Attention-Guided Music Generation with Variational Autoencoder and Latent Diffusion

Yuanxin Gan*

The Second High School Attached to Beijing Normal University, Beijing, China

*Corresponding author

Keywords: Music generation, Variational Autoencoder (VAE), Diffusion model, Attention regulation, Music restoration, Adaptive attention mechanism

Abstract: A pioneering two-phase music creation framework integrates Variational Autoencoders (VAE) and Conditional Diffusion Models (CDM). This model is designed to produce music segments akin to the initial input, while simultaneously fostering diversity and creativity via mechanisms that govern emotional expression and attention modulation. In the initial phase, the VAE processes the input music, distilling its fundamental attributes and the influence of attention into a compact, low-dimensional latent representation. This compression allows for a more efficient handling of the music's essence.The subsequent phase employs the diffusion model to fabricate novel music fragments through a sequential noise reduction procedure. This ensures that the synthesized music resonates with the original in terms of emotional tone, structural coherence, and attention management, yet introduces a balanced level of novelty and variety.Furthermore, the system includes an adaptive attention regulation component, which adjusts dynamically to the attentional dynamics within the music, enabling the generated compositions to preserve their structural soundness and musical integrity while profoundly impacting the listener's emotional engagement and focus.

1. Introduction

Music possesses an extraordinary capacity to stir emotions and captivate human attention, positioning it as a potent medium of expression and connection. To artists, music transcends mere performance; it becomes a reservoir of inspiration and motivation, driving creativity and innovation. The art of conveying and controlling emotion is central to this process. Composers, fueled by their personal sentiments, technical mastery, understanding of music theory, and insights into attention dynamics, meticulously craft and structure their compositions [1]. Their aim is to forge a meaningful link with the audience, drawing them into the auditory experience and engaging their focus.

Despite significant advancements in the field of next-generation music research, artificial intelligence continues to grapple with the intricacies of emotional resonance and the organic incorporation of the composer's intuition into the creative workflow. Presently, most neural network architectures rely on established musical content, neglecting the pivotal roles of emotion and attention in the production process[2][3]. This oversight constrains AI's potential to create music that fosters profound communicative bonds, orchestrates attention adeptly, and delivers authentic sensory experiences[4].

To address these limitations, our endeavor is centered around crafting a music model that harmonizes emotional nuances with attention management. By merging Variable Autoencoder (VAE) techniques with model dissemination strategies [5], the system is designed to replicate the essence of specific musical pieces while simultaneously evoking targeted emotional and attentional responses from listeners. This approach seeks to emulate the composer's creative journey more faithfully, enhancing the emotional depth and structural integrity of the music produced.

The significance of emotion and attention in music composition permeates every facet of the creative process, particularly in the conceptualization and training of deep learning algorithms.

  DOI: 10.25236/iwmecs.2024.006

While autoregressive models excel at encapsulating fundamental musical components, they often falter in providing clear, internalized guidance, complicating the interpretation of musical data. Conversely, adversarial networks adeptly discern aspects like pitch, rhythm, and texture [6], yet they struggle to encompass the dimensions associated with emotional and attentional dynamics, typically prioritizing analysis over musical finesse and compositional coherence [7].

In recent years, strides in natural language processing (NLP) have notably enhanced our understanding and creation of human language, primarily due to advancements in deep learning and large-scale pre-training models. While most NLP endeavors have traditionally concentrated on textual inputs, music, a rich and expressive 'language', deserves equal attention[8]. Much like words, music conveys emotions, directs attention, narrates stories, and exchanges ideas, all while adhering to its unique structure and rules. Our research aims to bridge the gap between textual and musical domains by leveraging NLP techniques to generate music from text prompts. This not only broadens the scope of NLP applications but also supports interdisciplinary exploration into the realms of language, music, and machine learning[9].

