



TVC-MusicGen: Time-Varying Structure Control for Background Music Generation via Self-Supervised Training

Chenyu Yang¹, Hangting Chen², Shuai Wang^{*,3}, Haina Zhu⁴, Haizhou Li^{1,5}

¹School of Data Science, SRIBD, The Chinese University of Hong Kong, Shenzhen, China

²Tencent AI Lab, Shenzhen, China ³Nanjing University, Suzhou, China

⁴X-LANCE Lab, Shanghai Jiao Tong University, Shanghai, China

⁵National University of Singapore, Singapore

chenyuyang2@link.cuhk.edu.cn, shuaiwang@nju.edu.cn

Abstract

Current text-to-music generation models typically are not involved in generating music with specific structures, thus not meeting some customized needs in practical applications. To address this limitation, we propose a self-supervised Time-Varying Control method (TVC-MusicGen). By providing the temporal boundary and text descriptions for each segment, it can effectively generate music adhering to the corresponding structures. TVC-MusicGen supports generation from both text (text-to-music) and existing music clips (music-to-music), enabling structure editing or local style transfer. Additionally, we propose a generation-based approach to bridge the gap between text and audio modalities in cross-modal models, which are typically used as feature extractors in text-to-music systems. Experiments on both language and diffusion-based models demonstrate that our approach achieves effective control without compromising overall quality. Audio samples are available in the demo page: https://cypress-yang.github.io/TVC-MusicGen_demo.

Index Terms: music generation, time-varying structure control, music segmentation, cross-modal alignment

1. Introduction

Significant progress has been made in the field of music generation. Modern generative models [1, 2, 3], particularly those leveraging audio language models, have advanced to the point where textual descriptions can be transformed into expressive music compositions, which have exhibited tremendous potential in practical applications.

While quality and musicality are usually emphasized in this task, there is an increasing need in real-world applications for finer control over the style of generated music along the timeline, especially in an easy-to-understand manner, to ensure that the generated music meets specific requirements and expectations. For instance, in the creation of background music for storytelling or poetry recitation, dynamic adjustments to the music's emotional shifts, rhythm changes, and harmonic transitions are necessary in response to the evolving storyline.

Although current text-to-music generation approaches can produce music in specific styles [1, 4], there is still an absence of control over the structure and timing of the music. They are typically restricted to adjusting the overall mood or theme of the piece without finer control over the specific musical structure and transitions.

Previous approaches were not designed to handle the precise control of musical structure during generation. One of the main reasons for this limitation is the lack of detailed structural

information in current datasets. These datasets typically include only music and overall text descriptions (captions), which are insufficient for enabling precise control of musical development. Manually annotating this structural information requires extensive expertise and is a time-consuming and labor-intensive process, making it challenging to produce high-quality annotated datasets.

In this paper, we propose Time-Varying Structure Control MusicGen (TVC-MusicGen), a novel approach to background music generation that incorporates time-varying structural control through temporal boundaries and detailed segmentation descriptions. The model is trained in a self-supervised manner, eliminating the reliance on additional human annotations. The primary contributions of this work can be summarized as follows:

- We propose a self-supervised paradigm that incorporates the clustering-based music segmentation algorithm into existing music generation models, enabling structural control and modification over the generated music.
- An additional text-to-audio adapter is proposed to enhance the semantic information of text embeddings, addressing the modality gap [5] caused by cross-modal alignment conditions. This adapter enables the model to achieve competitive performance when processing natural language descriptions, comparable to its performance with auditory prompts.

The proposed paradigm can be integrated into existing music generation models. Experiments on both autoregressive and non-autoregressive baselines have demonstrated the effectiveness of our methodology.

