Probing Token Spaces under Generator Shift in AI-Generated Music Detection

Joonyong Park [1 2] **Jungwoo Kim** [3 2] **Junyoung Koh** [3 2] **Yuki Saito** [1]

Abstract

AI-generated music detectors can appear robust on standard benchmark splits, yet their deployments require transfer to generator sources absent during training. We study this problem with source-restricted evaluation on MoM-OPEN, an open reconstruction of MoM-CLAM that replaces the non-redistributable real corpus with FMA and MTG-Jamendo while preserving the fake-generator protocol. To isolate the role of representation, we introduce CoMoE, a compact fixed classifier for comparing heterogeneous audio token spaces while keeping the downstream architecture and training recipe unchanged. Experiments show that standard and real-source-restricted splits are nearly saturated, whereas fake-source restriction exposes large differences between token spaces: X-Codec tokens are strongest when training on Udio alone, while MERT-derived tokens are stronger when training on Suno-v3.5 alone. These results suggest that codec-style discrete token spaces should be treated as a primary experimental axis under generator shift in AI-generated music detection. Our code and data are available at https://github.com/MAAP-LAB/CoMoE.

1. Introduction

AI-generated music detection aims to determine whether a music recording was produced by a human process or by a generative music system. The task is increasingly important as music generators can now produce full tracks with vocals, accompaniment, and near-release-quality production that are difficult to distinguish from human-made recordings (Rahman et al., 2025; Afchar et al., 2025; Cros Vila et al., 2025; Li et al., 2024b). Recent detectors based on spectrograms, raw waveforms, and self-supervised audio representations report strong benchmark performance (Batra et al., 2025; Rahman et al., 2025; Afchar et al., 2025). In deployment, however, a detector must flag outputs from generator sources that were absent during training, and standard benchmark splits may overstate this robustness when training and test sets share generator-specific artifacts (Batra et al., 2025; Rahman et al., 2025; Afchar et al., 2025). This motivates not only source-restricted evaluation, but also a closer examination of which audio representations still transfer when generator-specific artifacts change.

In this work, we examine codec-style discrete audio tokens as candidates for transferable representations under generator shift. First, they provide a forensic view that differs from continuous acoustic or semantic features. Neural audio codecs represent audio as codebook sequences with residual-quantization structure (Zeghidour et al., 2022; Défossez et al., 2023; Kumar et al., 2023), which may expose codebook usage, token-transition, and quantizer-level patterns that are not directly isolated by pooled continuous features. Second, codec tokens provide a compact interface for downstream detectors: once tokens are extracted, the classifier can operate on symbolic sequences rather than full-resolution waveforms. Such codec structure has been explored in speech deepfake detection (Li et al., 2024a; Wu et al., 2024; 2026), but music deepfake detection has mostly relied on waveform, spectrogram, or continuous representation detectors (Rahman et al., 2025; Batra et al., 2025; Afchar et al., 2025; Comanducci et al., 2025). Importantly, codec tokens do not define a single representation: different tokenizers induce different discrete spaces, with different codebooks, temporal rates, and quantization behavior.

This variability makes tokenizer choice a key experimental variable rather than a preprocessing detail, especially under source-restricted evaluation. To isolate this factor, we introduce Codec-Mixture-of-Experts (CoMoE), a compact fixed classifier for controlled tokenizer comparison. We keep the classifier architecture, training recipe, and evaluation protocol fixed, and replace only the input token space. We evaluate on MoM-OPEN, an open reconstruction of MoM-CLAM that replaces the non-redistributable YouTube-derived real corpus with FMA-medium and MTG-Jamendo while preserving the fake-generator protocol.

Our contributions are threefold: (i) we introduce CoMoE as a fixed classifier for comparing heterogeneous discrete

[1]The University of Tokyo, Tokyo, Japan [2]MAAP Lab [3]Yonsei University, Seoul, Republic of Korea. Correspondence to: Joonyong Park <joonyong-park@g.ecc.u-tokyo.ac.jp>.

Accepted at the ICML 2026 Workshop on Machine Learning for Audio.

