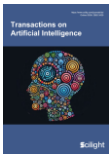
Scilight

Transactions on Artificial Intelligence
https://www.sciltp.com/journals/tai



Article

MIRTracks: A Large-Scale Multi-Dimensional Multi-Track Music Dataset

Yuehan Lee * and Yi Qin *

Music Engineering Department, Shanghai Conservatory of Music, Shanghai 200031, China
* Correspondence: hanklee508@gmail.com (Y.L.); qinyi@shcmusic.edu.cn (Y.Q.)

How To Cite: Lee, Y.; Qin, Y. MIRTracks: A Large-Scale Multi-Dimensional Multi-Track Music Dataset. *Transactions on Artificial Intelligence* **2025**, *1*(1), 282–290. https://doi.org/10.53941/tai.2025.100019

Received: 2 September 2025
Revised: 9 October 2025
Accepted: 27 October 2025
Published: 12 November 2025

Abstract: This paper presents MIRTracks, a large-scale dataset containing 240 h of royalty-free multi-track audio, aiming to address the limitations of traditional music source separation datasets, including single-dimensional annotation and semantic information gaps. By integrating multi-dimensional musical information annotation with a semi-automated annotation pipeline, MIRTracks achieves high-quality semantic annotation across rock, electronic, and pop music genres. Experiments demonstrate that fine-tuning a small-scale model on this dataset significantly improves beat detection accuracy from 66.2% to 80.1%, reaching 91.0% of the performance of large-scale models

Keywords: dataset; annotation; music information retrieval

1. Introduction

For core tasks of Music Information Retrieval (MIR), (such as beat detection) model performance relies heavily on fine-grained rhythmic annotations. However, before delving into the limitations of traditional datasets, it is worth noting that in the broader field of audio synthesis, generative AI models also face significant challenges. While generative AI models have demonstrated strong capabilities in synthesizing audio, their outputs can often lack in long-term structural coherence when compared to music composed by humans.

Unlike human-composed music, which follows clear formal logic (such as verse-chorus-verse structures or tempo modulations that drive emotional progression), AI-generated outputs often lack distinct sectional differentiation or consistent rhythmic development—an issue termed the “structural deficit”. Crucially, this deficit is not merely a limitation of model architectures but also stems from the lack of explicit high-level structural information in training data: traditional datasets rarely encode musical form, tempo transitions, or energy dynamics, leaving models to learn only low-level audio features (e.g., frequency spectra) rather than the “musical grammar” that underpin structural design.

This deficiency in annotation stems from the inherent challenges of MIR labeling: the dual demands of precision and scale render manual annotation highly specialized and labor-intensive, even for experts. Mainstream genres (pop/rock/electronic) intensify this issue—lacking scores and featuring diverse instruments, they force annotators to rely solely on audio analysis, exacerbating the complexity and cost of semantic annotation, thus widening the critical gap between the annotation requirements of advanced MIR models and the practical feasibility of annotation work.

This widening gap between the fine-grained, multi-dimensional annotation required by modern MIR models (to capture rhythmic dynamics, structural logic, and energy trajectories) and the practical limitations of manual labeling—especially for scoreless, instrument-dense mainstream genres like pop, rock, and electronic—has emerged as a critical bottleneck hindering progress in MIR research. Traditional manual workflows cannot scale to meet the data demands of advanced models, while fully automated annotation lacks the semantic accuracy needed to encode high-level musical structure.

To address this, we present MIRTracks, the first large-scale multi-dimensional multi-track dataset tailored for MIR tasks. By constructing a three-dimensional semantic annotation system (rhythmic patterns, structural segmentation, energy distribution) and integrating a semi-automated pipeline of "silence detection—wavelet feature extraction—human verification", we achieve high-quality annotations for mainstream genres including pop, rock, and electronic music. Experiments demonstrate that fine-tuning small-scale models on this dataset improves beat detection accuracy from 66.2% to 80.1%, providing critical data support for MIR research".

The paper is structured as follows: Section 2 surveys existing multi-track datasets. Section 3 details annotation methodologies. Section 4 describes MIRTracks' composition. Section 5 presents *small2* model fine-tuning results. Section 6 concludes and outlines future directions. Section 7 acknowledges contributions from mentors, institutions, and research projects.

