

# EMOTION-DRIVEN PIANO MUSIC GENERATION VIA TWO-STAGE DISENTANGLEMENT AND FUNCTIONAL REPRESENTATION

Jingyue Huang<sup>1</sup>      Ke Chen<sup>1</sup>      Yi-Hsuan Yang<sup>2</sup>

<sup>1</sup>Department of Computer Science and Engineering, UC San Diego

<sup>2</sup>Department of Electrical Engineering, National Taiwan University

{jih150, knutchen}@ucsd.edu, yhyangtw@ntu.edu.tw

## ABSTRACT

Managing the emotional aspect remains a challenge in automatic music generation. Prior works aim to learn various emotions at once, leading to inadequate modeling. This paper explores the disentanglement of emotions in piano performance generation through a two-stage framework. The first stage focuses on valence modeling of lead sheet, and the second stage addresses arousal modeling by introducing performance-level attributes. To further capture features that shape valence, an aspect less explored by previous approaches, we introduce a novel functional representation of symbolic music. This representation aims to capture the emotional impact of major-minor tonality, as well as the interactions among notes, chords, and key signatures. Objective and subjective experiments validate the effectiveness of our framework in both emotional valence and arousal modeling. We further leverage our framework in a novel application of emotional controls, showing a broad potential in emotion-driven music generation.

## 1. INTRODUCTION

With the recent advancements in symbolic music generation [1–6], there has been a growing interest in controlling high-level musical features throughout the generation process. Among these features, *emotion-driven music generation* [7–13] aims to generate music that conveys specific emotions, representing a crucial aspect for music appreciation and analysis. The downstream applications of such models have also been explored, such as music therapy for healthcare and educational purposes [14] and soundtrack generation for videos and movies [15].

Emotion could be represented in two dimensions from the literature [16]: *valence* and *arousal*. Valence refers to the positiveness of an emotion and arousal refers to energy or activation [17–19]. These two dimensions can be further divided into four quadrants (4Q), namely high valence high arousal (Q1), low valence high arousal (Q2), low valence low arousal (Q3), and high valence low arousal (Q4).

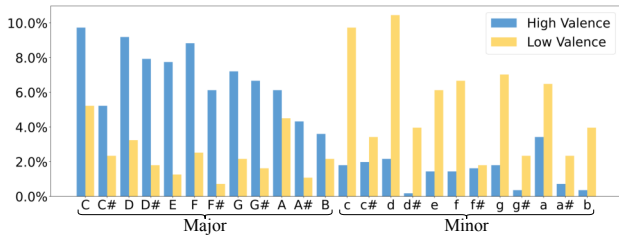
In this paper, we focus on the emotion-driven *piano performance generation* of these four quadrants. Throughout prior works, we observe crucial challenges from the perspectives of both model design and musical inductive bias.

First, previous emotion-driven piano performance generation models [7, 8] attempt to learn emotion quadrants and expressions in an *end-to-end* paradigm. In terms of model design, this approach poses training difficulty on the generation model, leading to the instability in achieving results of desired emotions. For example, many existing works [7, 9, 11] could effectively control the arousal levels of music, while their performance of *valence modeling*, especially in generating low valence (i.e., negative) music, is still poor. In terms of music, the creation process of music typically involves multiple stages, such as the *lead sheet composition* for melodies and chord progressions, and *performance generation* for textures and expressiveness. Consistently, emotion can be evoked through a combination of musical elements (e.g., melody, chord, texture). For example, major/minor chords have been found to seize different valence trends in psychological studies [20] and performance-level attributes like articulation, tempo, and velocity are more related to arousal [21, 22]. It is worth to explore the potential relation between the *disentanglement* of the generation process and the emotion expression.

Second, previous emotion-driven generation models have received limited attention regarding the influence of *tonality* on emotion modeling. It has been widely shown that major-minor tonality in composition is highly related to valence perception [22–25]. For example, as depicted in Figure 1, the histogram of musical keys derived from the emotion-labeled music dataset EMOPIA [7] supports the distribution skews to major keys for high valence clips and opposite trend for low valence ones. Furthermore, different tonalities may reveal similar patterns in the relative relationships between melodies and chords, while the distribution of melodies, chords, and tonalities can exhibit distinct shapes across different emotions. Current representations of symbolic music, such as REMI [2] and CP-Word [26], do not explicitly incorporate such interactions nor address its connection to emotion adequately. Therefore, it is necessary to consider a functional format of symbolic music representation considering the relationships between notes, chords and key signatures to better model the tonality in the emotion-driven music generation process.

In this paper, we contribute to combat above challenges:





**Figure 1.** Key histogram of high/low valence clips from the emotion-labeled piano music dataset EMOPIA [7].

- We employ a two-stage Transformer-based model on emotion-driven piano performance generation. The first stage focuses on valence modeling via lead sheet composition, while the second stage addresses arousal modeling by introducing performance-level attributes.
- We propose a novel functional representation for symbolic music, encoding both melody and chords with Roman numerals relative to musical keys, to consider the interactions among notes, chords and tonalities [27].
- Experiments demonstrate the effectiveness of our framework and representation on emotion modeling. Additionally, our method enables new capabilities to control the arousal levels of generation under the same lead sheet, leading to more flexible emotion controls.

