WEAK TOKENIZATION: A PRELIMINARY STUDY OF DYNAMIC AUDIO CHUNKING FOR IRREGULAR MUSIC GENERATION

Weixi Zhai
Quanzhou Normal University
wishzhai@gmail.com

ABSTRACT

Most large language models (LLMs) for music generation rely on strong tokenization, discretizing audio into fixed, uniform units. While effective for producing stylistically coherent outputs, such models struggle with genres like IDM and Glitch, where irregularity is central to the aesthetic. Inspired by tokenizer-free trends in NLP, we investigate the potential of an alternative framework combining: (1) a Dynamic Chunking mechanism that segments audio based on content similarity rather than fixed grids, and (2) the L-Score, a learnable complexity metric spanning timbral, rhythmic, and structural dimensions.

Preliminary results indicate that while the model captures some spectral features, it fails to produce rhythmic control—instead, it generates chaotic rather than deliberately irregular patterns. This limitation motivates future work on modeling controlled deviance in music generation—moving beyond statistical complexity toward learnable representations of aesthetic misdirection and expectation violation. (Our code is available at: https://github.com/wishzhai/WeakTOK)

1. INTRODUCTION

Technological innovation provides the tools for musical evolution, but cultural forces determine its creative trajectory. While technologies like synthesizers facilitated the rise of electronic music, their application is frequently a site of cultural negotiation. Intelligent Dance Music (IDM) serves as a prime example. Leveraging the same technologies as mainstream House and Techno, IDM consciously subverts the functional regularity of four-on-the-floor rhythms, instead embracing irregular time signatures, audio glitches, and non-linear forms. The genre's name is not an assertion of superiority but an ironic critique of its commercial counterparts' aesthetic predictability. Thus, IDM demonstrates how the cultural impulse of resisting homogeneity reappropriates technological means to expand aesthetic possibilities.

Today, large language models (LLMs) for music generation present a similar technological moment, marked by both creative potential and the risk of aesthetic flattening. These models demonstrate strong stylistic fluency, producing music aligned with established genres. While this enables coherent composition at scale, it also promotes uniformity, driven not only by training data but by how musical structure is tokenized.

Contemporary approaches across both symbolic and audio domains rely on discrete, uniform segmentation of musical content. In the symbolic domain, event-based representations like REMI [1] and MuMIDI [2], subword tokenization [3] techniques, and ABC notation adaptations such as MuPT [4] all impose regular structural assumptions. Similarly, in the audio domain, neural codecs discretize waveforms into acoustic tokens for autoregressive models. For instance, MusicGen [5] adopts EnCodec [6], while MusicLM [7] uses SoundStream for quantization. Even recent frameworks like CLAMP [8], which abstracts music into bar-level units, or JASCO [9], which aligns symbolic and acoustic tracks, maintain this fundamental dependence on grid-based organization.

IDM's aesthetic force emerges not from randomness, but from the intentional disruption of structural expectations. The genre relies on irregular time signatures, asymmetric phrase structures, and non-linear temporal development that resist conventional grid-based organization. Inspired by this challenge, we ask whether the concept of dynamic chunking can be transferred to the audio domain, enabling LLMs to reason about music in a manner analogous to IDM artists—through a deliberately anti-linear approach to structure. Our investigation focuses on two core components:

1. **A Dynamic Audio Chunking Framework for Weak Tokenization:** Inspired by tokenizer-free NLP models [10, 11], we investigate a generative architecture where segmentation is learned directly from audio features, enabling musically-aware, content-driven units rather than fixed temporal grids.

2. **A Learnable Objective for Cognitive Complexity:** We introduce the L-Score, a multi-dimensional complexity metric that spans timbral, rhythmic, and structural axes. By encouraging the model to match target distributions of L-Score, we aim to replace conventional notions of "pleasantness" with a learn-

able proxy for listener challenge and aesthetic tension.

2. METHOD

Our architecture combines a U-Net-inspired encoder–decoder with a Dynamic Audio Chunking module and a Transformer encoder, forming a hierarchical local-to-global modeling pipeline well-suited to IDM.

