MuseCoco: Generating Symbolic Music from Text

MuseCoco: 从文本生成象征性音乐

‡Microsoft Research Asia\

‡微软亚洲研究院

Peiling Lu‡∗, Xin XuZ∗, Chenfei Kang\∗, Botao Yu^∗, Chengyi Xing]∗, Xu Tan‡†, Jiang Bian‡Z^Zhejiang University,Shanghai Jiao Tong University Nanjing University,]Dalian University of Technology

吕佩玲ǂ *，辛旭 *，陈飞康 *，博涛宇 *，程毅星] *，徐坦ǂ †，江边ǂ z ^ 浙江大学，上海交通大学南京大学，] 大连理工大学

{peil, xuta, jiabia}@microsoft.com

{ peil，xuta，jiabia }@microsoft.com
xxucs@zju.edu.cn, chenfeikang314@gmail.com
Xxucs@zju.edu.cn，chenfeikang314@gmail.com
btyu@foxmail.com, xcyhbp@mail.dlut.edu.cn
Btyu@foxmail.com，xcyhbp@mail.dlut.edu.cn
https://github.com/microsoft/muzic
Https://github.com/microsoft/muzic

Abstract

摘要

Generating music from text descriptions is a user-friendly mode since the text is a relatively easy interface for user engagement. While some approaches utilize texts to control music audio generation, editing musical elements in generated audio is challenging for users. In contrast, symbolic music offers ease of editing, making it more accessible for users to manipulate specific musical elements. In this paper, we propose MuseCoco, which generates symbolic music from text descriptions with musical attributes as the bridge to break down the task into text-to-attribute understanding and attribute-to-music generation stages. MuseCoCo stands for Music Composition Copilot that empowers musicians to generate music directly from given text descriptions, offering a significant improvement in efficiency compared to creating music entirely from scratch. The system has two main advantages: Firstly, it is data efficient. In the attribute-to-music generation stage, the attributes can be directly extracted from music sequences, making the model training self-supervised. In the text-to-attribute understanding stage, the text is synthesized and refined by ChatGPT based on the defined attribute templates. Secondly, the system can achieve precise control with specific attributes in text descriptions and offers multiple control options through attribute-conditioned or text-conditioned approaches. MuseCoco outperforms baseline systems in terms of musicality, controllability, and overall score by at least 1.27, 1.08, and 1.32respectively. Besides, there is a notable enhancement of about 20% in objective control accuracy. In addition, we have developed a robust large-scale model with 1.2 billion parameters, showcasing exceptional controllability and musicality. Music samples generated by MuseCoco are available via this link 1, and the code is available at this link 2.

从文本描述生成音乐是一种用户友好的模式，因为文本是一个相对简单的用户参与界面。虽然一些方法利用文本来控制音乐音频的生成，但是编辑生成音频中的音乐元素对用户来说是一个挑战。相比之下，象征性音乐提供了易于编辑的功能，使用户更容易操作特定的音乐元素。在本文中，我们提出了 MuseCoco，它从文本描述生成符号音乐，

以音乐属性为桥梁，将任务分解为文本到属性的理解和属性到音乐的生成两个阶段。MuseCoCo 是音乐创作 Copilot 的缩写，它允许音乐家直接根据给定的文本描述创作音乐，与完全从零开始创作音乐相比，效率有了显著提高。这个系统有两个主要优势： 首先，它是数据高效的。在属性到音乐的生成阶段，可以直接从音乐序列中提取属性，使得模型训练具有自监督性。在文本到属性的理解阶段，基于定义的属性模板，通过 ChatGPT 对文本进行合成和细化。其次，系统可以实现对文本描述中特定属性的精确控制，并通过属性条件化或文本条件化的方式提供多种控制选项。MuseCoco 在音乐性、可控性和总分方面分别至少优于基线系统 1.27、1.08 和 1.32。此外，在客观控制准确性方面也有显著提高，约为 20% 。此外，我们还开发了一个具有 12 亿个参数的鲁棒大规模模型，展示了卓越的可控性和音乐性。MuseCoco 生成的音乐样本可以通过链接 1 获得，代码可以在链接 2 获得。

1 Introduction

1 简介

Text-to-music generation is an important task in automatic music generation because it allows users to generate music more easily and intuitively, using natural language as an interface. This makes it a user-friendly mode of music generation, particularly for those who do not have a background in music theory or composition. Previous work[1–4] generating musical audio from texts faces the

文本到音乐的生成是自动音乐生成中的一个重要任务，因为它允许用户使用自然语言作为界面，更容易和直观地生成音乐。这使得它成为一种用户友好的音乐生成模式，特别是对于那些没有音乐理论或作曲背景的人。以前的工作 [1-4] 从文本生成音乐音频面临

∗These authors contributed equally to this work.

* 这些作者对这项工作做出了同样的贡献。
†Correspondence: Xu Tan, xuta@microsoft.com
†通信： Xu Tan，xuta@microsoft.com
1https://ai-muzic.github.io/musecoco/
1https://ai-muzic. github.io/musecoco/

2https://github.com/microsoft/muzic/musecoco/

2https://github.com/microsoft/muzic/musecoco/

following challenges: 1) Limited adaptability: musical audio generated from texts is less adaptable than symbolic music. Once the audio is produced, it may be difficult to make significant changes to the music without starting the generation process over. 2) Lack of control: generating musical audio from text descriptions lacks control over specific aspects of the music such as tempo, meter, and rhythm since the generation process is not explicitly controlled. However, they can be easily handled by generating symbolic music from given texts. Symbolic music refers to a type of music notation that uses symbols and music language to represent specific music ideas, thus, it is more adaptable. Besides, specific musical attributes can be extracted from symbolic data, which can enable more precise control.

以下挑战： 1) 适应性有限： 由文本产生的音乐音频的适应性不如符号音乐。一旦音频被生成，如果不重新开始生成过程，可能很难对音乐做出重大改变。2) 缺乏控制： 根据文本描述生成音乐音频缺乏对音乐特定方面的控制，例如节奏、拍子和节奏，因为生成过程没有明确的控制。然而，它们可以通过从给定的文本生成象征性的音乐来轻松处理。符号化音乐是指用符号和音乐语言来表达特定音乐思想的一种音乐符号，具有更强的适应性。此外，还可以从符号数据中提取特定的音乐属性，从而实现更精确的控制。

There are some works[5–7] attempt to generate symbolic music from text descriptions, but they also face issues with unnatural text descriptions and poor performance. Limitations exist in certain approaches[5] where the model is restricted to specific textual inputs, hindering its ability to generalize text representations to a more user-friendly format. Besides, some work[5] can only control music generation based on limited music aspects, which limits the model's ability to capture the full range of musical creativity and does not satisfy the requirements of all users. While certain approaches[6, 7] allow for natural language inputs, their control accuracy and musical quality is limited due to the requirement for large amounts of paired text-music data. Mubert3 can produce editable MIDI compositions, but it merely retrieves and combines pre-existing music pieces rather than generating novel ideas, limiting its responsiveness to input prompts.

