

Let the Model Learn to Feel: Mode-Guided Tonality Injection for Symbolic Music Emotion Recognition

Haiying Xia^{1,2}, Zhongyi Huang^{1,2}, Yumei Tan^{1,2*}, Shuxiang Song^{1,2*}

¹Guangxi Key Laboratory of Brain-inspired Computing and Intelligent Chips,
School of Electronic and Information Engineering, Guangxi Normal University, Guilin 541004, China

²Key Laboratory of Integrated Circuits and Microsystems,
Education Department of Guangxi Zhuang Autonomous Region, Guilin 541004, China
xhy22@gxnu.edu.cn, zoeyhuang@stu.gxnu.edu.cn, tanyumei@gxnu.edu.cn, songshuxiang@gxnu.edu.cn

Abstract

Music emotion recognition is a key task in symbolic music understanding (SMER). Recent approaches have shown promising results by fine-tuning large-scale pre-trained models (e.g., MIDIBERT, a benchmark in symbolic music understanding) to map musical semantics to emotional labels. While these models effectively capture distributional musical semantics, they often overlook tonal structures, particularly musical modes, which play a critical role in emotional perception according to music psychology. In this paper, we investigate the representational capacity of MIDIBERT and identify its limitations in capturing mode-emotion associations. To address this issue, we propose a Mode-Guided Enhancement (MoGE) strategy that incorporates psychological insights on mode into the model. Specifically, we first conduct a mode augmentation analysis, which reveals that MIDIBERT fails to effectively encode emotion-mode correlations. Motivated by this observation, we further identify the MIDIBERT layer that shows the weakest emotion relevance and introduce a Mode-guided Feature-wise linear modulation injection (MoFi) framework to inject explicit mode features, thereby enhancing the model’s capability in emotional representation and inference. Extensive experiments on the EMOPIA and VGMIDI datasets demonstrate that our mode injection strategy significantly improves SMER performance, achieving accuracies of 75.2% and 59.1%, respectively. These results validate the effectiveness of mode-guided modeling in symbolic music emotion recognition.

Extended version —

<https://github.com/ZoeyHuang-paper/MoFi>

Introduction

Music emotion recognition in the symbolic domain plays a vital role in music understanding. It has diverse applications in music generation, music psychotherapy, music recommendation system and human-computer interaction. Compared with audio-based music emotion recognition, symbolic music provides explicit and structural information that can be processed similarly to words in natural language.

Since symbolic music consists of rule-governed symbolic sequences, advances in natural language processing (NLP)

*Corresponding author.

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

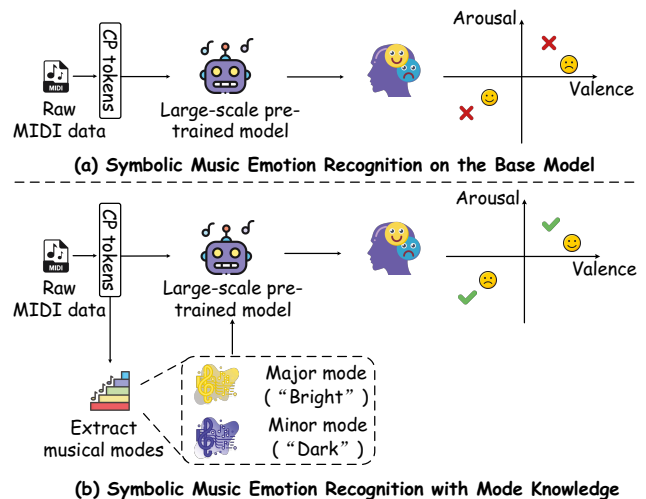


Figure 1: High-level overview of the proposed MoFi framework. We extract mode information and then injected into a large-scale pre-trained model. (a) illustrates the fine-tuning process of the base model without explicit mode guidance and shows the limitations of music recognition. (b) shows the model with mode knowledge.

have opened new opportunities for symbolic music analysis by enabling the use of language-modeling techniques. In recent years, the remarkable performance of Transformer-based pre-trained models (Vaswani et al. 2017) has inspired numerous efforts to adapt Transformer architectures (Huang et al. 2019) for symbolic music understanding tasks. The significant advancements in the BERT models (Devlin et al. 2019) trained on large-scale midi music datasets have enhanced the understanding of music (Zeng et al. 2021; Chou et al. 2024). Existing approaches have improved the capacity of models to capture richer musical information by designing diverse pre-training strategies and refining symbolic encoding formats (Zeng et al. 2021; Tian et al. 2024). However, due to the limited size of symbolic music emotion recognition datasets, these pre-trained models often fail to acquire emotion-related musical features effectively during fine-tuning. What’s more, emotions are a primary driver of human engagement with music, enabling people to ex-

perience and respond to a rich spectrum of affective states. In contrast, computational systems rely on intelligent algorithms to process and interpret music through fundamentally different mechanisms. While current models demonstrate strong capabilities in symbolic music understanding, they do not capture the intrinsic reasoning behind human emotional responses. As suggested in (Li et al. 2024), to better understand emotions, models need to explicitly capture features associated with human psychological responses to emotional triggers.

