Responding to the Call: Exploring Automatic Music Composition Using a Knowledge-Enhanced Model

Zhejing Hu[1], Yan Liu[1]*, Gong Chen[1], Xiao Ma[1], Shenghua Zhong[2], Qianwen Luo[1]

[1] Department of Computing, The Hong Kong Polytechnic University
[2] College of Computer Science and Software Engineering, Shenzhen University
zhejing.hu@connect.polyu.hk, yan.liu@polyu.edu.hk, gong-cg.chen@polyu.edu.hk, edward-xiao.ma@connect.polyu.hk, csshzhong@szu.edu.cn, qianwenluo@gmail.com

Abstract

Call-and-response is a musical technique that enriches the creativity of music, crafting coherent musical ideas that mirror the back-and-forth nature of human dialogue with distinct musical characteristics. Although this technique is integral to numerous musical compositions, it remains largely uncharted in automatic music composition. To enhance the creativity of machine-composed music, we first introduce the Call-Response Dataset (CRD) containing 19,155 annotated musical pairs and crafted comprehensive objective evaluation metrics for musical assessment. Then, we design a knowledge-enhanced learning-based method to bridge the gap between human and machine creativity. Specifically, we train the composition module using the call-response pairs, supplementing it with musical knowledge in terms of rhythm, melody, and harmony. Our experimental results underscore that our proposed model adeptly produces a wide variety of creative responses for various musical calls.

Introduction

Automatic music composition, a classic research topic in computational creativity (Carnovalini and Rodà 2020; Mateja and Heinzl 2021), has garnered significant research attention (Dong et al. 2018; Ferreira et al. 2022; Hsiao et al. 2021; Bretan et al. 2017; Jiang et al. 2020b) and applications (Hu et al. 2020; Wesseldijk, Mosing, and Ullén 2021; Engels, Tong, and Chan 2015). Compared to visual arts like painting, music is inherently abstract, lacking a tangible reference such as a definitive image. Studying this abstraction allows machines to emulate human musicians, offering promising advancements in genuine machine intelligence. Recent trends leverage sequence-to-sequence models to capture creativity in music composition (Ramesh et al. 2021; Li et al. 2022; Huang et al. 2018; Huang and Yang 2020; Ens and Pasquier 2020; Muhamed et al. 2021). The effectiveness of data-driven models largely depends on extensive training datasets. However, the specialized nature of music composition results in smaller datasets compared to those for text and images. This data limitation challenges the ability of these models to master complex creative composition techniques for evoking aesthetic and emotional depth.

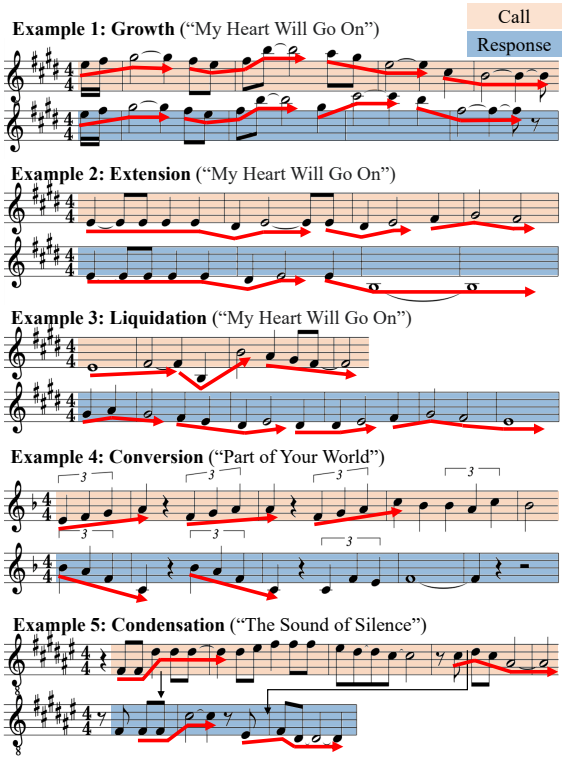


Figure 1: Different types of call-and-response effects to express creativity in human-composed music.

Call-and-response, a fundamental compositional technique in human-composed music, expresses creativity and is also known as antecedent and consequent phrases in musicology. In this technique, one musical phrase acts as the "call" and is "answered" by a corresponding phrase (Titon 2016; Benward 2018). Pivotal in improvisation and choral settings (Benetatos, VanderStel, and Duan 2020; Biles et al. 1994), this technique enriches compositional structures, adding layers of creativity, diversity, and narrative depth, especially in crafting verses or choruses (Schoenberg, Stein, and Strang 1967). Creativity shines through call-and-response with effects such as growth, extension, liquidation, conversion, and condensation (Figure 1). These effects

*Corresponding author.

guide human creativity, turning abstract ideas and emotions into tangible musical expressions by offering novel interpretations of the original melody and shaping new emotional landscapes.

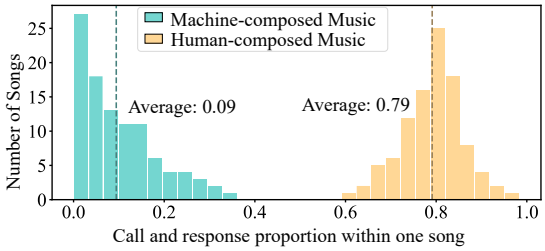


Figure 2: Comparative Analysis of Call-and-Response Proportions in Human-Composed and Machine-Composed Music: A Study of 100 Pop909 Songs (Wang et al. 2020) and 100 Machine-Composed Songs from (Agostinelli et al. 2023; Dhariwal et al. 2020; Payne 2019).

While call-and-response is foundational in human-composed music for expressing creativity, it is underrepresented in machine-composed pieces, including those produced by the most advanced music generation products (Agostinelli et al. 2023; Dhariwal et al. 2020; Payne 2019), as depicted in Figure 2[1]. Though creativity is an unbounded art, unrestricted freedom in automatic music composition might yield pieces beyond human comprehension and appreciation, diverging from music's essence as an accessible art form. As a result, employing the call-and-response technique to convey creativity remains a challenge for current automatic music composition models.

