CLAM: Safeguarding Authenticity and Addressing Implications for the Music Industry

Arnesh Batra[1†] **Krish Thukral**[2†] **Naman Batra**[3] **Dev Sharma**[1]
Ruhani Bhatia[1] **Aditya Gautam**[1]

[1]Indraprastha Institute of Information Technology Delhi (IIIT-Delhi), India
[2]Manipal University Jaipur, Rajasthan, India
[3]Netaji Subhas University of Technology (NSUT), Delhi, India

[†] Equal contribution

Abstract

The rapid rise of end-to-end AI music generation has introduced escalating challenges in detecting deepfake artist voice cloning, synthetic lyrics, and AI-altered acoustics, raising critical concerns around copyright, authenticity, and the preservation of human artistic expression. While early benchmarks like SONICS and models such as SpecTTTra offered foundational progress in identifying AI-generated songs, their reliance on a small number of generation sources limits generalization to newer models. In practice, we observe significant performance drops when applying existing detectors to out-of-distribution content from emerging generators such as Riffusion and Yue. To address this, we introduce Melody or Machine (MoM), a comprehensive benchmark dataset of 130,435 songs (6665.13 hours) synthesized using a diverse set of models and pipelines with controlled variations. MoM is curated to promote the development of robust, generalizable detection systems. Alongside this dataset, we introduce CLAM (Contrastive Learning for Audio Matching), a novel detection model that combines two pre-trained encoders: MERT for extracting expressive musical features and Wav2Vec2 for capturing vocal nuances. Their embeddings are fused through a learnable cross-aggregation module that extracts the learnable features from the intermediate layers. CLAM is trained using a combination of binary cross-entropy loss for detection and a triplet loss to align coherent music-vocal pairs in embedding space. CLAM achieves a state-of-the-art F1 score of 0.993 on the SONICS benchmark and 0.925 on our internal MoM dataset, setting a new state of the art in synthetic music forensics.

1 Introduction

The rise of AI-generated songs, often indistinguishable from real ones, has fueled their popularity on social media and raised concerns for reliable detection. Without effective solutions, this trend threatens artistic integrity, vocal authenticity, and development of new talent. While early Singing Voice Synthesis (SVS) and Conversion (SVC) works liu2021diffsinger, tae2021mlp, xu2022refineganuniversallygeneratingwaveform, along with datasets like SingFake [Zang et al., 2024b] and CtrSVDD [Zang et al., 2024a], initiated research in Singing Voice Deepfake Detection (SVDD), they suffered from short clips, karaoke artifacts, and limited diversity. The SONICS dataset [Rahman et al., 2025] addressed some issues with 97k long-form songs and the SpecTTTra model, but challenges such as male vocal bias and lack of multilingual coverage remain.

To overcome these gaps, we introduce **MoM**, a dataset of **64,960 AI-generated** and **65,475 real songs** spanning multiple languages and genres. Unlike prior work, MoM greatly expands the number

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

of fully synthetic tracks (53,922 vs. 2,173 in SONICS) and ensures out-of-distribution evaluation by separating train and test models.

We also propose **CLAM** (**C**ontrastive **L**earning for **A**udio **M**atching), a detection model that fuses pre-trained **MERT** [Li et al., 2024] and **Wav2Vec2** [Baevski et al., 2020] representations through cross-aggregation. Beyond binary classification, CLAM employs a contrastive alignment objective with triplet loss to capture human-like musical-vocal coherence. Trained at 24kHz, CLAM achieves state-of-the-art results with **F1 = 0.993** on SONICS and **0.931** on MoM.

Our main contributions are as follows:

- We introduce **MoM**, a large-scale and diverse benchmark with 130k+ real and AI-generated songs, offering significant improvements in scale, diversity, and out-of-distribution robustness over prior datasets.
- We propose **CLAM**, a novel dual-encoder detection model with contrastive alignment, achieving state-of-the-art results on both in-distribution and out-of-distribution benchmarks.
- Together, MoM and CLAM establish a strong foundation for future research in singing voice deepfake detection.

2 Impacts and Benefits

MoM and CLAM aim to increase transparency by providing tools to detect AI-generated content. These capabilities offer significant social and legal benefits, allowing listeners to understand the origins of a song while enabling rights holders to safeguard intellectual property. For musicians, it prevents synthetic media from being passed off as human-made, preserving human artistry and fair competition. However, this technology also presents critical ethical and practical challenges. There is a significant risk that the dataset could be used to train more sophisticated deep-fake generators. Furthermore, reliance on automated systems could lead to false positives and algorithmic bias, which would harm emerging artists or those with unconventional styles. Thus, careful deployment and oversight are essential to minimize these risks.

3 Methodology

3.1 CLAM: Contrastive Learning for Audio Matching

CLAM is a two-encoder architecture designed to distinguish between authentic and AI-generated music. The model's core hypothesis is that the organic alignment between vocal and instrumental components in real songs is difficult for generative models to replicate convincingly. In authentic recordings, a singer's phrasing adapts to the instrumental backing, and the accompaniment responds to the vocal dynamics, creating a perceptually unified piece. Synthetic compositions, in contrast, often exhibit subtle mismatches—such as vocals that are rhythmically displaced or emotionally detached from the music. CLAM is designed to capture these discrepancies by learning an embedding space where the components of real songs are closely aligned, while those of fake or mismatched songs are repelled.

3.1.1 Feature Extraction

To obtain rich, modality-specific representations, we leverage two powerful, self-supervised models as feature encoders.

Music Feature Extraction with MERT. For instrumental stems, we use MERT (Music undERstanding model with large-scale self-supervised Training). As a state-of-the-art model for musical understanding, MERT is pre-trained on a massive dataset using a masked-language-modeling style objective, enabling it to learn robust and generalizable features related to harmony, rhythm, and timbre.

Vocal Feature Extraction with Wav2Vec2. For vocal stems, we employ a fine-tuned **wav2vec2-base** model. This model was selected after an extensive ablation study comparing several alternatives.

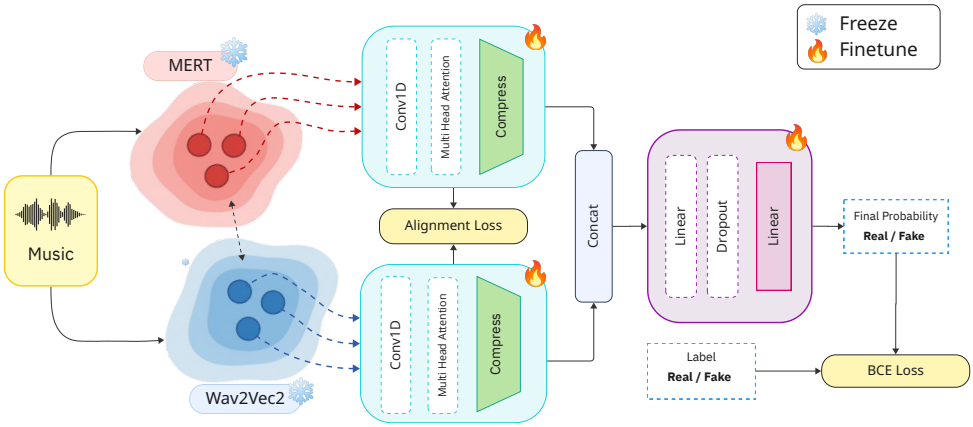


Figure 1: Overview of the CLAM (Contrastive Learning for Audio Matching) architecture. This two-encoder model processes instrumental and vocal embeddings (from sources like MERT and Wav2Vec2), incorporates a Weighted Cross-Aggregation (WCA) module, and a dual-loss objective.