In this work, we aim to enable the model to learn the key elements of music perception in a manner aligned with human understanding. To achieve this, we propose a Mode-Guided Enhancement (MoGE) strategy and a Mode-guided Feature-wise Linear Modulation injection (MoFi) framework to diagnose the limitations and enhance the emotional representation capabilities of large-scale pre-trained models. We adopt MIDIBERT (Chou et al. 2024), which serves as a benchmark model for symbolic music understanding, as the backbone. Existing literature in psychology and music theory has revealed a strong relationship between musical modes and emotional perception (Kastner and Crowder 1990a). Specifically, we first perform a mode augmentation analysis to investigate to what extent MIDIBERT has already understood the relationship between musical modes and emotional expression. Then, to better understand MIDIBERT’s capability, we probe each layer of the model to estimate how much emotional information it has already encoded. This analysis identifies the least emotionally informative layer as the target for MoFi. The MoFi framework extracts explicit mode features and injects them into this layer to improve MIDIBERT’s emotional representation and inference (see Figure 1).

With our design, our proposed method is equipped with mode knowledge, mimicking human perception of music. Moreover, our model inherits the robust music semantic understanding capabilities of the pre-trained MIDIBERT model. As a result, our method achieves remarkable results on two different scale datasets, EMOPIA (Hung et al. 2021) and VGMIDI (Ferreira and Whitehead 2019).

Our primary contributions are as follows:

- We tackle the challenge of capturing the intrinsic reasoning behind human emotional responses by employing the Mode-Guided Enhancement (MoGE) strategy, which consists of a targeted diagnostic experiment to analyze the model’s limitations.
- Recognizing that large-scale pre-trained model MIDIBERT neglects the relationship between mode features and emotion, we introduce a Mode-guided Feature-wise Linear Modulation Injection (MoFi) framework. This framework enables fine-grained and parameter-efficient incorporation of explicit music-theoretic priors into MIDIBERT, offering a principled and interpretable solution to its identified knowledge gaps.
- We conduct experiments on two different scale datasets, including EMOPIA and VGMIDI, and achieve superior or comparable results to state-of-the-art methods.

Related Work

Symbolic Music Emotion Recognition

Music emotion recognition (MER) is a core task in symbolic music understanding, requiring models to infer high-level affective states directly from the compositional structure of music.

Early approaches to Symbolic Music Emotion Recognition (SMER) primarily relied on hand-crafted features grounded in music theory (Wu et al. 2014; Malheiro et al. 2016), such as pitch histograms, rhythmic density, and harmonic complexity combined with traditional classifiers such as support vector machines (SVM) (Lin, Chen, and Yang 2013). While these methods laid important groundwork, they are constrained by the need for extensive feature engineering, which is both labor-intensive and often insufficient to capture the intricate and long-range temporal dependencies characteristic of musical expression. In recent years, with the advent of deep learning, particularly through Recurrent Neural Network (RNN) (Lin et al. 2017) and Convolutional Neural Network (CNN) (Fukushima 1980), marked a paradigm shift by enabling the automatic extraction of features from raw symbolic data (Simonetta et al. 2019).

More recently, the field of symbolic music understanding has been significantly advanced by the emergence of large-scale Transformer-based pre-trained models such as MIDIBERT (Chou et al. 2024), MusicBERT (Zeng et al. 2021) and PopMAG (Ren et al. 2020). Inspired by the success of BERT (Devlin et al. 2019) in natural language processing, these models adopt a two-stage framework consisting of pre-training followed by fine-tuning. During the pre-training phase, they utilize large-scale unlabeled MIDI corpora and self-supervised objectives, such as masked token prediction, to learn rich and contextualized musical representations. These representations are then adapted to downstream tasks through light-weight classifiers, achieving state-of-the-art performance on various challenges, including symbolic music emotion recognition (SMER). Despite their empirical success, the internal workings of these models remain largely opaque. The pre-training objective focuses on token-level reconstruction and offers no explicit incentive to learn abstract musical concepts such as tonal invariance or mode-dependent emotional structure. This raises a fundamental question: do these models perform well because they learn generalizable music-theoretic knowledge, or do they instead rely on superficial correlations present in the training data? Our work seeks to answer this question by examining whether these models have truly internalized essential principles of music theory, especially musical modes, that are critical for emotion recognition.

Musical Features and Emotion

The relationship between musical structure and human emotion is a foundational principle in both musicology and the cognitive psychology of music. For centuries, theorists and composers have emphasized the affective role of musical modes (Bowling 2013). Empirical studies have consistently demonstrated that the major/minor distinction serves as a primary cue for perceived emotional valence in Western mu-

music (Justus, Gabriel, and Pfaff 2018; Gagnon and Peretz 2003). Specifically, major-mode music is consistently associated with positive emotions such as happiness and joy, reflecting high emotional valence. In contrast, minor-mode music is reliably linked to negative emotions such as sadness and melancholy, indicating low valence (Krumhansl 1997). Literature (Kastner and Crowder 1990b) demonstrated that positive emotions are associated with music in major keys, whereas negative emotions tend to be evoked by music in minor keys. Similar findings have been reported by Gerardi and Gerken (Gerardi and Gerken 1995), Gregory et al. (Gregory, Worrall, and Sarge 1996), and Dalla Bella et al. (Dalla Bella et al. 2001). While this association is not strictly deterministic and may be influenced by other factors such as cultural background, its statistical robustness is well documented and constitutes a fundamental aspect of musical perception and literacy in human listeners.