Similar to text generation, music creation poses a formidable task, necessitating solutions across multiple abstraction levels. Current audio generation models employ techniques such as recurrent neural networks, generation counterpoint networks, autoencoders, and transformers[10]. Lately, the diffusion model has demonstrated exceptional success in computer vision and has been adapted for speech synthesis tasks[11][12]. However, only a select few models are currently equipped for music generation. Moreover, this domain still confronts persistent challenges, including the limitation of generated sequence length, model inefficiency, lack of musical diversity, and difficulty in controlling the generation process with text prompts.

In our study, we propose an innovative two-phase generation model that merges a Variational Autoencoder (VAE) with a diffusion model to synthesize music akin to a specified piece.

2. Model structure

We have innovated a two-phase generative framework that merges Variational Autoencoders (VAE) with diffusion models to craft music that aligns with a specified input and modulates human attention accordingly. In the first step, the VAE processes the input music, converting it into a compact latent code. This code captures essential aspects of the music, especially its influence on attention. The encoder condenses the input, while the decoder reconstructs segments of the music, ensuring the latent space holds onto the key structure, style, and attributes that regulate attention in the original piece.

Next, the latent code from the VAE is used by a diffusion model to generate new music. This happens through a gradual denoising process. The result is music that reflects the input's style and attention effects, while also allowing room for creative variation. By carefully adding adaptive noise and applying specific constraints, the diffusion model aligns the new composition with the input's structural and attention-regulating features, all while introducing innovative elements. An optimized 1D U-Net architecture, coupled with cross-attention mechanisms, is employed to augment the quality and efficiency of music generation. By synergizing the latent space learning of the VAE with the progressive generation of the diffusion model, this methodology is well-suited for a multitude of music generation tasks, producing outputs that are both faithful to the input and creatively enriched.

2.1. Innovative VAE Music Encoder Architecture

An innovative music encoder structure has been devised to attain the objective of generating music that precisely reflects an input while steering attention. This encoder adeptly extracts multi-channel features from the input music, encompassing its complex dimensions and attention dynamics, guaranteeing that the synthesized output closely matches the original in aspects of structure, style, and attention adjustment.

Based on sophisticated representation learning methodologies, the encoder's design boosts its expressive capacity via multi-scale latent space representation and amalgamated attention control

systems. This facilitates more accurate and subtle music creation, aligning with the nuances of the input data and enhancing the user's auditory experience.

2.1.1. Multi-Channel Music Feature Extraction and Representation

The input music is first transformed into a multi-channel feature representation matrix using Short-Time Fourier Transform (STFT). This matrix not only includes traditional spectral information but also captures key structural features of the music and its influence on attention, such as harmony and rhythm. Specifically, the feature matrix of the input music is denoted as $\mathbf{X}$, with dimensions $[C, F, T]$, where C represents the number of feature channels, and F and T correspond to the frequency and time dimensions, respectively.

Compared to traditional single-spectrum representation methods, our design extracts features across multiple channels to capture the complex patterns in music and their effects on attention regulation. These features are further enhanced through a cross-channel attention mechanism, ensuring effective integration and complementarity of information across channels. Ultimately, these enriched multi-channel features provide more detailed input for encoding into the latent space.

2.1.2. Adaptive Attention Flow Regulation

To capture and preserve the attention-related features of the input music, we introduce an adaptive attention flow regulation module within the encoder. This module analyzes the attention impact curve of the input music (e.g., arousal and focus levels) to generate attention feature vectors, which encapsulates the core attention information of the input music.

The design of the attention regulation module allows the model to adaptively adjust according to the dynamic changes in attention flow, ensuring that the latent space reflects the music's structure and its attention-related features.

2.1.3. Multi-Scale Latent Space Representation

As Figure 1 shows, to enhance the diversity and expressiveness of the latent representation, we employ a multi-scale latent space representation approach. Unlike traditional VAEs, our design compresses the attention-modulated feature representation $\mathbf{H}$ into multiple low-dimensional subspaces, with each subspace dedicated to capturing different musical features, such as melody, rhythm, harmony, and their impact on attention.