Wav2Vec2 demonstrated the strongest generalization to diverse vocal characteristics (e.g., timbre, articulation), making it highly effective for capturing the nuanced details of human singing.

3.1.2 Weighted Cross-Aggregation (WCA)

The Weighted Cross-Aggregation (WCA) module fuses multi-level features from the music (MERT) and vocal (Wav2Vec2) encoders to capture their dependencies.

First, a lightweight 1D convolution computes learnable weighted sums of the encoder layers, producing three compact feature maps for each modality. This yields two sets of aggregated representations: $\overline{\mathbf{M}}$ for music and $\overline{\mathbf{V}}$ for vocals.

Next, to preserve intra-modal structure, we apply separate multi-head self-attention on each set of aggregated features. This refines the representations without prematurely mixing modalities.

Finally, the attention-enhanced maps are averaged within each modality and projected through a linear layer, resulting in fixed-size embeddings $\mathbf{d}_{\text{music}}$ and $\mathbf{d}_{\text{vocal}}$.

3.1.3 Contrastive Alignment and Loss Objective

To effectively distinguish between real and AI-generated content, we train the model with a dual objective that combines a standard classification loss with a contrastive alignment loss. Contrastive learning is ideal for this task, as it teaches the model to create an embedding space where similar instances are clustered together and dissimilar ones are pushed apart.

After evaluating several alignment functions, we found that **Triplet Loss** yielded the best performance. The loss is defined as:

$$\mathcal{L}_{\text{triplet}}(e_a, e_p, e_n) = \max(0, D(e_a, e_p)^2 - D(e_a, e_n)^2 + \alpha) \tag{1}$$

where $D(\cdot, \cdot)$ is the Euclidean distance, $\alpha > 0$ is a predefined margin, and the squared distance is used for gradient stability. Triplets are constructed using an in-batch mining strategy for real songs:

- **Anchor (e_a):** The instrumental embedding of a real track, $E_I^{\text{real}}[i]$.
- **Positive (e_p):** The vocal embedding of the **same** real track, $E_V^{\text{real}}[i]$.
- **Negative (e_n):** A vocal embedding from a **different** real track in the batch, $E_V^{\text{real}}[j]$ (where $j \neq i$).

This process encourages the model to recognize and preserve the tight coupling of authentic vocal-instrumental pairs.

The final training objective is a weighted sum of the **Binary Cross-Entropy (BCE)** loss for classification and the Triplet Loss for alignment:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{BCE}} + \lambda \cdot \mathcal{L}_{\text{triplet}} \tag{2}$$

where λ is a scalar hyperparameter that balances the contribution of the two loss components.

Table 1: Performance comparison on MoM dataset.

Method	**Accuracy**	**F1 Score**	**Recall**	**Precision**
SpecTTTra				
Spectraa SpecTTTra-α (120 sec)	0.872	0.869	0.849	0.888
Unimodal Only				
MERT (No WCA)	0.835	0.813	0.734	0.918
Wav2Vec2 (No WCA)	0.827	0.802	0.712	0.891
MERT Only	0.861	0.853	0.768	0.943
Wav2Vec2 Only	0.831	0.846	0.746	0.925
Multimodal (No Alignment Loss)				
CLAM (No Alignment Loss)	0.913	0.906	0.832	0.993
Multimodal + Alignment Losses				
CLAM (MSE Loss)	0.916	0.908	0.837	0.993
CLAM (Huber Loss (δ=1.0))	0.918	0.911	0.841	0.994
CLAM (Triplet Loss)	**0.931**	**0.925**	**0.869**	**0.989**

4 Experiments and Results

All experiments were conducted using an NVIDIA RTX 4060 Ti 16GB GPU with 32 GB of RAM, totaling approximately 400 compute hours. Models were trained using the AdamW [Loshchilov and Hutter, 2019]optimizer with a learning rate of 1e-4. Each model (CLAM) uused 512-d embeddings and 4 attention heads. All audio samples were trimmed to 90 s and resampled to 24 kHz to balance retention and efficiency. Training was performed for 50 epochs with a batch size of 16, a validation split of 0.2, and a fixed seed of 42 for reproducibility. A curated test set with out-of-distribution (OOD) samples was used to assess generalization. For all cross-stem models employing alignment loss, the alignment weight was fixed at 0.5. We use the **F1 score** as the primary metric, given its ability to reflect both precision and recall, especially important for evaluating robustness under OOD conditions where accuracy alone can be misleading [Christen et al., 2023]. We compare unimodal baselines, a cross-stem model without alignment loss, and variants trained with different alignment objectives. As shown in Table 1, all alignment-based models outperform the no-alignment baseline. The Triplet-Loss model achieves 0.925 F1 (ours) and 0.993 on SONICS, showing strong cross-stream semantic representations. All the SpecTTTra models were trained with the hyperparameters described in the SONICS [Rahman et al., 2025] paper. We have provided all the training, testing, and validation code in the supplementary material.

5 Discussion and Conclusion

We introduced the MoM dataset and the CLAM detection model for AI-generated music. MoM comprises 130,435 songs (evenly split between real and synthesized), spanning multiple languages and gender-balanced content to address known gaps in dataset diversity. CLAM employs modality-specific encoders, a dual-loss training objective, and a Weighted Cross-Aggregation module to capture misalignment between vocals and instrumentals – a cue that distinguishes real from synthetic songs. Our experiments show that CLAM attains accuracy comparable to state-of-the-art music deepfake detectors while requiring significantly less computation. Through releasing the MoM dataset, we aim to catalyze more research focused on detecting AI involvement in music production. Across both OOD and in-domain settings, cross-stem models consistently outperform unimodal baselines such as MERT and Wav2Vec2.

References

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations, 2020. URL https://arxiv.org/abs/2006.11477.

Peter Christen, David J. Hand, and Nishadi Kirielle. A review of the f-measure: Its history, properties, criticism, and alternatives. *ACM Comput. Surv.*, 56(3), October 2023. ISSN 0360-0300. doi: 10.1145/3606367. URL https://doi.org/10.1145/3606367.

Michaël Defferrard, Kirell Benzi, Pierre Vandergheynst, and Xavier Bresson. FMA: A dataset for music analysis. In *18th International Society for Music Information Retrieval Conference (ISMIR)*, 2017. URL https://arxiv.org/abs/1612.01840.

Michaël Defferrard, Sharada P. Mohanty, Sean F. Carroll, and Marcel Salathé. Learning to recognize musical genre from audio. In *The 2018 Web Conference Companion*. ACM Press, 2018. ISBN 9781450356404. doi: 10.1145/3184558.3192310. URL https://arxiv.org/abs/1803.05337.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. URL https://arxiv.org/abs/1810.04805.

Yuan Gong, Yu-An Chung, and James Glass. Ast: Audio spectrogram transformer, 2021. URL https://arxiv.org/abs/2104.01778.

Google. Gemini 2.0: Flash, flash-lite and pro. https://developers.googleblog.com/en/gemini-2-family-expands/, Feb 2025. Accessed: May 15, 2025.

Yinlin Guo, Haofan Huang, Xi Chen, He Zhao, and Yuehai Wang. Audio deepfake detection with self-supervised wavlm and multi-fusion attentive classifier, 2024. URL https://arxiv.org/abs/2312.08089.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units, 2021. URL https://arxiv.org/abs/2106.07447.

Yizhi Li, Ruibin Yuan, Ge Zhang, Yinghao Ma, Xingran Chen, Hanzhi Yin, Chenghao Xiao, Chenghua Lin, Anton Ragni, Emmanouil Benetos, Norbert Gyenge, Roger Dannenberg, Ruibo Liu, Wenhu Chen, Gus Xia, Yemin Shi, Wenhao Huang, Zili Wang, Yike Guo, and Jie Fu. Mert: Acoustic music understanding model with large-scale self-supervised training, 2024. URL https://arxiv.org/abs/2306.00107.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019. URL https://arxiv.org/abs/1711.05101.