Critically, the association between mode and emotion is relatively invariant to key. Although transposition introduces slight variations, listeners generally perceive a piece in C major as conveying a similar joyful emotion to one in G major. It is the intervallic structure of the mode, rather than the specific key, that conveys the primary emotional character (BOYDEN 1994). Therefore, we consider to inject only two modes (major and minor) in our four-class classification task to reduce noise. Building on this foundation, we propose a Mode-Guided Enhancement (MoGE) strategy. First, MoGE investigates whether large-scale pre-trained models capture mode-emotion associations. Based on this analysis, we then inject explicit mode knowledge into the emotionally underperforming layers using the Mode-guided Feature-wise Linear Modulation Injection (MoFi) framework, enhancing MIDIBERT’s recognition capability.

Preliminary

Before injecting mode-related knowledge into a pre-trained symbolic music model, the first and foremost question is: has MIDIBERT already learned the emotional implications of musical mode? Answering this question is critical for determining whether such theoretical features are necessary and where they should be integrated within the model architecture.

To empirically investigate this question, we utilize EMOPIA (Hung et al. 2021) as a preliminary dataset. EMOPIA is a widely used dataset for SMER task, which contains MIDI files annotated with Russell’s 4Q (Russell 1980) emotional labels. Based on this setup, we conduct two preliminary studies as follows.

Data Augmentation-Based Diagnostic Study To assess whether MIDIBERT has internalized the music-theoretic association between mode and emotion, we design a diagnostic experiment based on mode-preserving pitch transposition. Specifically, we augment a preliminary dataset by applying mode-preserving pitch transpositions within a single octave and evaluate MIDIBERT’s performance on both the original and augmented data. A significant performance improvement after augmentation would indicate that the model fails to capture the affective features of musical mode, sug-

gesting insufficient integration of mode-related knowledge. Conversely, comparable performance across the two conditions would imply that such mode-emotional associations have already been implicitly encoded in the model’s learned representations.

Layer-Wise Representation Analysis In this study, we freeze all parameters of the pre-trained MIDIBERT model except for the self-attention layer before classification and softmax classification head. Our goal is to identify which layers retain affect-relevant information and which fail to encode such knowledge effectively. This analysis informs the selection of target layers for the subsequent injection of mode-related prior knowledge.

Effect of Data Augmentation

We perform mode-preserving data augmentation on EMOPIA by transposing each musical clip either upward or downward by a random number of semitones. The transposition range is restricted to a single octave and applied uniformly to all notes within a clip. As all notes in a clip are shifted together, the internal interval relationships remain unchanged, thereby preserving the underlying musical mode and its associated affective qualities. This augmentation technique leverages music-theoretic prior knowledge to synthesize new samples that maintain emotion-relevant structure, offering a simple yet effective way to enrich the training set with task-specific inductive bias.

We follow the official data split provided by MIDIBERT (Chou et al. 2024), maintaining a training:validation:test ratio of 8:1:1 and ensuring that all clips from the same song are placed in the same subset to prevent data leakage. To reduce the influence of training variance, we report the average performance over five runs with different random seeds.

Table 1 presents the classification accuracy of MIDIBERT trained on the original versus the augmented dataset. The results reveal a substantial improvement in recognition accuracy when using the augmented data, indicating that MIDIBERT has not fully internalized the affective regularities associated with musical mode.

Dataset	Original	Augmented
Accuracy	0.675	0.723

Table 1: Accuracy on the EMOPIA dataset with and without data augmentation.

Layer-wise Analysis of MIDIBERT on the SMER Task

Figure 2 shows the performance of individual layers within MIDIBERT on the SMER task, using the original dataset without augmentation. During fine-tuning, only one self-attention module used for weights average and the classification head are updated, while the remaining parameters are frozen. As a result, the performance variation across layers

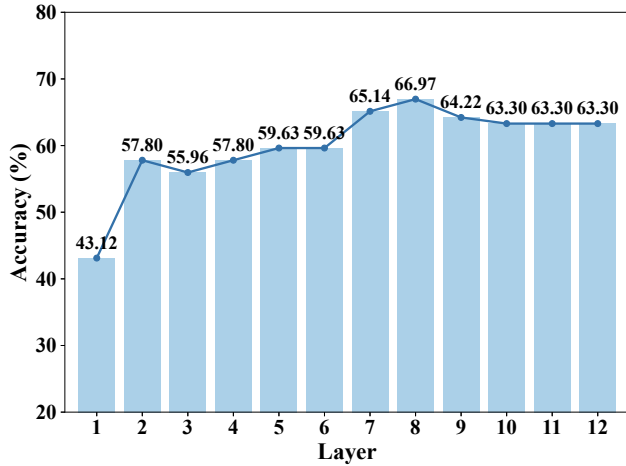


Figure 2: Performance of each MIDIBERT layer on EMOPIA dataset.

primarily reflects their respective contributions to emotion-relevant feature extraction in SMER task. Middle layers achieve the highest accuracy, whereas both lower and upper layers exhibit reduced effectiveness. This pattern aligns with previous findings in NLP, where intermediate transformer layers tend to encode the most semantically meaningful representations (Clark et al. 2019; Hewitt and Manning 2019).

Further analysis shows that the lower layers consistently yield the weakest performance, suggesting that they encode relatively little information relevant to emotional features. In contrast, the middle layers demonstrate significantly stronger performance, indicating their potential importance in capturing affective patterns in symbolic music. The upper layers also underperform, which may be attributed to their specialization in the model’s original pre-training purpose, namely masked token prediction based on contextual music information, rather than tasks centered on emotional understanding.