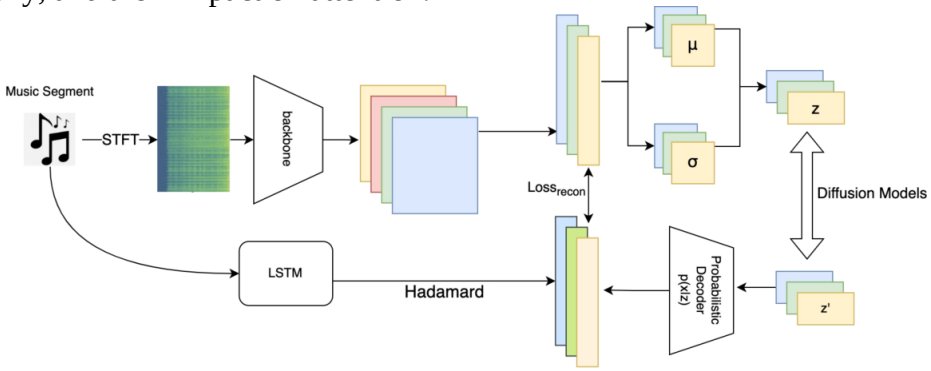

Figure 1: Architecture of the VAE-Diffusion Model for Music Generation and Attention Modulation

Specifically, the encoder generates a set of latent variables $\mathbf{z}_k$ in the latent space, defined as follows:

$$\mathbf{z}_k \sim \mathcal{N}(\mu_k(\mathbf{H}), \sigma_k^2(\mathbf{H})) \tag{1}$$

where μ_k and σ_k are the mean and variance functions of the k th subspace, respectively. The final latent representation is obtained by combining these multiple subspaces:

$$\mathbf{z} = concat[\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_K] \tag{2}$$

This multi-scale latent space representation method allows flexible control over the generation strategy across different feature dimensions, thereby enhancing the model's ability to capture complex musical structures and generate diverse outputs as well as their impact on attention.

2.1.4. Encoder Objective Function and Optimization

In VAE, our goal is to maximize the marginal likelihood of the data by learning the distribution of the latent variables. Given a data sample X and its corresponding latent variables z, this can be expressed as:

$$p(\mathbf{X}) = \int p(\mathbf{X} \mid \mathbf{z})\, p(\mathbf{z}) d\mathbf{z} \tag{3}$$

Directly optimizing this objective is very challenging; therefore, we approximate the posterior distribution $p(\mathbf{z} \mid \mathbf{X})$ by introducing a variational distribution $q(\mathbf{z} \mid \mathbf{X})$ and construct the Evidence Lower Bound (ELBO):

$$\log p(\mathbf{X}) = \mathbb{E}_{q(\mathbf{z}|\mathbf{X})}\left[\log \frac{p(\mathbf{X} \mid \mathbf{z})\, p(\mathbf{z})}{q(\mathbf{z} \mid \mathbf{X})}\right] + D_{KL}\left(q(\mathbf{z} \mid \mathbf{X}) \,\|\, p(\mathbf{z} \mid \mathbf{X})\right) \tag{4}$$

Where the KL divergence DKL is non-negative, thus:

$$\log p(\mathbf{X}) \geq \mathbb{E}_{q(\mathbf{z}|\mathbf{X})}\left[\log \frac{p(\mathbf{X} \mid \mathbf{z})\, p(\mathbf{z})}{q(\mathbf{z} \mid \mathbf{X})}\right] \tag{5}$$

This leads to the Evidence Lower Bound (ELBO):

$$\text{ELBO} = \mathbb{E}_{q(\mathbf{z}|\mathbf{X})}\left[\log p(\mathbf{X} \mid \mathbf{z})\right] - D_{KL}\left(q(\mathbf{z} \mid \mathbf{X}) \,\|\, p(\mathbf{z})\right) \tag{6}$$

The reconstruction loss $\mathcal{L}_{\text{recon}}$ is the expected log-likelihood, expressed as:

$$\mathcal{L}_{\text{recon}} = \mathbb{E}_{q(\mathbf{z}|\mathbf{X})}\left[\log p(\mathbf{X} \mid \mathbf{z})\right] \tag{7}$$

It assesses how accurately the input music and its attention effects are reconstructed from the latent representation.