Lam Pham, Phat Lam, Truong Nguyen, Huyen Nguyen, and Alexander Schindler. Deepfake audio detection using spectrogram-based feature and ensemble of deep learning models, 2024. URL https://arxiv.org/abs/2407.01777.

Orchid Chetia Phukan, Gautam Siddharth Kashyap, Arun Balaji Buduru, and Rajesh Sharma. Heterogeneity over homogeneity: Investigating multilingual speech pre-trained models for detecting audio deepfake, 2024a. URL https://arxiv.org/abs/2404.00809.

Orchid Chetia Phukan, Drishti Singh, Swarup Ranjan Behera, Arun Balaji Buduru, and Rajesh Sharma. Investigating prosodic signatures via speech pre-trained models for audio deepfake source attribution, 2024b. URL https://arxiv.org/abs/2412.17796.

Aref Farhadi Pour, Mohammad Asgari, and Mohammad Reza Hasanabadi. Gammatonegram based speaker identification. In *2014 4th International Conference on Computer and Knowledge Engineering (ICCKE)*, pages 52–55, 2014. doi: 10.1109/ICCKE.2014.6993383.

Md Awsafur Rahman, Zaber Ibn Abdul Hakim, Najibul Haque Sarker, Bishmoy Paul, and Shaikh Anowarul Fattah. SONICS: Synthetic or not - identifying counterfeit songs. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=PY7KSh29Z8.

Falguni Sharma and Priyanka Gupta. Deepfake detection of singing voices with whisper encodings, 2025. URL https://arxiv.org/abs/2501.18919.

DongJae Shin, HyeonSeok Lim, Inho Won, ChangSu Choi, Minjun Kim, SeungWoo Song, HanGyeol Yoo, SangMin Kim, and KyungTae Lim. X-llava: Optimizing bilingual large vision-language alignment. In *Findings of the Association for Computational Linguistics: NAACL 2024*, page 2463–2473. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-naacl.158. URL http://dx.doi.org/10.18653/v1/2024.findings-naacl.158.

Zhecan Wang, Junzhang Liu, Chia-Wei Tang, Hani Alomari, Anushka Sivakumar, Rui Sun, Wenhao Li, Md. Atabuzzaman, Hammad Ayyubi, Haoxuan You, Alvi Ishmam, Kai-Wei Chang, Shih-Fu Chang, and Chris Thomas. Journeybench: A challenging one-stop vision-language understanding benchmark of generated images, 2025. URL https://arxiv.org/abs/2409.12953.

Taiba Majid Wani, Syed Asif Ahmad Qadri, Danilo Comminiello, and Irene Amerini. Detecting audio deepfakes: Integrating cnn and bilstm with multi-feature concatenation. In *IHMMSec*, pages 271–276, 2024a. URL https://doi.org/10.1145/3658664.3659647.

Taiba Majid Wani, Syed Asif Ahmad Qadri, Danilo Comminiello, and Irene Amerini. Detecting audio deepfakes: Integrating cnn and bilstm with multi-feature concatenation. In *IHMMSec*, pages 271–276, 2024b. URL https://doi.org/10.1145/3658664.3659647.

Yuankun Xie, Jingjing Zhou, Xiaolin Lu, Zhenghao Jiang, Yuxin Yang, Haonan Cheng, and Long Ye. Fsd: An initial chinese dataset for fake song detection, 2023. URL https://arxiv.org/abs/2309.02232.

Rui Yang, Lin Song, Yanwei Li, Sijie Zhao, Yixiao Ge, Xiu Li, and Ying Shan. Gpt4tools: Teaching large language model to use tools via self-instruction, 2023. URL https://arxiv.org/abs/2305.18752.

Jiangyan Yi, Chenglong Wang, Jianhua Tao, Xiaohui Zhang, Chu Yuan Zhang, and Yan Zhao. Audio deepfake detection: A survey, 2023. URL https://arxiv.org/abs/2308.14970.

Yongyi Zang, Jiatong Shi, You Zhang, Ryuichi Yamamoto, Jionghao Han, Yuxun Tang, Shengyuan Xu, Wenxiao Zhao, Jing Guo, Tomoki Toda, and Zhiyao Duan. CtrSVDD: A benchmark dataset and baseline analysis for controlled singing voice deepfake detection. In *Proc. Interspeech*, pages 4783–4787, 2024a. doi: 10.21437/Interspeech.2024-2242.

Yongyi Zang, You Zhang, Mojtaba Heydari, and Zhiyao Duan. Singfake: Singing voice deepfake detection. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2024b.

A Technical Appendices and Supplementary Material

Table 2: SpecTTTra OOD F1 Scores

Dataset	Total Samples	F1 Score (%)
Riffusion	7,057	53.46
Yue	5,278	68.80
Voice clone	1,166	50.94
Suno 4	48	64.58
Total	13,549	**59.45**

Table 3: Qualitative Comparison of Datasets

Dataset	Fully Fake Songs	Text Lyrics Songs	Diverse Style Songs	Closed Source Models	Open Source Models	Multilingual Songs	
FSD [Xie et al., 2023]	-	-	-	-	-	✓	-
SingFake [Zang et al., 2024b]	-	-	-	✓	✓	✓	
CtrSVDD [Zang et al., 2024a]	-	-	-	✓	✓	✓	
SONICS [Rahman et al., 2025]	✓	✓	-	✓	-	-	
MoM (ours)	✓	✓	✓	✓	✓	✓	

A.1 Related Works

The rapid advancement of end-to-end generative models has led to a surge in AI-generated music, capable of producing entire songs with vocals, lyrics, instrumentation, and stylistic nuances. Detecting such synthetic compositions requires datasets and models that can capture both low-level audio artifacts and high-level musical coherence.

Early efforts in synthetic song detection relied on small, narrowly scoped datasets, but recent benchmarks have begun to emphasize greater scale and diversity.

SONICS It is the current State-of-the-Art Dataset consisting of large collection of over 97,000 full-length tracks(4751 hours of audio) evenly split between real recordings and synthetic songs generated by platforms such as Suno (v2–v3.5) and Udio (v32, v130). Crucially, SONICS's songs are quite long (average of 176 seconds), supporting the modeling of long-range musical and lyrical patterns. It also includes the text lyrics of the songs, which can aid future research. However, SONICS also has clear limitations since it consists only of English songs (mostly featuring male vocals) with audio processed at a fixed 16 kHz rate, and its synthetic tracks come exclusively from two AI models (Suno and Udio), with the Half Fake tracks being only generated from one model- which means one compositional scenario (real lyrics with Udio) is missing, narrowing the variety within that category.

Other Datasets SingFake offers 58 hours of paired real and synthetic vocal clips across five languages and 40 singers, using real instrumental backings to aid in artifact detection. However, it lacks fully synthetic lyrics or instrumentals. CtrSVDD builds upon this by providing 308 hours of controlled SVDD content (over 220,000 clips), enabling fine-grained manipulation through parameterized synthesis and metadata through it, centering on vocals rather than full-song generation. Meanwhile, FSD (Fake Song Detection) [Xie et al., 2023] is a Chinese language benchmark comprising 200 real and 450 fake songs produced via five modern synthesis and conversion techniques. Its difficulty for speech-based detectors underscores the need for music-specific modeling, though its limited linguistic and methodological scope hinders broader applicability.