2. Related Work

2.1. Music Segmentation and Structure Analysis

Prior works [6] have investigated structure analysis and music segmentation tasks. Given the challenges of human annotation, the development of unsupervised segmentation algorithms has emerged as a more practical approach. Conventional unsupervised methods typically divide the procedure into two steps, boundary detection [7, 8] followed by segment clustering [9, 10, 11, 12]. These approaches typically rely on the self-similarity matrix (SSM), which helps capture recurring patterns within the music and identifies potential segment boundaries. For example, [13] utilizes unsupervised features [14, 15, 16] to compute the SSM and adopts a checkerboard kernel [17] to produce a novelty function, which predicts the segment boundaries based on the SSM. Recent studies [18, 19, 20] propose directly predicting specific music structures, such as verse or chorus, from the waveform. These methods have achieved state-of-the-art performance. However, they are typically designed for song

* Corresponding author

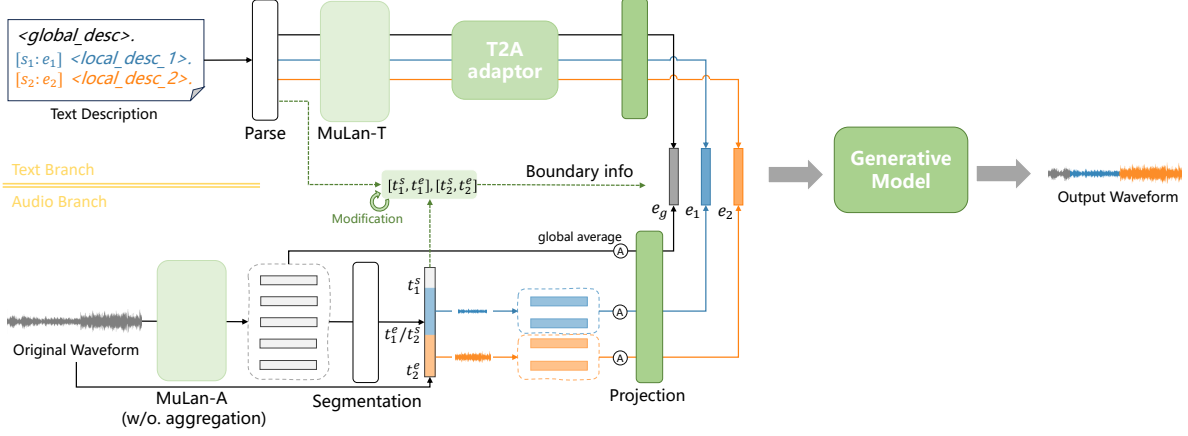


Figure 1: The overall architecture of TVC-MusicGen.

data, which generally exhibits more distinct and fixed structural characteristics, such as clear verse-chorus patterns.

We aim to leverage segmentation information to guide the structure and composition of background music generation. On the one hand, we strive to maintain high efficiency during training. On the other hand, we recognize that, unlike songs, the structure of background music should allow for relatively free variation. To address this, inspired by the conventional SSM-based approaches, we employ a spectral clustering-based algorithm to segment the intermediate feature sequence of audio.

2.2. Music generation

Music generation refers to the task of generating music according to the natural language descriptions. Existing approaches primarily utilize either language models [21, 1, 2] or diffusion models [4, 3]. Instead of directly generating mel-spectrograms or waveforms, current approaches tend to use discrete codec tokens [22] or latent vectors of variational autoencoders (VAE) as intermediate representations, enabling the synthesis of long-form and expressive music clips.

However, most of the models mentioned above solely rely on the global description of overall themes, styles, or moods, lacking finer control over time-varying transitions. Recent research efforts [23, 24, 25] have begun addressing this limitation by adding time-varying controls. For instance, Music Control-Net [24] proposes to leverage melody, dynamics, and rhythm signals to offer frame-level controls for pretrained diffusion models.

Despite these signal-based conditions offer precise control over the music’s progression, they still face practical challenges, (1) Musical signals must be extracted from existing tracks (2) Such features are highly abstract, making them difficult to modify or interpret.