To address this, we study the call-and-response technique to enhance the creativity of machine-composed music, aspiring to transition machine compositions from merely "correct" to truly "artistic". We introduce the Call-Response Dataset (CRD) with 19,155 annotated call-response pairs in music, establishing a structured resource for machines to learn and generate creative responses to human-composed music. Alongside, we present comprehensive objective evaluation metrics, further refining musical assessment in computational creativity. Additionally, our innovative Call-Response Generator (CRG) integrates a knowledge-enhanced mechanism that focuses on rhythm, melody, and harmony into the learning-based method, thereby bridging the gap between human and machine creativity. Preliminary results highlight the effectiveness of our model, adeptly producing creative and varied responses, thereby enriching the musical experience of machine-generated compositions.

Related Work

Automatic Music composition can be broadly divided into two primary streams. The first stream consists of learning-based methods (Briot, Hadjeres, and Pachet 2020; Ji, Luo, and Yang 2020) like MidiNet (Yang, Chou, and Yang 2017), MuseGAN (Dong et al. 2018), Music Transformer (Huang et al. 2018), and Pop Music Transformer (Huang and Yang 2020). These approaches leverage models such as CNN, GAN, and Transformer to derive patterns from data samples. While capable of producing minute-long compositions, these models often overlook central musical ideas and crucial techniques (Hernandez-Olivan and Beltran 2022). One solution could be to further increase the size of the music data, but this would require more well-structured data and a significant investment in time and resources for collection and pre-processing given the intricacy of composition. Even if more data were used, it doesn't ensure the capture of specific compositional techniques, which can lead to unpredictable outcomes (Wu and Yang 2020).

Recognizing that music is an art requiring both rule-based and learning-based approaches, researchers have designed a combination of learning-based and rule-based methods by incorporating music theory (Medeot et al. 2018; Jiang et al. 2020a; Roberts et al. 2018; Akama 2019). This approach reduces the reliance on vast data samples and closely emulates the human process of studying composition, which involves learning music theory and listening to numerous music pieces. Many studies have demonstrated that combining learning-based and rule-based methods can enhance the quality of automatic music composition (Dai et al. 2021, 2023; Lu et al. 2022). Some typical knowledge-enhanced automatic music composition methods involve incorporating knowledge such as genres (Mao, Shin, and Cottrell 2018), music themes (Shih et al. 2022), melody structures (Wu et al. 2020), harmonic features (Zhang et al. 2022), and internal graph structure (Zou et al. 2022) into machine learning models for composition. Recent studies have explored attention mechanisms for improved structure and coherence (Keerti et al. 2022; Hsiao et al. 2021; Ens and Pasquier 2020), as well as reinforcement learning algorithms for generating music with specific characteristics like motif and harmony (Guo, Xu, and Xu 2022).

Despite these advancements in automatic music composition, there remains a significant deficiency in reproducing the call-and-response technique, a cornerstone of numerous music genres for creativity. Therefore, a refined call-and-response generator is essential, highlighting a research gap in this domain.

Call-Response Dataset

Annotation Process

We enlisted three annotators with musical training backgrounds to manually identify each call-response pair. Two annotators are asked to annotate all music independently. All pairs preserve the information from the original dataset, which consists of three tracks: melody (lead melody), piano (primary accompaniment arrangement), and bridge (secondary melodies or lead instrument arrangements). Given the intricate nature of this task and our commitment to achieving precise and accurate results, we instituted a robust quality control process. The third annotator independently scrutinized the extracted call-response pairs, vigi-

[1]While products like MusicLM are prompt-based audio generation products, we include them in our comparison to highlight the gap between human-composed and machine-composed music.

lantly checking for any potential errors, discrepancies, or inconsistencies in the initial annotations. In situations where discrepancies were identified, all three annotators were engaged in deliberation to discuss and resolve the inconsistencies, thereby fostering consensus and maintaining the accuracy of the annotations. A detailed dataset collection process is described in Appendix A1.

Dataset Exploration

Objective Evaluation Metrics We evaluate the quality of calls and responses based on their rhythm, melody, and harmony. We use dynamic time warping (DTW) distance and maximum common subsequence (MCS). These metrics have been widely used in music composition for similarity comparisons (Collins 2012; Dai et al. 2023; Chikkamath and Nirmala 2021; Hu et al. 2022).

For rhythm and melody evaluation, we use DTW to measure two different attributes: note position and note duration. The DTW distance of note position and note duration between the call and the human-composed response is denoted as DTW_P and DTW_D. The DTW distance between the input call and generated response is denoted as $\hat{DTW}_P$ and $\hat{DTW}_D$. Rhythm Quality (RQ) is then computed as a normalized metric ranging from 0 to 1:

$$RQ = 1 - \frac{1}{2} \times \left(\frac{|DTW_P - \hat{DTW}_P|}{max(DTW_P, \hat{DTW}_P)} + \frac{|DTW_D - \hat{DTW}_D|}{max(DTW_D, \hat{DTW}_D)}\right) \quad (1)$$

If the generated response differs significantly from the human-composed response, RQ approaches 0; otherwise, it is closer to 1. A similar process is applied to measure Melody Quality (MQ) using the DTW metric, which is the distance of the melodic pitch:

$$MQ = 1 - \frac{|DTW_M - \hat{DTW}_M|}{max(DTW_M, \hat{DTW}_M)} \quad (2)$$

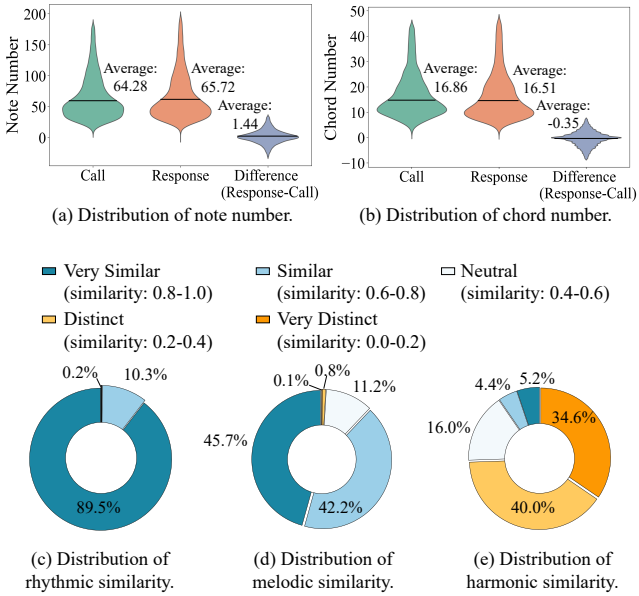
For harmony evaluation, we measure the MCS of chord progressions between two music pieces. Harmony Quality (HQ) is calculated as:

$$HQ = 1 - \frac{|MCS_H - \hat{MCS}_H|}{max(MCS_H, \hat{MCS}_H)} \quad (3)$$

We use an N-gram to measure the diversity of the generated responses. The N-gram measures the average number of times each n-gram appears in the generated music. This is widely used for assessing the diversity of text generated by language models (de Rosa and Papa 2021) and has been adopted in music composition (Zhang et al. 2022). The n-gram diversity of the pitch sequence within the generated response D_n is defined as:

$$D_n = \frac{|T|}{|G| - n + 1}, \quad (4)$$

where $|T|$ is the number of unique n-grams in all n-grams G, and $|G| - n + 1$ is the number of total n-grams that can be formed from G.



(a) Distribution of note number. (b) Distribution of chord number.

(c) Distribution of rhythmic similarity. (d) Distribution of melodic similarity. (e) Distribution of harmonic similarity.

Figure 3: CRD statistics. (a-b) Distributions of note and chord numbers. (c-e) Distributions of similarity between call-response pairs in terms of rhythm, melody, and harmony.

Dataset Statistics We extracted a total of 19,155 call-response pairs, totaling over 108 hours from Pop909 (Wang et al. 2020). Our CRD is the first to annotate the call-and-response technique in automatic music composition. Figure 3 (a) and (b) illustrate the distribution of note and chord numbers in calls, responses, and call-response differences. Most music pieces feature approximately 65 notes and 16 chords. The length difference is centered around 0, which suggests that the lengths of call-and-response are similar in most cases. Additionally, we demonstrate the rhythm, melody, and harmony similarity between each call-response pair in Figure 3 (c-e). For rhythm and melody, most call-response pairs exhibit similar rhythm and melody patterns. For harmony, most call-response pairs have distinct harmony patterns.

Method

Overview

In this section, we present the CRG, a model designed for call-and-response generation in music, as shown in Figure 4. CRG uses a music piece as input and produces a complementary response based on diverse music candidates from music knowledge for creative results. It comprises two main modules: a compositional encoder-decoder and a knowledge enhancement module.

Mathematically, let $\mathbf{X}$ be a two-dimensional matrix of size $L_1 \times A$, where L_1 represents the length of the music call sequence and A denotes the number of attributes. Each token of $\mathbf{X}$ is a natural number, so $\mathbf{X} \in \mathbb{N}^{L_1 \times A}$. We construct music knowledge $\mathcal{K}$ and design a model M_θ that generates response $\mathbf{Y} = M_\theta(\mathbf{X}, \mathcal{K})$. The response $\mathbf{Y} \in \mathbb{N}^{L_2 \times A}$ has a

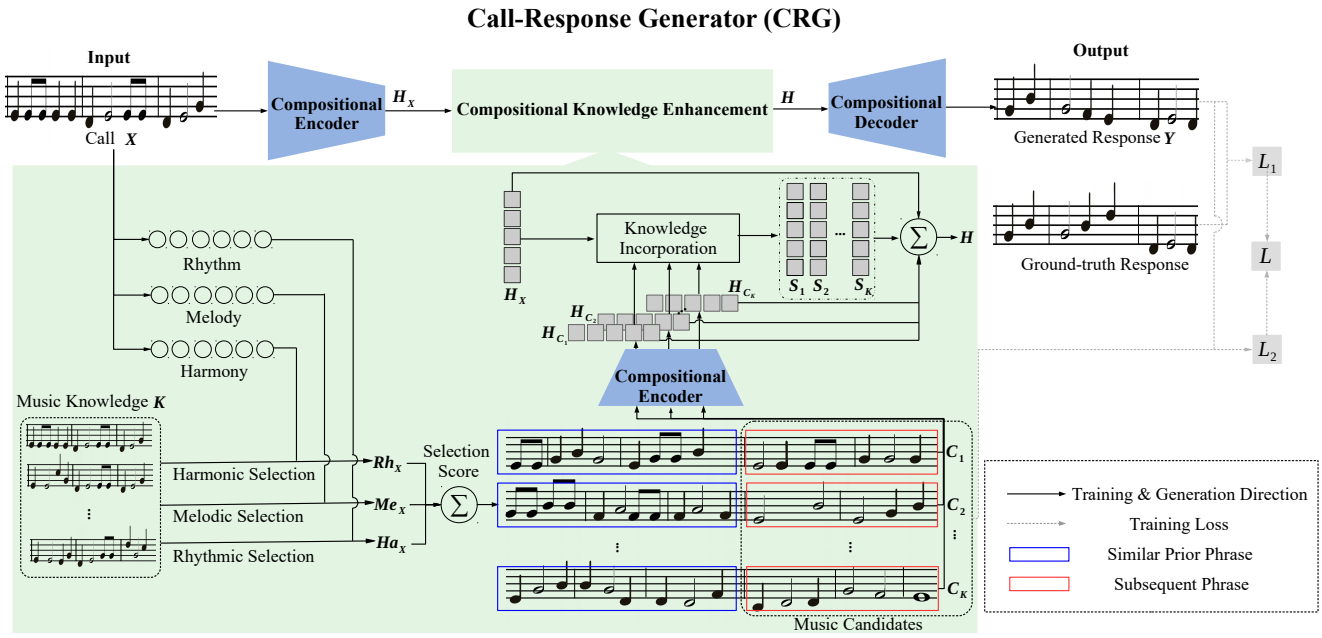


Figure 4: The framework for CRG. CRG takes a call as input and utilizes knowledge as a reference to generate high-quality responses.

length L_2 corresponding to the input.