2. Related Work

MUSDB18 HQ [1], as a benchmark dataset in the field of music source separation, comprises 150 multi-style complete tracks (100 for training/50 for testing), providing stereo mixes and raw multi-track data (such as isolated tracks for vocals, drums, bass, etc.). It adopts an uncompressed WAV format at 44.1 kHz, supporting the development and evaluation of classic models like OpenUnmix [2]. Its strengths lie in high audio quality and standardized design for separation tasks, with notable performance particularly in the separation of core instruments such as vocals and drum kits. For instance, the SCNet [3] model achieved a Signal-to-Distortion Ratio (SDR) of 9.0 dB on this dataset, significantly outperforming traditional methods.

MedleyDB [4], a high-annotation-density royalty-free multi-track recording dataset, focuses on melody extraction and other research to fill gaps in the field. It contains 196 audio files in 44.1 kHz/16 bit WAV format, with annotations covering dimensions such as melody f0, instrument activation, and music genre. While excelling in melody analysis, it is not designed for mainstream genres like pop, rock, and electronic music, and its total duration of only 80 h imposes limitations on large-scale studies of complex musical styles.

DeepBach [5] established a multi-dimensional annotation framework tailored for analyzing Baroque chorales, focusing on three core music theory dimensions: harmonic function annotation, which classified chords by their roles (e.g., tonic, dominant) and defined their functional relationships in progressions; voice movement trajectories, which mapped pitch intervals in vocal lines and flagged forbidden parallels; and cadence type recognition, which labeled phrase endings to link harmony with formal structure. However, this framework had notable limitations: it was strictly genre-specific to Bach's works, lacking rules for modern styles with complex harmonies or rhythms; it ignored critical dimensions like rhythm and structural segmentation, essential for tasks like beat detection; it required expert knowledge for manual annotation based on scores, limiting scalability to audio data; and it offered no links between music and emotional/cultural contexts, focusing only on technical rules. While pioneering theoretical annotation, its narrow scope and incomplete dimensions restricted broader use.

More recent large-scale audio models, such as OpenAI's Jukebox [6] and Google's MusicLM [7], produce audio with high timbral realism but exhibit variable control over long-term musical form. Their generative process can result in outputs that lack clear sectional differentiation. This suggests that increasing model scale and data volume alone may not be sufficient for models to learn high-level structural concepts efficiently. Our work is motivated by the hypothesis that providing data with explicit structural annotations, such as MIRTracks, can offer a more direct path toward this goal.

3. Detailed Specification of the Dataset

3.1. Dataset Overview and Licensing

The 240-h dataset is distributed under the Creative Commons BY-NC-SA 3.0 license.

3.2. Dataset Metadata

Each track includes artist, title, composer, and download link. Figure 1 shows genre distribution, Figure 2 shows track durations (3–5 min modal). Total multi-track duration is approximately 240 h.

3.3. Musical Annotation Information

The MIRTracks dataset addresses the limitations of single-dimensional annotations in traditional music source separation datasets by constructing a three-dimensional music annotation system. As shown in Figure 3, each track undergoes segment division through silent detection algorithms and manual cross-validation, followed

by automatic tempo/BPM detection using discrete wavelet transform (Daubechies-4 basis). Figure 4 illustrates the energy distribution quantization for each track based on short-time Fourier transform (STFT). Focusing on mainstream genres such as pop, rock, and electronic music (accounting for over 90% of the dataset), the system translates music theory rules into quantifiable annotations.