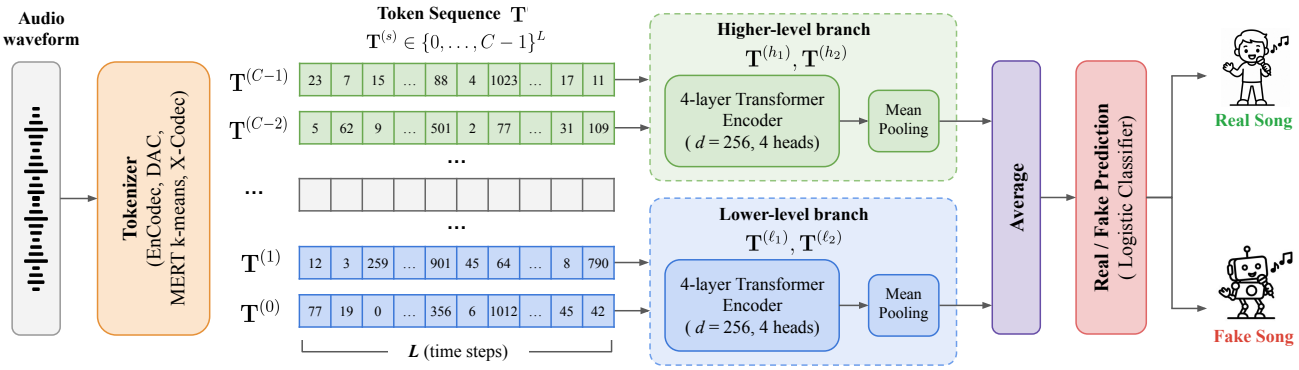

Figure 1. Architecture of CoMoE.

audio token spaces; (ii) construct MoM-open with source-restricted evaluation splits; and (iii) show that tokenizer choice is a primary experimental variable for cross-generator music deepfake detection.

2. Related Work

Neural audio codecs and forensic cues. Neural audio codecs compress waveforms into compact latent or discrete token sequences for high-fidelity reconstruction (Zeghidour et al., 2022; Défossez et al., 2023; Kumar et al., 2023). Many modern codecs use residual vector quantization (RVQ), where audio is represented by multiple codebook streams that capture different levels of acoustic detail. In speech deepfake detection, neural-codec representations and quantizer hierarchies have already been used as forensic cues (Li et al., 2024a; Wu et al., 2024; 2026). This suggests that codec tokens may reveal synthetic artifacts not directly exposed by continuous features.

Hybrid expert designs. Generated-content detectors often combine complementary views of the input. For example, AIDE uses both semantic and low-level artifact-sensitive branches for AI-generated image detection (Radford et al., 2021; Yan et al., 2025). This motivates branch-specialized designs for codec-token detection, where different codebook levels may carry different forensic information. For music deepfake detection, however, the unresolved question is not only how to design a classifier, but whether the token space itself controls robustness to unseen generators.

Music deepfake detection. Music deepfake detectors mostly rely on raw waveforms, spectrograms, or continuous self-supervised features. SONICS uses temporal and spectral tokenization over mel-spectrograms (Rahman et al., 2025), while CLAM uses continuous MERT and Wav2Vec2 streams (Batra et al., 2025). Other studies similarly evaluate raw-audio, spectrogram, or pretrained-representation baselines (Afchar et al., 2024; Comanducci et al., 2025). In contrast, codec-style discrete token spaces have not been systematically compared under cross-generator music deepfake evaluation.

3. CoMoE: A Controlled Token-Space Probe

Architecture. Figure 1 explains the structure of the model overall. The four streams consist of two lower-level and two higher-level token streams. This is a controlled interface rather than a theoretical constraint: for RVQ codecs, the streams are selected from early and late codebooks; for MERT k-means, they are selected from lower and upper self-supervised layers.