In summary, the data augmentation study suggests that MIDIBERT has not yet effectively captured the information of musical modes in SMER. The layer-wise performance analysis further reveals that the lower layers contribute the least to affective representation, suggesting that they are the most in need of enhancement. Taken together, these findings point to a promising optimization direction: injecting mode-related prior knowledge into the lowest layer of MIDIBERT (the first layer) to strengthen its affective modeling capabilities. The next section presents our approach to implementing this strategy.

Methodology

In this section, we incorporate prior knowledge of musical mode into the MIDIBERT model to enhance its affective representation capability. Guided by the findings in the previous section, we target the first layer of the model for knowledge integration. Given that each attention layer computes pairwise relationships among Compound Word (CP) tokens (Hsiao et al. 2021) and that prior studies in NLP

have shown the first layer to exhibit broad, undifferentiated attention patterns (Clark et al. 2019; Rogers, Kovalova, and Rumshisky 2021), we focus on modulating the attention computations in the first layer using tonality-related information. The details of the proposed approach are described below.

MIDIBERT and Compound Word

MIDIBERT (Chou et al. 2024) is a Transformer-based pre-trained benchmark model for symbolic music understanding. It is trained on a large-scale corpus of MIDI files to learn rich structural and contextual patterns, producing high-quality musical representations. We select MIDIBERT as the backbone encoder because it offers robust representational power and strong generalization. These qualities make it suitable for a variety of symbolic music understanding and generation tasks. It is widely adopted as a standard foundation model in the symbolic music domain.

Similar to words in natural language processing, symbolic music sequences (MIDI files) are first converted into a series of discrete CP events, which are then passed to MIDIBERT as input. This encoding scheme combines multiple musical attributes that occur at each time step, such as bar, position, pitch and duration, into a single token. Each CP token consists of four distinct sub-events, which together encode both the temporal (horizontal) and harmonic (vertical) dimensions of music structure. A detailed construction of CP words can be found in the Appendix.

- Bar event marks the beginning of a new bar and conveys information about metrical boundaries.
- Position event specifies the relative location of a note within a bar and captures rhythmic structure.
- Pitch event represents the absolute pitch of a note, including pitch class, octave, and accidentals.
- Duration event indicates the note’s length and contributes to the continuity of musical flow.

The core of MIDIBERT is a standard Transformer encoder stack, which adopts the architectural paradigm established by BERT in NLP. Each Transformer encoder layer comprises a multi-head self-attention mechanism and a feed-forward network.

Multi-Head Attention in MIDIBERT This mechanism allows the model to attend to all other Compound Word embeddings in the sequence when encoding each token, assigning weights based on contextual relevance. Given an input embedding sequence $\mathbf{E} \in \mathbb{R}^{L \times d_{\text{model}}}$, where L is the sequence length and d_{model} is the hidden dimension, the Multi-Head Self-Attention layer first projects the input into query (\mathbf{Q}), key (\mathbf{K}), and value (\mathbf{V}) matrices. These projections are computed using learnable parameter matrices $\mathbf{W}_i^{\mathbf{Q}}$, $\mathbf{W}_i^{\mathbf{K}}$, and $\mathbf{W}_i^{\mathbf{V}}$ for each attention head:

$$\text{MultiHead}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Concat}(\text{head}_1, \dots, \text{head}_h) \mathbf{W}^{\mathbf{O}}$$

$$\text{head}_i = \text{Attention}(\mathbf{Q}\mathbf{W}_i^{\mathbf{Q}}, \mathbf{K}\mathbf{W}_i^{\mathbf{K}}, \mathbf{V}\mathbf{W}_i^{\mathbf{V}}) \quad (1)$$