Although prior datasets have laid important groundwork for synthetic music detection, they often lack the diversity in generation pipelines and audio characteristics needed for models to generalize effectively to emerging AI content. Our contribution includes the Melody or Machine (MoM) dataset, specifically designed with extensive source and variation coverage to bridge this gap.

A.2 Modeling Approaches

Detecting AI-generated music spans a broad range of techniques, from classical signal transforms to deep neural architectures. **Early methods** [Pour et al., 2014, Wani et al., 2024a,b, Pham et al., 2024, Yi et al., 2023] convert raw waveforms into time–frequency representations such as STFT, CQT,

MFCCs, or gammatonegrams. These are typically fed into convolutional or recurrent networks to identify synthesis artifacts. While effective at highlighting low-level inconsistencies, these approaches often fail to capture long-term musical structure and are sensitive to genre and production variations.

To address these limitations, **recent methods incorporate long-context modeling** using attention mechanisms and spectro-temporal tokenization. For example, models like SpecTTTra [Rahman et al., 2025] (SOTA on SONICS) decompose spectrograms into patch-based tokens and apply transformers or memory-augmented RNNs to model global coherence in lyrics, rhythm, and instrumentation. These techniques show promise in detecting higher-order anomalies but risk overfitting to specific generator signatures, limiting their robustness across domains.

Another line of work leverages **self-supervised audio encoders** [Sharma and Gupta, 2025, Guo et al., 2024, Phukan et al., 2024a,b] such as MERT, wav2vec2, HuBERT [Hsu et al., 2021], and Audio Spectrogram Transformer (AST) [Gong et al., 2021]. These models are pretrained on large speech or music corpora to learn rich acoustic representations. When fine-tuned for deepfake detection, they often outperform purely supervised models. However, the gap between speech and music, characterized by wider pitch ranges, melodic structures, and vocal expression, necessitates domain-specific adaptation or joint pretraining across modalities.

Table 4: Quantitative Comparison of Fake Song Datasets

Dataset	Language	Avg. Length (sec)	# Algos	# Real Songs	# Fake Songs	Total Hours
FSD	Chinese	216.00	5	200	450	26
SingFake	Multi	13.75	–	634	671	58
CtrSVDD	Multi (no English)	4.87	14	32,312	188,486	307
SONICS	English	176.03	5	48,090	MoM (ours)	Multi (82% English)

A.3 Melody or Machine (MoM) Dataset

To support robust detection of AI-generated songs under diverse manipulations and model types, we present the Melody or Machine (MoM) Dataset—a large-scale benchmark reflecting the evolving landscape of song-level deepfakes. MoM spans three authenticity tiers: genuine recordings, synthetic audio with real lyrics/voice, and fully synthetic tracks, enabling evaluation across progressive levels of AI involvement. Unlike prior collections, MoM includes both open-source and proprietary music generators: **Suno (v2–v3.5)**, **Udio (v1.5)**, **Diffrythm**, **Yue**, **Voice Clones**, and **Riffusion (FUZZ-1.0)**.

A.3.1 Real Songs

The *Real Songs* subset includes 47,971 original human-created tracks sourced from YouTube using metadata from the Genius Lyrics Dataset. Spanning diverse genres, moods, and over 9,000 unique artists, these 2–3 minute songs serve as ground truth for classification tasks. Each is tagged with genre, tempo, and artist identifiers for fair comparison with synthetic data. To enrich diversity, we also include **17,504 high-quality human covers from YouTube**, introducing natural variation in timbre, arrangement, and style. **These real covers are especially important** in our setting as they challenge models to tolerate real-world creative variation while remaining sensitive to AI-induced artifacts. Together, these originals and covers help models distinguish authentic artistic variability from synthetic manipulations.

A.3.2 Fully Fake Songs

To generate **Fully Fake Songs**, we developed an exhaustive prompting pipeline consisting of three types of prompts to cover different dimensions of musical creativity. In contrast to other approaches that rely on randomness of topic, genre, and mood, potentially leading to irrelevant unnatural prompts, our pipeline ensures genre diversity, style richness, and ultimately a more realistic dataset of fake songs. **Detailed examples of all prompt types can be found in the Appendix.**

Type A: Prompts Inspired by Existing Songs with Genre Conditioning. These prompts aim to mimic human creativity by generating songs inspired by existing tracks but reimagined in a different

musical genre. This introduces grounded, reference-based creativity into the generation process. Here the genre is uniformly sampled from a curated list of 163 genres derived from the FMA music database [Defferrard et al., 2018, 2017].

Type B: Prompts Curated Using Musical Features and Attributes. We compiled a structured taxonomy of musical attributes—*genre, mood, tempo, key, instrumentation*, and *production style* inspired by leading AI music platforms. Thousands of prompts were auto-generated by randomly sampling values from each category.

Each prompt included: one genre, two mood descriptors, one tempo, one key, focal elements (e.g., "guitar solos", "synth layers"), and stylistic touches (e.g., "cinematic transitions", "vinyl crackle"). These elements were concatenated into raw prompts describing a song's sonic profile.

The raw text was refined by **Gemini 2.0 Flash** [Google, 2025], a large language model, into fluent paragraph-style prompts suitable for music generation. This produced expressive, musically grounded prompts used across multiple generations of systems.

Type C: Community-Sourced Random Prompts. To reflect organic user behavior and real-world prompt diversity, Type C prompts were sourced from high-frequency user interactions on state-of-the-art generative music platforms. These prompts capture practical edge cases and community-driven trends that emerge in unsupervised human-AI interaction scenarios. As such, they enhance prompt variety and model robustness.

A.3.3 Mostly Fake

These songs retain original lyrics but are paired with AI-generated vocals and instrumentals using Diffrythm, Yue, and other AI models, simulating realistic artist covers.

Type A: Real Lyrics + Generated Audio (Genre Inference) We sourced LRC-format files of existing commercial songs from RCLyricsBand.com, which include timestamped lyrics of the song. These LRC files were input into a BERT [Devlin et al., 2019] based model fine-tuned for few-shot classification to infer the most suitable musical style for the lyrical content. The predicted genre and corresponding LRC file were then combined and fed to Diffrythm and Yue to generate audio with stylistic coherence aligned to the lyrics.

Type B: Voice Cloning (Real Lyrics + Real Singer) We curated a special subset of approximately 1,000 samples mimicking real artists' vocal styles while modifying either the lyrics or the instrumental accompaniment. These samples were sourced from popular platforms such as YouTube and represent contemporary voice cloning practices with partial manipulation.

A.3.4 Dataset Evaluation

The MoM dataset improves on prior benchmarks like SONICS with broader genre, language, and vocal diversity, enabled by a more expressive prompt pipeline and a mix of open and closed-source models. It supports realistic, end-to-end song generation with coherent vocals and instrumentation, reflecting new AI music trends. Notably, it features new SOTA models like Riffusion, which delivers high-fidelity audio and holds the top ELO score in our web-based platform where listeners judge song realism. ELO scores, adapted from chess, rank models based on win/loss outcomes in head-to-head comparisons. Our dataset achieves a higher average ELO score (1016.35) compared to SONICS (894.60), indicating superior perceptual quality.

Our dataset emphasizes generalization by incorporating a curated test set composed of diverse out-of-distribution (OOD) samples, including *Riffusion (FUZZ-1.0)*, *Yue*, *Suno 3*, *Suno 4*, and *Voice Clones*. These sources provide a perceived quality comparable to the training set and are still being actively generated by other models. In contrast, prior works primarily evaluate on variants of similar models such as *Suno* and *Udio*, which leads to overly optimistic results that do not reflect real-world generalization as shown in Table 2.