Compositional Encoder and Decoder

The compositional encoder and decoder together constitute a sequence-to-sequence model. The encoder maps a variable-length input music sequence (call) into a fixed-length vector representation, while the decoder translates the vector representation into a variable-length output music sequence (response).

From a probabilistic standpoint, the encoder-decoder framework learns the conditional distribution of a variable-length sequence given another variable-length sequence:

$$P(\mathbf{Y}|\mathbf{X}) = P(y_1, \ldots, y_{L_2}|x_1, \ldots, x_{L_1}) = \prod_{t=1}^{L_2} p(y_t|\mathbf{X}, y_1, \ldots, y_{t-1}). \quad (5)$$

We denote the encoder process as:

$$\mathbf{H_X} = \text{ENCODER}(E(x_1), E(x_2), \ldots, E(x_{L_1})), \quad (6)$$

where $E(\cdot)$ is the embedding layer, and $\mathbf{H_X} \in \mathbb{R}^{L_1 \times H}$ is a sequence of hidden states $\{\mathbf{h}_i\}_{i=1}^{L_1}$ mapped from the input sequence. We use x_i for simplicity, but the actual process involves concatenating all hidden features of attributes at position i since music has more than one attribute.

The decoding function is formally represented as:

$$\mathbf{s}_t = \text{DECODER}(\mathbf{s}_{t-1}, E(y_{t-1})), \quad (7)$$

$$p(y_t|y_{t-1}, y_{t-2}, \ldots, y_1) = \text{softmax}(\text{FC}(\mathbf{s}_t)), \quad (8)$$

where $\mathbf{s}_t$ is the hidden state at time t and $\text{FC}(\cdot)$ is a fully connected layer. It should be noted that the initial hidden state $\mathbf{s}_0$ corresponds to the hidden representation of the input $\mathbf{H_X}$.

The compositional encoder and decoder can be trained as follows:

$$\mathcal{L}_0(\theta) = -log p_\theta(\mathbf{Y}|\mathbf{X}) = -\sum_{t=1}^{L_2} log(p_\theta(y_t|y_{<t}, \mathbf{X})). \quad (9)$$

Compositional Knowledge Enhancement

To effectively extract valuable information from knowledge, we develop a method for filtering and identifying relevant data. In the context of music theory, the phenomenon of musical call-and-response requires intelligibility, which leads to specific correlations between the call-and-response concerning rhythm, melody, and harmony (Schoenberg, Stein, and Strang 1967; Kachulis 2004).

We propose drawing insights from similar compositions in knowledge $\mathcal{K}$. The underlying assumption is that: if a previous phrase aligns closely with the input call, its subsequent music phrase can guide the generation of an appropriate response. The premise for this assumption lies in observing the patterns of music composition: if a phrase bears similarity to the "call" in other musical pieces, it is plausible that composers have employed similar techniques in crafting the following phrases. This pattern can serve as a learning basis for the model, enabling it to generate corresponding responses.

To ensure this, we define three selection criteria grounded in music theory, focusing on rhythm, harmony, and melody (Schoenberg, Stein, and Strang 1967; Kachulis 2004):

1. Rhythmic selection: Music phrases should possess a rhythmic pattern similar to the original input call.

2. Harmonic selection: Music phrases should display a chord progression similar to the original input call.
3. Melodic selection: Music phrases should showcase a melodic pattern similar to the original input call.

We extract the rhythm information of the input call and compare it with the rhythm information of music in knowledge $\mathcal{K}$. The comparison is executed by calculating the similarity between the rhythm of the input call and all phrases from $\mathcal{K}$ using Equation 1. We denote the similarity after rhythmic selection as $Rh_{\mathbf{X}} \in \mathbb{R}^N$, where N is the number of music pieces in $\mathcal{K}$ and the similarity calculation follows calculation described in Objective Evaluation Metrics. Similarly, we can extract the melody and harmony of the input call $\mathbf{X}$ and calculate the similarity, denoting them as $Me_{\mathbf{X}} \in \mathbb{R}^N$ and $Ha_{\mathbf{X}} \in \mathbb{R}^N$. In addition, we define the selection score as the summation of rhythm, melody, and harmony:

$$\text{Selection Score} = Rh_{\mathbf{X}} + Me_{\mathbf{X}} + Ha_{\mathbf{X}} \tag{10}$$

Next, we sort all music in knowledge in descending order based on their selection score. We select the top K elements from the knowledge and construct a music candidate list from their subsequent phrases $\mathbf{C} = \{\mathbf{C}_1,\ \mathbf{C}_2, ..., \mathbf{C}_K\}$, where $\mathbf{C}_k \in \mathcal{K}$.

For K music candidates, the hidden representation after the compositional encoder can be denoted as $\{\mathbf{H}_{\mathbf{C}_1}, \mathbf{H}_{\mathbf{C}_2}, \ldots, \mathbf{H}_{\mathbf{C}_K}\}$, following Equation 6. Given the input hidden representation $\mathbf{H}_{\mathbf{X}}$ and hidden representations from candidates, the knowledge incorporation module calculates the similarity between music candidates and the input music:

$$\mathbf{S}_k = \frac{FC(\mathbf{H}_{\mathbf{X}}) \cdot FC(\mathbf{H}_{\mathbf{C}_k})^T}{||FC(\mathbf{H}_{\mathbf{X}})|| * ||FC(\mathbf{H}_{\mathbf{C}_k})||^T}, \tag{11}$$

where $\cdot$ represents the dot product along the dimension of hidden space H, $*$ indicates element-wise multiplication, and $||\cdot||$ denotes the L2 norm.