有一些作品 [5-7] 试图从文本描述生成象征性的音乐，但他们也面临着不自然的文本描述和表现不佳的问题。在某些方法中存在局限性 [5] ，在这些方法中，模型被限制为特定的文本输入，阻碍了它将文本表示泛化为更加用户

友好的格式的能力。此外，[5] 的一些工作只能基于有限的音乐方面来控制音乐的生成，这限制了模型捕捉音乐创意的全方位能力，并不能满足所有用户的需求。虽然某些方法 [6,7] 允许自然语言输入，但是由于需要大量的成对文本 - 音乐数据，它们的控制精度和音乐质量受到限制。Mubert3 可以生成可编辑的 MIDI 作品，但它只是检索和组合已存在的音乐片段，而不是生成新颖的想法，这限制了它对输入提示的响应。

We propose MuseCoco, a system for generating symbolic music from text descriptions by leveraging musical attributes. MuseCoCo, stands for Music Composition Copilot, is a powerful tool that empowers musicians to generate music directly from provided text descriptions. By harnessing the capabilities of MuseCoCo, musicians experience a substantial increase in efficiency, eliminating the need to create music entirely from scratch. Our approach breaks down the task into two stages:text-to-attribute understanding and attribute-to-music generation stage. Musical attributes can be easily extracted from music sequences or obtained from existing attribute-labeled datasets, allowing the model in the attribute-to-music generation stage to be trained in a self-supervised manner. Musical attributes can also be inferred from natural languages in the text-to-attribute understanding stage. To synthesize paired text-to-attribute data, templates are created for each attribute, and a subset of these templates is combined and further refined into a coherent paragraph using ChatGPT's language generation capabilities.

我们提出 MuseCoco，一个利用音乐属性从文本描述生成符号音乐的系统。MuseCoCo 是音乐创作 Copilot 的缩写，是一个强大的工具，可以让音乐家直接从提供的文本描述中生成音乐。通过利用 MuseCoCo 的功能，音乐家们体验到了效率的大幅提升，完全不需要从头开始创作音乐。我们的方法将任务分为两个阶段： 文本到属性的理解和属性到音乐的生成阶段。音乐属性可以很容易地从音乐序列中提取出来，也可以从已有的属性标记数据集中获得，使得在属性 - 音乐生成阶段的模型能够以自监督的方式进行训练。在文本到属性的理解阶段，还可以从自然语言中推断音乐属性。为了合成成对的文本到属性的数据，为每个属性创建模板，并且使用 ChatGPT 的语言生成能力将这些模板的一个子集组合并进一步细化成一个连贯的段落。

With self-supervised training in attribute-to-music generation and supervised learning via the help of synthesized paired data in text-to-attribute understanding, a large amount of symbolic music data can

be leveraged without the need of providing textual prompts manually. This can help improve model performance by increasing data amount and model size simultaneously. Besides, by leveraging various music attributes, explicit control can be achieved in generating music across multiple aspects, enabling fine-grained manipulation and customization of the generated musical output. Moreover, due to the two-stage design, MuseCoco can support multiple ways of controlling. For instance, musicians with a strong knowledge of music can directly input attribute values into the second stage to generate compositions, while users without a musical background can rely on the first-stage model to convert their intuitive textual descriptions into professional attributes. Thus, MuseCoco allows for a more inclusive and adaptable user experience than those systems that directly generate music from text descriptions.

通过属性到音乐生成的自监督训练和文本到属性理解的合成配对数据的监督式学习，可以利用大量的符号化音乐数据，而不需要人工提供文本提示。这有助于通过同时增加数据量和模型规模来提高模型性能。此外，通过利用各种音乐属性，可以实现跨多个方面生成音乐的显式控制，从而可以对生成的音乐输出进行细粒度的操作和定制。此外，由于两阶段设计，MuseCoco 可以支持多种控制方式。例如，具有较强音乐知识的音乐家可以直接将属性值输入第二阶段生成作品，而没有音乐背景的用户可以依靠第一阶段模型将其直观的文本描述转换为专业属性。因此，MuseCoco 比那些直接从文本描述生成音乐的系统提供了更具包容性和适应性的用户体验。

The main contributions of this work are as follows:

这项工作的主要贡献如下：

• We introduce MuseCoco, a system that seamlessly transforms textual input into musically coherent symbolic compositions. This innovative approach empowers musicians and general users from diverse backgrounds and skill levels to create music more efficiently and with better control.

我们介绍 MuseCoco，一个无缝地将文本输入转换成音乐上连贯的符号作品的系统。这种创新的方法使音乐家和来自不同背景和技能水平的普通用户能够更有效地创作音乐，并有更好的控制。

• With this two-stage framework, a large amount of symbolic data can be used without the need for labeled text descriptions. It offers users two engagement options: directly specifying attribute values or leveraging text descriptions to control the music generation process.

有了这个两阶段的框架，大量的符号数据可以被使用，而不需要标签文本描

述。它为用户提供了两种参与选择： 直接指定属性值或利用文本描述来控制音乐生成过程。

• Subjective evaluation results have demonstrated that MuseCoco outperforms baseline systems in terms of musicality, controllability, and overall score, achieving a minimum improvement of 1.27,1.08, and 1.32, respectively. Additionally, there is a significant boost of about 20% in objective control accuracy, further affirming the system's enhanced performance.

・主观评价结果表明，MuseCoco 在音乐性、可控性和总分方面优于基准系统，分别取得了 1.27、1.08 和 1.32 的最小提升。此外，在客观控制精度方面有大约 20% 的显著提升，进一步肯定了系统的增强性能。

• We also extend our model to a large scale with 1.2 billion parameters, which exhibits notable controllability and musicality, further enhancing its performance.