2.1 U-Net for Local Feature Extraction.

We adopt a 3-layer U-Net-style convolutional backbone to extract time-frequency features from input spectrograms. The encoder progressively downsamples and compresses local information via strided convolutions, capturing transient patterns, percussive events, and glitch artifacts. The decoder mirrors this structure with upsampling and skip connections, ensuring that fine-grained local details are preserved during reconstruction. This structure allows the model to effectively model localized rhythmic and spectral features—a hallmark of IDM aesthetics.

2.2 Dynamic Chunking for Semantic Segmentation.

To handle IDM's non-metric and unpredictable temporal structure, we introduce a `DynamicChunkingLayer` that adaptively segments sequences based on learned content similarity. Operating on the U-Net encoder output $\mathbf{X} \in \mathbb{R}^{B\times T\times D}$ (with batch size B, time steps T, and feature dimension D), the layer discovers variable-length chunks aligned with perceptual discontinuities such as sudden kicks, silences, or glitchy transitions. The full process is illustrated in Figure 1, and comprises three stages:

(1) Boundary Detection via Cosine Similarity. We project frame-wise features into learnable query and key spaces:

$$\mathbf{Q} = \mathbf{X}\mathbf{W}_q, \quad \mathbf{K} = \mathbf{X}\mathbf{W}_k \tag{1}$$

$$s_t = \text{cosine\_similarity}(\mathbf{Q}_{t+1}, \mathbf{K}_t) \tag{2}$$

$$p_{\text{boundary},t} = 1.0 - s_t \tag{3}$$

Here, lower similarity implies a likely perceptual boundary—e.g., a transition from glitch to silence or a micro-cut.

(2) Differentiable Boundary Sampling. Using Gumbel-Softmax, we sample boundary indicators while maintaining differentiability:

$$\boldsymbol{\ell}_t = [p_{\text{boundary},t}, 1 - p_{\text{boundary},t}] \tag{4}$$

$$\mathbf{b}_t = \text{Gumbel-Softmax}(\boldsymbol{\ell}_t, \tau) \tag{5}$$

(3) Chunk-wise Pooling. Identified chunks are mean-pooled into embeddings, reducing sequence length while retaining semantic coherence:

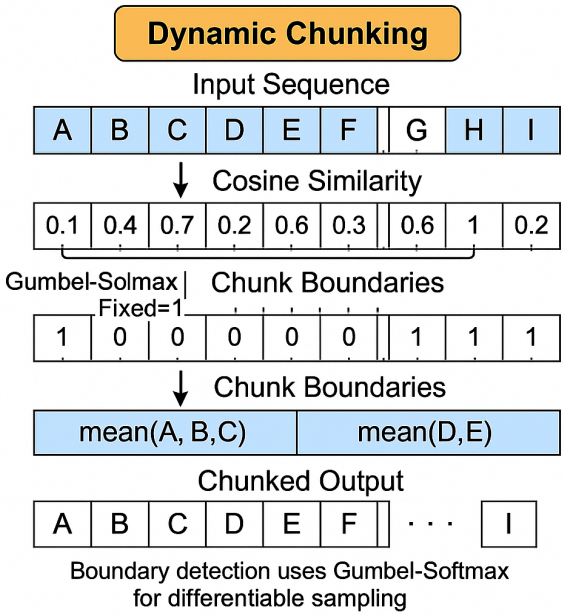

Figure 1. Dynamic Audio Chunking pipeline. Cosine similarity guides boundary prediction, Gumbel-Softmax sampling enables differentiable segmentation, and mean pooling produces compressed chunk embeddings.

$$c_t = \text{cumsum}(b_{t,0}) \tag{6}$$

$$\mathbf{O}_i = \frac{1}{|\{t : c_t = i\}|} \sum_{t:c_t=i} \mathbf{X}_t \tag{7}$$

This transformation condenses irregular temporal structures—such as micro-phrases and fragmented motifs—into discrete chunk-level representations.

2.3 Transformer for Global Structure Modeling.

The resulting chunk embeddings are passed to a Transformer encoder, which models long-range dependencies across segments. This enables the model to learn high-level temporal relationships such as motif repetition, structural contrast, and cross-time thematic links, which are crucial to IDM's layered and often non-linear form.