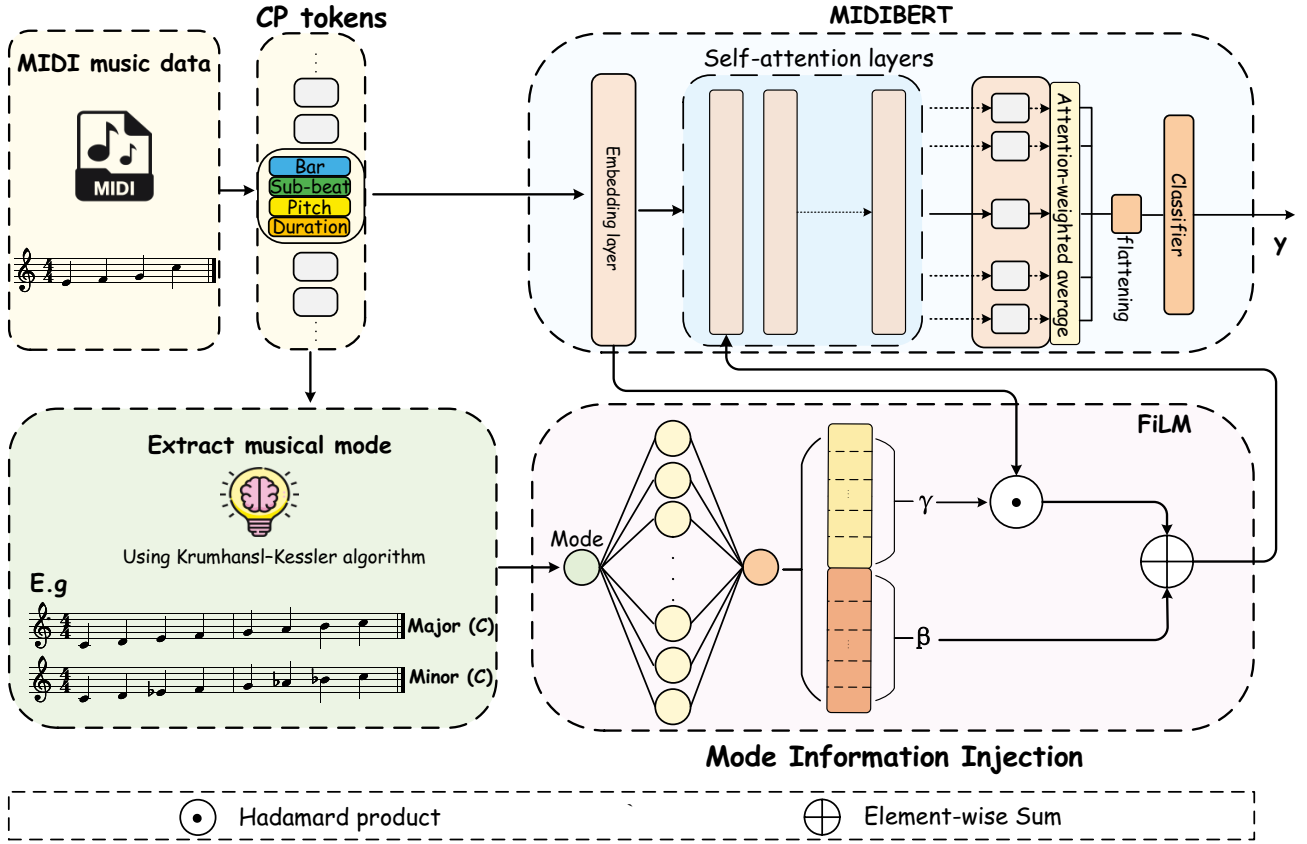


Figure 3: Architecture illustration of MoFi. Our method first extracts the mode knowledge (major/minor) of CP tokens using the Krumhansl–Kessler algorithm, which is then passed to the FiLM (Feature-wise Linear Modulation) method. This injected information is integrated into the first layer of the large-scale pre-trained model MIDIBERT, enhancing its capacity to capture emotional nuances in symbolic music. The mode knowledge injection enables the model to better adapt to emotional characteristics, while still retaining the benefits of the pre-trained MIDIBERT framework.

where \mathbf{W}_i^Q , \mathbf{W}_i^K , and \mathbf{W}_i^V are learnable projection matrices that map the input embeddings into lower-dimensional subspaces. The outputs from all heads are concatenated and then projected back to the original dimension d_{model} using the matrix \mathbf{W}^O . Each attention head is computed using the Scaled Dot-Product Attention mechanism:

$$\text{Attention}(Q, K, V) = \text{softmax} \left(\frac{QK^\top}{\sqrt{d_k}} \right) V \quad (2)$$

where d_k is the dimensionality of the Key vectors, and the softmax function normalizes the attention scores across the sequence. This mechanism allows the model to capture long-range dependencies and intricate structural patterns in symbolic music, which are essential for understanding high-level musical semantics.

Feed-Forward Network The output from the self-attention mechanism is passed through a two-layer position-wise feed-forward network, which applies nonlinear transformations independently at each sequence position. This component further refines the representations and enhances the model’s capacity to capture complex musical patterns.

MIDIBERT is pre-trained in a self-supervised manner using a masked music modeling objective on a large corpus of MIDI files. This training paradigm allows the model to learn rich and context-sensitive representations of symbolic music. On top of this pre-trained foundation, we incorporate external knowledge of musical mode to enhance the model’s ability to recognize affective content in music.

Mode Knowledge Extraction

An essential step in our Mode-guided Feature-wise Linear Modulation Injection (MoFi) framework is the extraction of accurate and transferable prior knowledge of musical mode from symbolic music data. This knowledge forms the basis for subsequent injection into the model. In Western music theory, major modes are typically linked to positive affective states such as brightness, cheerfulness, and joy, whereas minor modes are more frequently associated with negative or introspective emotions such as melancholy, sadness, and pensiveness (Horn and Huron 2015). This well-established correlation between mode and emotion constitutes a critical prior for symbolic music emotion recognition.

To capture essential mode cues while reducing data noise, we adopt a binary categorization scheme that distinguishes only between major and minor modes. Although other modal types such as Dorian and Lydian exist, they are relatively infrequent in symbolic music corpora and exhibit less consistent emotional profiles. This simplification preserves the most salient emotional distinctions related to mode while improving data balance and model generalizability. Furthermore, since major modes tend to convey brighter emotions while minor modes are associated with darker feelings, we adopt a binary mode categorization (major vs. minor) instead of using all 24 keys (defined by tonic) to reduce noise in our four-class classification task.

We extract mode information from symbolic MIDI data in an automatic and musically informed manner using the Krumhansl–Kessler (K-K) algorithm (Krumhansl 1997), a well-established technique in the field of Music Information Retrieval (MIR). Rooted in cognitive psychology, this algorithm closely approximates human perceptual processes of tonal recognition. We implement the method using the music21 toolkit. The extracted mode prior for each musical piece is encoded as a one-hot vector indicating either major or minor mode.

