Rhythmic Stability and Synchronization in Multi-Track Music Generation

Hongrui Wang*, **Fan Zhang***, **Zhiyuan Yu**, **Ziya Zhou**, **Xi Chen**, **Can Yang**†, **Yang Wang**†
The Hong Kong University of Science and Technology (HKUST)
{hwangfb, zyuaq, zzhoucp, xchengx}@connect.ust.hk,
{mafzhang, macyang, yangwang}@ust.hk

Abstract

Current paradigms for multi-track music generation and assessment notably overlook two perceptually crucial elements—beat stability and cross-track synchronization—despite their demonstrated impact on listener experience. This paper presents three novel evaluation metrics for multi-track music generation, establishing a comprehensive evaluation framework for multi-track music quality compared with existing standards. Moreover, we propose a synchronous multi-track waveform generation model, SyncTrack. Experiments demonstrate that SyncTrack achieves superior performance on both conventional and newly proposed metrics, validating the effectiveness of our model and the utility of our evaluation framework.

1 Introduction

Multi-track music generation has attracted researchers for its ability to enable individual track editing. However, beat stability and synchronization are not considered in the evaluation and generation of multi-track music. Temporal prediction errors caused by irregular beats or inter-track asynchrony lead to persistent violations of auditory expectations, inducing listener discomfort Mas-Herrero et al. (2018). Current works assess the quality of multi-track audio generation using the Fréchet Audio Distance (FAD) merely. FAD is designed to measure the similarity between generated and reference audio samples by compressing the file into VGGish embeddings Kilgour et al. (2018). Over-compression of temporal information renders it impossible to assess stability and synchronization.

In this paper, we propose three novel metrics, Inner-track Rhythm Stability (IRS), Cross-track Beat Synchrony (CBS) and Cross-track Beat Dispersion (CBD) to assess stability and synchronization. Specifically, ***IRS*** evaluates the rhythmic stability of an audio track based on the variance of its beat intervals. ***CBS*** quantifies the proportion of rhythmically aligned beats using a sliding tolerance window. ***CBD*** computes the timing errors between aligned beats, providing a more refined measurement of beat synchrony. Combining these metrics with FAD enables a more comprehensive and accurate evaluation of multi-track music generation quality.

Moreover, we propose a synchronous multi-track waveform music generation model named SyncTrack. To capture cross-track beat synchrony, we design two types of cross-track attention modules: 1) a cross-track attention module that calculates attention weights across distinct temporal segments and spectral bands; 2) a time-specific cross-track attention module that calculates attention weights across spectral bands within specific temporal segments. Intuitively, the time-specific module achieves finer-grained beat synchronization within localized temporal windows. To facilitate effective model training, we introduce shared parameters for cross-track knowledge transfer. Additionally, we design track-specific parameters enabled by conditioning to preserve track uniqueness.

* Equal contribution, † Corresponding author.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: AI for Music.

2 SyncTrack: Synchronous Multi-track Music Generation Model

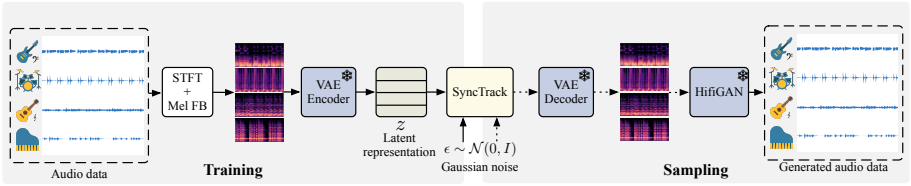

Figure 1: Illustration of Overall Framework of SyncTrack

Overall Framework. As shown in Fig 1, due to the highly structured nature, we compress the audio data $\{x^s\}_{s=1}^S$ into latent representation in two steps: 1) audio data to Mel-spectrogram using Short-Time Fourier Transform (STFT) and a Mel filter bank; 2) Mel-spectrogram to latent representation using a pre-trained Variational Autoencoder (VAE) Kingma, Welling (2013). The compression process of s-th track x^s is formulated as follows:

$$z_0^s := \text{VAE}_{\text{enc}}(\text{STFT\&MelFB}(x^s)) \in \mathbb{R}^{C\times T\times F}, \tag{1}$$

where T and F are the time and frequency bins. Temporal and frequency information are reserved .