Then, it incorporates the knowledge from music candidates into the input based on similarity. The final hidden representation fed into the decoder at time $t = 0$ is as follows:

$$\mathbf{H} = \mathbf{H}_{\mathbf{X}} + \sum_{k=1}^{K} \mathbf{H}_{\mathbf{C}_k} * (\mathbf{S}_k \cdot \mathbf{1}^\top), \tag{12}$$

where $\mathbf{1}$ represents a vector of ones with length H. Finally, $\mathbf{H}$ is fed into the compositional decoder to generate the output response $\mathbf{Y}$.

Knowledge-enhanced Response Generation

We first design two different objectives to train the model. Objective 1 is designed as a knowledge-enhanced objective, where the model is trained on $(\{\mathbf{C}_1, \ldots, \mathbf{C}_K, \mathbf{X}\}, \mathbf{Y})$ training examples. The objective function is:

$$\mathcal{L}_1(\theta) = -log p_\theta(\mathbf{Y}|\mathbf{X}, \mathcal{K}) = \\ -\sum_{t=1}^{L_2} log(p_\theta(y_t|y_{<t}, \mathbf{X}, \{\mathbf{C}_1, \ldots, \mathbf{C}_K\})), \tag{13}$$

where $\mathcal{K}$ represents knowledge, and $\mathbf{C}_i$ can be selected through knowledge selection.

Objective 2 is designed as an autoencoder-based knowledge-enhanced objective, trained on $(\{\mathbf{C}_1, \ldots, \mathbf{C}_K, \mathbf{X}\}, \mathbf{C}_i)$ training examples. The goal is to predict music candidates instead of the ground-truth response. The objective function is:

$$\mathcal{L}_2(\theta) = -\frac{1}{K} * \sum_{i=1}^{K} log p_\theta(\mathbf{C}_i|\mathbf{X}, \mathcal{K}) \\ = -\frac{1}{K} * \sum_{i=1}^{K} \sum_{t=1}^{L_2} log(p_\theta(c_{i,t}|c_{i,<t}, \mathbf{X}, \{\mathbf{C}_1, \ldots, \mathbf{C}_K\})). \tag{14}$$

The model is optimized based on Objectives 1 and 2 on call-response pairs. The goal can be expressed as follows:

$$\mathcal{L} = \lambda \mathcal{L}_1 + (1 - \lambda)\mathcal{L}_2, \tag{15}$$

where λ is the hyper-parameter.

During generation, the input call X is processed by CRG to identify K music candidates. The model then sequentially produces tokens until reaching the $\langle$EOM$\rangle$ (end of music) token. By utilizing these candidates, CRG ensures the response mirrors the rhythmic, melodic, and harmonic traits of the input call, offering diverse yet coherent musical outputs.

Experiments

Implementation Details

In our experiment, we convert symbolic music into a sequence of compound words using a predefined music attribute vocabulary (Hsiao et al. 2021). Each sequence is set to a length of 256 tokens. For knowledge selection, we employ DTW and MCS to calculate the similarity between input call and knowledge and normalize DTW between 0 and 1 by dividing the maximum possible DTW value for each pair of sequences. We use the Pop1K7 dataset (Hsiao et al. 2021) as music knowledge. During training, we set the hyper-parameters λ to 0.5. Additionally, we reserve call-response pairs from 40 songs for validation and testing purposes. This experimental setup enables us to evaluate the model's performance and its ability to generate high-quality musical responses.

Subjective Metrics and Participants

We carry out a subjective evaluation on social media. The listener evaluates the music in terms of quality, groove, coherence, and creativity (Zhang et al. 2022; Hu et al. 2022): **Overall Quality**: Please rate the overall quality of the response on a scale of 1 (lowest) to 5 (highest). **Groove**: Please rate the fluency of the response and the appropriateness of pauses on a scale of 1 (lowest) to 5 (highest). **Coherence**: Please rate the coherence of the call-and-response connection on a scale of 1 (lowest) to 5 (highest). **Creativity**: Please rate the creativity of the different responses on a scale of 1 (lowest) to 5 (highest). We collected 89 valid listener reports by distributing a survey on social media. A detailed subjective experiment design is described in Appendix A2.

Model	Objective						Subjective			
	RQ ↑	MQ ↑	HQ ↑	1-gram ↑	2-gram ↑	3-gram ↑	Quality ↑	Groove ↑	Coherence ↑	Creativity ↑
MT (Huang et al. 2018)	0.30	0.26	0.52	0.26	0.66	0.87	2.82	2.99	2.19	2.28
CPT (Hsiao et al. 2021)	0.29	0.29	0.61	0.19	0.62	0.87	2.61	2.33	2.09	2.09
HAT (Zhang et al. 2022)	0.41	0.29	0.57	0.24	0.65	0.87	2.79	3.08	1.90	2.66
CRG	**0.59**†	**0.57**†	**0.72**†	**0.31**†	**0.77**†	**0.92**†	**3.62**†	**3.76**†	**3.48**†	**3.91**†

Table 1: Model comparison: results of the objective and subjective evaluations. † indicates statistically significant improvement over MT.

Model	Loss			Objective						Subjective			
	$\mathcal{L}_0$	$\mathcal{L}_1$	$\mathcal{L}_2$	RQ ↑	MQ ↑	HQ ↑	1-gram ↑	2-gram ↑	3-gram ↑	Quality ↑	Groove ↑	Coherence ↑	Creativity ↑
CRG-Base	✓	✗	✗	0.44	0.29	0.57	0.24	0.66	0.87	3.01	2.75	2.90	2.75
CRG-K	✗	✓	✗	0.48	0.57	0.62	0.26	0.73	0.91	**3.68**	3.59	3.42	2.99
CRG-KM1	✓	✓	✗	0.57	0.55	0.71	0.30	0.75	0.91	3.58	3.60	3.45	2.96
CRG-KM2	✓	✗	✓	0.38	0.54	0.50	0.21	0.65	0.87	3.31	2.89	3.39	3.76
CRG-KM12	✓	✓	✓	**0.59**†	**0.57**†	**0.72**†	**0.31**†	**0.77**†	**0.92**†	3.62†	**3.76**†	**3.48**†	**3.91**†

Table 2: Ablation study: results of the objective and subjective evaluations. † indicates statistically significant improvement over CRG-Base.