3. TVC-MusicGen

3.1. Overall Architecture

The time-varying condition includes a global descriptor e_g and several local descriptors with corresponding boundaries, denoted as triplets (t_i^s, t_i^e, e_i) , where t_i^s and t_i^e represent the start and end time of segment i , respectively, and e_i is the corresponding feature embedding. The triplets can be extracted and, if necessary, adjusted from existing music tracks (**music-to-**

music) or derived from given text lines (**text-to-music**).

Figure 1 illustrates the overall architecture of TVC-MusicGen. The complete training procedure can be divided into three steps. First, a pretrained cross-modal alignment model is utilized. In this paper, we employ MuQ-MuLan [26], which features a two-tower audio-text encoder architecture, to encode the condition. We train the generative model with the output of the audio tower. Additionally, a text-to-audio embedding adapter is trained to map the text embeddings from MuLan into the same space as the audio embeddings.

3.2. Spectral Clustering for Music Segmentation

Assuming the internal frame sequence of the audio tower is $\hat{\mathbf{h}} \in \mathbb{R}^{(F \cdot L) \times D}$, where F represents the frame rate, L is the audio length in seconds, and D is the feature dimension, a mean pooling operation with a stride and window length of F is applied to obtain per-second embeddings $\mathbf{h} = \text{avg_pool}(\hat{\mathbf{h}}) \in \mathbb{R}^{L \times D}$, reducing the sequence length to L .

We use the cosine similarity to measure the pairwise similarity between embeddings, resulting in the self-similarity matrix $\mathbf{S} \in [-1, 1]^{L \times L}$, where $s_{ij} = \frac{\mathbf{h}_i^T \mathbf{h}_j}{\|\mathbf{h}_i\| \cdot \|\mathbf{h}_j\|}$ represents the cosine similarity between embeddings \mathbf{h}_i and \mathbf{h}_j . Finally, we construct the affinity matrix $\mathbf{R} \in [0, 1]^{L \times L}$ by applying a clamp operation on \mathbf{S} with a threshold δ [27]:

$$r_{ij} = \begin{cases} 0 & s_{ij} < \delta, \\ \frac{s_{ij} - \delta}{1 - \delta} & s_{ij} \geq \delta. \end{cases} \quad (1)$$

Here δ is empirically set to 0.

The eigendecomposition of \mathbf{R} is represented as:

$$\mathbf{R} = \mathbf{V} \mathbf{\Lambda} \mathbf{V}^{-1}, \quad \mathbf{\Lambda} = \text{diag}(\lambda_1, \dots, \lambda_L) \quad (2)$$

where eigenvector matrix $\mathbf{V} = [\mathbf{v}_1, \dots, \mathbf{v}_L]$ corresponds to descending eigenvalues $\lambda_1 \geq \dots \geq \lambda_L$.

Our purpose is to distinguish semantic components from tiny noises, and the eigenvalues indicate the size of clusters. There are multiple ways to determine the number of clusters based on the eigenvalues, such as maximum eigengap [10]. In this paper, we simply compute $N = \sum_{k=1}^L \mathbb{I}(\lambda_k > 1)$, counting all eigenvalues greater than 1. Subsequent clustering algorithm uses the determined N to obtain the segment boundaries, e.g., k-means on the embedding sequence. Worth noting is that N should be strictly greater than one, otherwise the setting of

segmentation is meaningless and we will ignore the following steps and output an empty set of segments. Apart from this, we also filter out all segments shorter than 3 seconds. The boundaries of segments are denoted as $\{(t_i^s, t_i^e) | i \in \{1, \dots, \tilde{N}\}\}$, where \tilde{N} is the final number of segments.