2.4 Local-to-Global Modeling Pipeline.

In summary, the architecture builds a local-to-global hierarchy: convolutional layers extract frame-level features, chunking segments these into musically coherent events, and the Transformer attends across chunks to model global structure. This layered composition is especially suited to IDM, which relies on intricate local textures (e.g., micro-glitches) as well as overarching structural aesthetics (e.g., rhythmic illusion, abrupt transitions, and broken repetition).

The complete flow is illustrated in Figure 2.

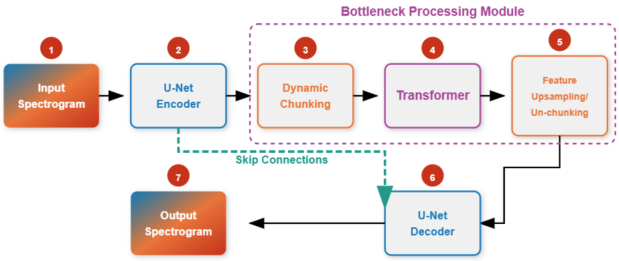

Figure 2. Overview of the model architecture. The model utilizes a U-Net structure where the core bottleneck module (within the purple dashed box) consists of three components: Dynamic Chunking, a Transformer, and a Feature Upsampling/Un-chunking layer. The output from the encoder is processed by this bottleneck module and then reconstructed into the final output spectrogram by the decoder, which is aided by skip connections.

3. L-SCORE: A COMPLEXITY-GUIDED TRAINING FRAMEWORK

To guide the generative process towards producing audio with specific, desirable characteristics of Intelligent Dance Music (IDM), we introduce the L-Score, a multi-dimensional complexity vector. Instead of relying solely on reconstruction accuracy, our training framework uses the L-Score to impose a statistical prior on the complexity of the generated audio, preventing mode collapse and enhancing musical structure.

3.1 L-Score Definition

The L-Score is a four-dimensional vector, $\mathbf{L} \in \mathbb{R}^4$, where each component quantifies a distinct aspect of musical complexity. The components are:

- **Timbral Complexity (L_T):** Measured by the spectral entropy of the generated spectrogram. This metric captures the richness and variability of the sound texture. As it is computed directly on the spectrogram, this component is fully differentiable and can be backpropagated through.
- **Rhythmic Density (L_{RD}):** Defined as the number of detected onsets per second in the reconstructed audio waveform. This component measures the overall rhythmic activity.
- **Rhythmic Irregularity (L_{RI}):** Calculated as the standard deviation of the inter-onset intervals (IOIs). This metric quantifies the predictability and complexity of the rhythm, distinguishing between regular, metronomic patterns and more syncopated, complex ones.
- **Structural Complexity (L_S):** Assessed using the off-diagonal self-similarity of a chroma-based feature representation. This component measures the degree of repetition and variation in the harmonic or melodic structure over time.

The latter three components (L_{RD}, L_{RI}, L_S) are computed on the reconstructed audio waveform and are therefore non-differentiable. They guide the training process through the distribution loss described below.

3.2 L-Score Distribution Loss

Rather than forcing the model to match a single, fixed complexity target, which could stifle creativity, we encourage it to generate audio whose complexity profile matches the statistical distribution of the training dataset. We define the L-Score Distribution Loss ($\mathcal{L}_{L-Score}$) as the L1 distance between the statistics (mean and standard deviation) of the L-Scores from a generated batch and the pre-computed target statistics from the entire dataset.

Let $\boldsymbol{\mu}_{target}$ and $\boldsymbol{\sigma}_{target}$ be the target mean and standard deviation vectors of the L-Score, and let $\boldsymbol{\mu}_{batch}$ and $\boldsymbol{\sigma}_{batch}$ be the corresponding statistics for a batch of generated samples. The loss is formulated as:

$$\mathcal{L}_{L-Score} = \lambda_\mu ||\boldsymbol{\mu}_{batch} - \boldsymbol{\mu}_{target}||_1 + \lambda_\sigma ||\boldsymbol{\sigma}_{batch} - \boldsymbol{\sigma}_{target}||_1 \quad (8)$$

where λ_μ and λ_σ are hyperparameters balancing the two statistical moments. This approach ensures that the generated audio exhibits a similar range and average complexity as the source material, promoting diversity and structural integrity.