Mode Injection via FiLM

To incorporate the extracted mode prior into the MIDIBERT model, we employ Feature-wise Linear Modulation (FiLM) (Perez et al. 2018), a conditioning mechanism originally developed in the field of computer vision. FiLM applies affine transformations to the intermediate features of a neural network, enabling the model to dynamically modulate its internal feature representations based on external inputs. Unlike other injection methods such as attention-based mechanisms, FiLM provides greater flexibility and parameter efficiency, making it particularly well-suited for integrating targeted domain knowledge into pre-trained architectures.

The core idea of FiLM is to utilize a conditioning vector c (representing mode knowledge) to generate two modulation parameters: a scaling factor γ and a shifting factor β . These parameters are applied to the sequence feature representations x through an element-wise affine transformation. Specifically, a parameter generation network f_{cond} maps the conditioning vector c into the parameter space:

$$[\gamma, \beta] = f_{\text{cond}}(c) \quad (3)$$

The FiLM operation then applies these parameters to the input feature representations x as follows:

$$\text{FiLM}(x, c) = \gamma \odot x + \beta \quad (4)$$

where \odot denotes the Hadamard product. In this formulation, the conditioning information c modulates the activation pattern of x in a fine-grained manner, thereby guiding the model to emphasize or attenuate specific feature dimensions based on prior knowledge.

As demonstrated by the preceding analysis, the first Transformer layer in MIDIBERT encodes the lowest level of mode-related emotional information. In the proposed MoFi

architecture, we introduce a FiLM-based conditioning module between the Compound Word embedding layer and the first Transformer encoder layer. This integration infuses the input representation with mode-aware conditioning prior to self-attention computation, so that all subsequent representations are informed by this critical music-theoretic prior. As shown in Figure 3.

Experiments

Setup

Datasets In our experiments, we evaluate our method on two widely used benchmark datasets: the small-scale VGMIDI and the relatively large-scale MIDI-based music emotion dataset EMOPIA.

EMOPIA The EMOPIA dataset (Hung et al. 2021) is a comprehensive collection of pop piano music clips constructed to support emotion recognition across symbolic and audio domains. It comprises 1,087 clips derived from 387 unique songs, each annotated with clip-level emotional labels. Emotions are categorized using Russell’s 4Q circumplex model (Russell 1980), defining four affective quadrants (Q1–Q4) (see Table 2). This dataset has been widely adopted in music emotion classification.

The original audio recordings were collected online using publicly available metadata. Corresponding MIDI files were generated through transcription using a high-fidelity piano transcription model (Kong et al. 2021). A subset of excerpts was randomly sampled and manually inspected by the dataset creators, who verified the accurate preservation of pitch, velocity, and duration. Tracks with engineered ambient effects were excluded from the final set due to their negative impact on transcription quality.

Taxonomy	Description
Happy (Q1)	High valence high arousal (HVHA)
Sad (Q2)	Low valence high arousal (LVHA)
Calm (Q3)	Low valence low arousal (LVLA)
Angry (Q4)	High valence low arousal (HVLA)

Table 2: Russell’s 4Q taxonomy.

VGMIDI The VGMIDI dataset (Ferreira and Whitehead 2019) comprises video game music tracks in MIDI format. It includes 200 MIDI compositions with corresponding emotion annotations, 97 aligned audio versions, and an additional 3,850 unlabeled pieces. In this study, we utilize only the labeled MIDI tracks. Each labeled clip was rated by 30 annotators using the Circumplex model of emotion, defined along the Valence–Arousal (VA) dimensions. To ensure consistency with EMOPIA, the VA annotations in VGMIDI were mapped to the Russell’s 4Q classification scheme used in EMOPIA (see Table 2). The dataset is divided into training (80%), validation (10%), and test (10%) subsets. Additional statistics for both EMOPIA and VGMIDI are reported in Table 4.

Models	EMOPIA		VGMIDI	
	Accuracy	F1	Accuracy	F1
SVM (Lin, Chen, and Yang 2013)	0.477	0.476	0.451	0.377
LSTM-Attn (Hung et al. 2021)	0.647	0.563	0.417	0.260
MIDIGPT (Ferreira, Lelis, and Whitehead 2020)	0.587	0.572	0.538	0.505
MT-MIDIGPT (Qiu, Chen, and Zhang 2022)	0.625	0.611	0.585	0.509
MT-MIDIBERT (Qiu, Chen, and Zhang 2022)	0.676	0.664	0.498	0.453
BiLMA (Xiao et al. 2024)	0.708	0.631	0.572	0.478
MIDIBERT-Piano(Baseline) (Chou et al. 2024)	0.634	0.628	0.473	0.432
Ours	0.752	0.751	0.591	0.587

Table 3: Comparison between existing midi models with our method.

INFO	EMOPIA	VGMIDI
Number of MIDI	1087	200
Train-valid-test splits	8:1:1	8:1:1
Source	Youtube	Video game soundtracks
Music Type	various	various
Single Duration	About 30s	NA

Table 4: Summary of EMOPIA and VGMIDI.

Training Details We adopt MIDIBERT as our backbone model, following the original architecture (Chou et al. 2024). The model comprises 12 Transformer layers, each with 12 attention heads, a hidden size of 768, and a total of 111 million parameters. We initialize our model using the pre-trained checkpoints released along with the MIDIBERT paper (Chou et al. 2024).