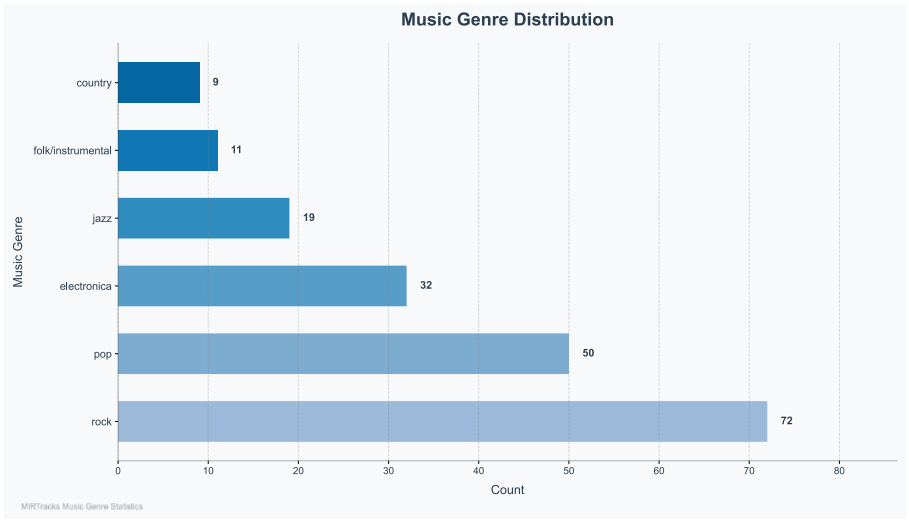


Figure 1. Distribution of music genre.

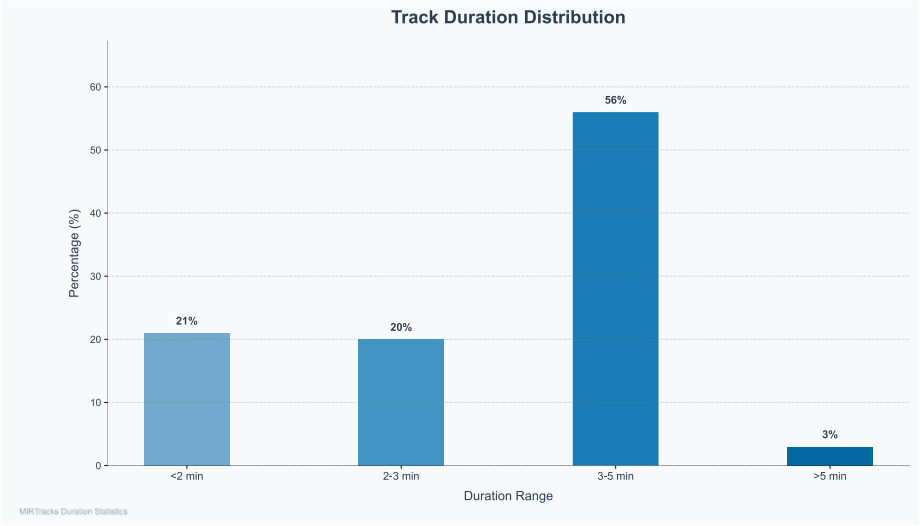


Figure 2. Distribution of music duration.

Notably, this three-dimensional annotation framework is deliberately designed to target the "structural deficit" of generative AI music highlighted in the introduction—i.e., by explicitly encoding the structural logic of human-composed music that traditional datasets omit.

For generative AI models, such data-driven structural annotations solve a key limitation of traditional training: instead of forcing models to infer vague structural patterns from low-level audio alone, MIRTracks provides explicit mappings between "what a section is" (e.g., chorus) and "what it sounds like" (high energy, stable tempo, full instrumentation). This directly addresses the semantic gap between low-level features and high-level structure noted in the introduction, enabling models to generate music with deliberate sectional differentiation—rather than disjointed texture—by learning from MIRTracks' quantified structural logic.

The MIRTracks dataset, including multi-track audio files, annotation files, and preprocessing scripts, is publicly available at https://github.com/bigblackLee123/MIRTracks.git.