As a minor contribution, we also refine key signature labels and extract lead sheet annotations for the EMOPIA dataset [7] to ensure the correct training of the two-stage framework. We share the data, open source our code<sup>1</sup> and present generation samples in the demo page.<sup>2</sup>

## 2. RELATED WORK

### 2.1 Emotion-driven Piano Performance Generation

Prior works apply emotion conditions on deep-learning models to guide the generation of piano performance [7, 8, 11], or develop searching methods to generate music of desired emotions [28, 29]. Musical elements via feature disentanglement [9] or supervised clustering [10] can further be regarded as a bridge between emotion labels and performances for generation. In contrast, our framework employs a two-stage generation approach to reduce the complexities of one-stage generation, fostering a more nature process of music creation as well as a better incorporation between emotion labels and generation results.

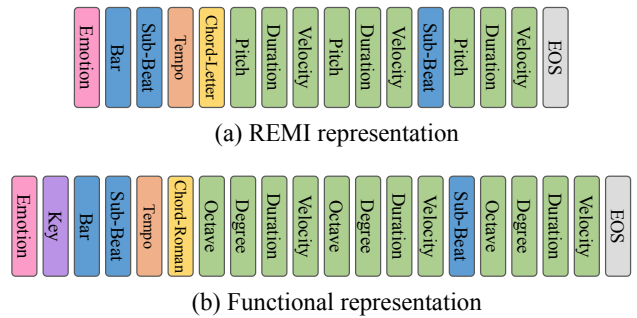
### 2.2 Tonality, Functional Harmony, and Emotion

Musical keys and functional harmony have been explored in the field of roman numeral analysis [30–32]. The analysis of how modes and tonalities relate to mid-level perceptual features (e.g., dissonance, tonal stability, minoriness) and affect the emotional perception of music pieces has also been discovered [22, 24].

While some music generation works attempted to combine key information into data representation [33], loss

<sup>1</sup> <https://github.com/Yuer867/EMO-Disentanger>

<sup>2</sup> <https://emo-disentanger.github.io/>



**Figure 2.** Illustration of (a) REMI [2], (b) the proposed functional representation, and their differences.

function [34] and text conditions [6], none of them explore the relation between musical keys and emotional perception. In this paper, we leverage both functional harmony knowledge and class-octave based pitch representation [35] to design a new data representation, incorporating the relationships between notes, chords and keys for emotion-driven music generation.

## 3. METHOD

In this section, we will first introduce the functional representation of symbolic music as the main generation unit. Then we introduce the two-stage model as the main component of the emotion disentanglement and generation.

### 3.1 Functional Representation

Figure 2 illustrates our proposed functional representation. Its design is initially based on REMI [2], a widely used event-based representation for symbolic music. We incorporate different note and chord events assisting to better learn the joint information of emotion and key signature.

#### 3.1.1 Emotion and Key Events

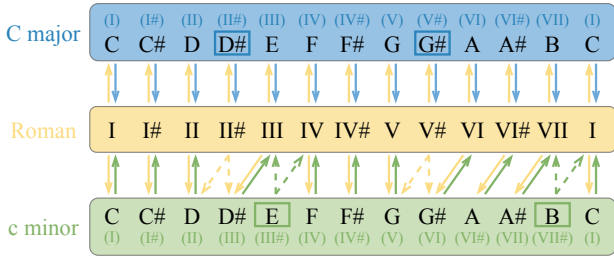
We follow CTRL [36] to set up the condition within the autoregressive generation process in Transformer architecture. To denote distinct emotions and affect overall properties, we begin the event sequence with `<Emotion_*>` event to indicate the emotion label of music clips. The `<Key_*>` event is appended after `<Emotion_*>` to provide the musical key property, with the total of 24 keys (12 tonic notes with two modes in EMOPIA [7]).

#### 3.1.2 Bar, Sub-Beat, Tempo and EOS Events

Similar to REMI, a `<Bar>` event denotes the new start of a bar; a `<Sub-Beat_*>` event denotes one of 16 possible discrete beat locations within a bar; a `<Tempo_*>` event denotes local tempo changes every four beats; and an `<EOS>` event denotes the end of sequence.

#### 3.1.3 Chord Events

A musical chord name typically consists of root note and chord quality. For example, `Fmaj` represents the chord F–A–C with root F and major quality. Such symbols describe correct note information in chord within the tonality,



**Figure 3.** The conversion between letters and Roman numerals in the cases of C major and c minor scales. Solid arrows denote strict one-to-one conversions, and dotted arrows denote optional one-to-either conversions.

but they overlook the variations in *chord functions* of the same chord across different tonalities. For example, while Fmaj serves the tonic function in F major scale, it serves the subdominant function in C major scale. Moreover, the chord progression follows these functional harmony rules to establish tonality and convey musical emotion [27].

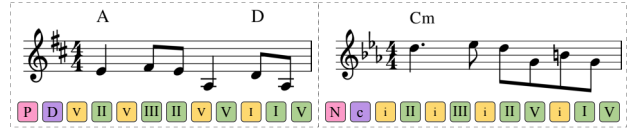
To introduce chord functions in the emotion modeling, we adopt Roman numerals from *Roman Numeral Analysis* [31] to notate chord roots in Figure 3. Given the  $\langle \text{Key}_* \rangle$  event, root notes in the absolute pitch are directly converted into Roman numerals based on their scale degrees relative to the key (i.e., relative pitch). For roots outside the scale, we employ a direct conversion for I#, II#, IV#, V# and VI# appearing in major keys, but randomly assign III# and VII#, which only appear in minor keys, as one of their neighboring degrees during the encoding and decoding process. This design ensures the notation to be key-independent and make every conversion of notes reasonable to the music theory. The notations of chord qualities remain unchanged, and the chord event  $\langle \text{Chord}_* \rangle$  appears every four beats.