此外，我们将模型扩展至包含 12 亿个参数的大规模模型，显示出显著的可控性和音乐性，进一步提升了模型的性能。

2 Related Work

2 相关工作

2.1 Text-to-Music Generation

2.1 文本到音乐的生成

Text-to-music generation based on deep learning has been an active research area. Though paired text-music data is scarce, there have been many works for generating audio music from input prompts. A database containing music created by musicians and sound designers with three tags(genres, moods and activities) is constructed in MubertAI4. It assigns the closest tags for the input prompt and generates combination of sounds from the database based on these tags. Riffusion5 leverages stable diffusion[8] to obtain images of music spectrograms paired with input text and produce audio from them. Meanwhile, both Moûsai[4] and ERNIE-Music[3] apply diffusion models with self-collected text-audio datasets to achieve text-audio music generation. Recently, Huang et al.[9] create a large audio-text dataset to train a joint embedding model linking music audio and natural language music descriptions, MuLan, which helps address the

absence of text-music paired data. For example, in Noise2Music[2], MuLan assigns captions to unlabeled audio clips to create music-text pairs. By directly using input text representation from MuLan during inference, MusicLM[1] can be trained to output audio music from Mulan audio representation on existing audio data to complete the text-to-music generation task without paired text-audio data. However, the challenge with musical audio generation from texts lies in its limited editability, which poses limitations in the composition process. By contrast, symbolic music, notated as a sequence of musical symbols, can be simply interpreted and manipulated by humans and machines.

基于深度学习的文本到音乐的生成一直是一个活跃的研究领域。虽然成对的文本 - 音乐数据很少，但是已经有很多从输入提示生成音频音乐的工作。Mubertai4 构建了一个数据库，其中包含音乐家和声音设计师创作的带有三个标签 (流派、情绪和活动) 的音乐。它为输入提示分配最接近的标签，并根据这些标签从数据库中生成声音组合。Riffusion5 利用稳定的扩散 [8] 获得与输入文本配对的音乐声谱图的图像，并从中产生音频。同时，Moûsai [4] 和 ERNIE-Music [3] 利用自收集的文本音频数据集应用扩散模型来实现文本音频音乐生成。最近，Huang 等 [9] 创建了一个大型的音频 - 文本数据集来训练一个连接音频和自然语言音乐描述的联合嵌入模型 —— 木兰，该模型有助于解决文本 - 音乐配对数据的缺失问题。例如，在 Noise2Music [2] 中，木兰将标题分配给未标记的音频剪辑以创建音乐 - 文本对。通过在推理过程中直接使用来自木兰的输入文本表示，可以训练 MusicLM [1] 在现有音频数据上从木兰音频表示输出音频音乐，完成无需成对的文本 - 音频数据的文本 - 音乐生成任务。然而，从文本生成音乐音频的挑战在于其有限的可编辑性，这在作曲过程中造成了限制。相比之下，作为音乐符号序列的象征性音乐可以被人类和机器简单地解读和操作。

Only a few works focus on generating symbolic music from text descriptions. BUTTER[5], a music-sentence representation learning framework, proposes music representation disentanglement and cross-modal alignment to firstly generate music presented by ABC notations6 from text including four musical keywords(the key, meter, style, and others). Limited to the folk song datasets and a few musical factors, it cannot generate symbolic music in many varieties. Instructed with musical input prompts, the large language model GPT-4[7] can also generate ABC notation music, however, without any nontrivial form of harmony[10]. Wu and Sun[6] also explores the text-to-music capability of pre-trained language models. Though fine-tuned with more than 200k text and ABC notation music pairs, the model cannot align the musical attribute values in loose text with the

generated music well enough and can only generate music with solo tracks.

只有少数作品注重从文本描述中生成象征性音乐。音乐 - 句子表征学习框架 BUTTER [5] 提出了音乐表征解缠和跨模态对齐，首先从包含四个音乐关键词 (键，节拍，风格等) 的文本中生成由 ABC 符号 6 呈现的音乐。受限于民歌数据集和少数音乐因素，无法生成多种类型的符号音乐。在音乐输入提示的指导下，大型语言模型 GPT-4 [7] 也可以生成 ABC 符号音乐，但是没有任何非平凡的和声形式 [10]。吴和孙 [6] 也探索了预先训练的语言模型的文本到音乐的能力。虽然微调了超过 200k 的文本和 ABC 符号音乐对，该模型不能将松散文本中的音乐属性值与生成的音乐进行足够好的对齐，并且只能生成独奏曲目的音乐。

Previous work directly generates music from text descriptions, which lacks explicit control over the generation process. In this paper, we propose MuseCoco, which can generate symbolic music from text descriptions with high control accuracy and good musicality. The generated music, in its symbolic format, offers easy editability and can be explicitly controlled through attribute values derived from text descriptions.

以前的工作直接从文本描述生成音乐，缺乏对生成过程的明确控制。本文提出了一种基于文本描述的符号化音乐生成算法 MuseCoco，该算法能够从文本描述中生成符号化音乐，具有较高的控制精度和良好的音乐性。生成的符号化音乐具有良好的可编辑性，并且可以通过文本描述的属性值进行显式控制。

2.2 Controllable Music Generation

2.2 可控的音乐生成

Controllable music generation refers to the ability to exert control over specific aspects or char-acteristics of the generated music. By applying conditional generative models(such as condi-tional VAE[11–13], GAN[14, 15] or diffusion models[2, 3]), previous work leverage different conditions to generate music: Some leverage descriptions like emotion[16–18], style[19–21] or structure[22, 23] as conditions to control music generation. The other applies music descriptions such as instrumentation[24, 25], chord[11, 26], note density[12, 13], etc. to generate music.

可控音乐生成是指对生成音乐的特定方面或特征施加控制的能力。通过应用条件生成模型 (如条件 VAE [11-13] ，GAN [14,15] 或扩散模型 [2,3]) ，以前

的工作利用不同的条件来生成音乐： 一些利用情绪 [16-18] ，风格 [19-21] 或结构 [22,23] 作为控制音乐生成的条件。另一种应用音乐描述，如乐器 [24,25]、和弦 [11,26]、音符密度 [12,13] 等来生成音乐。

However, general users usually have multiple requirements for generated music, and previous work with limited control over specific music aspects may result in limited expressiveness and diminished adaptability to different musical contexts. Besides, previous work can only control music generation with specified music attributes, which restricts the model ability to convey complex and nuanced musical ideas or concepts that can be effectively communicated through textual descriptions. Text-