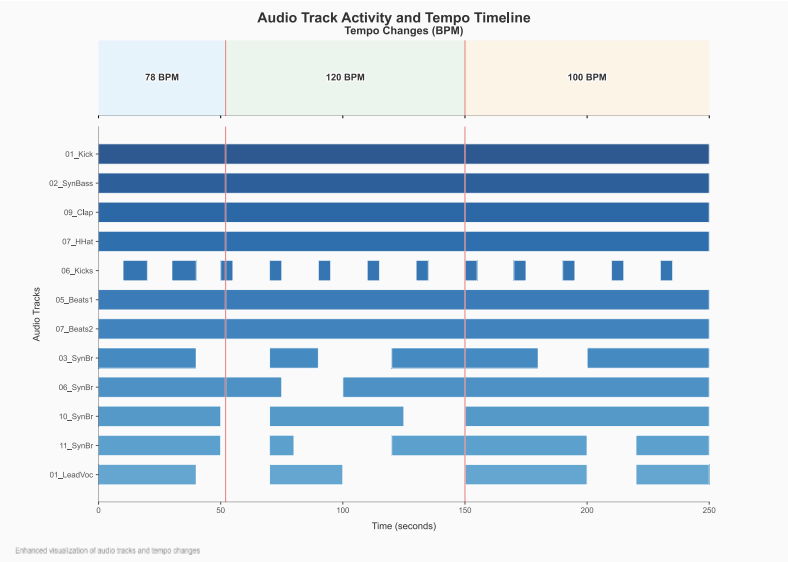


Figure 3. Example of segment division and BPM detection for a musical piece. The top section displays three tempo regions: 78 BPM (0–52 s), 120 BPM (52–150 s), and 100 BPM (150–250 s). The main section shows track activity as blue blocks, with rhythmic tracks maintaining consistent presence, synthesizer tracks exhibiting sectional patterns, and vocals displaying intermittent activity. Red vertical lines mark tempo transition points (52 s and 150 s), highlighting structural boundaries.

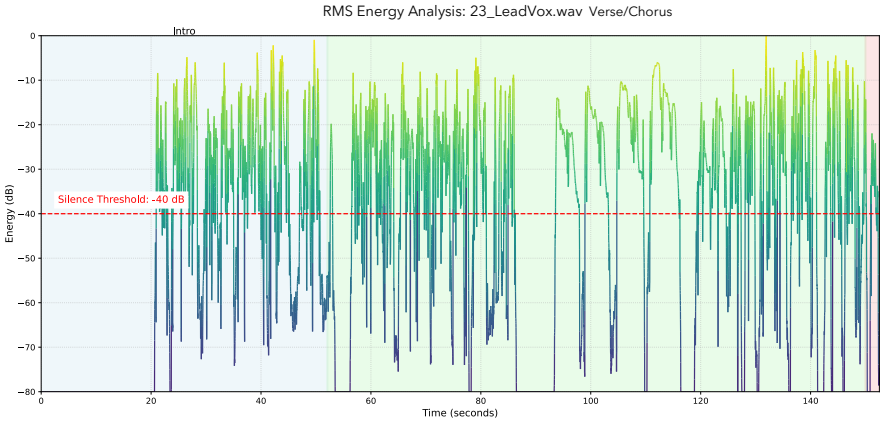


Figure 4. Example of vocal track energy distribution in a music piece. This graph depicts the vocal track's energy distribution over time. The yellow curve shows energy variations (−80 dB to 0 dB), with the red dashed line indicating the −40 dB silence threshold. Background colors correspond to Figure 1's tempo regions, establishing correlation between vocal energy patterns and rhythmic structure.

4. Annotation Task Definition

A core advantage of MIRTracks lies in formalizing key aspects of musical composition into actionable annotation tasks, paired with a semi-automated pipeline to balance accuracy and efficiency. Below is the technical definition of each annotation dimension and the detailed implementation process:

4.1. Annotation Task Definition

- Tempo: Defined as the number of beats per unit time (usually per minute), calculated by decomposing the audio signal into detail components of varying scales via wavelet transform. The tempo is reflected by frequency components corresponding to beats with significant periodic characteristics within specific frequency ranges of these detail components.
- Segment: In this study, segments are defined based on audio presence/absence segmentation. A segment is a continuous audio fragment divided by the presence or absence of sound signals, characterized by relatively independent musical features.

4.2. Automated Annotation Process

Segment information is extracted and aggregated via silent detection for each multi-track. Through frame-by-frame audio processing, continuous audio segments (active) and silent segments (inactive) are identified and labeled. For each track, this process sequentially marks the start and end times of segments, which are then aggregated to form the complete segmentation information for that track.

Audio segments corresponding to these time points are extracted, and discrete wavelet transform (DWT) is applied to calculate audio tempo.

Take the drum track in a multi-instrument mix (featuring drums, guitar, and vocals) as an example: during interludes, it has long silent segments; during verses/choruses, active segments with kick/snare beats drive rhythm.