3.3 Curriculum Learning Strategy

Directly optimizing for both reconstruction and L-Score loss from the beginning of training can be unstable, as the model may receive conflicting gradients before it has learned to produce coherent audio. To mitigate this, we employ a curriculum learning strategy. The total loss function is a weighted sum of a reconstruction loss ($\mathcal{L}_{recon}$, e.g., L1 loss on the spectrogram) and the L-Score loss:

$$\mathcal{L}_{total} = w_{recon}\mathcal{L}_{recon} + w_L\mathcal{L}_{L-Score} \quad (9)$$

The training is divided into two phases:

1. **Phase 1: Reconstruction Focus.** For an initial number of epochs, we set $w_L = 0$ and $w_{recon} = 1$. In this phase, the model learns to faithfully reconstruct audio from the latent representation, establishing a stable foundation for generation.
2. **Phase 2: Complexity Guidance.** Following the initial phase, we gradually introduce the L-Score loss by linearly annealing its weight w_L from 0 to its final value over a set number of epochs. Concurrently, w_{recon} can be held constant or annealed. This allows the model to first learn the basics of audio generation before being guided towards the more abstract and complex stylistic targets defined by the L-Score.

This curriculum-based framework stabilizes training and enables the model to effectively learn both the content and the complex structural properties of the target musical style.

4. EXPERIMENTS

4.1 Dataset

We use the Freeloop [12] dataset for our experiments. To curate a corpus aligned with our research focus, we selected all tracks tagged with the keywords "IDM" and "glitch". This filtering process yielded a specialized dataset of 373 tracks that embody the unconventional rhythmic and textural characteristics relevant to our study. The dataset was subsequently partitioned into training, validation, and test sets following a 70/15/15 split. This resulted in 261 tracks for training, 56 for validation, and 57 for testing.

4.2 Implementation Details

The experiments were conducted using an implementation of the Music Transformer, built with PyTorch and leveraging the Hugging Face transformers library [13].

5. EXPERIMENTAL RESULTS AND DISCUSSION

Our model was trained using a curriculum learning strategy, which proved effective for stable training. The process completed over 50 epochs without numerical instability, achieving a final test loss of 1.79. The curriculum consisted of two phases: an initial 10-epoch warmup focused solely on reconstruction (L-Score weight at 0.0), followed by a 40-epoch guidance phase where the L-Score weight was gradually increased as the reconstruction weight was annealed to 0.1.

We evaluated the final model's performance by quantitatively comparing the L-Score of generated audio against predefined targets and by qualitatively analyzing its musicality. The key results are summarized in Table 1.

L-Score Dimension	Target	Uncon	Seed
Timbral Complexity	0.4	**0.409**	0.502
Rhythmic Density	0.8	*2.962*	*6.667*
Rhythmic Irregularity	0.6	0.739	0.067
Structural Complexity	0.5	**0.496**	0.571

Table 1. Comparison of Target L-Scores with Generated Audio. Values that closely match the target are in bold. Values that critically fail to match are in *italics*.

5.1 Quantitative Analysis

As shown in Table 1, the model demonstrates partial success in aligning with the complexity targets. In the unconditional generation setting, both timbral complexity (0.409 vs. 0.4) and structural complexity (0.496 vs. 0.5) closely match the predefined targets. This suggests that our L-Score loss and curriculum strategy can effectively guide these specific spectral and structural attributes of the generated audio.

However, the results also reveal a critical failure in controlling rhythmic features. The rhythmic density in both unconditional (2.962) and seed-based (6.667) generation is dramatically higher than the target of 0.8. This indicates that our current loss function is insufficient for regulating the temporal characteristics of the output, leading to rhythmically oversaturated and chaotic results.

5.2 Qualitative Analysis and Key Limitations

The quantitative shortcomings of our method manifest as perceptual and musical limitations, which we highlight as critical directions for future research.

1. Limited Perceptual Coherence and Musicality: While the model matches certain statistical targets, the generated audio often lacks the aesthetic nuance and compositional intent characteristic of human-authored IDM. Outputs are frequently perceived as overly dense or noisy, suggesting that statistical complexity alone does not guarantee musical coherence.