For the SMER task, both datasets are split into training, validation, and test sets with an 8:1:1 ratio. For EMOPIA, we use the same split as in MIDIBERT (Chou et al. 2024). The batch sizes are set to 16 and 8 for EMOPIA and VGMIDI, respectively. The model is fine-tuned for up to 20 epochs on a single NVIDIA GeForce RTX 3090 GPU, resulting in a total training time of less than 30 minutes. Early stopping is employed when the validation performance does not improve for three consecutive epochs.

At the start of training, the FiLM layer is initialized with a scaling factor $\gamma = 1$ and a shifting factor $\beta = 0$, effectively preserving the pre-trained knowledge embedded in MIDIBERT. This initialization preserves training stability. It allows the model to start from the pretrained MIDIBERT embeddings and gradually incorporate mode-related information, avoiding disruption from randomly initialized parameters.

Results Analysis

Table 3 summarizes the overall performance of our proposed method in terms of accuracy and macro-F1 for SMER. The results align consistently with findings reported in prior work (Xiao et al. 2024). For comparison, we also report a traditional SVM (Lin, Chen, and Yang 2013) baseline and several previous symbolic music emotion recognition models. Notably, our Mode-guided Feature-wise Linear Modulation Injection (MoFi) framework achieves accuracy im-

provements of 11.8% on both the EMOPIA and VGMIDI datasets, and F1-score improvements of 12.3% and 15.5%, respectively, over the baseline. These gains are achieved with minimal architectural modifications.

Ablation Study

Table 5 presents an ablation study comparing the full model with and without mode injection. We observe that removing the FiLM-based mode injection module leads to a noticeable drop in performance. This result indicates that incorporating explicit mode information contributes positively to the model’s ability to recognize musical emotions. By injecting this music-theoretical prior at an early stage of representation learning, the model gains mode awareness that the original architecture fails to capture.

Models	EMOPIA		VGMIDI	
	Accuracy	F1	Accuracy	F1
Full Model	0.752	0.751	0.591	0.587
w/o Mode Knowledge	0.716	0.715	0.500	0.365

Table 5: Ablation study on mode injection via FiLM.

Conclusion

In this work, we introduce Mode-Guided Enhancement (MoGE) as an analytical strategy to assess MIDIBERT’s ability to capture music-theoretical knowledge of musical mode and uncover its limitations. Building upon this analysis, we propose Mode-guided Feature-wise Linear Modulation Injection (MoFi), a lightweight yet effective framework designed to inject mode-related priors into MIDIBERT. Rather than relying primarily on resource-intensive pre-training, our approach incorporates musical mode information to modulate input representations and guide the model’s attention. By bridging symbolic music modeling with music-theoretical priors, our MoFi framework demonstrates that even lightweight guidance rooted in domain knowledge can yield substantial improvements. These findings encourage further exploration of interpretable and theory-driven methods in AI music, potentially benefiting a broad range of tasks beyond emotion recognition.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grants 62366006 and 62366005, the Key Laboratory of AI and Information Processing, Education Department of Guangxi Zhuang Autonomous Region (Hechi University) under Grants 2024GXZDSY011, the Research Projects of Guangxi Key Laboratory of Precision Navigation Technology and Application under Grants DHKL2417, the Guangxi Natural Science Foundation for Youths under Grant 2025JJB170055 and the 2025 Joint Cultivation Project of the National Natural Science Foundation under Guangxi Normal University.