The regularization loss $\mathcal{L}_{\text{KL}}$ is the KL divergence between the approximate posterior distribution and the prior distribution, expressed as:

$$\mathcal{L}_{\text{KL}} = D_{KL}\left(q(\mathbf{z} \mid \mathbf{X}) \| \, p(\mathbf{z})\right) \tag{8}$$

The regularization loss reduces the KL divergence between the approximate posterior and the prior distribution, guiding the latent space towards a standard normal distribution. This is key for enhancing the model's ability to generate and generalize. By using a prior with zero mean and unit variance, the model gains a more uniform and decoupled representation space, allowing for more efficient sampling and resulting in diverse, creative music.

In our model, it's important that the generated music not only reflects the input's structure and style but also has a similar effect on attention. Ensuring the latent space is normal helps the model balance between retaining key features of the input and introducing variation. A standard normal latent space allows for generating new music segments that remain consistent with the input while managing attention-related aspects. In short, regularization helps maintain a stable latent space, letting the model fine-tune the attention effects of the generated music, which enhances the audience's experience.

The final objective function is:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{recon}} + \beta \cdot \mathcal{L}_{\text{KL}} \tag{9}$$

Where β is a weighting parameter used to balance the reconstruction loss and the regularization loss.

2.2. Conditional Diffusion Model for Generating Similar Music

In the diffusion phase, we generate music segments based on the latent space formed by the VAE. These segments aim to replicate the input music while introducing controlled variations, ensuring a similar influence on attention. To achieve this, we employ a layered denoising method using a Conditional Diffusion Model (CDM). This model operates by taking the latent representations and attention-related attributes from the VAE encoding as conditions. As a result, the generated music remains closely aligned with the original input, preserving its structure, attention dynamics, and emotional tone.

The CDM ultimately enables the creation of music that both mirrors the input and extends its attentional influence. This is done through a systematic noise-reduction process that maintains the original composition's nuances. The outcome is melodies that not only reflect the structural and emotional elements of the input but also offer a subtle enhancement to the music generation process.

2.2.1. Overview of the Diffusion Model

Diffusion models represent a potent generative technique, capable of reconstructing outputs akin to targeted datasets (e.g., music snippets) from pure randomness through a sequential addition and subsequent removal of Gaussian noise. We use a Conditional Diffusion Model (CDM) that combines the latent representation with conditions like the characteristics and attention effects of the input music during denoising. This creates new music segments that reflect the input's style and attention modulation. The core idea involves applying inverse Markov Chain steps to gradually remove noise and recover the original latent form. At each step, conditional data is added to ensure the generated music stays consistent with the input, preserving its features, structure, and attention properties.

2.2.2. Initial Noise Injection and Latent Space Corruption

The first step in the diffusion process is to inject noise into the latent representation z $\mathbf{z}_0$ generated by the VAE, gradually transforming it into a random noise representation z $\mathbf{z}_T$. This process is defined as follows:

$$q(\mathbf{z}_t | \ \mathbf{z}_{t-1}) = \mathcal{N}(\mathbf{z}_t; \sqrt{1-\beta_t}\mathbf{z}_{t-1}, \beta_t \mathbf{I}) \tag{10}$$

Where β_t is a positive number used to control the amount of noise added. This process gradually transforms the latent representation into pure noise. The latent representation at step t , $\mathbf{z}_t$ is obtained by adding Gaussian noise with variance β_t to $\mathbf{z}_{t-1}$ .We can directly generate the representation at step t from the initial representation $\mathbf{z}_0$ using the following formula:

$$\mathbf{z}_t = \sqrt{\bar{\alpha}_t}\mathbf{z}_0 + \sqrt{1-\bar{\alpha}_t}\varepsilon, \quad \varepsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \tag{11}$$

$$\alpha_t = 1-\beta_t, \bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i \tag{12}$$

This formula describes how the initial latent representation is gradually transformed into pure noise through the progressive injection of noise.

