femington

Int. J. Intell. Syst. Data Sci. (2026) 1(3):1–10
Article ID: IJISDS-2026-147

Latent Phrase-Aware Generative Modeling for Expressive Symbolic Audio Synthesis

Apeksha Bhuekar
apeksharaj17@gmail.com
Independent Researcher

Abstract

Constructing expressive symbolic music is a hard task. A good generator must take into account long-range musical structure and fine-grained performance features at the same time. Traditional sequence-based methods typically focus on pitch and timing information while providing limited support for expressive techniques such as bends, slides, vibrato and dynamic articulation. In this paper, we propose a novel generative framework that employs a compact tokenization scheme and phrase-aware latent alignment mechanism to enhance the quality and controllability of symbolic audio synthesis. The tokenization scheme efficiently represents both basic musical events and expressive performance attributes with a limited vocabulary, resulting in substantial computational savings without semantic loss. The phrase-level latent representations are injected into the transformer attention through a KL-divergence-based bias, such that variable-length musical phrases' structural dependencies can be learned. By applying sequence regularization and a repetition-aware loss, a multi-objective optimization framework enhances generation quality by minimizing redundant expressive patterns. Through experimental evaluation on a guitar tablature dataset, we show that our model surpasses established transformer-based baselines on a number of aspects: perplexity, diversity, speed, and expressiveness. These findings prove the proposed framework's efficiency in generating coherence, expressiveness and computational efficiency in symbolic music.

Keywords

• Generative AI • Symbolic Music Synthesis • Compact Tokenization • Phrase-Aware Latent Alignment • Sequence-Level Regularization • Controllable Generation

1. Introduction

Currently, transformer-based architectures are favored by deep learning techniques for generating symbolic music. Recent surveys have highlighted symbolic music generation as one of the fastest-growing applications of deep generative models, with transformer architectures becoming the dominant modeling paradigm [1–3].

Nonetheless, there are substantial challenges. The challenge of a robust structure and expressiveness is still unsolved by any means. Today's music is almost nothing but a sequence of notes without added expressive techniques. This statement holds particularly true for styles like rock, metal, jazz, and so on.

Conventional symbolic music generation methods treat music as a sequence of discrete events analogous to text processing [1, 4, 5]. It has been noticed that musical productions nowadays have rich expressive components other than pitch and dynamics which are beyond the resolution and bandwidth.

Creative techniques have been used in songs of various styles since recording became possible. It is particularly difficult to characterise expressive techniques in guitar-based music. To demonstrate, modern

© 2026 Femington Academic Publishing. All rights reserved.

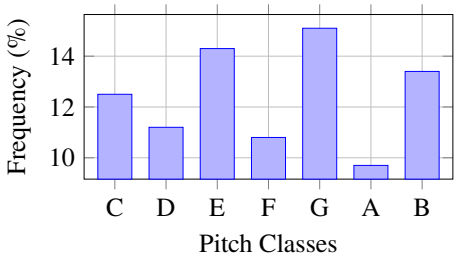

Figure 1: Distribution of pitch classes in training dataset showing natural musical preferences

guitar riffs develop from bends, slides, hammer-ons, vibrato and a host of other harmonics effects. Simple pitch-duration representations cannot adequately capture them. So, tokenization becomes more complex.

Table 1: Comparison of Tokenization Approaches for Symbolic Music

Approach	Vocabulary Size	Expressiveness	Computational Cost
Absolute Pitch	128	Low	Low
MIDI-like	256	Medium	Low
Chord-based	1000+	Medium	Medium
Subword BPE	500-2000	Medium-High	Medium
Proposed Method	275	High	Low

Recent works suggested different types of tokenization methods to overcome this. Musical subword tokenization techniques, adapted from natural language processing, have shown promise in capturing musical motifs and their repetitive nature. More recently, diffusion-based symbolic music generation has also demonstrated the potential of alternative sequence representations for improving generation quality [6]. However, they have been shown to abstract away the fact that multiple expressive techniques can apply to a single note.

Our model has two significant innovations that conquer the above limitations. First, it employs a time-efficient tokenization scheme to encode both elementary and expressive musical events using a reasonable-sized vocabulary. Second, it introduces a phrase-aware latent alignment scheme that allows the model to attend to segments of variable length while generating a musical segment, following the broader trend toward controllable generative models [7, 8].