After extracting audio segments corresponding to these time points, discrete wavelet transform (DWT) is applied to decompose the signal at multiple scales, separating high-frequency detail features (corresponding to beat periodicity) from low-frequency approximation features (reflecting overall tonal trends).

Given a musical signal $x(n)$, discrete wavelet transform decomposes it into approximation components $cAj(n)$ and detail components $cDj(n)$:

$$\begin{cases} cA_j(n) = \sum_k h(k) \cdot cA_{j-1}(2n-k) \\ cD_j(n) = \sum_k g(k) \cdot cA_{j-1}(2n-k) \end{cases} \tag{1}$$

This study employs 4-level decomposition ($j = 4$) using the Daubechies-4 wavelet basis, with low-pass filter coefficients:

$$h = [0.4829629131, 0.8365163037, 0.2241438680, -0.1294095225]$$

The autocorrelation function of the detail component $cD(n)$ is defined as:

$$R(\tau) = \sum_n cD(n) \cdot cD(n+\tau) \tag{2}$$

The *BPM* (beats per minute) is calculated as follows:

$$BPM = \frac{60 \cdot f_s}{2^{j-1} \cdot \Delta\tau} \tag{3}$$

STFT computes:

$$E(f) = \left| \sum_{\mathrm{n}} \mathrm{cD(n)} \cdot \mathrm{e}^{-\mathrm{i}2\pi \mathrm{fn}} \right|^2 \tag{4}$$

Energy distribution analysis in 40–220 *BPM* localizes the dominant rhythmic frequency and total energy.

4.3. Manual Annotation Process

Audio files are converted into waveform plots for visualization and editing, with segment information and wavelet algorithm detection results loaded simultaneously. Active but musically irrelevant segments (e.g., environmental sound samples) are removed from the audio.

Annotations are created by two annotators, both holding bachelor's degrees in music. Each annotation is evaluated by one annotator and verified by the other.

5. Benchmark Experiments

5.1. Experiment Overview

This study employs BEAT THIS as the benchmark model for beat recognition tasks. Innovatively removing traditional dynamic Bayesian network [8] postprocessing, the model achieves state-of-the-art (SOTA) performance (GTZAN [9] test set Beat F1 = 89.1%) on a multi-dimensional annotated dataset using a convolutional-Transformer hybrid architecture and displacement-tolerant loss function. Its open-source nature (code, pre-trained models) provides a feasible framework for small-scale model optimization. This experiment focuses on comparing BEAT THIS's *final0* (large model) and *small2* (small model) under the same framework,

evaluating the effectiveness of fine-tuning *small2* using MIRTracks through architecture, performance, differences between models of varying scales in this task. Core architectural differences are detailed in Table 1.

Table 1. Model difference.

Core Differences Between *Final0* and *Small2*		
Characteristic Model	*Final0*	*Small2*
Parameters	20 M	2 M
Architecture	conv + 6 × transformer	lite transformer
Input	128-bin Mel spectrogram	Simplified features

5.2. Experimental Design

- Model Configuration: The *small2* model from the BEAT THIS framework is selected for comparative experiments with the *final0* model (20 M parameters), aiming to explore performance differences between models of varying scales in this task.
- Training Data: A subset of the MIRTracks dataset is used, containing 100 songs with 146 kick drum tracks. These songs are selected in a style distribution of 42% rock, 28% electronic, and 25% pop to ensure diversity and representativeness.
- Validation Set: The DSD100 [10] dataset serves as the validation set, with 5 songs manually annotated for beats. These annotations provide a reliable reference for model performance evaluation.
- Evaluation Metric: Beat recognition accuracy is used as the core metric, calculated as the ratio of matched beat points to total beat points. This metric intuitively reflects the model's performance in beat recognition tasks
- Training Parameters:
  (1) Optimizer: The Adam optimizer is employed with a learning rate of 1×10^{-5}, leveraging its adaptive learning rate adjustment to enhance training efficiency.
  (2) Batch Size: A batch size of 16 is set to balance training efficiency and model generalization.
  (3) Epochs: The model is trained for 15 epochs to ensure sufficient feature learning through iterative training.
  (4) Data Augmentation: Two strategies are applied:
    (a) Time Stretching (±10%): Simulates tempo variations.
    (b) Frequency Masking (2 masks, 15% frequency coverage): Improves feature robustness.