Model Comparison

We compared our model's performance with three prominent automatic music composition models: **Music Transformer (MT)** (Huang et al. 2018), **CP-Transformer (CPT)** (Hsiao et al. 2021), and **HAT** (Zhang et al. 2022). Although none specifically target call-and-response, they represent varied approaches in the field.

Table 1 highlights the superior performance of our model in both objective and subjective experiments. This improved quality stems from our unique approach of selecting and incorporating music candidates to enhance the learning process, which provides additional and useful information for response generation. Furthermore, the diverse music candidates from our knowledge base amplify creativity. In contrast, other models rely on a standard reconstruction setting, learning solely from input music without external references, restricting their creativity. The coherence score further illustrates that these models often fail to produce responses that effectively echo the call, hindering artistic creativity in machine-generated music.

Ablation Analysis

We designed five model variants to assess our approach's components: **CRG-Base** uses Transformer settings, trained on call-response pairs with $\mathcal{L}_0$. **CRG-K**: Our main model with a knowledge mechanism trained on $\mathcal{L}_1$. **CRG-KM1**: Incorporates the knowledge mechanism, trained with $\mathcal{L}_0$ and then on $\mathcal{L}_1$. **CRG-KM2**: Uses the knowledge mechanism, trained on $\mathcal{L}_2$. **CRG-KM12**: Combines the knowledge mechanism and trains on both $\mathcal{L}_1$ and $\mathcal{L}_2$.

Table 2 indicates that the integration of knowledge and the utilization of multiple objectives enhance performance. This observation corroborates that the inclusion of musical knowledge improves model performance by providing additional valuable information. Incorporating $\mathcal{L}_0$ aids the model in learning the musical content within the sequence, while $\mathcal{L}_2$ guides the model to generate creative responses.

Knowledge Analysis

Knowledge Candidate Type Different types of knowledge can impact model performance, as they provide varying information to the model. In our experiment, we analyzed three knowledge types: **Random knowledge**: Candidates are randomly chosen. **Distinct knowledge**: Candidates have unique composition skills compared to the input call. **Similar knowledge**: Candidates' skills resemble input calls.

Tables 3 and 4 show that while all knowledge types improve performance, similar knowledge offers the most significant boost. This finding is consistent with music theory emphasizing the connection between calls and responses.

Type	Music Quality		
	RQ ↑	MQ ↑	HQ ↑
Random	0.55±0.19	0.54±0.39	0.71±0.22
Distinct	0.56±0.19	0.52±0.39	0.71±0.22
Similar	**0.57±0.20**	**0.57±0.39**	**0.72±0.22**

Table 3: Objective evaluation on different knowledge types: music quality.

Knowledge Candidate Number Figure 5 displays the music quality and diversity as the number of knowledge in-

Type	Music Diversity		
	1-gram ↑	2-gram ↑	3-gram ↑
Random	0.29±0.11	0.72±0.12	0.89±0.08
Distinct	0.30±0.10	0.73±0.11	0.90±0.07
Similar	**0.31±0.11**	**0.76±0.11**	**0.92±0.06**

Table 4: Objective evaluation on different knowledge types: music diversity.

creases. A trade-off between music diversity and music quality is observed when increasing the number of knowledge candidates.

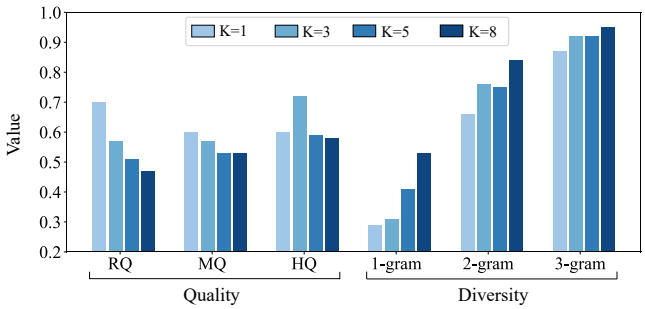


Figure 5: Objective evaluation on different knowledge candidate numbers: music quality and diversity.

Comparison with Human-Composed Music

A comparison between human-composed music and machine-composed music from CRG is shown in Figure 6. Although human-composed music received higher ratings than the proposed model in terms of quality, groove, and coherence, the subjects perceived greater creativity in the music generated by CRG. Figure 7 demonstrates the adaptability of CRG in generating a variety of outcomes tailored to a specific call. In a typical pop song, the human composer generally employs 1-3 response effects, and in this example, the human version only has one response type to maintain consistency and reinforce the main theme. The CRG can produce five different and creative response effects that have never been heard before. This characteristic is not meant to suggest that it generates music with greater overall creativity than human composers; rather, it highlights its ability to introduce creativity within the context of a single piece. This capacity for generating diverse and creative responses can serve as a valuable tool for arrangement or style transfer and has the potential to inspire and complement human creativity during the composition process. By offering different effects, the CRG serves as a valuable resource for composers seeking fresh ideas or novel perspectives while crafting their musical pieces.

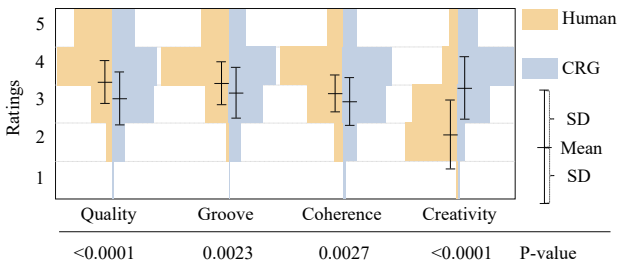


Figure 6: Subjective evaluation between human-composed music and machine-composed music (CRG).