然而，一般用户通常对生成的音乐有多种需求，以前的工作对特定音乐方面的控制有限，可能会导致有限的表现力和对不同音乐环境的适应性降低。此外，以往的工作只能控制具有特定音乐属性的音乐生成，这限制了模型传达能够通过文本描述有效沟通的复杂而微妙的音乐思想或概念的能力。文本

4https://github.com/MubertAI/Mubert-Text-to-Music

4https://github.com/MubertAI/Mubert-Text-to-Music

5https://www.riffusion.com/about

5 http://www.rifffusion.com/about

6https://abcnotation.com/

6 http://abcnotation.com/

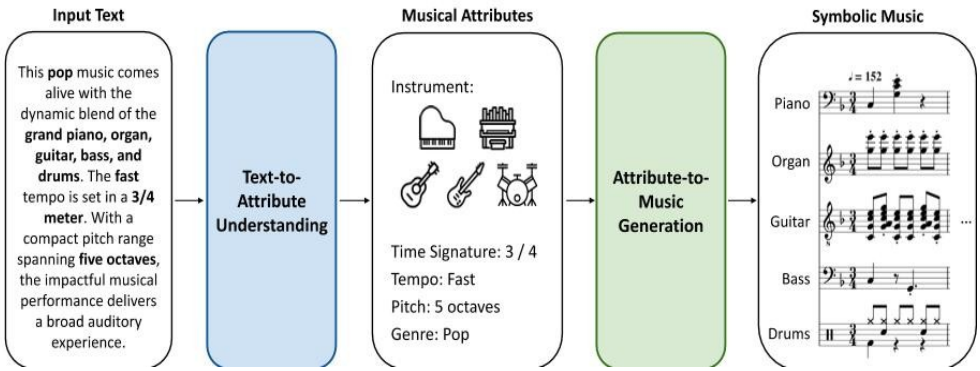


Figure 1: The two-stage framework of MuseCoco. Text-to-attribute understanding extracts diverse musical attributes, based on which symbolic music is generated through the attribute-to-music generation stage.

图 1: MuseCoco 的两阶段框架。文本到属性的理解提取不同的音乐属性，在此基础上，通过属性到音乐的生成阶段生成符号音乐。

based input is easily accessible and familiar to users, making it a practical and widely adopted choice for guiding generative tasks, such as text-to-image[8, 27–29] and text-to-music generation[2, 1, 6, 10]. Previous work has struggled to generate symbolic music directly from textual descriptions provided by users. MuseCoco maps text input to music attributes and then utilizes music attributes to control music generation, which can effectively solve the above problems.

基于输入的用户很容易访问和熟悉，使其成为指导生成任务 (如文本到图像 [8,27-29] 和文本到音乐生成 [2,1,6,10]) 的实用和广泛采用的选择。以前的工作一直在努力从用户提供的文本描述直接生成符号音乐。MuseCoco 将文本输入映射到音乐属性，然后利用音乐属性控制音乐生成，可以有效解决上述问题。

3 MuseCoco

3 MuseCoco

3.1 Overview

3.1 概述

To achieve text-to-music controllable generation, MuseCoco incorporates natural language and symbolic music into a two-stage framework that separates text-to-attribute understanding and attribute-to-music generation, which are trained independently. The pipeline of MuseCoco is shown in Figure 1. In this section, we will elaborate on their technical design and model architectures respectively.

为了实现文本到音乐的可控生成，MuseCoco 将自然语言和符号音乐整合到一个两阶段框架中，将文本到属性的理解和属性到音乐的生成分开，并分别进行训练。MuseCoco 的管道如图 1 所示。在本节中，我们将分别详细介绍它们的技术设计和模型架构。

3.2 Attribute-to-Music Generation

3.2 属性 - 音乐生成

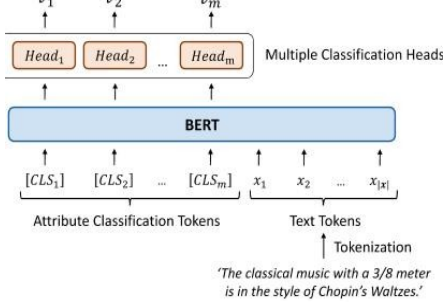
Most musical attributes can be easily obtained by extracting from

music sequences(Section 4.1), so the music generation model in the attribute-to-music stage can be trained in a self-supervised way. This method can leverage large amounts of unlabeled data, making it a highly data-efficient approach. Musical attributes(Table 1) can be classified into objective attributes like the tempo and meter and subjective attributes like the emotion and genre[30]. Objective attributes refer to quantifiable and measurable characteristics of musical elements, so they can be extracted from music sequences with pre-defined rules(please refer to Section 4.1 for more details). Subjective attributes refer to the qualities or characteristics of music that are based on personal interpretation, perception, or emotional response, which can be obtained from existing attribute-labeled datasets. After obtaining attributes from music sequences, we append those attribute tokens as prefix tokens into music sequences to provide explicit control over the music, which makes it easier to interpret and understand how the attributes are influencing the music. Different attribute tokens can be combined or sequenced to achieve complex musical expressions and transformations.

大多数音乐属性可以很容易地从音乐序列中提取出来 (4.1 节) ，因此在属性 - 音乐阶段的音乐生成模型可以通过自监督的方式进行训练。这种方法可以利用大量的未标记数据，使其成为一种高效的数据处理方法。音乐属性 (表 1) 可以分为节奏和节拍等客观属性以及情感和流派等主观属性 [30]。客观属性是指音乐元素的可量化和可测量的特征，因此它们可以通过预定义的规则从音乐序列中提取出来 (详见 4.1 节)。主观属性是指基于个人理解、感知或情感反应的音乐质量或特征，可以从已有的属性标注数据集中获得。在从音乐序列中获取属性后，我们将这些属性标记作为前缀标记添加到音乐序列中，以提供对音乐的显式控制，从而更容易解释和理解属性是如何影响音乐的。不同的属性标记可以通过组合或排序来实现复杂的音乐表达和转换。

Specifically, given dataset{V,Y}, where Y is the symbolic music dataset and V is the set of attribute values, MuseCoco transforms attribute values into prefix tokens to control music generation, as shown in Figure 2(b). Using prefix tokens to guide the generation process is an effective method for directing the output towards a particular direction[31–34]. We pre-define m musical attributes and their values as shown in Appendix A. For each music sequence y=[y1, y2,..., yn] ∈ Y, and its attribute values v={v1, v2,..., vm} ∈ V, we consider the following distribution:

具体而言，给定数据集 {v，y} ，其中 y 是符号音乐数据集，v 是属性

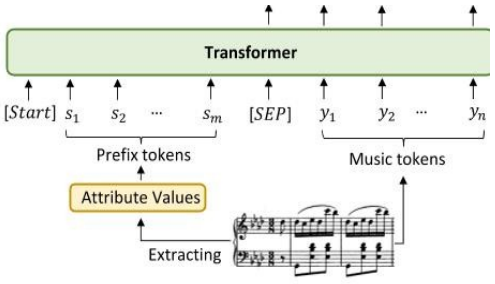


值的集合，MuseCoco 将属性值转换为前缀标记以控制音乐生成，如图 2 (b) 所示。使用前缀标记来指导生成过程是将输出指向特定方向的有效方法 [31-34]。对于每个音乐序列 y = [y1，y2，... ，yn]∈ y，其属性值 v = {v1，v2，... ，vm}∈ v，我们考虑以下分布：

$$p(\boldsymbol{y}|\boldsymbol{v}) = \prod_{i=1}^{n} p(y_i|y_{<i}, v_1, v_2, ..., v_m). \quad (1)$$

¥···[End]
... [结束]

(b) Attribute-to-music generation



(b) 属性到音乐的生成
(a) Text-to-attribute understanding
(a) 文本到属性的理解

Figure 2: During training, each stage is independently trained. During inference, the text-to-attribute understanding stage firstly extracts music attribute values, based on which the attribute-to-music generation stage secondly generates symbolic music.

图 2: 在培训期间，每个阶段都是独立培训的。在推理过程中，文本到属性的理解阶段首先提取音乐属性值，在此基础上，属性到音乐的生成阶段再生成符号音乐。

We encode each attribute value vj into a prefix token sj. Position embeddings for such prefix tokens{sj}mj=1 are not used. Since it is not usual for users to provide all attribute values in real-world scenarios, we introduce a special token, represented by sNAj, to exclude this attribute(i.e. vj) that is not specified in inputs from influencing the music generation process. Then the input sequence is encoded into:

我们将每个属性值 vj 编码为前缀标记 sj。对于这样的前缀标记 {sj} mj = 1 不使用位置嵌入。由于用户在现实场景中不常提供所有属性值，我们引入了一个特殊的令牌，以 sNAj 为代表，以排除在输入中未指定的这个属性 (即 vj) 影响音乐生成过程。然后，输入序列被编码为：

$$s_1, s_2, s_3, ..., s_m, \texttt{[SEP]}, y_1, y_2, ..., y_n. \quad (2)$$

During training, some attribute tokens(e.g., s2, s3) are randomly replaced with special tokens(i.e.,sNAj) to enable adaptation to various attribute combinations. During inference, attribute values are provided directly from inputs. Any attributes that are absent in the inputs are represented by special tokens, which are combined with the other prefix tokens to effectively control the music generation process as required.

在训练过程中，一些属性标记 (如 s2、 s3) 被随机替换为特殊标

记 (如 sNAj) ，以适应不同的属性组合。在推断期间，属性值直接从输入提供。输入中缺少的任何属性都由特殊的标记表示，这些标记与其他前缀标记组合在一起，以便根据需要有效地控制音乐生成过程。

3.3 Text-to-Attribute Understanding

3.3 文本到属性的理解

Control information within input text comprises different musical attributes. Therefore, the text-to-attribute task is required to extract musical attribute values from plain text. These attribute values will be used in the attribute-to-music generation stage to generate desired music. As shown in Figure 2(a), this stage can be denoted as{X,V}, where X is the input text set, V is the value set of m pre-defined musical attributes. In the dataset, each instance x ∈ X is paired with a combination of mattribute values v={vi}mi=1, v ∈ V. Given a pre-trained language model M, BERTlarge[35], x is converted by the tokenizer of M into corresponding tokens{x1, x2,..., x|x|}. To adapt to multipleattribute classification, we prepend m attribute classification tokens[CLSi]mi=1 to input text tokens as m attribute classification heads and the encoded[CLSi] head is used to compute the probabilitydistribution over the i class set with a softmax classifier. Position embeddings of[CLSi]mi=1 are not used to be consistent with the pre-training stage. Specifically, the input x will be encoded to hidden vectors:

输入文本中的控制信息包含不同的音乐属性。因此，需要通过文本到属性的任务从纯文本中提取音乐属性值。这些属性值将用于属性到音乐的生成阶段，以生成所需的音乐。如图 2 (a) 所示，这个阶段可以表示为 {x，v} ，其中 x 是输入文本集，v 是 m 个预定义音乐属性的值集。在数据集中，每个实例 x ∈ x 与一组 matttribute 值 v = {vi} mi = 1，v ∈ v 配对。给定一个预先训练好的语言模型 m，BERTlarge [35] ，x 被 m 的 tokenizer 转换成相应的 token {x1，x2，... ，x | x |}。为了适应多属性分类，我

们预先设置 m 个属性分类标记 [CLSi] mi = 1 来输入文本标记作为 m 个属性分类头，并使用编码的 [CLSi] 头用 softmax 分类器计算 i 类集合上的概率分布。[CLSi] mi = 1 的位置嵌入不用于与预训练阶段一致。具体而言，输入 x 将被编码为隐藏向量：

$$\mathbf{h}_{[\mathrm{CLS}_1]}, \mathbf{h}_{[\mathrm{CLS}_2]}, \ldots, \mathbf{h}_{[\mathrm{CLS}_m]}, \mathbf{h}_{x_1}, \mathbf{h}_{x_2}, \ldots, \mathbf{h}_{x_{|\infty}}$$

Table 1: Musical attribute descriptions.

表 1: 音乐属性描述。

Type 类型	Attribute 属性	Description 描述
Objective 目的	Instrument 乐器	played instruments in the music clip 在音乐剪辑中演奏的乐器
Subjective 主观	Pitch 沥青	the number of octaves covering all pitches in one music clipwhether the piece sounds danceable 一个音乐片段中包含所有音高的八度音数是否可以跳舞
	Rhythm Danceability 节奏舞蹈能力	
	Rhythm Intensity 节奏强度	the intensity of the rhythm 节奏的强

		度
	Bar 酒吧	the total number of bars in one music clip 一个音乐剪辑中的小节总数
	Time Signature 时间签名	the time signature of the music clip 音乐剪辑的时间签名
	Key 钥匙	the tonality of the music clip 音乐片段的调性
	Tempo 节奏 Time 时间： ArtistGenre 艺术家流派 Emotion 情绪	the tempo of the music clip 音乐片段的节奏 the approximate time duration of the music clip 音乐片段的大致持续时间 the artist(style) of the music

		clipthe genre of the music clip 音乐片段的艺术家(风格) 音乐片段的类型 the emotion of the music clip 音乐片段的情感

Table 2: An example of synthesizing a text-attribute pair. We randomly select a template from the available templates for each attribute. Here shows two of each. Then the templates are refined by ChatGPT and then filled in values.