### 3.1.4 Note-related Events

A note is denoted by  $\langle \text{Pitch}_* \rangle$ ,  $\langle \text{Duration}_* \rangle$  and  $\langle \text{Velocity}_* \rangle$  events, where  $\langle \text{Pitch}_* \rangle$  event indicates the onset of pitches from A0 to C8. Inspired by [35, 37], we decompose  $\langle \text{Pitch}_* \rangle$  into  $\langle \text{Octave}_* \rangle$  and  $\langle \text{Degree}_* \rangle$  events according to the note octave and degree in the certain key scale. The conversion rule from  $\langle \text{Pitch}_* \rangle$  to  $\langle \text{Degree}_* \rangle$  is the same as that of chord roots in Figure 3. For example, pitch D#4 is decomposed into  $\langle \text{Octave}_4 \rangle$  and  $\langle \text{Degree}_{III} \rangle$  in c minor scale, but  $\langle \text{Degree}_I \rangle$  in D# major scale. Such degree-octave pitch representation narrows the difference between melodies, thus improves the learning of connections between emotions, chords, and melodies, as demonstrated in Figure 4.

## 3.2 Two-stage Emotion Disentanglement

We use the idea of Compose & Embellish [38] to generate music in two stages: lead sheet first, and then piano performance. While Compose & Embellish is emotion-agnostic, we extend it so that the lead sheet model involves *valence modeling* and the performance model *arousal modeling*.



**Figure 4.** Two lead sheet examples from different songs in EMOPIA. In our functional representation, they have the same melody events (green), but different chord events (yellow) by different emotions (Positive and Negative by pink) and keys (D major or c minor by purple).

### 3.2.1 Valence Modeling

The top left section of Figure 5 denotes the first stage, where only emotion events  $\langle \text{Emotion\_Positive} \rangle$  and  $\langle \text{Emotion\_Negative} \rangle$  are considered as conditions. The former includes music pieces of Q1 and Q4 (high valence) and the latter includes those of Q2 and Q3 (low valence). The lead sheet model first predicts a key event  $k$  conditioned on the given emotion event  $e$ , and then generates the lead sheet sequence  $M = \{m_1, \dots, m_T\}$  of length  $T$ , as melody and chord progression, conditioned on previous tokens step-by-step:

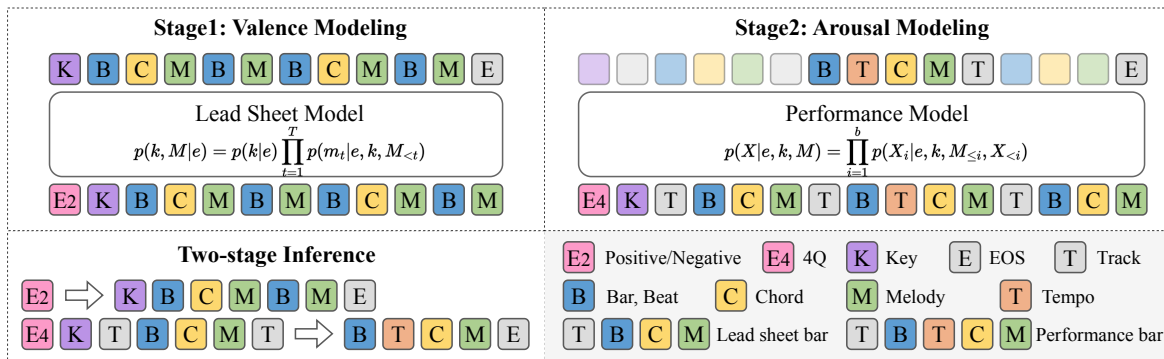
$$p(k, M|e) = p(k|e) \prod_{t=1}^T p(m_t|e, k, M_{<t}), \quad (1)$$

where  $p(k|e)$  and  $p(m_t|e, k, M_{<t})$  are jointly learned through the Transformer-based generation model [26, 38]. Performance-related events  $\langle \text{Velocity}_* \rangle$  and  $\langle \text{Tempo}_* \rangle$  are removed in the first stage (i.e., lead sheet generation), as we mainly focus on the contributions of key, pitch and chord for valence perception.