To prevent information leakage, the local MuLan embedding is derived directly from the original waveform and segment boundaries. The embedding is computed as $e_i = \text{MuLan}(x^{wav}[t_i^s \times sr : t_i^e \times sr])$, where x^{wav} denotes the waveform, and sr is the sample rate. The input sequence of generative models can be represented as:

$$z_{cond} = [e_g] \oplus \left[\underbrace{0, \dots, e_1, \dots, e_1, \dots, e_{\tilde{N}}, \dots, e_{\tilde{N}}, \dots, 0}_{\substack{L \\ t_1^s \sim t_1^e \quad t_{\tilde{N}}^s \sim t_{\tilde{N}}^e}} \right],$$

where for each segment (t_i^s, t_i^e, e_i) , we set $z_{cond}[t_i^s : t_i^e] = e_i$. Additionally, the global embedding e_g is derived in the same way over the entire waveform and prepended to the sequence.

3.3. Text-to-Audio Embedding Adaption

The cross-modal feature-based music generation model can utilize both existing music and natural language descriptions as input conditions. However, some inherent limitations can also be observed. As studied in [5], multi-modal contrastive representation learning can lead to features of different modalities being confined to distinct, narrow cones in the embedding space. Moreover, current music captions generally fail to cover all aspects of corresponding music pieces. Natural language descriptions have significantly lower information density compared to music, which can negatively impact the quality of the generated audio when cross-modal inputs are used. Additionally, for fine-grained structural control, since the overall theme tends to remain consistent, local structures often exhibit subtle differences. Experiments show that semantic differences might be disrupted by the modality gap, resulting in suboptimal performance when the model accepts natural language descriptions as conditions during inference.

Previous approaches typically set the feature embedding $\mathbf{e} := \mathbf{e}^{wav}$ for training and $\mathbf{e} := \mathbf{e}^{text}$ for inference, where $\mathbf{e}^{wav}, \mathbf{e}^{text} \in \mathbb{R}^D$ represent the outputs of MuLan towers for audio and text modalities, respectively. In this paper, we introduce a small generative model to bridge the gap between audio and text, which transforms the text embedding \mathbf{e}^{text} into the auditory hidden space. Specifically, we train a DDPM-based [28] diffusion model that generates a pseudo \mathbf{e}^{wav} by predicting $p_\theta(\mathbf{e}^{wav} | \mathbf{e}^{text}, \mathbf{h}^{text})$. Here, $\mathbf{h}^{text} \in \mathbb{R}^{S \times D}$ is the text embedding sequence after passing through the RoBERTa frontend which provides more precise detailed characteristics, and S is the length of sentences.

To encourage diversity during training, we further replace $p_\theta(\mathbf{e}^{wav} | \mathbf{e}^{text}, \mathbf{h}^{text})$ with $p_\theta(\mathbf{z} | \mathbf{e}^{text}, \mathbf{h}^{text})$, where $\mathbf{z} = \{\mathbf{z}_k\}_{k=1}^K$ is a series of MuLan embeddings derived from different clips taken from the same track, with K representing the total number of embeddings. In our experiment, we intercept the audio at a fixed length, such that $\mathbf{z}_k = \text{MuLan}(x^{wav}[k \times stride : k \times stride + win_size])$. These clips correspond to the same text but exhibit slight differences. During inference, we compute \mathbf{e}^{wav} by calculating the average of predicted \mathbf{z} .

4. Experiments

4.1. Baselines

MusicGen: MusicGen [1] is an auto-regressive decoder-only transformer that processes natural language descriptions as input. In our experiments, we replace the T5 text encoder [29] with our MuLan implementation. The Encodec [22] frontend remains unchanged from the original paper.

MusicLDM: For the diffusion-model baseline, we utilize the Latent Diffusion Model framework [30, 4, 31]. Specifically, the DiT [32] serves as the backbone of our framework. A modified VAE¹ from the Descript-Audio-Codec [33] is used for audio compression. This model is capable of reconstructing audio at a 44.1 kHz sampling rate.

4.2. Configuration Details

The training dataset includes the Pond5 and Shutterstock music collections referenced in [1], along with an additional internal music dataset. Our Pond5 and Shutterstock collections contain 250K instrument-only tracks, while the internal dataset consists of 50K tracks that have undergone vocal separation to remove possible vocals, totaling approximately 15K hours of music. Each sample is 30 seconds long.