表 2: 合成文本 - 属性对的示例。我们从每个属性的可用模板中随机选择一个模板。下面是每个属性的两个模板。然后通过 ChatGPT 对模板进行细化，并填入值。

Attribute 属性		Value 价值	Template 模板
Key 钥匙	Major 主要	This music is composed in the[KEY] key. 这首曲子是用 [KEY] 键谱写的。	
			This music's use of[KEY] key creates a distinct atmosphere. 这种音乐使用 [KEY] 键创造了一种独特的氛围。
Bar 酒吧	13~16 13 ~ 16	The song is composed of approximately[NUM_BARS] bars. 这首歌大约由 [NUM _ bars] 小节组成。	
			The song comprises[NUM_BARS] bars.

			这首歌由 [NUM _ bars] 小节组成。
Emotion 情绪	Happiness 幸福	The music is imbued with[EMOTION]. 音乐充满了 [情感]。	
	4/4 4/4	The music has a[EMOTION] feeling. 音乐有一种情感。 The[TIME_SIGNATURE] time signature is used in the music. 在音乐中使用 [TIME _ signature] 时间签名。 The music is in[TIME_SIGNATURE]. 音乐在 [TIME _ signature] 中。	
Time Signature 时间签名	Refine via ChatGPT and fill in place- 通过 ChatGPT 提炼并填写 -		The music is imbued with happiness, and the major key in this music 这首音乐充满了快乐，是这首音乐的主调
holders with values: 持有者的价值观：			provides a powerful and memorable sound. The song progresses 提供了一个强有力的和令人难忘的声音
{Key, Bar, Emotion, Time Signature} { Key，Bar，Emotion，Time Signature }			through 13~ 16 bars, with 4/4 as the meter of the music. 通过 13 ~ 16 小节，以 4/4 作为音乐的节拍。

The probability distribution of the value of i-th attribute is pi(vi|x)= Softmax(Wih[CLSi]+ bi), where Wi and bi are learnable parameters. The cross-entropy loss of i-th attribute is

第 i 个属性值的概率分布是 pi (vi | x) = Softmax (with [CLSi] + bi)，其中 Wi 和 bi 是可学习的参数。第 i 个属性的

交叉熵损失是

$$\mathcal{L}_i = -\frac{1}{|\mathcal{X}|}\sum_{\boldsymbol{x}\in\mathcal{X}}\log p_i(v_i|\boldsymbol{x})$$

During training, all learnable parameters are fine-tuned by minimizing the sum of attribute cross-entropy losses L=m i=1Li on{X,V}.

在训练过程中，通过最小化 {x，v} 上的属性交叉熵损失之和 l = mi = 1Li 来微调所有可学习的参数。

3.4 Data Construction

3.4 数据构建

By using musical attributes to break down the text-to-music generation task into the attribute-to-music generation stage and the text-to-attribute understanding stage, we can leverage large amounts of symbolic music data without text descriptions. In the attribute-to-music generation stage, attributes can be extracted from music sequences with rules or obtained from attribute-labeled datasets(detailed in Section 4.1). The system only requires paired data in the text-to-attribute stage, we synthesize these text-attribute pairs in the following steps:

通过使用音乐属性将文本到音乐的生成任务分解为属性到音乐的生成阶段和文本到属性的理解阶段，我们可以利用大量的符号化音乐数据而不需要文本描述。在属性到音乐的生成阶段，可以通过规则从音乐序列中提取属性，或者从属性标记的数据集中获取属性 (详见 4.1 节)。系统在文本到属性阶段只需要成对的数据，我们在以下步骤中合成这些文本 - 属性对：

1. Write templates for each attribute: As shown in Table 2, we write several templates as a set for each attribute, where its values are represented with a placeholder. By utilizing this placeholder, we can accommodate diverse combinations of attribute values without requiring exact values.

1.为每个属性编写模板： 如表 2 所示，我们为每个属性编写几个模板作为一个集合，其中它的值用占位符表示。通过利用这个占位符，我们可以容纳不同的属性值组合，而不需要精确的值。

2. Create attribute combinations and concatenate their templates as paired texts: The generation process is usually controlled by multiple attributes together. Hence, constructing various different combinations of attribute values and paired text is necessary. Because of the long-tail distribution per attribute value in real-world data, to enrich the diversity of paired text-attribute training data and avoid the long-tailed issue, we stochastically create v per instance based on pre-defined musical attributes and their values on our own to ensure the number of instances including vi is balanced, i.e., each value of each attribute is sampled equally. And then, the paired texts of these created combinations are synthesized by simply concatenating their corresponding templates, randomly chosen from template sets of each attribute.

2.创建属性组合并将它们的模板连接成对的文本： 生成过程通常由多个属性共同控制。因此，构建各种不同的属性值和成对文本的组合是必要的。由于实际数据中每个属性值都具有长尾分布，为了丰富成对文本 - 属性训练数据的多样性，避免长尾问题，我们根据预先定义的音乐属性及其值随机创建每个实例的 v，以保证包含 vi 的实例数量是平衡的，即每个属性的每个值都被等量采样。然后，从每个属性的模板集合中随机选取相应的模板串联，合成这些组合的成对文本。

3. Refine concatenated templates via ChatGPT: Since simply concatenated templates are less fluent than real users' input, ChatGPT7 is utilized to refine them as shown in Table 2.

3.通过 ChatGPT 完善连接模板： 由于简单的连接模板不如真实用户的输入流畅，因此使用 chatgpt7 来完善它们，如表 2 所示。

4. Fill in placeholders: Finally, attribute values or their synonyms fill in placeholders, ensuring that the resulting text effectively conveys the intended meaning and maintains a consistent narrative structure.

4.填充占位符： 最后，属性值或它们的同义词填充占位符，确保结果文本有效地传达预期的含义，并保持一致的叙事结构。

Through these steps, we can independently construct the datasets for either of the two stages without the need for paired text-music data.