Model	Train / Validation	Test
Suno 3.5	23695	–
Udio 1.5	19500	–
Diffrythm	4606	–
Suno 2	110	–
Suno 1	–	48
Suno 3	–	3512
Riffusion	–	7057
Yue	–	5278
Voice Clones	–	1166
Total	**47911**	**17061**

Table 5: **Dataset composition across synthetic audio sources.**
This table shows the number of audio samples used for training/validation and testing, organized by model.
Models marked in **red** are **closed-source** (e.g., Suno, Udio, Riffusion, Voice Clone).
Models marked in **blue** are **open-source** (e.g., Yue and Diffrythm).

Model	ELO Score
Riffusion (ours)	1105.58
Udio (ours)	1093.34
Real	1032.84
Suno (ours)	1013.76
Voice Clones (ours)	1007.23
Yue (ours)	958.37
Diffrythm (ours)	934.14
Suno (SONICS)	901.75
Udio (SONICS)	887.44
Average (Ours)	**1016.35**
Average (SONICS)	**894.60**

Table 6: **Human evaluation results using ELO-based ranking.**
This leaderboard ranks models by perceived audio quality based on human preference. Participants compared outputs for clarity, realism, musicality, and lyrical coherence on our website; more details are provided in the Appendix.

A.4 Dataset Description

For our experiments, we curated a diverse dataset comprising both real and synthetic songs sourced from several state-of-the-art generative audio models. The dataset spans a wide variety of generation techniques, quality levels, and stylistic characteristics, enabling comprehensive evaluation of detection and generalization performance.

A.4.1 Composition and Distribution

The synthetic portion of the dataset includes samples from the following models:

- **Suno v2** – 110 samples
- **Suno v3** – 3,512 samples *(Test only)*
- **Suno v3.5** – 23,695 samples
- **Suno v4** – 48 samples *(Test only)*
- **Udio v1.5** – 19,500 samples
- **Diffrythm** – 4,606 samples
- **Riffusion** – 7,057 samples *(Test only)*
- **Yue** – 5,278 samples *(Test only)*
- **Voice Cloning** – 1,166 samples *(Test only)*

Fake-Type Categorization. We further categorize the synthetic songs based on the degree of AI synthesis:

- **Fully Fake:** Suno (v2, v3, v3.5, v4), Udio v1.5, Riffusion
- **Mostly Fake:** Yue, Diffrythm, Voice Cloning

This structured labeling allows systematic evaluation under varying levels of generative realism.

A.5 Human-AI Perceptual Benchmark

To benchmark our detection models against human perceptual judgment, we developed a blind listening evaluation platform.

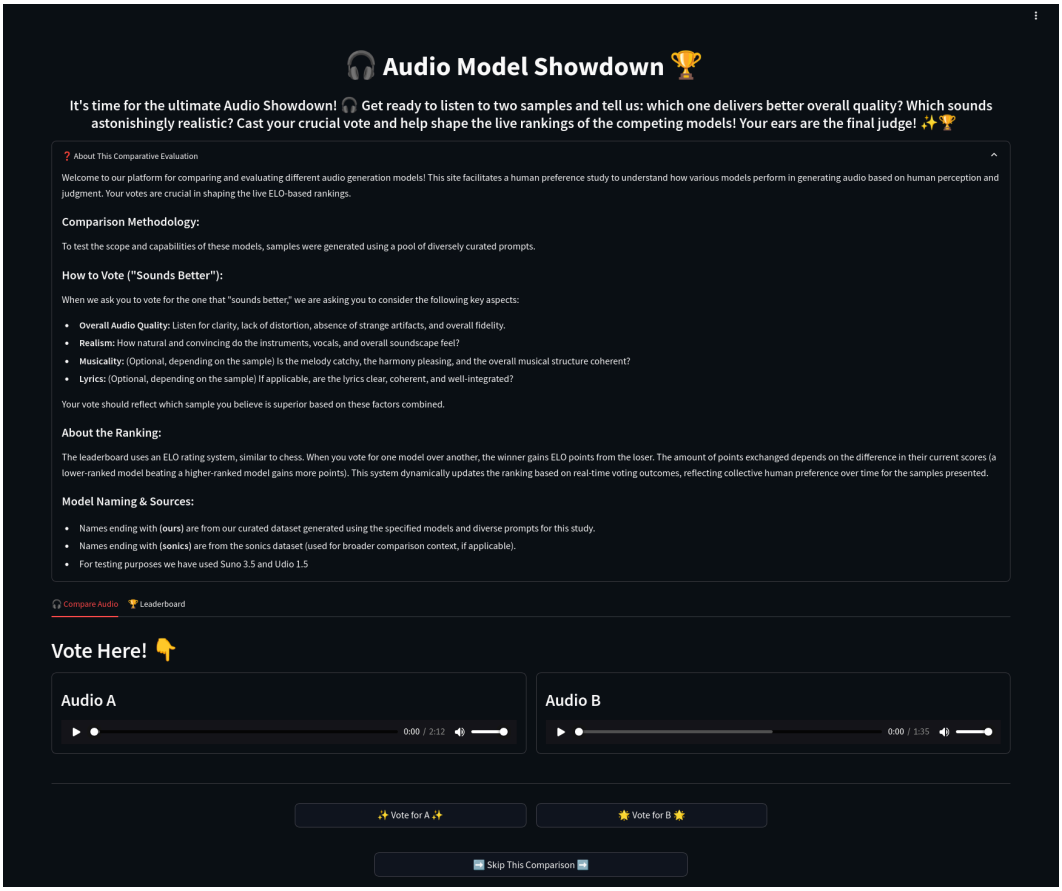
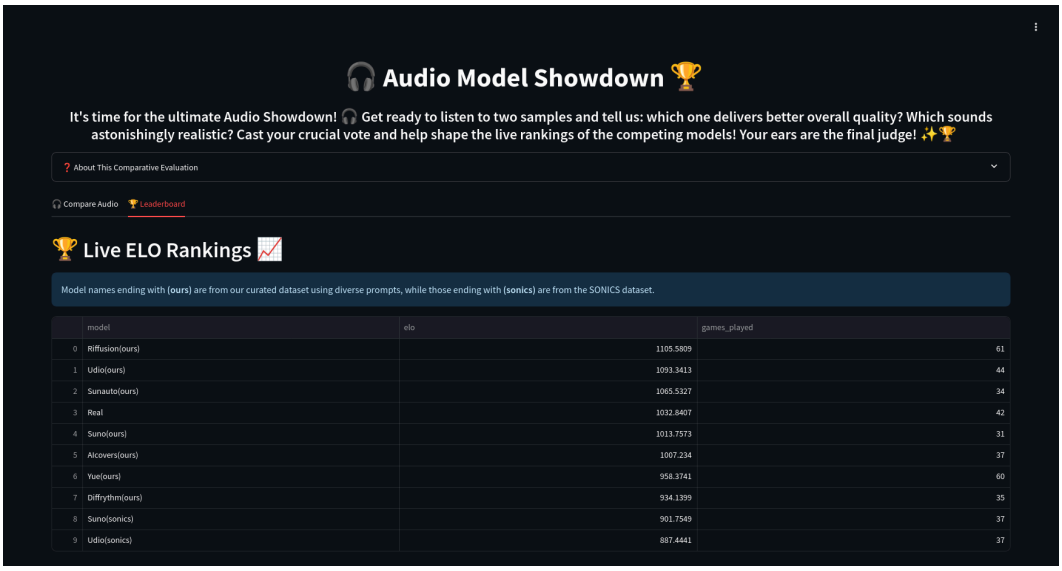


Figure 2: Interface of the Song Arena platform on Huggingface.



	model	elo	games_played
0	Riffusion(ours)	1105.5809	61
1	Udio(ours)	1093.3413	44
2	Sunauto(ours)	1065.5327	34
3	Real	1032.8407	42
4	Suno(ours)	1013.7573	31
5	Aicovers(ours)	1007.234	37
6	Yue(ours)	958.3741	60
7	Diffrythm(ours)	934.1399	35
8	Suno(sonics)	901.7549	37
9	Udio(sonics)	887.4441	37

Figure 3: Leaderboard based on Elo scores from human evaluations on the Song Arena platform. Higher scores indicate stronger preference by listeners.