For MusicGen models, we utilize the LLaMA-based [34] architecture. For MusicLDM models, we adopt the implementation of DiT in [35]. In both cases, the models consist of 24 layers with a hidden dimension of 1536.

All models were trained by the AdamW [36] optimizer with a learning rate of 1e-4. During training, each segment was dropped out randomly with a probability of 0.2. We trained each model for approximately 200K steps, and the batch size was set to 128. During inference, the classifier-free guidance was adopted with a coefficient of 3.0. For MusicLDM models, the Euler sampler is applied for 250 inference steps.

4.3. Evaluation Metrics

We use the Fréchet Audio Distance (FAD) [37] to assess the overall quality of the generated samples. Additionally, for music-to-music tasks, we evaluate structural similarity using the cosine similarity of second-level averaged embeddings computed with MERT [38] (SIM). For text-to-music tasks, we measure the similarity between music and text using the MuLan Cycle Consistency (MCC) in [21]. Since in our experiments, a sample may have multiple descriptions (both global and local), we compute a weighted average over the scores of each description and its corresponding span of music based on their lengths.

For subjective evaluation, the mean opinion score (MOS) listening test is conducted. The test mainly focuses on two aspects: overall musicality and the consistency between generated music and original inputs (either waveform or multi-line text descriptions). The consistency score also accounts for segment-level similarity, ensuring that local coherence within the generated music is properly assessed. 10 raters are employed to attach a score to each sample, ranging from 1 to 5.

5. Results

5.1. Music-to-Music Generation

In addition to the baselines and our proposed models, we also tested a structure control method based on fixed-length seg-

¹<https://github.com/innnky/descript-audio-vae>

Model	Objective		Subjective	
	FAD↓	SIM↑	Musicality↑	Consistency↑
Ground truth	-	-	4.14	4.80
MusicGen	5.11	0.86	3.15	2.31
FX-MusicGen	3.18	0.89	3.36	3.62
TVC-MusicGen	2.52	0.91	3.42	4.26
MusicLDM	3.72	0.86	3.17	3.10
TVC-MusicLDM	2.00	0.90	3.33	4.26

Table 1: *Music-to-music generation.* A subset of the *MagnaTagTune* dataset [39] which includes 100 instrument-only 30-second samples is used as the test set. For the subjective evaluation part, we further selected 20 samples randomly for the listening test.

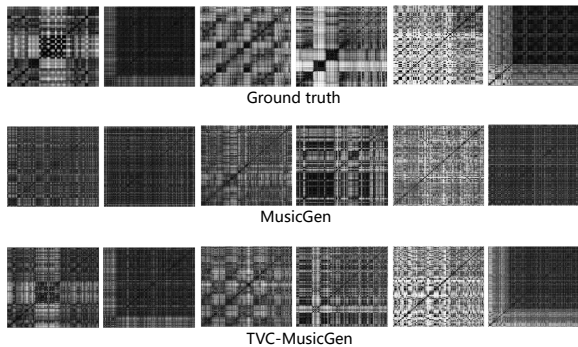


Figure 2: *Chroma-based self-similarity matrices.*

mentation (FX-MusicGen), which divides the audio into three equal-length segments, each lasting 10 seconds. This baseline model also provides segment-level information, allowing us to isolate the effect of auxiliary information and the potential temporal information that may exist in MuLan embeddings. As shown in Table 1, our models significantly outperform baselines in both objective and subjective metrics. Although part of the improvement can be attributed to the introduction of segment-level information, a proper boundary detection method remains crucial, as it further enhances performance, particularly in ensuring consistency between generated music and expectations.

To further demonstrate the structure control capability of our methods, Figure 2 illustrates the chroma-based self-similarity matrices for different models. The proposed TVC-MusicGen model generates music that closely matches the ground truth in terms of SSM, while MusicGen produces results that tend to be more monotonous and lack temporal variation.