2. Insufficient Rhythmic Control: The most prominent failure mode lies in the model's inability to regulate rhythmic density. Generated audio tends to exhibit uncontrolled bursts of micro-events, resulting in chaotic textures rather than the precise, deliberate irregularity that defines IDM's rhythmic identity.

3. High Dependence on Seed Material: Unconditional generations show significant degradation in structure and coherence compared to seed-conditioned outputs. This suggests the model operates more effectively as a transformation layer than a generative composer, relying heavily on structural priors from the input.

In sum, while our framework shows initial promise in targeting spectral and structural properties, its limitations in rhythmic control and generative autonomy underscore the challenge of modeling anti-functional musical aesthetics. These findings invite further exploration into architectures that balance content-awareness with controlled deviation, particularly in genres that deliberately resist regularization.

6. REFERENCES

[1] Y.-S. Huang and Y.-H. Yang, "Pop music transformer: Beat-based modeling and generation of expressive pop piano compositions," in *ACM MM 2020*, 2020, pp. 1180–1188.

[2] Y. Ren, J. Liu, X. Chen, and et al., "Popmag: Pop music accompaniment generation," in *ACM MM 2020*, 2020, pp. 1198–1206.

[3] A. Kumar and P. Sarmento, "From words to music: A study of subword tokenization techniques in symbolic music generation," 2023. [Online]. Available: https://arxiv.org/abs/2304.08953

[4] X. Qu, Y. Bai, Y. Ma, Z. Zhou, K. M. Lo, J. Liu, R. Yuan, L. Min, X. Liu, T. Zhang, X. Du, S. Guo, Y. Liang, Y. Li, S. Wu, J. Zhou, T. Zheng, Z. Ma, F. Han, W. Xue, G. Xia, E. Benetos, X. Yue, C. Lin, X. Tan, S. W. Huang, J. Fu, and G. Zhang, "Mupt: A generative symbolic music

pretrained transformer," 2024. [Online]. Available: https://arxiv.org/abs/2404.06393

[5] J. Copet, F. Kreuk, I. Gat, T. Remez, D. Kant, G. Synnaeve, Y. Adi, and A. Défossez, "Simple and controllable music generation," 2024. [Online]. Available: https://arxiv.org/abs/2306.05284

[6] A. Défossez, J. Copet, G. Synnaeve, and Y. Adi, "High fidelity neural audio compression," *arXiv preprint arXiv:2210.13438*, 2022.

[7] A. Agostinelli, T. I. Denk, Z. Borsos, J. Engel, M. Verzetti, A. Caillon, Q. Huang, A. Jansen, A. Roberts, M. Tagliasacchi, M. Sharifi, N. Zeghidour, and C. Frank, "Musiclm: Generating music from text," 2023. [Online]. Available: https://arxiv.org/abs/2301.11325

[8] S. Wu, D. Yu, X. Tan, and M. Sun, "Clamp: Contrastive language-music pre-training for cross-modal symbolic music information retrieval," 2023.

[9] O. Tal, A. Ziv, I. Gat, F. Kreuk, and Y. Adi, "Joint audio and symbolic conditioning for temporally controlled text-to-music generation," 2024. [Online]. Available: https://arxiv.org/abs/2406.10970

[10] B. Deiseroth, M. Brack, P. Schramowski, K. Kersting, and S. Weinbach, "T-free: Subword tokenizer-free generative llms via sparse representations for memory-efficient embeddings," 2025. [Online]. Available: https://arxiv.org/abs/2406.19223

[11] A. T. Owodunni, O. Ahia, and S. Kumar, "Flexitokens: Flexible tokenization for evolving language models," 2025. [Online]. Available: https://arxiv.org/abs/2507.12720

[12] A. Ramires, F. Font, D. Bogdanov, J. B. L. Smith, Y.-H. Yang, J. Ching, B.-Y. Chen, Y.-K. Wu, H. Wei-Han, and X. Serra, "The freesound loop dataset and annotation tool," in *Proc. of the 21st International Society for Music Information Retrieval (ISMIR)*, 2020.

[13] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush, "Huggingface's transformers: State-of-the-art natural language processing," 2020. [Online]. Available: https://arxiv.org/abs/1910.03771