Formally, CoMoE consumes four discrete token streams,

$$\mathbf{T} = \left(\mathbf{T}^{(\ell_1)}, \mathbf{T}^{(\ell_2)}, \mathbf{T}^{(h_1)}, \mathbf{T}^{(h_2)}\right), \quad \mathbf{T}^{(s)} \in \{0, \ldots, C-1\}^L, \tag{1}$$

where C is the codebook size, L is the fixed token sequence length after truncation or padding, and s indexes one of the four streams. The superscripts ℓ and h denote lower- and higher-level streams, respectively. The two lower-level streams $\mathbf{T}^{(\ell_1)}, \mathbf{T}^{(\ell_2)}$ and the two higher-level streams $\mathbf{T}^{(h_1)}, \mathbf{T}^{(h_2)}$ are processed by separate Transformer encoders $f^{(\ell)}$ and $f^{(h)}$ with identical architecture. Each encoder has four layers, hidden size $d = 256$, and four attention heads.

The encoder outputs are mean-pooled over time to obtain two branch representations,

$$\mathbf{h}^{(\ell)} = \mathrm{Pool}\left(f^{(\ell)}\left(\mathbf{T}^{(\ell_1)}, \mathbf{T}^{(\ell_2)}\right)\right), \quad \mathbf{h}^{(h)} = \mathrm{Pool}\left(f^{(h)}\left(\mathbf{T}^{(h_1)}, \mathbf{T}^{(h_2)}\right)\right), \tag{2}$$

where $\mathbf{h}^{(\ell)}, \mathbf{h}^{(h)} \in \mathbb{R}^d$. The two branch representations are averaged and fed to a binary logistic classifier:

$$\mathbf{z} = \frac{1}{2}\left(\mathbf{h}^{(\ell)} + \mathbf{h}^{(h)}\right), \quad \hat{y} = \sigma\left(\mathbf{w}^\top \mathbf{z} + b\right), \tag{3}$$

where $\mathbf{w} \in \mathbb{R}^d$ and $b \in \mathbb{R}$ are trainable classifier parameters. All CoMoE variants use the same four-stream classifier,

Table 1. MoM-OPEN composition. Real audio is drawn from two openly redistributable music corpora; fake audio follows the $\mathcal{F}_T$/$\mathcal{F}_O$ designation of the original benchmark (Batra et al., 2025).

Class	Source	Clips
Real $\mathcal{R}$ (train+test)	FMA-medium (Defferrard et al., 2017)	24,979
	MTG-Jamendo (Bogdanov et al., 2019)	52,501
Fake $\mathcal{F}_{\mathrm{T}}$ (train)	Suno-v2 (Suno, Inc., 2024)	660
	Suno-v3.5 (Suno, Inc., 2024)	28,611
	Udio (Udio Inc., 2024)	19,500
	DiffRhythm (Ning et al., 2025)	4,594
Fake $\mathcal{F}_{\mathrm{O}}$ (OOD)	Riffusion (Forsgren & Martiros, 2022)	7,043
	Suno-v3 (Suno, Inc., 2024)	3,116
	Suno-v4 (Suno, Inc., 2024)	27
	YuE (Yuan et al., 2026)	5,278
Total		146,309

Table 2. Split definitions. The held-out target additionally contains the base-split OOD real set so AUC is computed in the standard binary sense.

Split	Train	Test
base	$\mathcal{R}$ train $\cup$ $\mathcal{F}_{\mathrm{T}}$	$\mathcal{R}$ test $\cup$ $\mathcal{F}_{\mathrm{O}}$
Real-FMA	**FMA** $\cup$ $\mathcal{F}_{\mathrm{T}}$	$\mathcal{F}_{\mathrm{O}}$ $\cup$ Jamendo
Real-Jamendo	**Jamendo** $\cup$ $\mathcal{F}_{\mathrm{T}}$	$\mathcal{F}_{\mathrm{O}}$ $\cup$ FMA
Fake-Suno3.5	$\mathcal{R}$ train $\cup$ **Suno-v3.5**	$\mathcal{R}$ test $\cup$ ($\mathcal{F} \setminus$ Suno-v3.5)
Fake-Udio	$\mathcal{R}$ train $\cup$ **Udio**	$\mathcal{R}$ test $\cup$ ($\mathcal{F} \setminus$ Udio)

so differences among CoMoE rows reflect the input token space rather than changes in the downstream classifier.