5.2. Text-to-Music Generation

In this section, we evaluate the performance of different approaches for text-to-music generation. Due to the absence of segment-level annotated text-music data for evaluation, we leverage the music samples mentioned earlier. These samples are first segmented and then processed by MU-LLaMA [40], a music understanding language model, to produce pseudo-annotations, which are subsequently used for text-to-music generation.

Table 2 compares the performance of our model before and after applying time-varying control. As shown in the table, when local descriptions are provided, there is a significant boost in MCC_{seg} , indicating that our approach effectively manages finer control over the structure. On the other hand, the

Model	Objective			Subjective	
	FAD↓	MCC_{all} ↑	MCC_{seg} ↑	Musicality↑	Consistency↑
MusicGen	6.44	0.36	0.18	3.29	3.38
TVC-MusicGen	5.53	0.26	0.29	3.43	4.02
- w/o local	5.28	0.35	0.17	3.47	3.71
MusicLDM	5.24	0.37	0.16	3.17	3.33
TVC-MusicLDM	5.35	0.31	0.30	3.39	3.79
- w/o local	4.95	0.38	0.17	3.56	3.43

Table 2: *Text-to-music generation.* The MCC score is calculated for both entire samples (MCC_{all}) and detailed segments (MCC_{seg}). Additionally, we tested the performance of TVC-based methods with only the global description (w/o local).

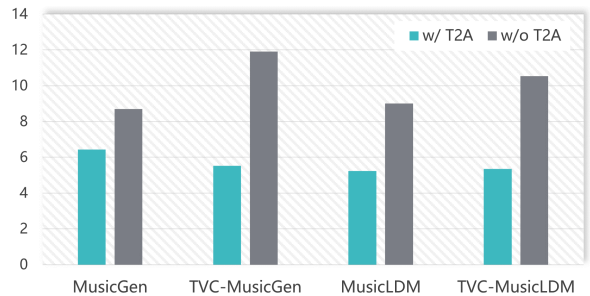


Figure 3: *Comparison of FAD scores (↓) for different models with or without the text-to-audio adaptation module.*

performance of our models does not degrade compared to baselines when no segment-level descriptions are provided, showing competitive performance in terms of quality. Subjective evaluations further show that TVC-based methods improve both musicality and consistency, with segment-level descriptions providing a significant boost to consistency.

An ablation study is also conducted to assess the text-to-audio adaptation module. Figure 3 shows the FAD scores for different models. Introducing the text-to-audio embedding adaptation significantly decreases the FAD score, indicating improved overall quality. Additionally, introducing TVC conditions results in a higher FAD score without the text-to-audio adapter. As there are multiple MuLan embeddings present, the model becomes more sensitive to the modality gap, making adaptation crucial for our models.

6. Conclusion and Discussion

In this paper, we introduced TVC-MusicGen, a versatile plugin for music generation models that enables time-varying structure control through natural language descriptions. This approach can be applied to both music generation and editing, such as changing the style or length of specific segments. However, there are still some limitations. Despite the text-to-audio adaptation module, the differences between segments are not as pronounced as in music-to-music tasks, indicating a need for further enhancement or data augmentation of MuLan.

7. Acknowledgement

- Shenzhen Science and Technology Program (Shenzhen Key Laboratory, Grant No. ZDSYS20230626091302006)
- Shenzhen Science and Technology Research Fund (Fundamental Research Key Project, Grant No. JCYJ20220818103001002)

- Program for Guangdong Introducing Innovative and Entrepreneurial Teams, Grant No. 2023ZT10X044