Through this process, the model gradually transforms the initial latent representation into pure noise. Although the original feature information is progressively covered by noise at this stage, the

process provides a starting point for the subsequent reverse diffusion process. During the reverse diffusion, the model will gradually denoise, recovering and reconstructing music fragments similar to the input music in terms of features and attention regulation.

2.2.3. Conditional Diffusion Process and U-Net Denoising Network

In the reverse diffusion process, the model gradually denoises to recover the initial latent representation $\mathbf{z}_0$. This process is modeled as:

$$p_\theta(\mathbf{z}_{t-1} | \mathbf{z}_t, \mathbf{x}_0) = \mathcal{N}(\mathbf{z}_{t-1}; \mu_\theta(\mathbf{z}_t, t, \mathbf{x}_0), \Sigma_\theta(\mathbf{z}_t, t)) \tag{13}$$

Where $\mu_\theta(\mathbf{z}_t, t, \mathbf{x}_0)$ and $\Sigma_\theta(\mathbf{z}_t, t)$ are the conditional mean and variance, depending on the time step t and the conditional information $\mathbf{x}_0$ (input music features and attention impact). The conditional mean is given by:

$$\mu_\theta(\mathbf{z}_t, t, \mathbf{x}_0) = \frac{1}{\sqrt{\alpha_t}} \left(\mathbf{z}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}} \epsilon_\theta(\mathbf{z}_t, t, \mathbf{x}_0) \right) \tag{14}$$

Where $\epsilon_\theta(\mathbf{z}_t, t, \mathbf{x}_0)$ is the noise residual predicted by the U-Net denoising network. The conditional variance σ_t^2 is given by:

$$\sigma_t^2 = \beta_t (1-\alpha_{t-1}) / (1-\alpha_t) \tag{15}$$

Through the above formula, the model is able to gradually denoise during the reverse diffusion process, ultimately recovering the noise-free latent representation. To ensure that the generated music accurately reflects the features of the input music and its impact on attention, we employ a Conditional Diffusion Model (CDM), introducing conditional information—namely, the latent representation and attention features obtained during the first-stage encoding—into the denoising steps.

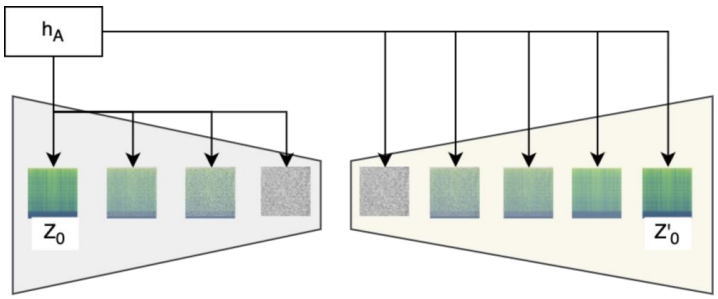

Figure 2: Conditional Diffusion Model for Music Generation with Attention Modulation

As Figure 2 shows, we use a U-Net architecture as the denoising network. U-Net is a symmetric convolutional neural network with skip connections, effectively preserving the integrity of the input information during the denoising process. To further enhance the denoising effect, we introduce a type of conditioning mechanisms within the U-Net.