Token front-ends. All tokenizers are mapped to the fixed four-stream interface defined above, with codebook size $C = 1024$ and fixed sequence length L after truncation or padding. For each tokenizer, we construct two lower-level streams $\mathbf{T}^{(\ell_1)}, \mathbf{T}^{(\ell_2)}$ and two higher-level streams $\mathbf{T}^{(h_1)}, \mathbf{T}^{(h_2)}$.

EnCodec 24 kHz (Défossez et al., 2023) provides acoustic RVQ codebook streams[1]. We map codebooks $q = 0, 1$ to $\mathbf{T}^{(\ell_1)}, \mathbf{T}^{(\ell_2)}$ and codebooks $q = 6, 7$ to $\mathbf{T}^{(h_1)}, \mathbf{T}^{(h_2)}$.

DAC 44 kHz (Kumar et al., 2023) is used as a second acoustic codec[2]. We apply the same early/late rule and map codebooks $q = 0, 1$ to $\mathbf{T}^{(\ell_1)}, \mathbf{T}^{(\ell_2)}$ and codebooks $q = 7, 8$ to $\mathbf{T}^{(h_1)}, \mathbf{T}^{(h_2)}$.

X-Codec mini (Ye et al., 2025) is a music-trained semantic-aware codec checkpoint[3]. X-Codec mini provides twelve RVQ codebook streams; we map codebooks $q = 0, 1$ to $\mathbf{T}^{(\ell_1)}, \mathbf{T}^{(\ell_2)}$ and codebooks $q = 10, 11$ to $\mathbf{T}^{(h_1)}, \mathbf{T}^{(h_2)}$.

To compare neural audio codec tokens with self-supervised discrete units, we also construct **MERT k-means** tokens from MERT-v0-public hidden states (Li et al., 2024c)[4]. We use layers $\{0, 1, 11, 12\}$, cluster frame features with Mini-Batch k-means (Sculley, 2010), and emit four streams of discrete units. Layers $0, 1$ are mapped to $\mathbf{T}^{(\ell_1)}, \mathbf{T}^{(\ell_2)}$, and layers $11, 12$ are mapped to $\mathbf{T}^{(h_1)}, \mathbf{T}^{(h_2)}$.

Continuous MERT ablation. To separate the effect of MERT representations from the effect of discretization, we also evaluate a continuous-input ablation. This model uses the same low/high Transformer backbone as CoMoE, but replaces the token embedding lookup with a linear projection of continuous MERT-v0 frame features. We use the same four MERT layers, $\{0, 1, 11, 12\}$, mapping layers $0, 1$ to the lower-level branch and layers $11, 12$ to the higher-level branch. This variant is not a discrete-token CoMoE model; it is included only to test whether the MERT k-means result is due to discretization or to the underlying MERT representation.

Baselines. We include two non-CoMoE baselines. MLP (MERT) uses mean-pooled MERT-v0-public features followed by a small multilayer perceptron. CLAM (Batra et al., 2025) is the dual-rate reference detector from the original benchmark, using MERT and Wav2Vec2 streams with weighted cross-attention. The MERT-MLP and CLAM baselines follow their respective recipes.

Training. All CoMoE variants are trained with the same recipe: 12 epochs of AdamW (Loshchilov & Hutter, 2019), learning rate 2×10^{-4}, label smoothing 0.05, seed 42, and a single H100 GPU. The MERT-MLP and CLAM baselines follow their respective baseline recipes.

4. MoM-OPEN and Source-Restricted Splits

Dataset. We construct MoM-OPEN, an open reconstruction of MoM-CLAM. Since the original benchmark relies on YouTube-derived real audio that is difficult to redistribute or reliably rebuild, we replace the real half with FMA-medium and MTG-Jamendo while keeping the original fake-generator protocol. These corpora have been widely used for music information retrieval and audio-based music analysis tasks, including tagging, genre analysis, and popularity prediction (Defferrard et al., 2017; Bogdanov et al., 2019; Lee & Lee, 2018). Table 1 summarizes the resulting 146,309 clips. All clips are normalized to a shared audio representation by standardizing duration, channel configuration, sampling rate, codec, and metadata handling.