通过这些步骤，我们可以独立地为两个阶段中的任何一个构建数据集，而不需要成对的文本 - 音乐数据。

4 Experiments

4 个实验

4.1 Experiment Setup

4.1 实验设置

Datasets To train the attribute-to-music generation stage and evaluate our proposed method, we collect an assortment of MIDI datasets from online sources. Ta-ble 3 lists all of the used datasets along with their respec-tive counts of valid MIDI files. Specifically, the MMD dataset[36] consists of many datasets collected from the internet8. The Emotion-gen dataset is generated by our internal emotion-controllable music generation system, and the others are all publicly released datasets. We did the necessary data filtering to remove dupli-cated and poor-quality samples, and there are 947,659MIDI samples remaining. From each MIDI file, we ran-domly extracted 3 clips within 16 bars. The attributes described in Table 1 were then extracted from each clip.

数据集为了训练属性到音乐的生成阶段并评估我们提出的方法，我们从网上收集了一系列 MIDI 数据集。表 3 列出了所有使用的数据集，以及它们各自的有效 MIDI 文件的计数。具体而言，MMD 数据集 [36] 由从互联网收集的许多数据集组成 8。Emotion-gen 数据集是由我们的内部情绪可控音乐生成系统生成的，其他数据集都是公开发布的数据集。我们做了必要的数据过滤去除重复的和低质量的样本，还剩下 947,659 个 MIDI 样本。从每个 MIDI 文件中，我们随机提取了 16 个条内的 3 个剪辑。然后从每个剪辑中提取表 1 中描述的属性。

Table 3: Statistics of the used datasets.

表 3: 所用数据集的统计。

Dataset 数据集	#MIDI 图片来源：
MMD[36] MMD [36]	1,524,557 1,524,557
EMOPIA[16] 艾美利亚 [16]	1,078 1,078
MetaMidi[37] MetaMidi [37]	612,088 612,088
POP909[38] 流行音乐 909	909
Symphony[39] 交响乐 [39]	46,360 46,360 美元
Emotion-gen Emotion-gen 情绪	25,730 25730
Total(after filtering) 总计 (过滤后)	947,659 947,659

The objective attribute values used in the training are directly extracted from MIDI files and the subjective attribute values are obtained from some of the datasets(details in Appendix A).

训练中使用的客观属性值是直接从 MIDI 文件中提取的，而主观属性值是从一些数据集中获得的 (详见附录 a)。

System Configuration For the attribute-to-music generation stage, we use a REMI-like[40] representation method to convert MIDI into token sequences. We apply Linear Transformer[41] as the backbone model, which consists of 16 layers with causal attention and 12 attention heads. The hidden size is 1024 and FFN hidden size is 4096, yielding an approximate parameter count of 203million. The max length of each sample is 5120, covering at most 16-bar music segments. During training, the batch size is 64. The dropout rate is set to 0.1. We use Adam optimizer[42] withβ1= 0.9, β2= 0.98 and e= 10−9. The learning rate is 2 × 10−4 with warm-up step 16000 and an invert-square-root decay. In text-to-attribute understanding, we leverage BERTlarge9 as the backbone model and the max sequence length of it is set to 256, which covers common user input. We use1,125 thousand samples for fine-tuning with a train/valid portion of 8:1. During training, the batch size is 64 and the learning rate is 1 × 10−5.

系统配置对于属性到音乐的生成阶段，我们使用类似 remi 的 [40] 表示方法将 MIDI 转换为令牌序列。我们应用线性变压器 [41] 作为骨干模型，它由 16 层因果关注和 12 个关注头组成。隐藏大小为 1024，FFN 隐藏大小为 4096，产生大约 2.03 亿的参数计数。每个样本的最大长度为 5120，最多覆盖 16 小节音乐片段。在训练期间，批量大小是 64。辍学率设置为 0.1。我们使用 Adam 优化器 [42] ，β1 = 0.9，β2 = 0.98 和 e = 10.9。学习速率为 2 × 10-4，预热步骤为 16000，反平方根衰减。在文本到属性的理解中，我们利用 bertlarge9 作为骨干模型，并将其最大序列长度设置为 256，这涵盖了常见的用户输入。我们使用 1,125,000 个样本进行微调，火车 / 有效部分为 8:1。在训练期间，批量大小为 64，学习率为 1 × 10-5。

Evaluation Dataset and Metrics To evaluate MuseCoco, we construct a standard test set includ-ing 5,000 text-attribute pairs in the same way in Section 3.4. Musical attributes of each test sample originated from real music in the test set of the attribute-to-music stage, instead of creating them on our own to make sure musical rationality and all values of the attributes are covered in the test set for thorough testing. Meanwhile, in order to accord with usual user inputs, we randomly assign the NA value(meaning the attribute is not mentioned in the text) to some attributes per sample to synthesize text prompts with different lengths.

评估数据集和度量为了评估 MuseCoco，我们构建了一个标准测试集，包括 5000 个文本 - 属性对，方法与第 3.4 节相同。在属性 - 音乐阶段的测试集中，每个测试样本的音乐属性来源于真实音乐，而不是自己创建它们，以确保音乐的合理性和所有属性的值都覆盖在测试集中进行彻底测试。同时，为了与通常的用户输入相一致，我们为每个样本的某些属性随机分配 NA 值 (即文本中没有提到的属性) ，以合成不同长度的文本提示。

To evaluate the models objectively, we propose the metric called Average Sample-wise Accuracy(ASA), which is calculated by determining the proportion of correctly predicted attributes in each sample, followed by the calculation of the average prediction accuracy across the whole test set. To conduct a subjective evaluation of MuseCoco's performance, we employ a user study. We invite individuals with musical backgrounds to fill out the questionnaires(details in Appendix B.1) Participants are asked to rate the following metrics on a scale of 1(lowest) to 5(highest):

为了客观地评价模型，提出了平均样本精度 (ASA) 的度量方法，通过确定每个样本中正确预测属性的比例，然后计算整个测试集的平均预测精度。为了对 MuseCoco 的表现进行主观评估，我们采用了用户研究。我们邀请有音乐背景的个人填写调查问卷 (详见附录 B.1)。参与者被要求在 1 (最低) 到 5 (最高) 的范围内对以下指标进行评分：

• Musicality: This metric assesses the degree to which the generated music exhibits qualities akin to the artistry of a human composer.

音乐性： 这个指标评估生成的音乐在多大程度上表现出类似于人类作

曲家的艺术性。

• Controllability: This metric measures how well the samples adhere to the musical attribute values specified in the text descriptions.