We denote SyncTrack as ϵ_θ, which utilizes a U-Net backbone to learn the distribution of z_0^s from z_l^s (l is uniformly sampled from $\{1, \cdots, L\}$). The training objective is as follows:

$$\mathcal{L}(\theta) = \mathbb{E}_{\epsilon\sim\mathcal{N}(0,I),\{z^s\}_{s=1}^S,l} \left\| \epsilon - \epsilon_\theta(\{z_l^s\}_{s=1}^S, l) \right\|^2. \tag{2}$$

SyncTrack utilizes two designed cross-track attention modules to better capture the relationships between multiple tracks, enhancing beat synchronization. Besides, we design shared and track-specific parameters to capture the cross-track common information and track-wise uniqueness.

In the sampling phase, we utilize the trained SyncTrack to approximate the distribution of z_0^s. Then the samples $\hat{z}_0^s$ are decoded into audio by VAE decoder and HiFi-GANKong et al. (2020) vocoder:

$$\hat{x}^s = \text{HiFiGAN}\left(\text{VAE}_{\text{dec}}(\hat{z}_0^s)\right). \tag{3}$$

More details regarding the architecture of SyncTrack can be referred in Appendix A.

Cross-Track Information Extraction. We leverage the 2D U-Net as the backbone. However, 2D U-Net only conduct inner-track attention. Thus, we design two types of cross-track attention modules, named **cross-track attention** and **time-specific cross-track attention**.

1) Cross-Track Attention. To capture global synchrony and interactions across tracks, such as beat alignment and harmonic coordination, we introduce a cross-track attention module. Take the representation $z_{t,f}^s \in \mathbb{R}^C$ in the f-th frequency bin, t-th time bin of the s-th track as an example. We aggregate information from all time bin and frequency bin of all tracks:

$$\text{Attn}_{\text{cross}}(z_{t,f}^s) = \text{Attn}(W^Q z_{t,f}^s, W^K z_{1:T,1:F}^{1:S}, W^V z_{1:T,1:F}^{1:S}). \tag{4}$$

2) Time-Specific Cross-Track Attention. While global cross-track attention enables broad coordination, finer-grained rhythmic alignment—such as precise beat synchronization—may require context localized in time. To this end, we introduce a time-specific cross-track attention module, which computes attention across all tracks and spectral bands for each time step t:

$$\text{Attn}_{\text{time\_cross}}(z_{t,f}^s) = \text{Attn}(W^Q z_{t,f}^s, W^K z_{t,1:F}^{1:S}, W^V z_{t,1:F}^{1:S}). \tag{5}$$

Here, the query, key, and value are restricted to features at the same time index t, but across all tracks and frequencies. This mechanism emphasizes beat-level synchronization by allowing the model to fuse information between instruments in localized temporal windows.

Learnable Instrument Prior. We design a learnable instrument prior to serve as track-specific parameters. First, we leverage the one-hot vectors V to represent different tracks. Then, we feed the vectors into SyncTrack as extra input.

$$\hat{\epsilon} = \text{SyncTrack}(\{z_l^s\}_{s=1}^S, l, V). \tag{6}$$

V is first encoded via positional encoding Mildenhall et al. (2021) and subsequently transformed by a two-layer neural network. Finally, the embedding of V is concatenated with time embedding of n. Our experiments show that it allows for robust generalization and performance improvement.

3 Evaluation Metrics for Rhythmic Stability and Synchronization

To address the rhythmic stability and synchronization issue for multi-track music generation, we provide three different metrics. These metrics directly capture whether beats are synchronized across tracks S for all samples N, which provide a reproducible and interpretable way to assess multi-track music generation results across different methods.

Inner-Track Rhythm Stability (IRS). For music with stable rhythm, beat intervals should remain consistent within each track. IRS quantifies temporal consistency by averaging standard deviation of Inter-Beat IntervalDannenberg (1987); Robertson (2012) across all samples for each track s:

$$\text{IRS} = \mathbb{E}_{s,n}\left[std(I_n^s)\right], \tag{7}$$

where I_n^s denotes the beat intervals for for the track s in sample n, whose element is defined as time difference between two consecutive beats.