Source-restricted splits. Table 2 defines the evaluation splits. The base split follows the original fake-generator partition. Real-source restriction tests whether detectors rely on FMA- or Jamendo-specific real-corpus cues, while fake-source restriction tests whether a detector trained on one fake generator source transfers to unseen fake sources.

[1] huggingface.co/facebook/encodec_24khz

[2] github.com/descriptinc/descript-audio-codec

[3] huggingface.co/m-a-p/xcodec_mini_infer

[4] huggingface.co/m-a-p/MERT-v0-public

Table 3. OOD AUC (%) on MOM-OPEN across the base split, real-source-restricted splits, and fake-source-restricted splits. Split names indicate the source retained in training. Values in parentheses are absolute changes from the corresponding base AUC in percentage points.

Model	base	REAL-FMA	REAL-JAMENDO	FAKE-SUNO3.5	FAKE-UDIO
CLAM	99.92	**99.71** (−0.2)	**99.85** (−0.1)	**97.72** (−2.2)	66.51 (−33.4)
MLP (MERT)	99.77	99.07 (−0.7)	99.47 (−0.3)	86.87 (−12.9)	67.45 (−32.3)
COMOE (X-Codec)	**99.93**	99.62 (−0.3)	99.73 (−0.2)	86.97 (−13.0)	**89.04** (−10.9)
COMOE (DAC)	99.82	98.98 (−0.8)	99.51 (−0.3)	88.33 (−11.5)	77.28 (−22.6)
COMOE (EnCodec)	96.44	95.64 (−0.8)	94.76 (−1.7)	85.15 (−11.3)	58.64 (−37.8)
COMOE (MERT k-means)	99.83	99.14 (−0.7)	99.53 (−0.3)	92.22 (−7.6)	73.26 (−26.6)
MERT-CONTINUOUS (same backbone)	99.87	99.01 (−0.9)	99.57 (−0.3)	93.84 (−6.0)	71.91 (−28.0)

Table 4. Held-out-fake detection rate (%) under the validation-selected threshold.

Model	FAKE-SUNO3.5	FAKE-UDIO
CLAM	**71.0**	2.6
MLP (MERT)	60.1	26.0
COMOE (X-Codec)	38.7	**45.1**
COMOE (EnCodec)	43.8	23.5
COMOE (DAC)	61.4	29.2
COMOE (MERT k-means)	51.9	17.3
MERT-CONTINUOUS	49.9	7.8

5. Results

Metrics and validation. For each condition, validation examples are drawn only from the sources retained in the training split; held-out fake sources are never used for threshold selection. We report AUC and held-out-fake detection rate. The latter uses a threshold $\tau^\star$ selected by maximizing validation F1 and then applied unchanged to each held-out generator source.

Base and real-source-restricted splits are nearly saturated. Table 3 shows that the base split is close to saturated for all strong detectors: CLAM, MLP (MERT), and most COMOE variants reach AUCs near 99.8–99.9%, except for the lower but still high EnCodec-token COMOE. Real-source restriction is also mild, with much smaller drops than the fake-source-restricted conditions.

Fake-source restriction exposes large model differences. The rightmost two columns of Table 3 are much more discriminative than the base or real-source-restricted splits. In FAKE-SUNO3.5, CLAM remains strongest. In FAKE-UDIO, however, CLAM drops sharply, while COMOE with X-Codec tokens becomes the strongest configuration.

Token-space identity is the dominant factor among fixed-architecture COMOE variants. Because all COMOE rows in Table 3 use the same classifier, their differences isolate the input token space. Under FAKE-UDIO, EnCodec drops to 58.64%, DAC improves over EnCodec but remains below X-Codec, and X-Codec reaches 89.04%. MERT k-means is strongest among COMOE variants on FAKE-SUNO3.5, whereas X-Codec is strongest on FAKE-UDIO.