### 3.2.2 Arousal Modeling

The top right section of Figure 5 denotes the second stage. Given the lead sheet  $M$ , the performance model generates performance  $X$  conditioned on the true emotion label (Q1 to Q4). As the valence aspect has already been modeling in the first stage, this stage focuses on the generation of musical textures for the lead sheet, and more importantly, on how to perform it through variations of tempo, velocity, articulation, and other performance-level attributes that largely influence perceived arousal [21, 22]. During the training and inference phases, with the positions of  $\langle \text{Bar} \rangle$  events,  $M$  and  $X$  are further segmented into  $\{M_1, \dots, M_b\}$  and  $\{X_1, \dots, X_b\}$ , where  $b$  is the number of bars. The segmented sequences are “interleaved” in the form of  $\{\dots \langle \text{Track}_M \rangle, M_i, \langle \text{Track}_X \rangle, X_i \dots\}$  with additional  $\langle \text{Track}_* \rangle$  events to distinguish  $M$  and  $X$  tracks. In that, the target performance bar  $X_i$  is appended to its corresponding conditions  $M_i$ , as mapping each lead sheet segment to its corresponding performance segment [26]. With the emotion condition and key event from lead sheet as prefix tokens, the performance model is summarized as

$$p(X|e, k, M) = \prod_{i=1}^b p(X_i|e, k, M_{\leq i}, X_{<i}). \quad (2)$$



**Figure 5.** The two-stage framework of emotion-driven piano performance generation. Squares with transparent background denote the tokens that are not included in the loss computation during the training phase.

### 3.2.3 Training Objectives

Lead sheet and performance models are trained separately by both optimizing the negative log-likelihood loss of the sequence. Since existing emotion-labeled music datasets are not large, we leverage large-scale music datasets without emotion annotations to *pretrain* both models for better music understanding. During pretraining, the emotion event is marked as `<Emotion_None>`. We then finetune two models on the emotion-labeled dataset (detail in Section 4) to learn composition and performance styles specific to different emotion contexts.

### 3.2.4 Two-stage Inference

The left bottom section of Figure 5 denotes the inference process of both models. In the first stage, the lead sheet model predicts the key event and generates the lead sheet sequence step-by-step given `<Emotion_Positive>` or `<Emotion_Negative>` event, creating a musical motif for the specific valence preference. Even though our framework has the capability to generate any-key music of specific emotions, we observe that some generation results, such as a high-valence and high-arousal song with a minor key scale, may go beyond the current definition of emotion in [16], where the valence naturally has a strong correlation to the major-minor tonality (Figure 1). Therefore, we limit major keys to `<Emotion_Positive>` and minor keys to `<Emotion_Negative>` during the inference stage. We acknowledge that this can be overly simplifying. Since this paper focuses mainly on the valence-arousal disentanglement during the generation process, we leave this exploration of generating any emotion within any key as an advanced topic for future research.

In the second stage, the performance model generates piano performance with desired valence and arousal combination given the lead sheet from the first stage. For example, to generate a music piece of Q3, a “Negative” lead sheet and a “Q3” emotion event are selected as conditions. Additionally, this two-stage framework enables the flexibility to generate different arousal levels of piano performance under the same lead sheet, delivering some scenarios when the music need to shift quickly to complement the scenes in movies or daily videos (detail in Section 4 and the demo page).

## 4. EXPERIMENTS

### 4.1 Datasets and Preprocessing

As presented in Table 1, we collect different datasets for pretraining and finetuning phases as mentioned in Section 3.2.3. For pretraining the lead sheet model, we use the HookTheory dataset [39,40], where we choose 18,206 lead sheets with high-quality and human-transcribed melody, chord and key annotations in 4/4 time signature. We simplify 249 chord quality classes into 11 types<sup>3</sup> as the same set in the other datasets below. For pretraining the performance model, we use the Pop1k7 dataset [26], consisting of 1747 transcribed pop piano performances. Since Pop1k7 does not contain lead sheet annotations, we refer [38] to extract melodies using the skyline algorithm [41], recognize chords using the chorder library [42], and detect key signatures using [43] in MIDI Toolbox [44].

For finetuning the models with emotion conditions, we use the EMOPIA dataset [7], consisting of 1,071 music clips with human-annotated emotion labels. Similar to Pop1k7, we obtain the lead sheets of EMOPIA by extracting melodies using the algorithm in [45] and recognizing chords using the algorithm in [46]. Empirically, we observe that specifically in the EMOPIA dataset, melodies and chords extracted by these alternative algorithms are more correct compared to the skyline algorithm and the chorder library. Additionally, we found the key signature labels in EMOPIA are not fully correct since they are also obtained by the detection algorithm with error rates. Since the valence modeling is strongly related to the musical keys and modes, we manually correct the key annotations of 367 clips in EMOPIA to ensure a high quality of lead sheets.

All datasets are randomly divided into respective training and validation sets at the ratio of 9:1. As a result in our functional representation, the vocabulary size of events is 215 for lead sheet and 324 for piano performance.

### 4.2 Model Settings

The lead sheet model is a 12-layer Transformer Decoder [47] with 8 heads, 512 hidden dimensions and relative positional encoding [48]. The performance model is

<sup>3</sup> Major, minor, augment, diminish, suspend2, suspend4, major7, minor7, dominant7, diminish7, half-diminish7

Dataset	# clips (major)	# bars	# events
HookTheory [40]	18,206 (9,737)	10.84	282.81
Pop1k7 [26]	1,747 (1,264)	104.82	6794.86
EMOPIA(L) [7]	1,071 (618)	16.94	435.22
EMOPIA [7]	1,071 (618)	17.09	1311.47

**Table 1.** The datasets. (major) denotes the number of clips in major key (and the left is in minor key). The #bars and #events are average numbers across a dataset. EMOPIA(L) refers to EMOPIA lead sheets.

similar to the lead sheet model except with Performer attention [49]. The total parameter sizes are 41 million and 38 million respectively.