Attention Flow Conditioning Injection: We introduce an attention regulation module to generate attention features $\mathbf{h}_A$. These features are injected into different layers of the U-Net to guide the model in generating music fragments that are similar to the input music in terms of attention impact. This process can be expressed as:

$$\hat{\mathbf{z}}_{t-1} = f_\theta(\mathbf{z}_t, \mathbf{h}_A, t) \tag{16}$$

Where $\hat{\mathbf{z}}_{t-1}$ is the denoised result combined with the attention features, the denoising function is

given by f_θ, where $\mathbf{h}_A$ represents the attention features and t denotes the time step.

$$f_\theta(\mathbf{z}_t, \mathbf{h}_A, t) = \gamma_t \cdot \phi_\theta(\mathbf{z}_t) + \lambda_t \cdot \psi_\theta(\mathbf{h}_A, t) \tag{17}$$

Through the combination of a nonlinear transformation sub-function and a dynamic adjustment sub-function, the denoising process is provided with enhanced adaptability. The design of the nonlinear transformation sub-function needs to consider the characteristics of the input latent representation. To better capture the complex feature interactions within the latent representation, we adopt the following structure:

$$\phi_\theta(\mathbf{z}_t) = \sigma\left(\mathbf{W}_2 \cdot \mathrm{GELU}(\mathbf{W}_1 \cdot \mathbf{z}_t + \mathbf{b}_1) + \mathbf{b}_2\right) \tag{18}$$

The core of the dynamic adjustment sub-function $\psi_\theta(\mathbf{h}_A, t)$ lies in modeling the interaction between the attention features $\mathbf{h}_A$ and the time step t. We achieve this through a self-attention mechanism:

$$\psi_\theta(\mathbf{h}_A, t) = \mathrm{Softmax}\left(\frac{\mathbf{Q}(\mathbf{h}_A) \cdot \mathbf{K}(\mathbf{h}_A)^\top}{\sqrt{d_k}}\right) \cdot \mathbf{V}(\mathbf{h}_A) \tag{19}$$

$\mathbf{Q}$, $\mathbf{K}$, $\mathbf{V}$ are the query, key, and value matrices in the attention mechanism, respectively, all of which are obtained by applying linear transformations to the attention features $\mathbf{h}_A$. In this way, the model can dynamically adjust the focus on different features during the denoising process to meet the requirements of different time steps.

To achieve flexible handling of different time steps, the introduced coefficients γ_t and λ_t can be dynamically adjusted using the following formulas:

$$\gamma_t = \frac{1}{1 + \exp(-k_1 \cdot (t - t_0))}, \quad \lambda_t = 1 - \gamma_t \tag{20}$$

This adjustment allows the model to flexibly balance the reliance on the latent representation and attention features at different stages of denoising. Our denoising function can flexibly adjust the dependency on the latent representation and attention features at different stages, providing strong nonlinear transformation capabilities while dynamically adjusting the weights of different features through the self-attention mechanism. This ensures that, at each denoising step, the generated music not only maintains its structural consistency but also faithfully reflects the input music's impact on attention. By injecting the latent representation and attention features into the U-Net denoising network, the model effectively integrates these conditional pieces of information throughout the denoising process, thereby preserving the features, structure, and attention regulation effects of the input music in the final generated music.

2.2.4. DDIM Sampling and Denoising Process

We employ the Denoising Diffusion Implicit Models (DDIM) sampler for the denoising process. The DDIM sampler accelerates the denoising process through a set of simplified equations, enabling us to generate high-quality music fragments within fewer time steps. The core steps of the DDIM denoising process are as follows: Calculate the noise residual for the current step:

$$\hat{\epsilon}_t = \mathbf{z}_t - \sqrt{\bar{\alpha}_t}\mathbf{z}_0 \tag{21}$$

Use the noise residual to compute the next step's latent representation.

$$\mathbf{z}_{t-1} = \sqrt{\bar{\alpha}_{t-1}}\left(\frac{\mathbf{z}_t - \sqrt{1-\bar{\alpha}_t}\hat{\epsilon}_t}{\sqrt{\bar{\alpha}_t}}\right) + \sqrt{1-\bar{\alpha}_{t-1}}\epsilon_\theta(\mathbf{z}_t, t) \tag{22}$$

At each denoising step, the DDIM sampler gradually optimizes the latent representation, moving towards the noise-free original latent representation until it recovers the final music fragment similar to the input music. Throughout this process, the conditional information (latent representation and attention flow) ensures that the denoising process consistently preserves the features, structure, and attention impact of the input music.