5.3. Experimental Procedures

Data Preprocessing: 100 songs were randomly selected from the MIRTracks dataset according to the predefined style distribution (42% rock, 28% electronic, and 25% pop), and their kick drum tracks were extracted. The preprocess module in the beat_this framework was applied to generate spectrogram data in .npy format (i.e., NumPy arrays storing time-frequency energy values, with their visualization presented in Figure 5), which converts audio signals into time-frequency representations for model input.

- signals into time-frequency representations for model input. Subsequently, the previously described data augmentation strategies were applied to the spectrograms, and the processed data were packaged into a fine-tuning dataset.
- Model Training: The pre-trained *small2* model was loaded as the initial model for fine-tuning. The model was trained on the fine-tuning dataset for 15 epochs while recording metrics such as loss and accuracy. Training curves (Figure 6) visually demonstrated performance trends during training, aiding in convergence and stability analysis.
- Performance Evaluation: The trained *final0* model, untuned *small2* model, and fine-tuned *small2* model were applied to 5 manually annotated songs in the DSD100 validation set for beat recognition. Beat recognition accuracy was calculated by comparing model predictions with manual annotations. Results are presented in Table 1, clearly showing performance differences across models.

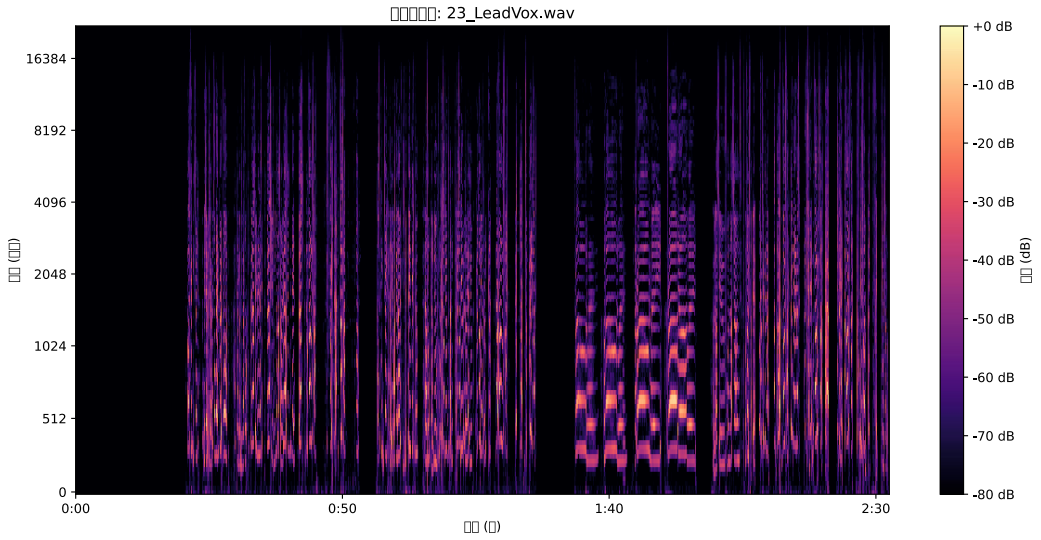


Figure 5. Spectrograms in .npy format.