Both models are trained with the batch size of 4, the maximum sequence length of 512 (lead sheet model) or 3072 (performance model), and the Adam optimizer with  $\beta = (0.99, 0.9)$ . We adopt a 200-step warm-up to achieve the maximum learning rate of  $1e-4$  for pretraining and  $1e-5$  for finetune. All models are implemented by PyTorch and trained on one NVIDIA Tesla V100 GPU. The lead sheet model took around 180,000 steps to converge and the performance model took around 200,000 steps. Nucleus sampling [50] is employed in the inference phase. We referred [38, 51] to choose the sampling hyperparameters  $\tau = 1.2$ ,  $p = 0.97$  for the lead sheet model and  $\tau = 1.1$ ,  $p = 0.99$  for the performance model.

### 4.3 Baseline and Ablations

We consider the emotion-driven piano performance generation model in EMOPIA [7] as our baseline, which generates music in an end-to-end paradigm instead of two stages. To ensure the fair comparison of generation performance, we trained the baseline model under the same datasets in both pretraining and finetune phases, and replaced the original CP-Word representation with REMI as the former usually yields better generation performance and more comparable to our proposed functional representation. Two other related works [8, 9] are not included in comparison due to the main reason that we focus more on the evaluation of the two-stage framework in valence and arousal modeling; and the partial reason that they are not open-source or releasing the reproducible model weights.

We conduct a comprehensive ablation study to evaluate if each proposed design benefits the emotion modeling of music generation. Specially, these designs include: 1) the two-stage generation, 2) the functional representation, and 3) the dataset pretraining. In the following sections, models are denoted as **<representation(stage)>**. For example, REMI(one) denotes the one-stage generation model with REMI representation as the baseline, and REMI(two) denotes the two-stage generation as one variant.

### 4.4 Objective Evaluation and Results

Even though previous studies [7–9] employ metrics, such as Pitch Range (PR) and Number of Pitch Classes (NPC), to evaluate the generation performance, they do not pro-

	M	C	M+C	P
REMI+key (two)	0.465	0.065	0.075	0.418
–w/o. pretrain	0.350	0.105	0.130	0.343
functional (two)	<b>0.505</b>	<b>0.700</b>	<b>0.735</b>	<b>0.548</b>
–w/o. pretrain	0.400	0.570	0.625	0.430
Real data	0.578	0.695	0.746	0.812

**Table 2.** Key consistency calculated across all components, including melody (M), chord (C), lead sheet (M+C) and performance (P).

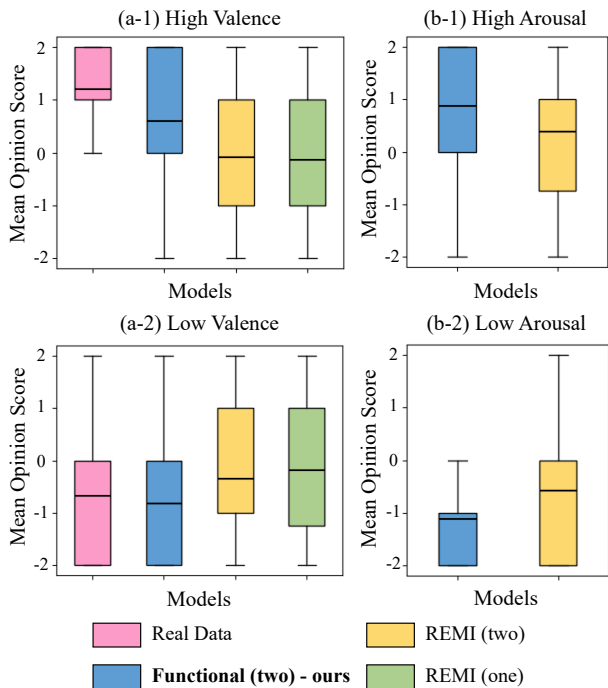
vide any evidences on the superiority of melody development, chord progression, and texture arrangement of music. Therefore, a model with more similar PR and NPC values to those of the target dataset does not necessarily promise a better generation quality than others.

Instead of using such metrics, we wish to evaluate the consistency between the input musical conditions and the generation results. We introduce **key consistency** to assess if a model can generate music pieces that adhere to the desired input key signatures, which is highly correlated to the lead sheet development and valence modeling. Specifically, key consistency measures the match between the key condition  $\langle \text{Key}_* \rangle$  and the actual key detected in the generation using the algorithm [44] with an 81% accuracy rate. We compare REMI and our functional representation to determine if the functional representation can improve the key consistency via more close and interactive designs on key, melody, and chord. Since this metric requires the key as conditions, we add the  $\langle \text{Key}_* \rangle$  event in REMI after  $\langle \text{Emotion}_* \rangle$  (as REMI+key in Table 2) when training the model. The non-pretrained versions are also included for comparison. Each model generates 200 lead sheets (100 high and 100 low valence) and 400 performance samples (100 per emotion quadrant) for evaluation.

From Table 2, the functional representation outperforms REMI (i.e., REMI+key) across all components and even achieves compatible accuracy to real data over the lead sheet component. This demonstrates the effectiveness of the functional representation, by representing notes and chord roots relative to key events for key modeling. In contrast, REMI struggles with associating chord events with keys due to the ignorance of chord labels serving different functions in different key scales. Moreover, pretraining process introduces musical priors to enhance the learning of key relationships with other musical elements, improving key consistency for both representations.