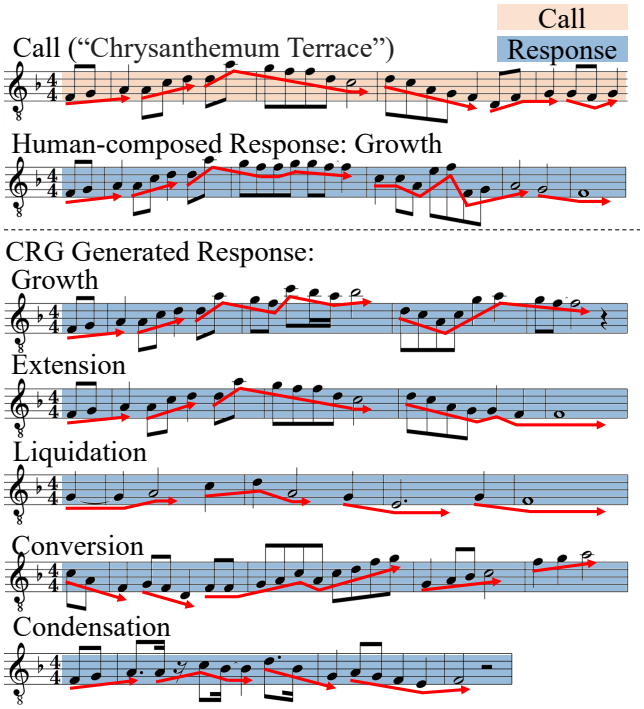


Figure 7: Demonstration of responses.

Conclusion

To the best of our knowledge, this is the first work that employs the call-and-response technique to improve computational creativity in music. We introduce a new dataset consisting of 19,155 call-response pairs. In addition, we define three objective metrics for musical evaluation. Based on this dataset, we propose a novel CRG with a knowledge-enhanced mechanism that incorporates more creative training data, in addition to the ground-truth call-response labels. Compared to existing learning-based models, our proposed model has significantly improved the quality of machine-composed music in terms of rhythm, melody, and harmony. In addition, our model shows promising results in generating creative responses given a call. In the future, we plan to explore bias in evaluation methods (Déguernel and Sturm 2023) and enhance our approach by including the control process of response effects. The code and dataset are available at https://github.com/hu-music/Call-Response.

Acknowledgements

This work is supported by The Hong Kong Polytechnic University for the Project: P0033738 Artificial Intelligence and Robotics (AIR) Group: Artifical Intelligence Art. The author also would like to thank group members who helped with proofreading the manuscript.

References

Agostinelli, A.; Denk, T. I.; Borsos, Z.; Engel, J.; Verzetti, M.; Caillon, A.; Huang, Q.; Jansen, A.; Roberts, A.; Tagliasacchi, M.; et al. 2023. Musiclm: Generating music from text. *arXiv preprint arXiv:2301.11325*.

Akama, T. 2019. Controlling Symbolic Music Generation based on Concept Learning from Domain Knowledge. In *Proceedings of the 20th International Society for Music Information Retrieval Conference*, 816–823.

Benetatos, C.; VanderStel, J.; and Duan, Z. 2020. BachDuet: A Deep Learning System for Human-Machine Counterpoint Improvisation. In *20th International Conference on New Interfaces for Musical Expression, NIME*, 635–640.

Benward, B. 2018. *Music in Theory and Practice vol. 2*.

Biles, J.; et al. 1994. GenJam: A genetic algorithm for generating jazz solos. In *ICMC*, volume 94, 131–137.

Bretan, M.; Oore, S.; Engel, J.; Eck, D.; and Heck, L. 2017. Deep music: Towards musical dialogue. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 5081–5082.

Briot, J.-P.; Hadjeres, G.; and Pachet, F.-D. 2020. *Deep learning techniques for music generation*, volume 1.

Carnovalini, F.; and Rodà, A. 2020. Computational creativity and music generation systems: An introduction to the state of the art. *Frontiers in Artificial Intelligence*, 3: 14.

Chikkamath, S.; and Nirmala, S. 2021. Melody generation using LSTM and BI-LSTM Network. In *2021 International Conference on Computational Intelligence and Computing Applications (ICCICA)*, 1–6.

Collins, N. 2012. Automatic composition of electroacoustic art music utilizing machine listening. *Computer Music Journal*, 36(3): 8–23.

Dai, S.; Jin, Z.; Gomes, C.; and Dannenberg, R. B. 2021. Controllable deep melody generation via hierarchical music structure representation.

Dai, S.; Ma, X.; Wang, Y.; and Dannenberg, R. B. 2023. Personalised popular music generation using imitation and structure. *Journal of New Music Research*, 1–17.

de Rosa, G. H.; and Papa, J. P. 2021. A survey on text generation using generative adversarial networks. *Pattern Recognition*, 119: 108098.

Déguernel, K.; and Sturm, B. L. 2023. Bias in Favour or Against Computational Creativity: A Survey and Reflection on the Importance of Socio-cultural Context in its Evaluation. In *International Conference on Computational Creativity*.

Dhariwal, P.; Jun, H.; Payne, C.; Kim, J. W.; Radford, A.; and Sutskever, I. 2020. Jukebox: A generative model for music. *arXiv preprint arXiv:2005.00341*.

Dong, H.-W.; Hsiao, W.-Y.; Yang, L.-C.; and Yang, Y.-H. 2018. Musegan: Multi-track sequential generative adversarial networks for symbolic music generation and accompaniment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 34–41.

Engels, S.; Tong, T.; and Chan, F. 2015. Automatic real-time music generation for games. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, volume 11, 220–222.

Ens, J.; and Pasquier, P. 2020. Mmm: Exploring conditional multi-track music generation with the transformer. *arXiv preprint arXiv:2008.06048*.

Ferreira, L. N.; Mou, L.; Whitehead, J.; and Lelis, L. H. 2022. Controlling perceived emotion in symbolic music generation with monte carlo tree search. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, volume 18, 163–170.

Guo, X.; Xu, H.; and Xu, K. 2022. Fine-Tuning Music Generation with Reinforcement Learning Based on Transformer. In *2022 IEEE 16th International Conference on Anti-counterfeiting, Security, and Identification (ASID)*, 1–5.

Hernandez-Olivan, C.; and Beltran, J. R. 2022. Music composition with deep learning: A review. *Advances in Speech and Music Technology: Computational Aspects and Applications*, 25–50.