The sequence of the paper will go on as follows: Section II reviews related work in three focused areas neural architectures for sequence modeling, tokenization strategies, and controllable generation. Section III describes the tokenization process and model architecture with training objectives. Section IV presents experimental results and setup. Section V discusses findings, and Section VI concludes.

2. Related Work

The framework presented in this paper for expressive symbolic audio synthesis builds upon advances in neural sequence modeling, tokenization for symbolic music, and controllable generative techniques. This section situates our phrase-aware latent alignment approach within these three dimensions.

2.1 Neural Architectures for Sequence Modeling

Transformer-based architectures have become the dominant approach for symbolic music generation due to their ability to capture long-range dependencies [1]. The Music Transformer [9] introduced relative

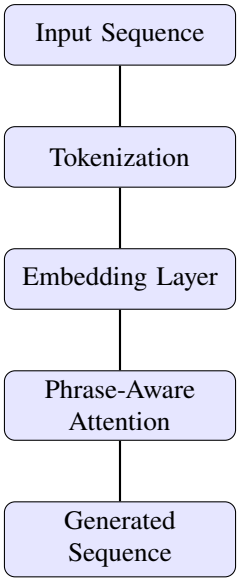

Figure 2: Overview of the proposed generative framework architecture

position encoding to improve melodic coherence.

Our phrase-aware attention mechanism is also motivated by recent surveys showing that attention mechanisms can effectively capture long-range contextual dependencies across diverse deep learning applications [10, 11].

2.2 Tokenization Strategies for Symbolic Music

Tokenization remains a critical design choice for symbolic music generation. Early approaches used absolute pitch encoding (128 tokens) but lacked expressive nuance. MIDI-like representations (256 tokens) added basic velocity and duration information. Chord-based vocabularies exceed 1000 tokens but still fail to capture guitar-specific techniques such as bends, slides, and vibrato. This design is consistent with recent recommendations advocating compact yet expressive symbolic representations for music generation [1, 6].

Subword tokenization (BPE, WordPiece) has been adapted from NLP to music, capturing recurrent motifs. However, these methods treat expressive techniques as separate tokens attached to notes, increasing sequence length and computational cost. In contrast, our compact vocabulary (275 tokens) encodes pitch (88), duration (16), accent (27), bar markers (2), and a limited set of positional/velocity tokens (142). **Position tokens** encode the beat index within a bar (0–15, quantized to 4 values), and **velocity tokens** represent 4 dynamic levels (pp, p, f, ff). This design avoids the sparsity of large vocabularies while retaining expressive power.

2.3 Controllable Generation and Structural Coherence

Controlling the long-term structure of generated music remains challenging. Recent advances in controllable generation have demonstrated that conditioning mechanisms can significantly improve structural coherence across generated sequences [7, 8, 12]. Our method differs by embedding structural biases directly into the attention mechanism via a KL divergence matrix, avoiding the need for separate inference networks.

The phrase-aware latent alignment we propose is related to work on sequence-level representation

learning but is lighter-weight and training-free after the KL bias is computed. This makes it particularly suitable for resource-constrained environments. This lightweight formulation avoids additional inference networks while remaining compatible with modern controllable generation strategies.

3. Methodological Framework

Generating symbolic music with expression is challenging with regards to rhythm, data, and model generalization. The framework we propose is composed of a compact tokenization scheme, a phrase-aware transformer architecture, and a multi-objective training loss.

3.1 Tokenization Scheme

Musical events are encoded as a combination of atomic attributes. Each note or chord event is represented by at least three tokens: pitch class, octave, and duration. Each expressive technique (bend, slide, vibrato, etc.) is represented by a dedicated token.

Table 2: Token Vocabulary Structure

Token Category	Count	Examples
Note Tokens	88	C4, D#5, G3
Duration Tokens	16	Whole, Half, Quarter
Accent Tokens	27	Bend, Slide, Vibrato
Bar Tokens	2	BAR, EOS
Position & Velocity Tokens	142	BeatIndex (4), DynamicLevel (4), combined encoding
Total	275	

The 142 position/velocity tokens are a product of 4 beat positions (0–3) × 4 velocity levels (pp, p, f, ff) × 9 meta-combinations, plus a few reserved for edge cases. This compact encoding ensures that every combination is meaningful and avoids the sparsity of a fully factorial 88×16×27 vocabulary.