Cross-Track Beat Synchrony (CBS). CBS measures rhythmic alignment among multiple tracks. Inspired by the tolerance window concept in beat tracking Dixon (2001), we divide the timeline into multiple windows and compute the proportion of tracks that contain at least one beat within each window. Only tracks with non-empty content are considered. CBS is defined as:

$$\text{CBS} = \mathbb{E}_n\left[\frac{\sum_{i=1}^{T} r_{i,n}}{\sum_{i=1}^{T}\mathbb{I}(r_{i,n}>0)}\right], \tag{8}$$

where $r_{i,n}$ is the ratio of tracks containing at least one beat within the i-th window, and we utilize $\mathbb{I}(r_{i,n} > 0)$ to exclude windows where no beat occurs in any track..

Cross-Track Beat Dispersion (CBD).The CBD metric quantifies rhythmic consistency in multi-track music by measuring the dispersion of beat alignment across all pairs of tracks. Drawing inspiration from GOTO's method Goto, Muraoka (1997), which evaluates alignment errors between estimated and reference beats using beat error sequences, CBD extends this concept to multi-track scenarios.

We select each track as reference in turn and compute the beat error sequence with respect to all other tracks. For track s in sample n, let $b_{n,t}^s$ denote the t-th beat in the reference track. For each $b_{n,t}^s$, we find the matching beats in the other tracks and extract error sequence, denoted as $e(b_{n,t}^s)$. The CBD metric is defined as the mean or other statistics of the beat error sequence:

$$\text{CBD(mean)} = \mathbb{E}_{s,n,t}\left[e(b_{n,t}^s)\right]. \tag{9}$$

Note that, to eliminate the influence of tempo variations, we use beat interval to normalize $e(b_{n,t}^s)$. Since matching beats of $b_{n,t}^s$ are within the two intervals $\left[-I_{n,t-1}^s/2 + b_{n,t}^s, b_{n,t}^s\right]$ and $\left[b_{n,t}^s, b_{n,t}^s + I_{n,t}^s/2\right]$, we divided $e(b_{n,t}^s)$ by the corresponding interval length $I_{n,t-1}^s/2$ or $I_{n,t}^s/2$.

4 Experiments

Experimental Setup. We adopt the Slakh2100 dataset Manilow et al. (2019), following the common subset of four tracks: bass, drums, guitar, and piano. All audios are resampled to 16 kHz and segmented into 10.24-second clips. We use MusicLDM Chen et al. (2024) as our backbone and adopt the publicly available checkpoints to initialize the shared parameters. We extract beat for each track using madmom library Böck et al. (2016). The detailed comparisons of different parameter settings in beat extraction tools and their impact are provided in the Appendix.

Analysis of Inner Track Stability. We first evaluate the rhythm stability of each generated track. High-quality music must possess a stable beat, especially for rhythmic instruments such as drums. To this end, our proposed IRS metric quantified the variance of inter-beat intervals for each track.

As shown in Table 1, all instruments exhibit low IRS values in ground truth, with percussion (drums) tracks achieving the lowest IRS. This aligns with the intuition that rhythmic instruments are expected to demonstrate greater beat stability. SyncTrack outperforms MSG-LD and MSDM in IRS across all

Table 1: Track-wise IRS and FAD Scores

Track	Metrics	Ground Truth	SyncTrack	MSG-LD	MSDM
Bass	IRS↓	0.015	**0.021**	0.041	0.050
	FAD↓	-	**0.682**	1.050	6.304
Drum	IRS↓	0.005	**0.011**	0.040	0.036
	FAD↓	-	**0.698**	0.980	6.721
Guitar	IRS↓	0.016	**0.024**	0.039	0.034
	FAD↓	-	**1.388**	1.830	4.259
Piano	IRS↓	0.015	**0.023**	0.039	0.046
	FAD↓	-	**1.011**	2.040	5.563

tracks. Notably, the improvement is most significant for drums, demonstrating SyncTrack's superior modeling of rhythmic patterns.

Although FAD compresses detailed information and does not cover all dimensions of music quality, it remains indicative of the overall gap between generated and real music. We observe that lower IRS is consistently accompanied by lower FAD scores.

Analysis of Cross-track Synchronization. To analyze rhythmic consistency across tracks in multi-track music generation. We employ CBS and CBD as evaluation metrics. As shown in Table 2, SyncTrack achieves the best performance, indicating tighter rhythmic synchronization across tracks.