A.5.1 Evaluation Setup

Participants are presented with randomized audio pairs and asked: *"Which song sounds more realistic?"* Each pair falls into one of two categories:

- **Intra-Dataset:** Both songs are drawn from the same dataset but different generation models.
- **Inter-Dataset:** Songs are drawn from entirely different datasets (e.g., MoM vs. external benchmarks like SONICS).

A.5.2 Model Elo Ranking from our website

To quantitatively assess perceived audio realism from human evaluations, we compute an Elo rating based on aggregated pairwise comparisons our platform. Higher scores indicate greater preference by listeners. Our proposed Riffusion-based model achieves the top human realism score.

Table 7: Elo Scores from Human Evaluation

Model	Elo Score
Riffusion (ours)	1105.58
Udio (ours)	1093.34
Real	1032.84
Suno (ours)	1013.76
Voice Clones (ours)	1007.23
Yue (ours)	958.37
Diffrythm (ours)	934.14
Suno (SONICS benchmark)	901.75
Udio (SONICS benchmark)	887.44

Interpretation: The Elo-based analysis reveals that multiple synthetic models (notably our versions of Riffusion and Udio) surpassed human-composed songs in perceived realism. This demonstrates the increasing difficulty of AI song detection and motivates the need for sophisticated modality-alignment-based detection methods.

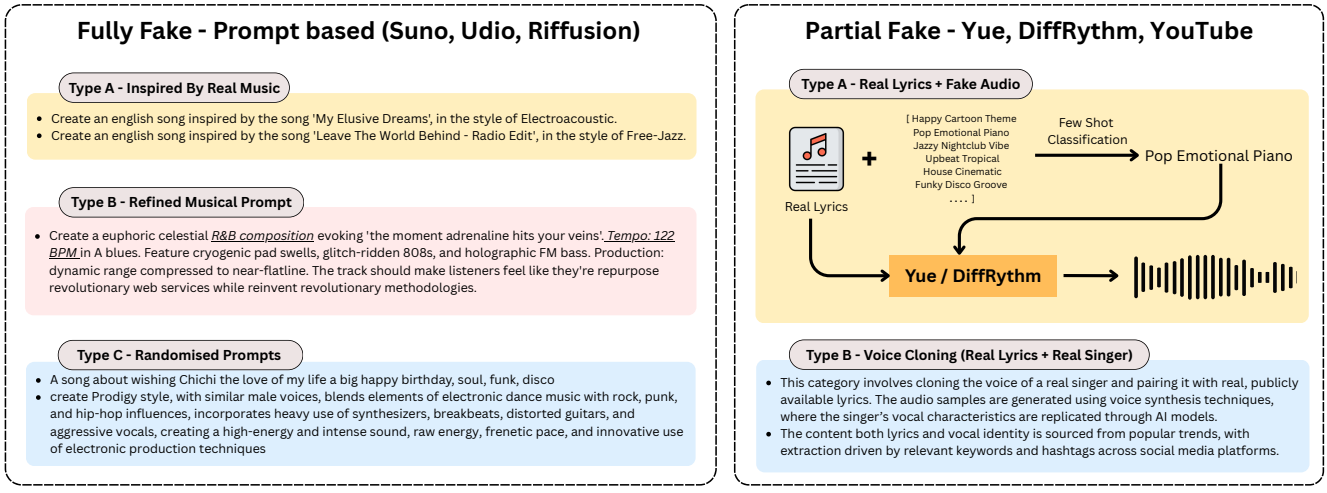


Figure 4: Distribution of AI-generated songs in our dataset.

A.6 Prompt Metadata and Style Attributes

Each generated song is associated with prompt metadata and stylistic attributes which provide a rich representation of musical intent. These structured tags help in conditioning models, analyzing generalization, and understanding detection robustness.

A.6.1 Genre List

pop, rock, hip-hop, R&B, EDM, jazz, blues, country, metal, punk, folk, classical, ambient, lo-fi, trap, drum and bass, dubstep, synthwave, vaporwave, house, techno, trance, reggae, dancehall, afrobeats, k-pop, j-pop, gospel, funk, soul, indie rock, indie pop, math rock, psychedelic, experimental, industrial, noise, electroacoustic, bossa nova, samba, latin pop, grime, phonk, drill, hardstyle, gabber, post-rock, post-punk, new wave, retrowave, dream pop, shoegaze, dark ambient, minimal techno, chillwave, trip-hop, future bass, glitch hop, electropop, synth-pop, neo soul, alternative metal, progressive metal, black metal, death metal, sludge metal, djent, folk metal, jazz fusion, smooth jazz, big band, bebop, cool jazz, bluegrass, americana, celtic, singer-songwriter, anime opening, game music, cinematic, soundtrack, orchestral, epic score, gregorian chant, chorale, opera, baroque, romantic era, modern classical, contemporary classical

A.6.2 Mood (Vibe) Tags

energetic, melancholic, uplifting, dark, dreamy, romantic, rebellious, introspective, nostalgic, aggressive, peaceful, mysterious, epic, groovy, funky, sultry, eerie, haunting, hypnotic, playful, somber, joyful, ambient, chill, moody, dramatic, whimsical, ethereal, gritty, raw, emotional, spiritual, majestic, cinematic, futuristic, retro, vintage, spacey, glitchy, lo-fi, high-energy, slow-burning, minimalist, maximalist, experimental, avant-garde, traditional, modern, classic, innovative, bold, subtle, intense, light-hearted, serene, chaotic, structured, free-form, rhythmic, melodic, percussive, harmonic, dissonant, consonant, layered, sparse

A.6.3 Tempo Range

60 BPM, 65 BPM, 70 BPM, 75 BPM, 80 BPM, 85 BPM, 90 BPM, 95 BPM, 100 BPM, 105 BPM, 110 BPM, 115 BPM, 120 BPM, 125 BPM, 130 BPM, 135 BPM, 140 BPM, 145 BPM, 150 BPM, 155 BPM, 160 BPM, 165 BPM, 170 BPM, 175 BPM, 180 BPM

A.6.4 Key Signatures

C major, G major, D major, A major, E major, B major, F# major, C# major, F major, Bb major, Eb major, Ab major, Db major, Gb major, Cb major, A minor, E minor, B minor, F# minor, C# minor, G# minor, D# minor, A# minor, D minor, G minor, C minor, F minor, Bb minor, Eb minor, Ab minor

A.6.5 Focal Points

haunting female vocals, Spanglish verses, melodic rap, funky basslines, Afrobeats rhythms, Latin percussion, ambient synth layers, dynamic drum patterns, side-chaining production, pop-drop structure, Afropiano elements, Arabic melodic influences, vintage 80s synths, orchestral string flourishes, glitch effects, lo-fi textures, trap hi-hats, 808 bass, guitar solos, piano arpeggios, saxophone riffs, violin sections, choir harmonies, turntable scratches, beatboxing, synth arpeggios, modular synths, field recordings, spoken word segments, call-and-response vocals

A.6.6 Extra Descriptors

a retro-futuristic atmosphere, cinematic transitions, subtle glitchy textures, introspective lyrics, 808-style percussion, dreamlike vocal layering, lo-fi textures, neon-lit arcade vibes, space-themed effects, underwater ambiance, forest soundscapes, urban street noise, vintage radio samples, tape hiss, vinyl crackle, crowd chants, live concert feel, studio ambiance, rain sounds, wind effects, birdsong, ocean waves, city traffic, subway sounds, clock ticking, heartbeat rhythm, typewriter clicks, camera shutters

A.7 Prompt Examples of Fully Fake songs

Prompt Type A - Prompts Inspired by Existing Songs with Genre Conditioning (10 samples)

Prompt: Create an English song inspired by the song 'On The Alamo', in the style of Noise.