### 4.5 Subjective Evaluation and Results

We leverage an online listening test to assess the emotion modeling ability of models. The test was conducted to collect user responses on three parts: 1) valence modeling, 2) arousal modeling, and 3) 4Q emotion modeling. During this test, the quality of the generated music has also been assessed implicitly as it is a prerequisite to the emotion expression in the music. 22 participants were engaging in this test, 5 with less than 2 years of musical training, 8 with 2-5 years, 3 with 5-10 years, and 6 with more than 10 years.



**Figure 6.** The mean opinion score performance on the valence-oriented and arousal-oriented listening tests. For (a-1) and (b-1), the higher score the better performance; for (a-2) and (b-2), the lower score the better performance.

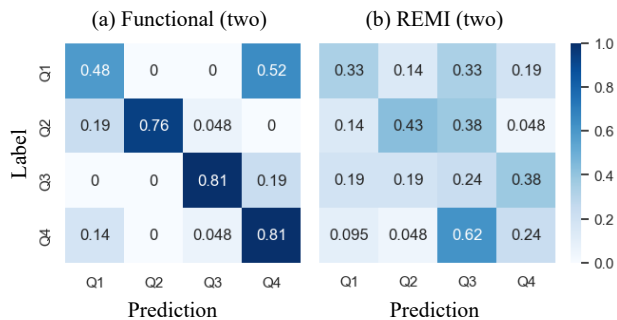
#### 4.5.1 Valence Modeling

In this part, each participant listened to 16 generated tracks of piano performance from four models [four tracks (two high valence and two low valence) per model]: 1) Real data; 2) REMI (one); 3) REMI (two); 4) Functional (two). For each track, participants rated its positiveness from  $-2$  (low valence) to  $2$  (high valence) with the step size 1.

The left of Figure 6 (‘a’) presents the mean opinion scores for the valence-oriented test, where the Functional (two) model significantly outperforms both REMI (two) and REMI (one) models. The REMI (two) model shows a slight improvement over REMI (one) due to its two-stage design. Our proposed Functional (two) model even marginally exceeds real data in low valence scores, which could be due to the potential subjective biases in the negative emotion as discussed in [10]. And our model achieves both great performance in high valence and low valence results, demonstrating a good balance in valence modeling.

#### 4.5.2 Arousal Modeling

In the second part, the functional (two) and REMI (two) models are chosen to compare their arousal modeling performance. Specifically, we wish to explore whether they can generate piano performance with either high or low arousal under the same lead sheet based on the given conditions (Q1 and Q4 for positive lead sheets, Q2 and Q3 for negative ones). Two pairs of generated tracks are randomly drawn for each model and each valence level, where every pair includes two tracks of different arousal conditions. For each track, participants rated its arousal level from  $-2$  (low arousal) to  $2$  (high arousal) with the step size 1.



**Figure 7.** The confusion matrices on the 4Q listening tests.

The right of Figure 6 (‘b’) presents the results. The Functional (two) model surpasses REMI (two) by an average of 0.5 point, highlighting its superior ability to differentiate between the musical features of the two arousal levels through performance. Additionally, it is rare for Functional (two) to be incorrectly identified as high arousal under low arousal conditions (Figure 6 (b-2)).

#### 4.5.3 4Q Emotion Judgement

In the last part, participants needed to choose the best option from four options (4Q) for each track, with 8 tracks in total for the two models the last section (4 tracks per model and 1 track per emotion).

Figure 7 presents the confusion matrices of two models. The Functional (two) model achieves the higher overall accuracy than that of REMI (two) (71.5% vs. 31.0%). When examining each emotion category, Functional(two) demonstrates superior performance in Q3 and Q4 than Q1 and Q2. Furthermore, music pieces generated from it with high valence conditions are misidentified almost based on their arousal levels; for instance, pieces intended for Q1 are almost mistaken for Q4 and vice versa. In contrast, for REMI(two), the misclassifications are across all categories, demonstrating its limitations in modeling the four emotion classes although through two-stage generation.

All above evaluations support that the combination of two-stage framework and functional representation is effective in controlling the emotion of the music it generates to a certain extent.

## 5. CONCLUSION AND FUTURE WORK

In this paper, we first explore emotion disentanglement through a two-stage Transformer-based framework for emotion-driven piano performance generation. Then we propose a novel functional representation for symbolic music to capture the interactions among musical keys, modes, chords, and melodies in relation to the emotion contexts. An objective metric is designed to qualify the key modeling of the proposed method, and subjective evaluations further confirm its ability to convey desired emotional perception. In the future, we wish to focus on enhancing the flexibility of emotional music generation across all musical keys and investigating new applications fostered by our framework, such as the controls of valence and arousal attributes under similar music motifs.

## 6. ACKNOWLEDGMENT

The work is supported by a grant from the National Science and Technology Council of Taiwan (NSTC 112-2222-E-002-005-MY2). This work also benefits from the review comments of an unpublished work [52], which is an early version of the present study. While the concept of functional representation has been mentioned in that previous work [52], its focus was on emotion-driven melody harmonization and the effectiveness of the functional representation in that context was unclear, as the melody was pre-given and fixed. In the current work, melody and chords are generated together, allowing more flexibility for emotion control with the proposed functional representation. The source code of that previous work can be found at the GitHub repo: [https://github.com/Yuer867/EMO\\_Harmonizer](https://github.com/Yuer867/EMO_Harmonizer).

