining the core attributes and creative essence of the original music during the style transfer process.

Interestingly, we observed that models without LoRA exhibited greater preservation capabilities than those with LoRA. While LoRA enhances flexibility and adaptability in style transfer, this result suggests that it may come at the cost of some loss in originality.

When examining the impact of different instruments during the style transfer process, the piano showed the highest performance in preserving the original music, followed by drums and guitars. This superior performance of the piano could be attributed to its broader tonal range and its role as

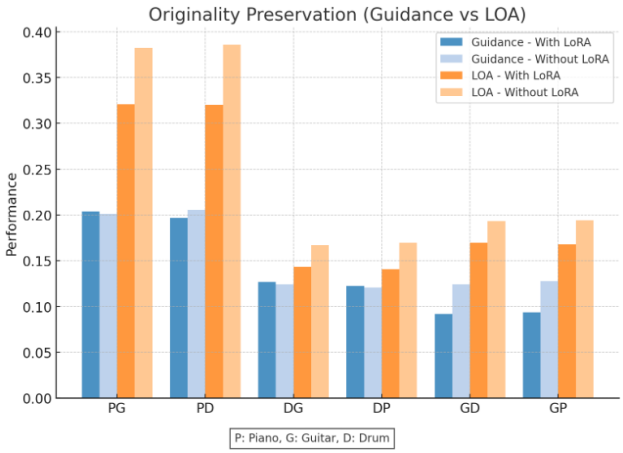


Figure 4. Comparison of originality preservation performance between LOA and Guidance methods, with and without LoRA, across three instruments (Piano, Guitar, and Drum). The LOA method generally outperforms the Guidance method in preserving originality. Interestingly, models without LoRA show better preservation compared to those with LoRA. Among the instruments, piano exhibits the highest preservation performance, followed by drums and guitars.

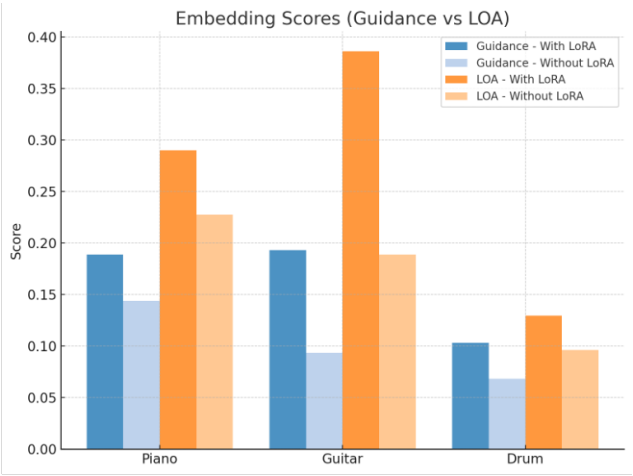


Figure 5. Comparison of embedding scores between LOA and Guidance methods, with and without LoRA, across three instruments (Piano, Guitar, and Drum). The LOA method consistently outperformed the Guidance method, achieving higher scores. Models with LoRA applied demonstrated improved alignment with the desired instrument. Among the instruments, Piano and Guitar achieved the highest performance, while Drums exhibited slightly lower scores.

a foundational instrument in many musical compositions.

Figure 5 displays the results of our evaluation on instrumental characteristic retention. In comparisons with instrument embeddings, the LOA method consistently achieved higher scores than the Guidance method. This result highlights LOA's effectiveness in accurately reflecting the desired instrumental characteristics during the transformation process.

Furthermore, models with LoRA applied demonstrated a higher degree of alignment with the target instrument, suggesting that LoRA enhances the model's ability to adapt to specific instrumental styles. This finding underscores LoRA's value in achieving precise stylistic transformations.

Finally, when analyzing performance across different instruments, we observed that Piano and Guitar outperformed drums, generating music that more closely matched the intended instrument embeddings. In contrast, Drums exhibited slightly lower performance, which could be attributed to their rhythm-focused nature, making them more challenging to encapsulate within the embedding framework.

In summary, the LOA method, particularly when combined with LoRA, demonstrates superior alignment with instrumental characteristics, while balancing the trade-offs in originality preservation remains an area for further improvement.

## 5. Conclusion

In this study, we explored methods to improve instrument-level controllability in autoregressive music composition through LoA-based and Guidance-based style transfer techniques, while also evaluating the impact of LoRA integration. We experimented with four configurations: LoA, Guidance, LoA with LoRA, and Guidance with LoRA. Our results demonstrate that LoA-based methods outperformed Guidance in preserving the original music structure and enhancing instrument similarity. Furthermore, integrating LoRA into both approaches led to notable improvements, LoA with LoRA achieving the highest overall instrument-level similarity results.

Despite these advancements, challenges remain. The autoregressive generation process ensures consistency across multiple music layers but comes at the expense of longer generation times. As part of our future work, we aim to address these limitations by investigating diverse style-transfer techniques leveraging alternative foundation models. Additionally, optimizing the multi-layered autoregressive generation process will be a focus to reduce computation time while maintaining consistency and quality in the generated music.

## References

[1] Musical Instrument's Sound Dataset. https://www.kaggle.com/datasets/soumendraprasad/musical-instruments-sound-dataset, 2022. 4

[2] Suno: Make a song about anything. https://suno.com, 2024. 2

[3] Udio: Make your music, Create any song. Just describe it. https://udio.com, 2024. 2

[4] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 4

[5] 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. 1

[6] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53, 2024. 4

[7] Prafulla Dhariwal, Heewoo Jun, Christine Payne, Jong Wook Kim, Alec Radford, and Ilya Sutskever. Jukebox: A generative model for music. *arXiv preprint arXiv:2005.00341*, 2020. 2

[8] SeungHeon Doh, Keunwoo Choi, Jongpil Lee, and Juhan Nam. Lp-musiccaps: Llm-based pseudo music captioning. *arXiv preprint arXiv:2307.16372*, 2023. 4

[9] Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen. Clotho: An audio captioning dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 736–740. IEEE, 2020. 3

[10] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 2, 3

[11] Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. Audiocaps: Generating captions for audios in the wild. In *NAACL-HLT*, 2019. 3

[12] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *Advances in neural information processing systems*, 33:17022–17033, 2020. 5

[13] Chang Li, Ruoyu Wang, Lijuan Liu, Jun Du, Yixuan Sun, Zilu Guo, Zhenrong Zhang, and Yuan Jiang. Quality-aware masked diffusion transformer for enhanced music generation. *arXiv preprint arXiv:2405.15863*, 2024. 4

[14] 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. 2, 4

[15] Xinhao Mei, Xubo Liu, Qiushi Huang, Mark D Plumbley, and Wenwu Wang. Audio captioning transformer. *arXiv preprint arXiv:2107.09817*, 2021. 3

[16] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 5

[17] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 5

[18] Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, et al. Tacotron: Towards end-to-end speech synthesis. *arXiv preprint arXiv:1703.10135*, 2017. 1