Prompt: Create an english song inspired by the song 'There'll be Some Changes Made', in the style of Rap.

Prompt: Create an english song inspired by the song 'Let Me Roll It - Remastered 2010', in the style of Polka.

Prompt: Create an english song inspired by the song 'No Mother In This World Today', in the style of Black-Metal.

Prompt: Create an English song inspired by the song 'Fly Out', in the style of Interview.

Prompt: Create an English song inspired by the song 'I Want It All (feat. Mack 10)', in the style of South Indian Traditional.

Prompt: Create an English song inspired by the song 'You Are My Everything', in the style of Techno.

Prompt: Create an English song inspired by the song 'Once In a Blue Moon', in the style of Poetry.

Prompt: Create an English song inspired by the song 'Somewhere', in the style of Musique Concrete.

Prompt: Create an English song inspired by the song 'It Gets Better', in the style of Goth.

Prompt Box B - Prompts Curated Using Musical Features and Attributes (10 samples)

Prompt: Post-rock fusion: Uplifting yet somber G minor at 95 BPM. Piano arpeggios meet vintage synths and Afrobeats rhythms. Subway ambience weaves through cinematic transitions. Lyrical themes of urban loneliness.

Prompt: Craft a Latin Pop track: sparse yet modern. Tempo: 135 BPM, D major. Instrumentation: glitchy beats, ambient synths, lo-fi textures. Mood: dreamy, cinematic. Lyrical themes: longing, nostalgia. Arrangement: cinematic transitions, layered vocals.

Prompt: Trip-hop track; romantic yet high-energy. 80 BPM, G# minor. Vintage synths, ambient pads, guitar solo. Retro-futuristic feel with rain. Melancholy layered with driving rhythm; a dance in the downpour. Lyrical themes: longing, passion.

Prompt: Craft an experimental track (romantic chaos) at 120 BPM in A major. Imagine saxophone cries amidst synth arpeggios, grounded by Latin percussion. Vinyl crackle dusts cinematic transitions. Lyrical themes touch upon both love and destruction.

Prompt: Craft a DnB track at 155 BPM in D major, blending cinematic grandeur with avant-garde grit. Think: Afropiano-infused pop-drop, lo-fi textures, and colossal 808 drums. Dreamlike, layered vocals float over the chaos, lyrically exploring [Lyrical themes].

Prompt: Craft a bold, gritty funk track (175 BPM, C Major). Haunting female vocals meet lo-fi textures and trap hi-hats. Dreamlike vocal layers interweave with typewriter clicks,

adding a unique, unsettling ambiance. Lyrical themes: defiance, longing.

Prompt: Genre: Classical/Funk Fusion. Instrumentation: Orchestra, 808, vocals. Mood: Ethereal, energetic. Tempo: 175 BPM. Key: C minor. Lyrical Themes: Oceanic exploration, freedom. Arrangement: Free-form classical structure with funky 808 bass, call-and-response vocals, layered with ocean waves underwater ambiance.

Prompt: Craft a neo-classical track; genre: Classical. Instrumentation: Violin sections, field recordings. Mood: Uplifting, layered. Tempo: 125 BPM. Key: C Major. Arrangement: Tape hiss, neon arcade vibes, creating an energetic yet nostalgic soundscape.

Prompt: Craft a rebellious yet groovy drum and bass track (175 BPM, E minor). Instrumentation: pounding drums, pulsating bass, piano arpeggios, and Afrobeats rhythms. Mood: introspective, rebellious. Lyrical themes: rebellion, rain, reflection. Arrangement: rain intro builds to an explosive, introspective lyrical drop.

Prompt: Craft a modern classical piece (80 BPM, A minor). Imagine joyful, harmonic melodies soaring over funky basslines, punctuated by Spanglish verses. Retro-futuristic synths and spacey effects create an otherworldly atmosphere.

A.8 Modality Gap Detection via Embedding Alignment

We propose a novel audio detection model, **CLAM** (**C**ontrastive **L**earning for **A**udio **M**atching), that employs dual pre-trained encoders:

- **MERT**: Captures rich **musical representations**
- **Wav2Vec2**: Captures **vocal nuances**

These are fused via a **weighted cross-aggregation module**. The model is trained using a hybrid objective:

- **Binary Classification Loss** (`BCEWithLogitsLoss`) to distinguish real vs. AI-generated samples
- **Triplet Margin Loss** to **contrastively align real instrumental and vocal embeddings**, helping the model learn natural semantic coherence

To effectively distinguish between AI-generated and real music, aligning the instrumental and vocal embedding spaces for real music samples is crucial. This alignment ensures that the multimodal representations of real music exhibit a consistent and predictable relationship in the learned embedding space, which can then be leveraged by the classifier to identify deviations present in synthetic content. Unlike unimodal analysis, comparing the aligned multimodal embeddings allows the detection of subtle inconsistencies characteristic of generative processes.

Contrastive learning approaches are particularly well-suited for this task as they focus on learning a metric or an embedding space where a notion of similarity is encoded by distance. The objective is to learn a mapping $\phi : \mathcal{X} \to \mathbb{R}^d$ such that for any two samples x_i, x_j, their distance in the embedding space, $D(\phi(x_i), \phi(x_j))$, reflects their semantic similarity. Among various contrastive methods, Triplet Loss was explored for aligning the instrumental and vocal embeddings of real music samples.

The core idea behind Triplet Loss is to enforce a relative distance constraint. For a given Anchor sample a, a Positive sample p (which is semantically similar to a), and a Negative sample n (which is semantically dissimilar to a), the loss minimizes the distance between the Anchor and the Positive ($D(e_a, e_p)$) while simultaneously maximizing the distance between the Anchor and the Negative ($D(e_a, e_n)$). This is done such that the distance to the positive is smaller than the distance to the negative by at least a predefined margin α. For embeddings $e_a = \phi(a)$, $e_p = \phi(p)$, and $e_n = \phi(n)$, the Triplet Loss $\mathcal{L}_{\text{triplet}}$ is defined as:

$$\mathcal{L}_{\text{triplet}}(e_a, e_p, e_n) = \max(0, D(e_a, e_p)^2 - D(e_a, e_n)^2 + \alpha)$$

where $D(\cdot, \cdot)$ denotes a distance metric, and $\alpha > 0$ is the margin. The squared Euclidean distance, $\|e_i - e_j\|_2^2$, is commonly used for $D(e_i, e_j)^2$ due to its computational efficiency and differentiability. The $\max(0, \cdot)$ function, also known as the hinge loss, ensures that no penalty is incurred if the distance constraint is already satisfied. The optimization objective of minimizing $\mathcal{L}_{\text{triplet}}$ encourages the model to learn an embedding function ϕ such that for any valid triplet (a, p, n), the following inequality holds in the embedding space:

$$\|e_a - e_p\|_2^2 + \alpha \leq \|e_a - e_n\|_2^2$$

This inequality directly promotes a structured embedding space where embeddings of similar items are clustered together, and clusters of dissimilar items are pushed apart by a significant margin.