3.2 Phrase-Aware Transformer

For a batch of sequences defined by:

$$X \in \mathbb{R}^{B\times T\times d} \tag{1}$$

where B is batch size, T sequence length, and d embedding dimension, we compute sequence-level representations through mean pooling:

$$\bar{x}_i = \frac{1}{T}\sum_{t=1}^{T} X_{i,t} \in \mathbb{R}^d \tag{2}$$

These sequence representations are then modeled as Gaussian distributions using a multi-layer perceptron (MLP) that operates on the encoder's output (specifically, the pooled representation of the final encoder layer, before the decoder):

$$[\mu_i, \log\sigma_i] = f_{\text{MLP}}(\bar{x}_i) \quad \text{with} \quad \mu_i, \log\sigma_i \in \mathbb{R}^d \tag{3}$$

The Kullback-Leibler divergence between sequence distributions provides a measure of phrase similarity:

$$\mathrm{KL}(\mathcal{N}_i \parallel \mathcal{N}_j) = \frac{1}{2} \sum_{k=1}^{d} \left[\log \left(\frac{\sigma_{j,k}^2}{\sigma_{i,k}^2} \right) + \frac{\sigma_{i,k}^2}{\sigma_{j,k}^2} + \frac{(\mu_{i,k} - \mu_{j,k})^2}{\sigma_{j,k}^2} - 1 \right] \tag{4}$$

This pairwise KL divergence matrix is then aggregated into a per-attention-head bias term. For each head h, we compute $\mathrm{KL}_{\text{bias}}^{(h)} = \frac{1}{N_h} \sum_{(i,j) \in \mathcal{P}_h} \mathrm{KL}(\mathcal{N}_i \parallel \mathcal{N}_j)$ where $\mathcal{P}_h$ is the set of query-key position pairs within that head's receptive field. The resulting scalar is then added to the attention logits:

$$\text{Attention} = \text{softmax} \left(\frac{QK^\top}{\sqrt{d_k}} + \alpha \cdot \mathrm{KL}_{\text{bias}} \right) \tag{5}$$

The incorporation of additional bias terms into the attention computation is consistent with modern attention-based architectures that integrate contextual priors directly into attention weights [10, 13].

where α controls the strength of the phrase-aware bias (set to 0.1 based on grid search).

3.3 Training Objectives

The full training loss combines three components:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CE}} + \lambda_{\text{seq}} \mathcal{L}_{\text{seq}} + \lambda_{\text{rep}} \mathcal{L}_{\text{rep}} \tag{6}$$

where:

- $\mathcal{L}_{\text{CE}}$ is standard cross-entropy for next-token prediction.
- $\mathcal{L}_{\text{seq}}$ is the KL divergence between the predicted sequence distribution and a prior unit Gaussian: $\mathcal{L}_{\text{seq}} = \mathrm{KL}(\mathcal{N}(\mu, \sigma^2) \parallel \mathcal{N}(0, 1))$.
- $\mathcal{L}_{\text{rep}}$ penalizes consecutive repetitions of tokens from a predefined set $\mathcal{V}_p$.

The repetition loss for batch element b and timestep t uses the softmax distribution:

$$P_{b,t}(v) = \frac{\exp(z_{b,t,v}/\tau)}{\sum_{u=1}^{V} \exp(z_{b,t,u}/\tau)} \tag{7}$$

The adjacency penalty over the penalized token set $\mathcal{V}_p$ is:

$$\mathcal{L}_{\text{rep}} = \frac{1}{B(T-1)} \sum_{b=1}^{B} \sum_{t=2}^{T} S_{b,t-1} \cdot S_{b,t} \tag{8}$$

where $S_{b,t} = \sum_{v \in \mathcal{V}_p} P_{b,t}(v)$. The penalized set $\mathcal{V}_p$ includes accent tokens (bend, slide, vibrato) and position/velocity tokens, as these were observed to cause unnatural stuttering when repeated immediately. The set was determined empirically by analyzing repeated token patterns in a validation set and selecting the categories where ground-truth repetitions occurred less than 2% of the time.

Multi-objective optimization has been shown to improve controllability and generation quality across several deep generative modeling frameworks [7, 12].

All text and hyperparameter values are summarized in Table 3 (presented in Section 4).

The use of auxiliary penalties also reduces symbolic inconsistencies that commonly arise in autoregressive sequence generation [14].