Table 2: Cross-track Synchronization Metrics Scores

Metrics	Ground Truth	SyncTrack	MSG-LD	MSDM
CBS (w=0.15)↑	0.5740	**0.5206**	0.3861	0.4694
CBD (mean)↓	0.2412	**0.2681**	0.3714	0.3127
CBD (std)↓	0.1578	**0.2131**	0.2642	0.2217
CBD (median)↓	0.2066	**0.2258**	0.3545	0.2811
Paired Ratio↑	0.5643	0.7059	**0.7182**	0.3487

A key challenge in multi-track evaluation is the trade-off between synchronization and content richness. Some models (e.g., MSDM) generate many empty tracks, which can artificially improve alignment metrics but at the cost of musical richness. To address this, we introduce the Paired Ratio metric. SyncTrack achieve Paired Ratio of 0.7059, close to MSG-LD (0.7182), and much higher than MSDM (0.3487), demonstrating SyncTrack balances synchronization with generation diversity.

Analysis of Mixture Music Quality. In multi-track music generation, the most widely used metric for evaluating overall audio quality is the FAD computed on the mixture of all tracks. The mixture FAD is closely related to our proposed metrics. When individual tracks are rhythmically stable and well synchronized, the resulting mixture is more musically coherent, leading to lower FAD scores. This trend is consistent with our previous analysis of rhythm stability and synchronization. As shown in Table 3, SyncTrack achieves the lowest FAD.

Table 3: FAD Scores of Mixture Music

Metrics	MSDM	MSG-LD	SyncTrack w/o cross-track attention	SyncTrack
FAD↓	6.55	1.31	1.74	**1.26**

5 Conclusion

Current multi-track audio generation and evaluation systems face significant challenges, particularly in addressing beat stability and synchronization. Our work introduce three specialized assessment metrics, thereby completing the quality evaluation system for generated multi-track music compositions. Furthermore, we introduce a synchronous multi-track waveform music generation model called SyncTrack, significantly improving the quality of multi-track audio generation in terms of the completed evaluation system.

References

Böck Sebastian, Krebs Florian, Widmer Gerhard. Joint Beat and Downbeat Tracking with Recurrent Neural Networks. // ISMIR. 2016. 255–261.

Chen Ke, Wu Yusong, Liu Haohe, Nezhurina Marianna, Berg-Kirkpatrick Taylor, Dubnov Shlomo. Musicldm: Enhancing novelty in text-to-music generation using beat-synchronous mixup strategies // ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2024. 1206–1210.

Dannenberg Roger B. Following an improvisation in real-time // Proc. of ICMC, 1987. 1987.

Dixon Simon. Automatic extraction of tempo and beat from expressive performances // Journal of New Music Research. 2001. 30, 1. 39–58.

Goto Masataka, Muraoka Yoichi. Issues in evaluating beat tracking systems // Working Notes of the IJCAI-97 Workshop on Issues in AI and Music-Evaluation and Assessment. 1997. 9–16.

Kilgour Kevin, Zuluaga Mauricio, Roblek Dominik, Sharifi Matthew. Fr\'echet audio distance: A metric for evaluating music enhancement algorithms // arXiv preprint arXiv:1812.08466. 2018.

Kingma Diederik P, Welling Max. Auto-encoding variational bayes // arXiv preprint arXiv:1312.6114. 2013.

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

Manilow Ethan, Wichern Gordon, Seetharaman Prem, Le Roux Jonathan. Cutting music source separation some Slakh: A dataset to study the impact of training data quality and quantity // 2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA). 2019. 45–49.

Mas-Herrero Ernest, Dagher Alain, Zatorre Robert J. Modulating musical reward sensitivity up and down with transcranial magnetic stimulation // Nature human behaviour. 2018. 2, 1. 27–32.

Mildenhall Ben, Srinivasan Pratul P, Tancik Matthew, Barron Jonathan T, Ramamoorthi Ravi, Ng Ren. Nerf: Representing scenes as neural radiance fields for view synthesis // Communications of the ACM. 2021. 65, 1. 99–106.

Robertson Andrew. Decoding Tempo and Timing Variations in Music Recordings from Beat Annotations. // ISMIR. 2012. 475–480.