Pooled MERT features alone are not sufficient. The MLP (MERT) baseline in Table 3 tests whether mean-pooled continuous music-SSL features alone explain the robustness gains. Although it is strong on the base and real-source-restricted splits, it drops substantially under fake-source restriction, especially FAKE-UDIO. Thus, the X-Codec result cannot be explained simply by using a music-pretrained representation; sequential token structure also matters.

Discretization alone does not explain AUC, but affects operating-point stability. The MERT-CONTINUOUS ablation in Table 3 uses the same low/high Transformer backbone as MERT k-means, but replaces discrete units with continuous MERT frame features. It improves AUC on FAKE-SUNO3.5, but is slightly worse on FAKE-UDIO; thus, AUC differences are not explained by discretization alone. However, Table 4 shows a larger operating-point gap: under FAKE-UDIO, MERT k-means retains 17.3% held-out-fake detection rate, while MERT-continuous drops to 7.8%.

AUC and operating-point behavior diverge. Table 4 shows that validation-selected thresholds do not always transfer to held-out fake sources. The clearest case is CLAM: under FAKE-UDIO, it retains non-random AUC in Table 3, but its held-out-fake detection rate drops to 2.6%. In contrast, COMOE with X-Codec tokens gives the best FAKE-UDIO detection rate and the smallest cross-condition gap, suggesting that fake-source restriction should be evaluated with both ranking and operating-point metrics.

6. Conclusion

We presented a controlled study of cross-generator AI-generated music detection in which the downstream classifier is fixed and only the audio token space is varied. Experiments on MOM-OPEN show that standard and real-source-restricted splits are nearly saturated, while fake-source restriction reveals large differences between token spaces. These results suggest that codec-style discrete token spaces should be treated as a primary experimental axis in music deepfake detection, rather than as a preprocessing detail. However, the study has some limitations: MOM-OPEN is an open reconstruction, and X-Codec mini is not lineage-free with respect to YuE-related tooling. Future work should evaluate more generator sources, control training-pool size, and test calibration or fusion methods under generator shift.

Acknowledgement. This work was supported by JSPS KAKENHI Grant Number 26KJ0771.

References

Afchar, D., Meseguer-Brocal, G., and Hennequin, R. Detecting music deepfakes is easy but actually hard. *arXiv preprint arXiv:2405.04181*, 2024.

Afchar, D., Meseguer-Brocal, G., and Hennequin, R. AI-generated music detection and its challenges. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2025.

Batra, A., Sharma, D., Thukral, K., Bhatia, R., Batra, N., and Gautam, A. Melody or machine: Detecting synthetic music with dual-stream contrastive learning. *Transactions on Machine Learning Research*, 2025. ISSN 2835–8856.

Bogdanov, D., Won, M., Tovstogan, P., Porter, A., and Serra, X. The MTG-Jamendo dataset for automatic music tagging. In *Machine Learning for Music Discovery Workshop at the International Conference on Machine Learning (ICML)*, 2019.

Comanducci, L., Bestagini, P., and Tubaro, S. FakeMusicCaps: A dataset for detection and attribution of synthetic music generated via text-to-music models. *Journal of Imaging*, 11(7), 2025.

Cros Vila, L., Sturm, B. L. T., Casini, L., and Dalmazzo, D. The AI music arms race: On the detection of AI-generated music. *Transactions of the International Society for Music Information Retrieval*, 8(1):179–194, 2025.

Defferrard, M., Benzi, K., Vandergheynst, P., and Bresson, X. FMA: A dataset for music analysis. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pp. 316–323, 2017.

Défossez, A., Copet, J., Synnaeve, G., and Adi, Y. High fidelity neural audio compression. *Transactions on Machine Learning Research*, 2023. ISSN 2835–8856.

Forsgren, S. and Martiros, H. Riffusion: Stable diffusion for real-time music generation. https://riffusion.com/about, 2022.

Kumar, R., Seetharaman, P., Luebs, A., Kumar, I., and Kumar, K. High-fidelity audio compression with improved RVQGAN. In *Advances in Neural Information Processing Systems*, volume 36, pp. 27980–27993, 2023.