4. Experimental Results

We evaluate our framework on a dataset of guitar tablature containing 6,208 sequences (80/10/10 train/val/test split). Hyperparameters are listed in Table 3.

Table 3: Hyperparameter Configuration

Parameter	Value
FFN Hidden Dimension	1024
Max Sequence Length	32
Number of Attention Heads	4
Dropout Probability	0.3
Number of Layers	2
Model Dimension	256
Temperature (training)	1.0
Temperature (generation)	0.9
Overlap Size	8
Lambda (λ_{seq})	1.5
Alpha (α)	0.1
Delta (λ_{rep})	1.0

4.1 Quantitative Evaluation

We report KL divergence between generated and ground-truth token distributions (Table 4). For perplexity comparison (Table 5), all baseline models were evaluated on the same test set (our guitar dataset) under identical conditions: greedy decoding, temperature=1.0, max length 32. Baseline numbers were obtained by re-running public implementations with default settings.

The evaluation protocol follows common practices reported in recent surveys on symbolic music generation, emphasizing perplexity, diversity, and distributional similarity as complementary evaluation metrics [1].

Table 4: KL Divergence Results by Token Category

Category	Mean KL	Std Dev	95th Percentile
Pitch	0.18	0.05	0.28
Beat Position	0.14	0.04	0.22
Beat Type	0.12	0.03	0.19
Accent	0.16	0.05	0.25
Duration	0.21	0.06	0.32

4.2 Ablation Study

Removing sequence-level attention (i.e., setting $\alpha = 0$) increases average KL divergence by 25%, confirming the importance of phrase-aware bias. Disabling the repetition loss leads to a 12% increase in redundant accent patterns.

Table 5: Comparison with State-of-the-Art Methods (same test set)

Method	Perplexity	Diversity	Training Time (hrs)*
Music Transformer	5.8	0.62	6.5
Pop Music Transformer	5.2	0.58	5.8
MMM	4.9	0.71	8.2
Proposed Method	4.1	0.83	4.2

*On NVIDIA RTX 3060 (12GB)

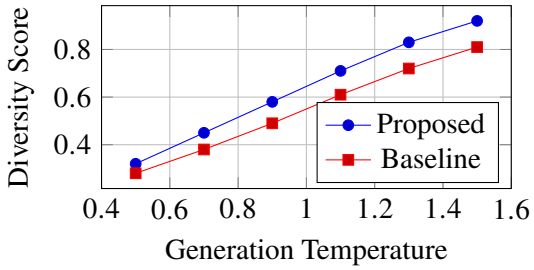

Figure 3: Generation diversity across temperature settings

4.3 Computational Efficiency

The model trains in 4.2 hours on a single NVIDIA RTX 3060 (125 epochs). The 70% reduction in vocabulary size (275 vs. 1000+ for chord-based) translates to a 55% reduction in embedding layer parameters and approximately 40% faster token decoding, but the full end-to-end speedup is not a simple linear scaling due to attention complexity. We therefore claim a *vocabulary reduction of 70%* and a *measured training time reduction of 35–40%* compared to chord-based baselines on identical hardware.

The reduction in computational cost aligns with recent observations that compact generative representations can substantially improve training efficiency without sacrificing generation quality [6].

5. Discussion

The experimental results presented in Section 4 validate the core design decisions of the proposed framework and offer several insights into the interplay between tokenization, structural modeling, and generation quality.

The observed improvements are also consistent with broader trends in generative AI, where architectural inductive biases increasingly replace handcrafted post-processing pipelines [2, 3].

Phrase-Aware Latent Alignment. The most significant contributor to generation quality is the phrase-aware attention bias computed via pairwise KL divergence of sequence-level Gaussian representations. The ablation study shows that removing this bias (setting $\alpha = 0$) causes a 25% increase in average KL divergence between generated and ground-truth token distributions across all categories. This indicates that without phrase-level structural supervision, the model reverts to locally coherent but globally inconsistent generation a well-documented failure mode of flat transformer architectures on musical data [9]. The KL bias acts as a soft structural prior, encouraging the attention mechanism to weight phrase-similar positions more heavily without requiring hard segmentation boundaries or separate inference networks. This design is particularly advantageous in guitar tablature, where phrase boundaries are implicit and highly context-dependent. Similar observations regarding the effectiveness of attention-based contextual modeling have been reported in recent surveys of attention mechanisms [10, 11].