References

- Bowling, D. L. 2013. A vocal basis for the affective character of musical mode in melody. *Frontiers in psychology*, 4: 464.
- BOYDEN, D. 1994. *Manual of Counterpoint Based*. CARL FISCHER. ISBN 9780825827648.
- Chou, Y.-H.; Chen, I.-C.; Ching, J.; Chang, C.-J.; and Yang, Y.-H. 2024. Midibert-piano: Large-scale pre-training for symbolic music classification tasks. *Journal of Creative Music Systems*, 8(1).
- Clark, K.; Khandelwal, U.; Levy, O.; and Manning, C. D. 2019. What Does BERT Look at? An Analysis of BERT’s Attention. In Linzen, T.; Chrupala, G.; Belinkov, Y.; and Hupkes, D., eds., *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, BlackboxNLP@ACL 2019, Florence, Italy, August 1, 2019*, 276–286. Association for Computational Linguistics.
- Dalla Bella, S.; Peretz, I.; Rousseau, L.; and Gosselin, N. 2001. A developmental study of the affective value of tempo and mode in music. *Cognition*, 80(3): B1–B10.
- Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Burstein, J.; Doran, C.; and Solorio, T., eds., *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics.
- Ferreira, L.; Lelis, L.; and Whitehead, J. 2020. Computer-generated music for tabletop role-playing games. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, volume 16, 59–65.
- Ferreira, L. N.; and Whitehead, J. 2019. Learning to Generate Music with Sentiment. *ISMIR’19*.
- Fukushima, K. 1980. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological cybernetics*, 36(4): 193–202.
- Gagnon, L.; and Peretz, I. 2003. Mode and tempo relative contributions to “happy-sad” judgements in equitone melodies. *Cognition and emotion*, 17(1): 25–40.
- Gerardi, G. M.; and Gerken, L. 1995. The development of affective responses to modality and melodic contour. *Music Perception*, 12(3): 279–290.
- Gregory, A. H.; Worrall, L.; and Sarge, A. 1996. The development of emotional responses to music in young children. *Motivation and Emotion*, 20(4): 341–348.
- Hewitt, J.; and Manning, C. D. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4129–4138.
- Horn, K.; and Huron, D. 2015. On the changing use of the major and minor modes 1750–1900. *Music Theory Online*, 21(1): 1–11.
- Hsiao, W.-Y.; Liu, J.-Y.; Yeh, Y.-C.; and Yang, Y.-H. 2021. Compound word transformer: Learning to compose full-song music over dynamic directed hypergraphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 178–186.
- Huang, C. A.; Vaswani, A.; Uszkoreit, J.; Simon, I.; Hawthorne, C.; Shazeer, N.; Dai, A. M.; Hoffman, M. D.; Dinculescu, M.; and Eck, D. 2019. Music Transformer: Generating Music with Long-Term Structure. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.
- Hung, H.; Ching, J.; Doh, S.; Kim, N.; Nam, J.; and Yang, Y. 2021. EMOPIA: A Multi-Modal Pop Piano Dataset For Emotion Recognition and Emotion-based Music Generation. In Lee, J. H.; Lerch, A.; Duan, Z.; Nam, J.; Rao, P.; van Kranenburg, P.; and Srinivasamurthy, A., eds., *Proceedings of the 22nd International Society for Music Information Retrieval Conference, ISMIR 2021, Online, November 7-12, 2021*, 318–325.
- Justus, T.; Gabriel, L.; and Pfaff, A. 2018. Form and meaning in music: Revisiting the affective character of the major and minor modes. *Auditory Perception & Cognition*, 1(3-4): 229–247.
- Kastner, M. P.; and Crowder, R. G. 1990a. Perception of the major/minor distinction: IV. Emotional connotations in young children. *Music Perception*, 8(2): 189–201.
- Kastner, M. P.; and Crowder, R. G. 1990b. Perception of the major/minor distinction: IV. Emotional connotations in young children. *Music Perception*, 8(2): 189–201.
- Kong, Q.; Li, B.; Song, X.; Wan, Y.; and Wang, Y. 2021. High-resolution piano transcription with pedals by regressing onset and offset times. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29: 3707–3717.
- Krumhansl, C. L. 1997. An exploratory study of musical emotions and psychophysiology. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, 51(4): 336.
- Li, J.; Lin, Z.; Wang, L.; Si, Q.; Cao, Y.; Yu, M.; Fu, P.; Wang, W.; and Zhou, J. 2024. Think out Loud: Emotion Deducing Explanation in Dialogues. *arXiv preprint arXiv:2406.04758*.

Lin, Y.; Chen, X.; and Yang, D. 2013. Exploration of Music Emotion Recognition Based on MIDI. In *ISMIR*, 221–226.

Lin, Z.; Feng, M.; dos Santos, C. N.; Yu, M.; Xiang, B.; Zhou, B.; and Bengio, Y. 2017. A STRUCTURED SELF-ATTENTIVE SENTENCE EMBEDDING. In *International Conference on Learning Representations*.

Malheiro, R.; Panda, R.; Gomes, P.; and Paiva, R. P. 2016. Emotionally-relevant features for classification and regression of music lyrics. *IEEE Transactions on Affective Computing*, 9(2): 240–254.

Perez, E.; Strub, F.; De Vries, H.; Dumoulin, V.; and Courville, A. 2018. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Qiu, J.; Chen, C.; and Zhang, T. 2022. A novel multi-task learning method for symbolic music emotion recognition. *arXiv preprint arXiv:2201.05782*.

Ren, Y.; He, J.; Tan, X.; Qin, T.; Zhao, Z.; and Liu, T.-Y. 2020. PopMAG: Pop Music Accompaniment Generation. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, 1198–1206. New York, NY, USA: Association for Computing Machinery. ISBN 9781450379885.

Rogers, A.; Kovaleva, O.; and Rumshisky, A. 2021. A primer in BERTology: What we know about how BERT works. *Transactions of the association for computational linguistics*, 8: 842–866.

Russell, J. A. 1980. A circumplex model of affect. *Journal of personality and social psychology*, 39(6): 1161.

Simonetta, F.; Chacón, C. E. C.; Ntalampiras, S.; Widmer, G.; et al. 2019. A Convolutional Approach to Melody Line Identification in Symbolic Scores. In *ISMIR 2019*, 1–8. ISMIR.

Tian, J.; Li, Z.; Li, J.; and Wang, P. 2024. N-gram unsupervised compounding and feature injection for better symbolic music understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 15364–15372.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L. u.; and Polosukhin, I. 2017. Attention is All you Need. In Guyon, I.; Luxburg, U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Wu, B.; Zhong, E.; Horner, A.; and Yang, Q. 2014. Music emotion recognition by multi-label multi-layer multi-instance multi-view learning. In *Proceedings of the 22nd ACM international conference on Multimedia*, 117–126.

Xiao, Y.; Ruan, H.; Zhao, X.; Jin, P.; Tian, L.; Wei, Z.; Cai, X.; Wang, Y.; and Liu, L. 2024. An Efficient Bi-modal Fusion Framework for Music Emotion Recognition. *IEEE Transactions on Affective Computing*.

Zeng, M.; Tan, X.; Wang, R.; Ju, Z.; Qin, T.; and Liu, T.-Y. 2021. MusicBERT: Symbolic Music Understanding with Large-Scale Pre-Training. In Zong, C.; Xia, F.; Li, W.; and Navigli, R., eds., *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 791–800. Online: Association for Computational Linguistics.