Lee, J. and Lee, J.-S. Music popularity: Metrics, characteristics, and audio-based prediction. *IEEE Transactions on Multimedia*, 20(11):3173–3182, 2018.

Li, X., Li, K., Zheng, Y., Yan, C., Ji, X., and Xu, W. SafeEar: Content privacy-preserving audio deepfake detection. In *Proceedings of the 2024 ACM SIGSAC Conference on Computer and Communications Security*, pp. 3585–3599, 2024a.

Li, Y., Milling, M., Specia, L., and Schuller, B. W. From audio deepfake detection to ai-generated music detection–a pathway and overview. *arXiv preprint arXiv:2412.00571*, 2024b.

Li, Y., Yuan, R., Zhang, G., Ma, Y., Chen, X., Yin, H., Xiao, C., Lin, C., Ragni, A., Benetos, E., Gyenge, N., Dannenberg, R. B., Liu, R., Chen, W., Xia, G., Shi, Y., Huang, W., Wang, Y., Guo, Y., and Fu, J. MERT: Acoustic music understanding model with large-scale self-supervised training. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024c.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2019.

Ning, Z., Chen, H., Jiang, Y., Hao, C., Ma, G., Wang, S., Yao, J., and Xie, L. Diffrhythm: Blazingly fast and embarrassingly simple end-to-end full-length song generation with latent diffusion. *arXiv preprint arXiv:2503.01183*, 2025.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning*, volume 139, pp. 8748–8763, 2021.

Rahman, M. A., Hakim, Z. I. A., Sarker, N. H., Paul, B., and Fattah, S. A. SONICS: Synthetic or not – identifying counterfeit songs. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025.

Sculley, D. Web-scale K-means clustering. In *Proceedings of the 19th International Conference on World Wide Web*, pp. 1177–1178, 2010.

Suno, Inc. Suno AI music generator. https://suno.com, 2024.

Udio Inc. Udio AI music platform. https://udio.com, 2024.

Wu, H., Tseng, Y., and Lee, H.-y. CodecFake: Enhancing anti-spoofing models against deepfake audios from codec-based speech synthesis systems. In *Interspeech*, pp. 1770–1774, 2024.

Wu, J., Pan, Z., Zhang, Q., Bhupendra, S. H., and Mondal, S. Quantizer-aware hierarchical neural codec modeling for speech deepfake detection. *arXiv preprint arXiv:2603.16914*, 2026.

Yan, S., Li, O., Cai, J., Hao, Y., Jiang, X., Hu, Y., and Xie, W. A sanity check for AI-generated image detection. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025.

Ye, Z., Sun, P., Lei, J., Lin, H., Tan, X., Dai, Z., Kong, Q., Chen, J., Pan, J., Liu, Q., Guo, Y., and Xue, W. Codec does matter: Exploring the semantic shortcoming of codec for audio language model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 25697–25705, 2025.

Yuan, R., Lin, H., Guo, S., Zhang, G., Pan, J., Zang, Y., Liu, H., Liang, Y., Ma, W., Du, X., Du, X., Ye, Z., Zheng, T., Jiang, Z., Ma, Y., Liu, M., Tian, Z., Zhou, Z., Xue, L., Qu, X., LI, Y., Wu, S., Shen, T., Ma, Z., Zhan, J., Wang, C., Wang, Y., Chi, X., Zhang, X., Yang, Z., XiangzhouWang, Liu, S., Mei, L., Li, P., Wang, J., Yu, J., Pang, G., Li, X., Wang, Z., Zhou, X., Yu, L., Benetos, E., Chen, Y., Lin, C., Chen, X., Xia, G., Zhang, Z., Zhang, C., Chen, W., Zhou, X., Qiu, X., Dannenberg, R., Liu, J., Yang, J., Huang, W., Xue, W., Tan, X., and Guo, Y. Yue: Scaling open foundation models for long-form music generation. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2026.

Zeghidour, N., Luebs, A., Omran, A., Skoglund, J., and Tagliasacchi, M. Soundstream: An end-to-end neural audio codec. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:495–507, 2022.