8. References

- [1] J. Copet, F. Kreuk, I. Gat, T. Remez, D. Kant, G. Synnaeve, Y. Adi, and A. Défossez, “Simple and controllable music generation,” *NeurIPS*, vol. 36, 2024.
- [2] M. W. Lam, Q. Tian, T. Li, Z. Yin, S. Feng, M. Tu, Y. Ji, R. Xia, M. Ma, X. Song *et al.*, “Efficient neural music generation,” *NeurIPS*, vol. 36, pp. 17 450–17 463, 2024.
- [3] J. Melechovsky, Z. Guo, D. Ghosal, N. Majumder, D. Herremans, and S. Poria, “Mustango: Toward controllable text-to-music generation,” in *NAACL*, 2024, pp. 8293–8316.
- [4] K. Chen, Y. Wu, H. Liu, M. Nezhurina, T. Berg-Kirkpatrick, and S. Dubnov, “Musicldm: Enhancing novelty in text-to-music generation using beat-synchronous mixup strategies,” in *ICASSP*, 2024, pp. 1206–1210.
- [5] V. W. Liang, Y. Zhang, Y. Kwon, S. Yeung, and J. Y. Zou, “Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning,” *NeurIPS*, vol. 35, pp. 17 612–17 625, 2022.
- [6] O. Nieto, G. J. Mysore, C.-i. Wang, J. B. Smith, J. Schlüter, T. Grill, and B. McFee, “Audio-based music structure analysis: Current trends, open challenges, and applications,” *Transactions of the International Society for Music Information Retrieval*, vol. 3, no. 1, 2020.
- [7] B. McFee and D. P. Ellis, “Learning to segment songs with ordinal linear discriminant analysis,” in *ICASSP*, 2014, pp. 5197–5201.
- [8] J. Serra, M. Müller, P. Grosche, and J. L. Arcos, “Unsupervised music structure annotation by time series structure features and segment similarity,” *IEEE Transactions on Multimedia*, vol. 16, no. 5, pp. 1229–1240, 2014.
- [9] M. Levy and M. Sandler, “Structural segmentation of musical audio by constrained clustering,” *IEEE transactions on audio, speech, and language processing*, vol. 16, no. 2, pp. 318–326, 2008.
- [10] B. McFee and D. Ellis, “Analyzing song structure with spectral clustering,” in *ISMIR*, 2014, pp. 405–410.
- [11] O. Nieto and T. Jehan, “Convex non-negative matrix factorization for automatic music structure identification,” in *ICASSP*, 2013, pp. 236–240.
- [12] F. Kaiser and T. Sikora, “Music structure discovery in popular music using non-negative matrix factorization,” in *ISMIR*, 2010, pp. 429–434.
- [13] M. C. McCallum, “Unsupervised learning of deep features for music segmentation,” in *ICASSP*, 2019, pp. 346–350.
- [14] T. Grill and J. Schlüter, “Music boundary detection using neural networks on spectrograms and self-similarity lag matrices,” in *23rd European Signal Processing Conference (EUSIPCO)*, 2015, pp. 1296–1300.
- [15] J. Schlüter and S. Böck, “Improved musical onset detection with convolutional neural networks,” in *ICASSP*, 2014, pp. 6979–6983.
- [16] K. Ullrich, J. Schlüter, and T. Grill, “Boundary detection in music structure analysis using convolutional neural networks,” in *ISMIR*, 2014, pp. 417–422.
- [17] J. Foote, “Automatic audio segmentation using a measure of audio novelty,” in *ICME*, vol. 1, 2000, pp. 452–455.
- [18] T. Kim and J. Nam, “All-in-one metrical and functional structure analysis with neighborhood attentions on demixed audio,” in *2023 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2023, pp. 1–5.
- [19] J.-C. Wang, Y.-N. Hung, and J. B. Smith, “To catch a chorus, verse, intro, or anything else: Analyzing a song with structural functions,” in *ICASSP*, 2022, pp. 416–420.
- [20] J.-C. Wang, J. B. Smith, J. Chen, X. Song, and Y. Wang, “Supervised chorus detection for popular music using convolutional neural network and multi-task learning,” in *ICASSP*, 2021, pp. 566–570.