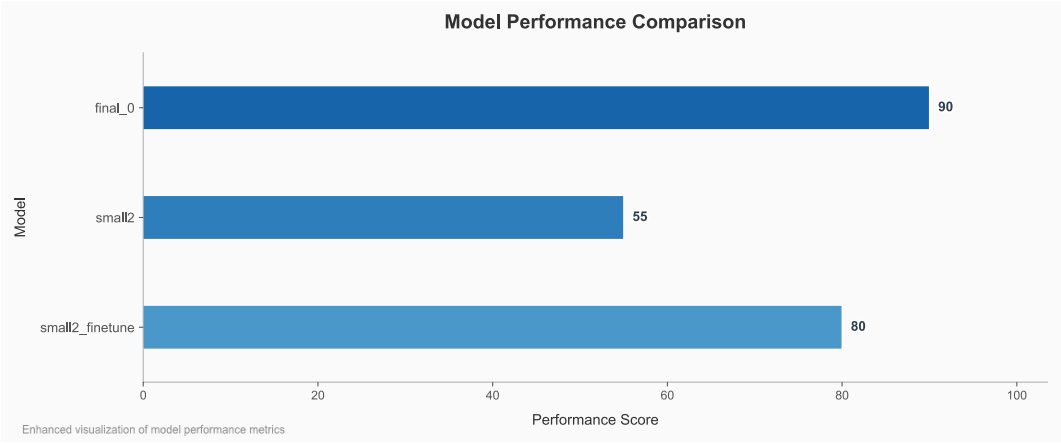


Figure 6. Accurancy of models.

5.4. Result Analysis

The experimental results demonstrate that fine-tuning the *small2* model with MIRTracks significantly improves its beat recognition performance (visualization shown in Figure 6). The fine-tuned *small2* achieves an accuracy of 80.1%, a 13.9% increase from its original 66.2% and 91.0% of the *final0* model's accuracy (88.0%)

- Substantial Performance Improvement: The fine-tuned *small2* model achieved 91.0% of *final0*'s beat recognition accuracy, validating the effectiveness of using MIRTracks' multi-track audio for small model optimization. This highlights the rich information in multi-track data for enhancing small model performance in beat recognition tasks.
- Generalization Validation: On the DSD100 validation set (unused during training), the fine-tuned *small2* demonstrated a significantly reduced standard deviation (0.9 vs. 2.5), indicating superior generalization stability across unseen datasets.
- Efficiency Advantage: With only 14% of *final0*'s parameters, the fine-tuned *small2* achieved near-state-of-the-art performance, proving that small models combined with appropriate training data and strategies can balance efficiency and performance for resource-constrained applications.

6. Conclusions and Future Work

6.1. Conclusions

Generative AI models often excel at local musical texture synthesis but struggle with long-term structural coherence—a structural deficit' linked to the lack of explicit musical form encoding in traditional datasets. This study introduces MIRTracks, a large-scale dataset addressing the limitations of traditional music source separation

datasets through its three-dimensional musical semantic annotation framework (tempo/BPM, segmentation, energy distribution).

Experimental results validate the dataset's effectiveness in improving small model performance: fine-tuning *small2* on MIRTracks increases beat recognition accuracy from 66.2% to 80.1%, achieving 91.0% of the *final0* model's performance (88%). This improvement not only demonstrates the dataset's ability to enhance low-level rhythm processing but also validates that structural annotations enable models to learn context-aware musical grammar—reducing performance variance (from 2.5 to 0.9) by moving beyond surface-level pattern replication. Key innovations include:

- Annotation System Innovation: First integration of rhythmic morphology and structural analysis into a computable three-dimensional framework, providing theoretical constraints for music generation models.
- Workflow Optimization: Semi-automated "silence detection-DWT feature extraction-human verification" process ensures annotation quality while reducing manual effort.
- Style Coverage Expansion: Focus on mainstream genres (pop/rock/electronic) with 240 h of data—three times the scale of MedleyDB—supporting large-scale music analysis.

These innovations collectively address the semantic gap in music representation: the three-dimensional framework encodes high-level musical form, the semi-automated workflow balances quality and scalability, and expanded style coverage ensures relevance to contemporary music practices.

6.2. Future Work

Despite these contributions, several directions for improvement remain:

- Dimension Expansion: Add harmonic progressions, melodic contours, and instrumentation details to complement existing rhythmic, structural, and energy annotations—forming a holistic music-theoretic framework that aligns with human compositional logic.
- Algorithm Enhancement: Reduce DWT-based rhythm detection errors (10.2%) using Transformer-based deep learning methods. This refinement aims to reduce errors in tempo variation detection, a critical component of rhythmic structure annotation that directly impacts models' ability to track musical section transitions.
- Multimodal Integration: Explore associations between lyrics, emotions, and multi-track data for cross-modal music understanding.
- Generalization Validation: Verify dataset effectiveness in additional MIR tasks (e.g., source separation, chord recognition).
- Ecosystem Development: Expand coverage to underrepresented genres (classical/jazz) through community-driven data collection. ensuring MIRTracks evolves as a inclusive resource that supports structural analysis across diverse musical traditions.