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*, 178–186.

Hu, Z.; Liu, Y.; Chen, G.; Zhong, S.-h.; and Zhang, A. 2020. Make your favorite music curative: Music style transfer for anxiety reduction. In *Proceedings of the 28th ACM International Conference on Multimedia*, 1189–1197.

Hu, Z.; Ma, X.; Liu, Y.; Chen, G.; and Liu, Y. 2022. The Beauty of Repetition in Machine Composition Scenarios. In *Proceedings of the 30th ACM International Conference on Multimedia*, 1223–1231.

Huang, C.-Z. A.; Vaswani, A.; Uszkoreit, J.; Shazeer, N.; Simon, I.; Hawthorne, C.; Dai, A. M.; Hoffman, M. D.; Dinculescu, M.; and Eck, D. 2018. Music transformer. *arXiv preprint arXiv:1809.04281*.

Huang, Y.-S.; and Yang, Y.-H. 2020. Pop music transformer: Beat-based modeling and generation of expressive pop piano compositions. In *Proceedings of the 28th ACM International Conference on Multimedia*, 1180–1188.

Ji, S.; Luo, J.; and Yang, X. 2020. A comprehensive survey on deep music generation: Multi-level representations, algorithms, evaluations, and future directions. *arXiv preprint arXiv:2011.06801*.

Jiang, J.; Xia, G. G.; Carlton, D. B.; Anderson, C. N.; and Miyakawa, R. H. 2020a. Transformer vae: A hierarchical model for structure-aware and interpretable music representation learning. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 516–520.

Jiang, N.; Jin, S.; Duan, Z.; and Zhang, C. 2020b. Rl-duet: Online music accompaniment generation using deep reinforcement learning. In *Proceedings of the AAAI conference on artificial intelligence*, 710–718.

Kachulis, J. 2004. *The Songwriter's Workshop: Harmony*.

Keerti, G.; Vaishnavi, A.; Mukherjee, P.; Vidya, A. S.; Sreenithya, G. S.; and Nayab, D. 2022. Attentional networks for music generation. *Multimedia Tools and Applications*, 81(4): 5179–5189.

Li, B.; Zhao, Y.; Zhelun, S.; and Sheng, L. 2022. Danceformer: Music conditioned 3d dance generation with parametric motion transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 1272–1279.

Lu, P.; Tan, X.; Yu, B.; Qin, T.; Zhao, S.; and Liu, T.-Y. 2022. MeloForm: Generating melody with musical form based on expert systems and neural networks. *arXiv preprint arXiv:2208.14345*.

Mao, H. H.; Shin, T.; and Cottrell, G. 2018. DeepJ: Style-specific music generation. In *2018 IEEE 12th International Conference on Semantic Computing (ICSC)*, 377–382.

Mateja, D.; and Heinzl, A. 2021. Towards machine learning as an enabler of computational creativity. *IEEE Transactions on Artificial Intelligence*, 2(6): 460–475.

Medeot, G.; Cherla, S.; Kosta, K.; McVicar, M.; Abdallah, S.; Selvi, M.; Newton-Rex, E.; and Webster, K. 2018. StructureNet: Inducing Structure in Generated Melodies. In *Proceedings of the 19th International Society for Music Information Retrieval Conference*, 725–731.

Muhamed, A.; Li, L.; Shi, X.; Yaddanapudi, S.; Chi, W.; Jackson, D.; Suresh, R.; Lipton, Z. C.; and Smola, A. J. 2021. Symbolic music generation with Transformer-GANs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 408–417.

Payne, C. 2019. MuseNet. *OpenAI Blog*, 3.

Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; and Sutskever, I. 2021. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, 8821–8831. PMLR.

Roberts, A.; Engel, J.; Raffel, C.; Hawthorne, C.; and Eck, D. 2018. A hierarchical latent vector model for learning long-term structure in music. In *International conference on machine learning*, 4364–4373.

Schoenberg, A.; Stein, L.; and Strang, G. 1967. *Fundamentals of musical composition*.

Shih, Y.-J.; Wu, S.-L.; Zalkow, F.; Muller, M.; and Yang, Y.-H. 2022. Theme Transformer: Symbolic Music Generation with Theme-Conditioned Transformer. *IEEE Transactions on Multimedia*, 1–1.

Titon, J. T. 2016. *Worlds of music: an introduction to the music of the world's peoples*.

Wang, Z.; Chen, K.; Jiang, J.; Zhang, Y.; Xu, M.; Dai, S.; and Xia, G. 2020. POP909: A Pop-Song Dataset for Music Arrangement Generation. In *Proceedings of the 21th International Society for Music Information Retrieval Conference, ISMIR 2020, Montreal, Canada, October 11-16, 2020*, 38–45.

Wesseldijk, L. W.; Mosing, M. A.; and Ullén, F. 2021. Why is an early start of training related to musical skills in adulthood? A genetically informative study. *Psychological Science*, 32(1): 3–13.

Wu, J.; Liu, X.; Hu, X.; and Zhu, J. 2020. PopMNet: Generating structured pop music melodies using neural networks. *Artificial Intelligence*, 286: 103303.

Wu, S.; and Yang, Y. 2020. The Jazz Transformer on the Front Line: Exploring the Shortcomings of AI-composed Music through Quantitative Measures. In *Proceedings of the 21st International Society for Music Information Retrieval Conference*, 142–149.

Yang, L.; Chou, S.; and Yang, Y. 2017. MidiNet: A Convolutional Generative Adversarial Network for Symbolic-Domain Music Generation. In *Proceedings of the 18th International Society for Music Information Retrieval Conference*, 324–331.

Zhang, X.; Zhang, J.; Qiu, Y.; Wang, L.; and Zhou, J. 2022. Structure-enhanced pop music generation via harmony-aware learning. In *Proceedings of the 30th ACM International Conference on Multimedia*, 1204–1213.

Zou, Y.; Zou, P.; Zhao, Y.; Zhang, K.; Zhang, R.; and Wang, X. 2022. MELONS: generating melody with long-term structure using transformers and structure graph. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 191–195.