- [21] A. Agostinelli, T. I. Denk, Z. Borsos, J. Engel, M. Verzett, A. Caillon, Q. Huang, A. Jansen, A. Roberts, M. Tagliasacchi *et al.*, “Musiclm: Generating music from text,” *arXiv preprint arXiv:2301.11325*, 2023.
- [22] A. Défossez, J. Copet, G. Synnaeve, and Y. Adi, “High fidelity neural audio compression,” *arXiv preprint arXiv:2210.13438*, 2022.
- [23] L. Lin, G. Xia, J. Jiang, and Y. Zhang, “Content-based controls for music large language modeling,” *arXiv preprint arXiv:2310.17162*, 2023.
- [24] S.-L. Wu, C. Donahue, S. Watanabe, and N. J. Bryan, “Music controlnet: Multiple time-varying controls for music generation,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 2692–2703, 2024.
- [25] Y.-H. Lan, W.-Y. Hsiao, H.-C. Cheng, and Y.-H. Yang, “Musicongen: Rhythm and chord control for transformer-based text-to-music generation,” *arXiv preprint arXiv:2407.15060*, 2024.
- [26] H. Zhu, Y. Zhou, H. Chen, J. Yu, Z. Ma, R. Gu, Y. Luo, W. Tan, and X. Chen, “Muq: Self-supervised music representation learning with mel residual vector quantization,” *arXiv preprint arXiv:2501.01108*, 2025.
- [27] S. Horiguchi, S. Watanabe, P. García, Y. Xue, Y. Takashima, and Y. Kawaguchi, “Towards neural diarization for unlimited numbers of speakers using global and local attractors,” in *ASRU*, 2021, pp. 98–105.
- [28] J. Ho, A. Jain, and P. Abbeel, “Denoising diffusion probabilistic models,” *NeurIPS*, vol. 33, pp. 6840–6851, 2020.
- [29] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, “Exploring the limits of transfer learning with a unified text-to-text transformer,” *Journal of machine learning research*, vol. 21, no. 140, pp. 1–67, 2020.
- [30] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, “High-resolution image synthesis with latent diffusion models,” in *CVPR*, 2022, pp. 10 684–10 695.
- [31] 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,” in *ICML*, 2023, pp. 21 450–21 474.
- [32] W. Peebles and S. Xie, “Scalable diffusion models with transformers,” in *ICCV*, 2023, pp. 4195–4205.
- [33] R. Kumar, P. Seetharaman, A. Luebs, I. Kumar, and K. Kumar, “High-fidelity audio compression with improved rvqgan,” *NeurIPS*, vol. 36, 2024.
- [34] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale *et al.*, “Llama 2: Open foundation and fine-tuned chat models,” *arXiv preprint arXiv:2307.09288*, 2023.
- [35] Z. Evans, J. D. Parker, C. Carr, Z. Zukowski, J. Taylor, and J. Pons, “Stable audio open,” *arXiv preprint arXiv:2407.14358*, 2024.
- [36] I. Loshchilov and F. Hutter, “Decoupled weight decay regularization,” *arXiv preprint arXiv:1711.05101*, 2017.
- [37] K. Kilgour, M. Zuluaga, D. Roblek, and M. Sharifi, “Fréchet audio distance: A reference-free metric for evaluating music enhancement algorithms,” in *Interspeech*, 2019, pp. 2350–2354.
- [38] Y. Li, R. Yuan, G. Zhang, Y. Ma, X. Chen, H. Yin, C. Xiao, C. Lin, A. Ragni, E. Benetos *et al.*, “Mert: Acoustic music understanding model with large-scale self-supervised training,” *ICLR*, 2024.
- [39] E. Law, K. West, M. I. Mandel, M. Bay, and J. S. Downie, “Evaluation of algorithms using games: The case of music tagging,” in *ISMIR*, 2009, pp. 387–392.
- [40] S. Liu, A. S. Hussain, C. Sun, and Y. Shan, “Music understanding llama: Advancing text-to-music generation with question answering and captioning,” in *ICASSP*, 2024, pp. 286–290.