3. Conclusion

Our proposed two-stage generative architecture, combining Variational Autoencoder (VAE) encoding with Conditional Diffusion Models (CDM), achieves a harmonious balance between fidelity and diversity in music creation. The CDM segment preserves the core attributes and attention dynamics of the input music, while simultaneously infusing creativity and variation through strategic attention adjustments and condition-based inputs. This design ensures the necessary coherence for music generation tasks, yet also opens avenues for adaptable and exploratory creativity, laying a solid groundwork for future advancements in the field of attention-driven music synthesis.

References

[1] I.Copet, F. Kreuk, I. Gat, T. Remez, D. Kant, G. Synnaeve, Y. Adi, and A. Défossez, "Simple and controllable music generation," arXiv, 2023.

[2] Q.Kumar, P. Seetharaman, A. Luebs, I. Kumar, and K. Kumar, "High-fidelity audio compression with improved rvqgan," Advances in Neural Information Processing Systems (NeurIPS), 2024.

[3] B.Elizalde, S. Deshmukh, M. A. Ismail, and H. Wang, "Clap: Learning audio concepts from natural language supervision," arXiv, 2022.

[4] H.F. Garcia, P. Seetharaman, R. Kumar, and B. Pardo, "VampNet: Music generation via masked acoustic token modeling," arXiv, 2023.

[5] G. Cideron, S. Girgin, M. Verzetti, D. Vincent, M. Kastelic, Z. Borsos, B. McWilliams, V. Ungureanu, O. Bachem, O. Pietquin, M. Geist, L. Hussenot, N. Zeghidour, and A. Agostinelli, "MusicRL: Aligning music generation to human preferences," arXiv, 2024.

[6] H.Liu, Z. Chen, Y. Yuan, X. Mei, X. Liu, D. Mandic, W. Wang, and M. D. Plumbley, "AudioLDM: Text-to-audio generation with latent diffusion models," arXiv, 2023.

[7] T.Pascual, G. Bhattacharya, C. Yeh, J. Pons, and J. Serrà, "Full-band general audio synthesis with score-based diffusion," IEEE Int. Conf. on Acoustics, Speech and Sig. Proc. (ICASSP), 2023.

[8] Q. Huang, D. S. Park, T. Wang, T. I. Denk, A. Ly, N. Chen, Z. Zhang, Z. Zhang, J. Yu, C. Frank, J. Engel, Q. V. Le, W. Chan, Z. Chen, and W. Han, "Noise2music: Text-conditioned music generation with diffusion models," arXiv, 2023.

[9] F.Schneider, Z. Jin, and B. Schölkopf, "Moûsai: Text-to-music generation with long-context latent diffusion," arXiv, 2023.

[10] N.Zeghidour, A. Luebs, A. Omran, J. Skoglund, and M. Tagliasacchi, "Soundstream: An end-to-end neural audio codec," IEEE/ACM Trans. on Audio, Speech, and Language Processing, 2021.

[11] Z.Evans, C. Carr, J. Taylor, S. Hawley, and J. Pons, "Fast timing-conditioned latent audio diffusion," arXiv, 2024.

[12] AA.Borsos, R. Marinier, D. Vincent, E. Kharitonov, O. Pietquin, M. Sharifi, D. Roblek, O. Teboul, D. Grangier, M. Tagliasacchi et al., "AudioLM: a language modeling approach to audio generation," IEEE/ACM Trans. on Audio, Speech, and Language Processing, 2023.