In the context of aligning instrumental and vocal embeddings for real music samples within a training batch, the triplet can be constructed using an in-batch mining strategy. For each real music sample i in the batch, we define the triplet as follows:

- Anchor (e_a): The instrumental embedding of real music track i from the batch ($E_I^{\text{real}}[i]$).
- Positive (e_p): The vocal embedding of the **same** real music track i from the batch ($E_V^{\text{real}}[i]$).
- Negative (e_n): The vocal embedding from a **different** real music track j within the **same batch** ($E_V^{\text{real}}[j]$ where $j \neq i$).

The objective is to make the vocal embedding of the correct track closer to its corresponding instrumental embedding than any vocal embedding from other tracks present in the batch, by a margin α. This specific triplet configuration directly enforces that the learned embedding space respects the natural pairing of instrumental and vocal streams within authentic music.

The process for calculating the in-batch Triplet Loss for alignment, considering all valid triplets formed by pairing each Anchor-Positive pair with all other available Negatives within the batch, is formally described in Algorithm 1. This in-batch mining strategy is a practical approximation to using all possible triplets in the dataset and is effective in practice by providing a diverse set of negative examples during training.

The overall training objective for CLAM utilizes a dual-loss formulation: the Binary Cross-Entropy (BCE) loss for the final real/fake classification and the alignment loss to structure the embedding space for real samples. Specifically, the total loss is a weighted sum of the BCE loss ($\mathcal{L}_{\text{BCE}}$) and the Triplet Loss ($\mathcal{L}_{\text{triplet}}$) calculated as described above:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{BCE}} + \lambda \cdot \mathcal{L}_{\text{triplet}}$$

where λ is the `alignment_loss_weight` hyperparameter controlling the influence of the alignment objective. By minimizing this combined loss, the model learns to simultaneously classify the music's origin and structure the embedding space such that real multimodal pairs are distinguishable from other combinations.

In our ablation study comparing various alignment loss functions, including Mean Squared Error (MSE), L1 Loss, Huber Loss, Cosine Similarity Loss, and Triplet Loss, using this dual-loss objective, the Triplet Loss achieved the best overall performance in terms of classification accuracy and F1-score on the validation set. This empirical finding supports the hypothesis that explicitly enforcing a relative distance margin between positive (aligned real pairs) and negative (misaligned real pairs) in the embedding space is highly effective for learning discriminative representations for AI-generated music detection. The structured embedding space learned via Triplet Loss enhances the downstream classifier's ability to identify the subtle, unnatural discrepancies present in synthetic content.

Algorithm 1: In-Batch Triplet Loss Calculation for Alignment

Input: Batch of data containing instrumental embeddings E_I, vocal embeddings E_V, and labels L.
Output: Average Triplet Loss for the real samples in the batch.
$E_I^{\text{real}} \leftarrow \{e_i \in E_I \mid L_i = 0\}$
$E_V^{\text{real}} \leftarrow \{e_v \in E_V \mid L_v = 0\}$
$N \leftarrow |E_I^{\text{real}}|$ Number of real samples in the batch
$\mathcal{L}_{\text{batch}} \leftarrow 0$
$count \leftarrow 0$
if $N > 1$ **then**
  for $i \leftarrow 0$ *to* $N - 1$ **do**
    Indices start from 0 in code $e_a \leftarrow E_I^{\text{real}}[i]$
    Anchor: Instrumental embedding of real sample i $e_p \leftarrow E_V^{\text{real}}[i]$
    Positive: Vocal embedding of real sample i
    for $j \leftarrow 0$ *to* $N - 1$ **do**
      if $i \neq j$ **then**
        $e_n \leftarrow E_V^{\text{real}}[j]$
        Negative: Vocal embedding of real sample j $d_{pos}^2 \leftarrow \|e_a - e_p\|_2^2$
        Squared L2 distance $d_{neg}^2 \leftarrow \|e_a - e_n\|_2^2$
        $\mathcal{L}_{triplet\_ij} \leftarrow \max(0, d_{pos}^2 - d_{neg}^2 + \alpha)$
        $\mathcal{L}_{\text{batch}} \leftarrow \mathcal{L}_{\text{batch}} + \mathcal{L}_{triplet\_ij}$
        $count \leftarrow count + 1$
      end
    end
  end
end
if $count > 0$ **then**
  $\mathcal{L}_{\text{batch}} \leftarrow \mathcal{L}_{\text{batch}}/count$
  Average loss over valid triplets
end
return $\mathcal{L}_{batch}$

A.9 Broader Impacts

The MoM dataset and the CLAM model create noteworthy positive societal outcomes through a greater capacity for the detection of AI-created music. For example, content platforms, rights holders, as well as forensic analysts, gain the ability to identify synthetic tracks. This supports the security of artist intellectual property and the maintenance of trust in music distribution. MoM's detailed annotations plus song representations with attention to multilingualism and gender balance give support to the creation of moderation tools in the future. These tools can offer more fairness and inclusivity across the music industry. At the same time, a large, good corpus of actual plus synthetic songs could help in the training of deepfake generators that appear more real.

The focal point of the project is to label music on popular platforms, so that the music is presented to the listeners with full transparency about the production of the music they are consuming, making them aware of authenticity as well as the ability of AI. Music artists may benefit from this labeling as well, as it prevents faking AI generated music to be posed as human made which provides an unfair advantage by displaying a level of human element to the listener, this is similar to the issues faced by graphic designers who had their livelihood affected by the advancements in AI image generators which would have been prevented to an extent, had a labelling system been introduced sooner.

Full dependency on automated detectors presents a risk of false positives or unfair treatment of genres or vocal styles not well represented in the data as well as potentially harmful to smaller artists if their music is incorrectly labelled as synthetic. Another point of potential issues affects artists who either unknowingly use samples generated with AI or those who are physically incapable of some part of production- either vocals or instrumentals and want to use AI to complete their work. To lower those risks, a release strategy for the full dataset with controls is a good idea.

A.10 Limitations and Future Work

The domain of AI-generated audio is undergoing rapid advancement, with sophisticated new open-source models continuously emerging. The open-source models central to our study are very recent, representing some of the first to genuinely rival the quality of leading closed-source systems such as Suno and Udio. As these models rapidly improve, approaching the finesse and quality of professionally produced music, the datasets used for training risk becoming outdated. Our dataset is particularly susceptible to this, and it is important to note that it is predominantly English, comprising 92% of the content. Therefore, updating this dataset on a regular basis and periodically retraining our model are crucial steps necessary for keeping pace with the state-of-the-art in song generation. A further limitation of our work is that we prioritized model performance over computational complexity, a design choice appropriate given that this task does not require real-time application.

A.11 Ethics Statement

In our dataset, Real songs and Mostly Fake Type B (voice cloning) samples are sourced from YouTube. To respect copyright and usage policies, we provide only the original YouTube links instead of hosting the audio directly. In the Fully Fake set, Type A prompts draw inspiration from real song titles, while Type B prompts are refined using Gemini. Mostly Fake Type A lyrics are from public LRC files, with AI-generated vocals and instrumentals. Only AI-generated songs will be released on Hugging Face under a **CC BY-NC 4.0** license. Although the generative models used to create our dataset may have been trained on copyrighted material, including lyrics and stylistic metadata, this practice aligns with prior research practices, as seen in datasets such as LLaVA-Instruct-158K [Shin et al., 2024], Gpt4tools [Yang et al., 2023] and JourneyBench [Wang et al., 2025] Our research complies with the NeurIPS Code of Ethics, focusing on responsible data sourcing, copyright and fair use adherence, transparency in dataset construction, and principles of responsible AI use.