## 7. REFERENCES

- [1] C.-Z. A. Huang, A. Vaswani, J. Uszkoreit, I. Simon, C. Hawthorne, N. Shazeer, A. M. Dai, M. D. Hoffman, M. Dinculescu, and D. Eck, "Music Transformer: Generating music with long-term structure," in *Proc. ICLR*, 2019.
- [2] Y.-S. Huang and Y.-H. Yang, "Pop Music Transformer: Beat-based modeling and generation of expressive pop piano compositions," in *Proc. ACM Multimed.*, 2020.
- [3] D. von Rütte, L. Biggio, Y. Kilcher, and T. Hoffmann, "FIGARO: Generating symbolic music with fine-grained artistic control," in *Proc. ICLR*, 2023.
- [4] K. Chen, C. Wang, T. Berg-Kirkpatrick, and S. Dubnov, "Music sketchnet: Controllable music generation via factorized representations of pitch and rhythm," in *Proc. ISMIR*, 2020.
- [5] S.-L. Wu and Y.-H. Yang, "MuseMorphose: Full-song and fine-grained piano music style transfer with one Transformer VAE," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 31, pp. 1953–1967, 2023.
- [6] P. Lu, X. Xu, C. Kang, B. Yu, C. Xing, X. Tan, and J. Bian, "MuseCoco: Generating symbolic music from text," *CoRR*, vol. abs/2306.00110, 2023.
- [7] H. Hung, J. Ching, S. Doh, N. Kim, J. Nam, and Y.-H. Yang, "EMOPIA: A multi-modal pop piano dataset for emotion recognition and emotion-based music generation," in *Proc. ISMIR*, 2021.
- [8] P. L. T. Neves, J. Fornari, and J. B. Florindo, "Generating music with sentiment using Transformer-GANs," in *Proc. ISMIR*, 2022.
- [9] S. Ji and X. Yang, "MusER: Musical element-based regularization for generating symbolic music with emotion," in *Proc. AAAI*, 2024.
- [10] C. Kang, P. Lu, B. Yu, X. Tan, W. Ye, S. Zhang, and J. Bian, "EmoGen: Eliminating subjective bias in emotional music generation," *CoRR*, vol. abs/2307.01229, 2023.
- [11] L. Ferreira and J. Whitehead, "Learning to generate music with sentiment," in *Proc. ISMIR*, 2019.
- [12] E. Choi, Y. Chung, S. Lee, J. Jeon, T. Kwon, and J. Nam, "YM2413-MDB: A multi-instrumental FM video game music dataset with emotion annotations," in *Proc. ISMIR*, 2022.
- [13] W. Cui, P. Sarmiento, and M. Barthelet, "MoodLoopGP: Generating emotion-conditioned loop tablature music with multi-granular features," in *Proc. EvoMUSART*, 2024.
- [14] K. Zheng, R. Meng, C. Zheng, X. Li, J. Sang, J. Cai, J. Wang, and X. Wang, "EmotionBox: A music-element-driven emotional music generation system based on music psychology," *Frontiers in Psychology*, vol. 13, 2022.
- [15] M. T. Haseeb, A. Hammoudeh, and G. Xia, "GPT-4 driven cinematic music generation through text processing," in *Proc. ICASSP*, 2024.
- [16] J. A. Russell, "A circumplex model of affect," *Journal of Personality and Social Psychology*, 1980.
- [17] Y.-H. Yang, Y.-C. Lin, Y.-F. Su, and H. H. Chen, "A regression approach to music emotion recognition," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 2, pp. 448–457, 2008.
- [18] —, "Machine recognition of music emotion: A review," *ACM Trans. Intelligent Systems and Technology*, vol. 3, no. 3, 2012.
- [19] J. S. G. Cañón, E. Cano, T. Eerola, P. Herrera, X. Hu, Y.-H. Yang, and E. Gómez, "Music emotion recognition: Toward new, robust standards in personalized and context-sensitive applications," *IEEE Signal Process. Magazine*, vol. 38, no. 6, pp. 106–114, 2021.
- [20] D. R. Bakker and F. H. Martin, "Musical chords and emotion: major and minor triads are processed for emotion," *Cognitive, Affective, & Behavioral Neuroscience*, 2015.
- [21] Y.-C. Wu and H. H. Chen, "Generation of affective accompaniment in accordance with emotion flow," *IEEE Trans. Audio, Speech, Lang. Process.*, 2016.
- [22] S. Chowdhury and G. Widmer, "On perceived emotion in expressive piano performance: Further experimental evidence for the relevance of mid-level perceptual features," in *Proc. ISMIR*, 2021.
- [23] R. Panda, R. Malheiro, and R. P. Paiva, "Audio features for music emotion recognition: A survey," *IEEE Trans. Affective Computing*, 2020.