Future research will focus on:

- Developing contrastive learning frameworks for aligning music theory rules with audio features.
- Building interpretable music generation models guided by MIRTracks' structural annotations.

These directions aim to translate MIRTracks' structural annotations into actionable controls for AI music generation, enabling models to produce long-form pieces with deliberate formal development—directly addressing the 'structural deficit' identified in the introduction.

Through continuous refinement, MIRTracks aims to provide a universal tool for music information retrieval and AI composition, fostering deeper integration between music technology and artistic creation. In summary, MIRTracks is more than just a large dataset; it represents a paradigm shift from purely audio-centric data collection to a structured, semantic approach. By offering dense, multi-dimensional annotations for mainstream music, we provide the foundational data infrastructure necessary to unlock the next generation of structurally coherent AI models, fundamentally advancing both Music Information Retrieval and the creative capabilities of AI composition.

Author Contributions

Y.Q.: conceptualization, supervision; Y.L.: software, data curation, writing—original draft preparation, methodology. All authors have read and agreed to the published version of the manuscript.

Funding

This research was supported by the project entitled "Multimodal Emotion Analysis and Mapping Based on Music-Visual Synesthesia".

Data Availability Statement

All methods for obtaining research data and annotation tools have been fully open-sourced at: https://github.com/bigblackLee123/MIRTracks.git.

Conflicts of Interest

The authors declare no conflict of interest.

Use of AI and AI-Assisted Technologies

During the preparation of this work, the authors used Cursor to assist with code writing, and used Doubao to assist with translation work. After using these tools/services, the authors reviewed and edited the content as needed and takes full responsibility for the content of the published article.

References

1. Rafii, Z.; Liutkus, A.; Stöter, F.R.; et al. MUSDB18—A corpus for music separation. *arXiv preprint* **2017**, arXiv:1710.11192.
2. Stöter, F.R.; Uhlich, S.; Liutkus, A.; et al. Open-Unmix: A Reference Implementation for Music Source Separation. *J. Open Source Softw.* **2019**, *4*, 1667.
3. Tong, W.; Zhu, J.; Chen, J.; et al. SCNet: Sparse compression network for music source separation. *arXiv preprint* **2024**, arXiv:2401.13276.
4. Bittner, R.M.; Salamon, J.; Tierney, M.; et al. A Multitrack Dataset for Annotation-Intensive MIR Research. In Proceedings of the 15th International Society for Music Information Retrieval Conference (ISMIR), Taipei, Taiwan, 27–31 October 2014; pp. 155–160.
5. Hadjeres, G.; Pachet, F.; Nielsen, F. DeepBach: A Steerable Model for Bach Chorales Generation. In Proceedings of the 34th International Conference on Machine Learning (PMLR), Sydney, Australia, 6–11 August 2017; pp. 1362–1371.
6. Dhariwal, P.; Jun, H.; Payne, C.; et al. Jukebox: A generative model for music. *arXiv e-print* **2020**, arXiv:2005.00341.
7. Agostinelli, A.; Denk, T.I.; Borsos, Z.; et al. MusicLM: Generating music from text. *arXiv preprint* **2023**, arXiv:2301.11325.
8. Foscarin, F.; Schlüter, J.; Widmer, G. Beat this! Accurate beat tracking without DBN postprocessing. *arXiv preprint* **2024**, arXiv:2407.21658.
9. Sturm, B.L. The GTZAN dataset: Its contents, its faults, their effects on evaluation, and its future use. *arXiv preprint* **2013**, arXiv:1306.1461.
10. Liutkus, A.; Stöter, F.-R.; Rafii, Z.; et al. The 2016 Signal Separation Evaluation Campaign. In *Latent Variable Analysis and Signal Separation, Proceedings of the 12th International Conference, LVA/ICA 2015, Liberec, Czech Republic, 25–28 August 2015; pp. 323–332*; Tichavský, P., Babaie-Zadeh, M., Michel, O.J.J.; et al., Eds.; Springer International Publishing: Cham, Switzerland, 2017. https://doi.org/10.1007/978-3-319-19544-2_31.