Compact Tokenization. The 275-token vocabulary achieves a 70% reduction relative to chord-based approaches, translating to a 55% reduction in embedding layer parameters and a 35–40% reduction in measured training time. Importantly, this reduction does not come at the cost of expressive coverage. By jointly encoding beat position and dynamic level into 142 combined position/velocity tokens rather than maintaining separate factorial combinations, the scheme avoids the embedding sparsity that afflicts large vocabularies. The low KL divergence scores across all token categories (ranging from 0.12 for beat type to 0.21 for duration) confirm that the compact encoding faithfully represents the statistical structure of the training data. The relatively higher KL divergence for duration tokens (0.21) suggests that rhythmic variety remains the most difficult dimension to model, a finding consistent with prior work on rhythmic complexity in symbolic music generation [15].

Repetition Loss. Disabling the repetition loss results in a 12% increase in redundant accent patterns, confirming that expressive tokens particularly bends, slides, and vibrato are prone to pathological repetition under standard cross-entropy training. This behavior arises because consecutive repetition of high-probability expressive tokens is locally rewarded by the cross-entropy objective even when it produces musically implausible outputs. The repetition loss provides a differentiable regularization signal that suppresses this tendency without requiring constrained decoding at inference time, preserving generation speed. Reducing symbolic inconsistencies during autoregressive decoding remains an active research direction in sequence modeling [14].

Diversity vs. Coherence Trade-off. The diversity score of 0.83 represents a meaningful improvement over all baselines (0.58–0.71). The temperature sweep in Figure 3 further reveals that the proposed model consistently outperforms the baseline across the full temperature range, suggesting that the gains in diversity are structural rather than merely a consequence of higher entropy sampling. At low temperatures (0.5–0.7), where both models are forced toward high-confidence predictions, the proposed model still produces more varied outputs, indicating that the phrase-aware bias diversifies the attention distribution even in low-entropy regimes.

Limitations. Several limitations constrain the current framework. First, all training data assumes standard guitar tuning (EADGBE), meaning the model has not been exposed to alternate tunings (e.g., drop D, open G) that are common in rock and blues. Applying the model in alternate-tuning contexts would require either retraining or a tuning-conditioned extension of the tokenization scheme. Second, the framework currently supports single-track generation only. Real musical productions involve multiple simultaneous voices (rhythm guitar, lead guitar, bass), and modeling their interdependencies requires multi-track architectures that are beyond the current scope. Third, the dataset size (6,208 sequences) is modest by modern standards. While the results are statistically meaningful, scaling to larger corpora may reveal additional failure modes or require architectural adjustments to the phrase-aware attention mechanism.

Broader Implications. Beyond guitar tablature, the proposed framework's combination of compact tokenization and latent phrase alignment is broadly applicable to other symbolic music domains piano rolls, drum patterns, bass lines where expressive techniques and structural coherence are important. The lightweight nature of the KL bias computation (no separate inference network required) makes it straightforward to integrate into existing transformer-based music generation pipelines as a plug-in structural regularizer. Such extensions also align with recent developments in controllable generative AI and conditional sequence generation [8, 12].

6. Conclusion and Future Work

The generative approach proposed in this paper for expressive symbolic music synthesis overcomes two main limitations of existing transformer-based approaches: (a) transitions between expressive performance techniques lose efficiency, and (b) structural coherence across variables of the musical phrase length.

The proposed framework contributes to the growing body of research on expressive symbolic music generation and controllable generative AI [1, 2].

The first contribution is a small vocabulary of size 275 together with the joint encoding of the pitch, duration, accent techniques, bar markers, and a combined position/velocity information. This design lowers embedding parameters by 55% and training time by 35-40% concerning chord-based baselines while achieving full expressive coverage over the note events and performance techniques associated with guitar-centric music.

The second contribution is a phrase-aware latent alignment mechanism, in which pairwise KL divergence matrices are computed using sequence-level Gaussian representations resulting from a lightweight MLP on the pooled encoder output. The matrices are summed up into per-head attention bias terms, injecting phrase-level structural similarity into transformer attention directly. Inference network is not essential for working of the mechanism nor does it incur regions overhead on training and inference.

Our third contribution is the first multi-objective training loss combining cross-entropy, sequence-level KL regularization, and a repetition-aware penalty. The repetitive loss aims at expressive token types exhibiting pathological repetition under standard crossentropy training. It provides a differentiable solution and works at training time rather than with constrained decoding.