- [24] A. Aljanaki and M. Soleymani, “A data-driven approach to mid-level perceptual musical feature modeling,” in *Proc. ISMIR*, 2018.
- [25] Y. Hong, R. K. Mo, and A. Horner, “The effects of mode, pitch, and dynamics on valence in piano scales and chord progressions,” in *Proc. ICMC*, 2018.
- [26] W. Hsiao, J. Liu, Y. Yeh, and Y.-H. Yang, “Compound Word Transformer: Learning to compose full-song music over dynamic directed hypergraphs,” in *Proc. AAAI*, 2021.
- [27] T. Chen and L. Su, “Functional harmony recognition of symbolic music data with multi-task recurrent neural networks,” in *Proc. ISMIR*, 2018.
- [28] L. N. Ferreira, L. Mou, J. Whitehead, and L. H. S. Lelis, “Controlling perceived emotion in symbolic music generation with monte carlo tree search,” in *Proc. of AAAI (AIIDE Workshop)*, 2022.
- [29] L. N. Ferreira, L. H. S. Lelis, and J. Whitehead, “Computer-generated music for tabletop role-playing games,” in *Proc. of AAAI (AIIDE Workshop)*, 2020.
- [30] N. N. López, M. Gotham, and I. Fujinaga, “Augmentednet: A roman numeral analysis network with synthetic training examples and additional tonal tasks,” in *Proc. ISMIR*, 2021.
- [31] G. Micchi, M. Gotham, and M. Giraud, “Not all roads lead to Rome: Pitch representation and model architecture for automatic harmonic analysis,” *TISMIR*, 2020.
- [32] E. Karystinaios and G. Widmer, “Roman numeral analysis with graph neural networks: Onset-wise predictions from note-wise features,” in *Proc. ISMIR*, 2023.
- [33] B. L. Sturm, J. F. Santos, O. Ben-Tal, and I. Korshunova, “Music transcription modelling and composition using deep learning,” *CoRR*, vol. abs/1604.08723, 2016.
- [34] Y. Yeh, W. Hsiao, S. Fukayama, T. Kitahara, B. Genchel, H. Liu, H. Dong, Y. Chen, T. Leong, and Y.-H. Yang, “Automatic melody harmonization with triad chords: A comparative study,” *CoRR*, vol. abs/2001.02360, 2020.
- [35] Y. Li, S. Li, and G. Fazekas, “An comparative analysis of different pitch and metrical grid encoding methods in the task of sequential music generation,” *CoRR*, vol. abs/2301.13383, 2023.
- [36] N. S. Keskar, B. McCann, L. R. Varshney, C. Xiong, and R. Socher, “CTRL: A conditional transformer language model for controllable generation,” *CoRR*, vol. abs/1909.05858, 2019.
- [37] M. Mongeau and D. Sankoff, “Comparison of musical sequences,” *Computers and the Humanities*, vol. 24, no. 3, pp. 161–175, 1990.
- [38] S.-L. Wu and Y.-H. Yang, “Compose & Embellish: Well-structured piano performance generation via a two-stage approach,” in *Proc. ICASSP*, 2023.
- [39] “HookTheory,” <https://www.hooktheory.com/> [Accessed: (September 1, 2023)].
- [40] C. Donahue, J. Thickstun, and P. Liang, “Melody transcription via generative pre-training,” in *Proc. ISMIR*, 2022.
- [41] A. L. Uitdenbogerd and J. Zobel, “Manipulation of music for melody matching,” in *Proc. ACM Multimed.*, 1998.
- [42] J. Chang, “Chorders,” <https://github.com/joshuachang2311/chorder>.
- [43] C. L. Krumhansl, “Cognitive foundations of musical pitch,” *Oxford University Press*, 2001.
- [44] P. Toiviainen and T. Eerola, “MIDI toolbox 1.1,” <https://github.com/miditoolbox/>, 2016.
- [45] “Midi\_Toolkit,” [https://github.com/RetroCirce/Midi\\_Toolkit](https://github.com/RetroCirce/Midi_Toolkit) [Accessed: (September 1, 2023)].
- [46] Z. Wang, K. Chen, J. Jiang, Y. Zhang, M. Xu, S. Dai, and G. Xia, “POP909: A pop-song dataset for music arrangement generation,” in *Proc. ISMIR*, 2020.
- [47] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” in *Proc. NeurIPS*, 2017.
- [48] Z. Dai, Z. Yang, Y. Yang, J. G. Carbonell, Q. V. Le, and R. Salakhutdinov, “Transformer-XL: Attentive language models beyond a fixed-length context,” in *Proc. ACL*, 2019.
- [49] K. M. Choromanski, V. Likhoshesterov, D. Dohan, X. Song, A. Gane, T. Sarlós, P. Hawkins, J. Q. Davis, A. Mohiuddin, L. Kaiser, D. B. Belanger, L. J. Colwell, and A. Weller, “Rethinking attention with Performers,” in *Proc. ICLR*, 2021.
- [50] A. Holtzman, J. Buys, M. Forbes, and Y. Choi, “The curious case of neural text degeneration,” in *Proc. ICLR*, 2019.
- [51] Z. Fu, W. Lam, A. M. So, and B. Shi, “A theoretical analysis of the repetition problem in text generation,” in *Proc. AAAI*, 2021.
- [52] J. Huang and Y.-H. Yang, “Emotion-driven melody harmonization via melodic variation and functional representation,” *CoRR*, vol. abs/2407.20176, 2024.