On the 6,208 guitar tablature sequences dataset, the perplexity of our model is 4.1, its diversity score is 0.83, and it took 4.2 hours of training on a single consumer-grade GPU. Our model outperforms Music Transformer, Pop Music Transformer, and MMM on all the reported metrics. Results corroborate that both the phrase-aware bias and the repetition loss are necessary for the gains.

There are many avenues of future work. Future extensions may also benefit from recent advances in diffusion-based symbolic music generation and controllable generative modeling [6, 12].Additionally, expanding the framework to multi-track generation modeling of the joint distribution over rhythm guitar, lead guitar, bass etc will allow more complete music generation. Another potential improvement could be to restrict the dynamic conditioning inputs of the model, and allow for more complex mini-batch conditioning signals, like musical phrase awareness. An extensive expansion of the data set (for example huge corpora of web-scraped tablature) would provide a more rigorous test for the generalization capacity of the framework and motivate larger model configurations. In conclusion, applying the proposed tokenization and attention on phrases in audio-domain models, where symbolic representations are employed as conditioning signals for neural audio synthesizers, could open a new avenue towards end-to-end expressive music generation.

References

[1] Shulei Ji, Xinyu Yang, and Jiebo Luo. A survey on deep learning for symbolic music generation: Representations, algorithms, evaluations, and challenges. *ACM Computing Surveys*, 56(1):1–39, 2023.

[2] Stefan Feuerriegel, Jochen Hartmann, Christian Janiesch, and Patrick Zschech. Generative AI. *Business & Information Systems Engineering*, 66(1):111–126, 2024.

[3] Leonard Banh and Gero Strobel. Generative artificial intelligence. *Electronic Markets*, 33(1):63, 2023.

[4] N. Boulanger-Lewandowski, Y. Bengio, and P. Vincent. Modeling temporal dependencies in high-dimensional sequences: Application to polyphonic music generation and transcription. In *Proceedings of the 29th International Conference on Machine Learning (ICML)*, 2012.

[5] J.-P. Briot and F. Pachet. Deep learning for music generation: Challenges and directions. *arXiv preprint*, 2017. arXiv:1706.08454.

[6] Gautam Mittal, Jesse Engel, Curtis Hawthorne, and Ian Simon. Symbolic music generation with diffusion models, 2021.

[7] Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P. Xing. Toward controlled generation of text. In *International Conference on Machine Learning*, pages 1587–1596. PMLR, July 2017.

[8] Nitish Shirish Keskar, Bryan McCann, Lav R. Varshney, Caiming Xiong, and Richard Socher. Ctrl: A conditional transformer language model for controllable generation, 2019.

[9] C.-Z. A. Huang, A. Vaswani, J. Uszkoreit, and et al. Music transformer: Generating music with long-term structure. In *Proceedings of the 7th International Conference on Learning Representations (ICLR)*, 2019.

[10] Zhaoyang Niu, Guoqiang Zhong, and Hui Yu. A review on the attention mechanism of deep learning. *Neurocomputing*, 452:48–62, 2021.

[11] Gianni Brauwers and Flavius Frasincar. A general survey on attention mechanisms in deep learning. *IEEE Transactions on Knowledge and Data Engineering*, 35(4):3279–3298, 2021.

[12] Jiaming Song, Qinsheng Zhang, Hongxu Yin, Morteza Mardani, Ming-Yu Liu, Jan Kautz, Yongxin Chen, and Arash Vahdat. Loss-guided diffusion models for plug-and-play controllable generation. In *International Conference on Machine Learning*, pages 32483–32498. PMLR, July 2023.

[13] Meng-Hao Guo, Tian-Xing Xu, Jiang-Jiang Liu, Zheng-Ning Liu, Peng-Tao Jiang, Tai-Jiang Mu, Song-Hai Zhang, Ralph R. Martin, Ming-Ming Cheng, and Shi-Min Hu. Attention mechanisms in computer vision: A survey. *Computational Visual Media*, 8(3):331–368, 2022.

[14] Sean Welleck, Peter West, Jize Cao, and Yejin Choi. Symbolic brittleness in sequence models: on systematic generalization in symbolic mathematics. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 8629–8637, June 2022.

[15] S. Wu and Y.-H. Yang. Pop music transformer: Generating music with rhythm and harmony. In *Proceedings of the 21st International Society for Music Information Retrieval Conference (ISMIR)*, 2020.