可控性： 这个指标衡量样本对文本描述中指定的音乐属性值的坚持程度。

• Overall: This metric quantifies the overall quality of this generated music considering both its musicality and controllability.

总体： 这个指标量化了生成音乐的整体质量，同时考虑了它的音乐性和可控性。

8We obtained the dataset for this work with the help of the authors, as it was not publicly available.9https://huggingface.co/bert-large-uncased

8 我们在作者的帮助下获得了这项工作的数据集，因为它还没有公开。9 https://huggingface.co/bert-large-uncase

4.2 Comparison with Baselines

4.2 与基线的比较

Baselines In this study, we compare our method to two existing works for generating symbolic music from 21 text descriptions randomly selected from the standard test set:

基线在这项研究中，我们比较了我们的方法和两个现有的从标准测试集中随机选择的 21 个文本描述生成符号音乐的工作：

• GPT-4: GPT-4[7] is a large-scale language model that demonstrated its capabilities in various domains, including music. Following Bubeck et al.[10], we instruct GPT-4 to generate ABC notation music with the task-specific prompts(in Appendix B.5) using the official web page10manually.

GPT-4: GPT-4 [7] 是一个大规模的语言模型，展示了它在包括音乐在内的各个领域的能力。继 Bubeck 等 [10] 之后，我们指示 gpt-4 使用官方网页 10 手动生成具有任务特定提示 (在附录 B.5 中) 的 ABC 符号音乐。

• BART-base: Wu and Sun[6] release a language-music BART-base11, which shows a solid perfor-mance. Text descriptions are fed into this model and guide it to generate ABC notation music for comparison.

・BART-base: Wu 和 Sun [6] 发布了一个语言音乐 BART-base11，表现出良好的性能。文本描述被输入到这个模型中，并引导它生成用于比较的 ABC 符号音乐。

For a fair comparison, we convert ABC notation music generated by baselines into MIDI music using music2112. As for the subjective evaluation, well-designed questionnaires including generated music from baselines and our method are distributed to individuals, who are all in music backgrounds and required to score the subjective metrics described in Section 4.1(details in Appendix B.3). Meanwhile, to objectively compare the model ability, we calculate the average sample-wise accuracy of generated music for both baselines and our method(details in Appendix B.2).

为了进行公平的比较，我们使用 music2112 将由基线生成的 ABC 符号音乐转换为 MIDI 音乐。至于主观评估，包括从基线生成的音乐和我们的方法在内的精心设计的问卷被分发给个人，这些个人都在音乐背景中，并且需要对 4.1 节中描述的主观指标进行评分 (详见附录 B.3)。同时，为了客观地比较模型的能力，我们计算了基线和我们的方法生成的音乐的平均样本精度 (详见附录 B.2)。

Main Results Table 4 reports the main results of the comparison. In terms of musicality, MuseCoco achieves the highest score(mean 4.06 out of 5), indicating its ability to generate music closely resembling compositions by humans and approaching the quality of real-world music. As for the conditional generation ability, MuseCoco outperforms all baselines in terms of controllability and the average sample-wise accuracy at 1.08 and 19.95% respectively, which illustrates the effectiveness of controlling musical attributes with the two-stage framework. GPT-4 also can generate more coherent music with input prompts than BART-base due to its powerful ability of language understanding. Meanwhile, the best overall score(mean 4.13 out of 5) of MuseCoco shows our method is capable of generating the most favorite and fair-sounding music through the auditory test. Music samples generated by MuseCoco are available via

this link13.

主要结果表 4 报告了比较的主要结果。在音乐性方面，MuseCoco 获得了最高分 (平均分为 4.06 分，满分为 5 分) ，这表明它有能力生成与人类作品非常相似的音乐，并接近真实世界音乐的质量。在条件生成能力方面，MuseCoco 在可控性和平均样本准确率方面均优于所有基准，分别为 1.08% 和 19.95% ，说明了两阶段框架控制音乐属性的有效性。Gpt-4 由于其强大的语言理解能力，与 BART-base 相比，gpt-4 还可以通过输入提示生成更加连贯的音乐。同时，MuseCoco 的最佳综合得分 (平均 4.13 分) 表明我们的方法能够通过听觉测试生成最喜欢的和听起来不错的音乐。MuseCoco 生成的音乐样本可以通过这个链接获得。

Table 4: Comparison between MuseCoco, GPT-4, and BART-base. ASA stands for the average sample-wise accuracy.

表 4: MuseCoco，gpt-4 和 BART-base 之间的比较。 ASA 代表平均样本精度。

	Musicality 音乐性	Controllability 可控性	Overall 总体而言	ASA (%) 生化需氧量 (%)
MuseCoco MuseCoco 博物馆	4.06 ± 0.82 4.06 ± 0.82	4.15 ± 0.78 4.15 ± 0.78	4.13 ± 0.75 4.13 ± 0.75	77.59
GPT-4[7] GPT-4[7] BART-base[6] BART-base [6]	2.79 ± 0.97 2.79 ± 0.97 2.19 ± 1.14 2.19 ±	3.07 ± 1.05 3.07 ± 1.05 2.02 ± 1.09 2.02 ±	2.81 ± 0.97 2.81 ± 0.97 2.17 ± 1.03 2.17 ±	57.6431.9

	1.14	1.09	1.03	8

4.3 Method Analysis

4.3 方法分析

In this section, we conduct analysis experiments on the two stages respectively.

在本节中，我们将分别对这两个阶段进行分析实验。

4.3.1 Analysis on Text-to-Attribute Understanding

4.3.1 文本到属性的理解分析

Attribute Comprehension To evaluate the ability to extract each attribute from text, we test the text-to-attribute model on the standard test set and show the classification accuracy of each attribute in Appendix A. Each accuracy consistently surpasses 99%, which proves that the model exhibits exceptional performance on all attributes and showcases the high accuracy and reliability of the text understanding ability.

属性理解为了评估从文本中提取每个属性的能力，我们在标准测试集上测试了文本到属性的模型，并在附录 a 中显示了每个属性的分类准确率。每个准确率始终超过 99% ，这证明了该模型在所有属性上都表现出优异的性能，展示了文本理解能力的高准确性和可靠性。

Different classification heads We explore the effectiveness of using multiple classification heads and report the ASA shown in Table 5. The model with multiple classification heads outperforms the

不同的分类头我们探讨了使用多个分类头的有效性，并报告了表 5 所示的 ASA。使用多个分类头的模型优于

10https://chat.openai.com/
10https://chat.openai.com/
11https://huggingface.co/sander-wood/text-to-music

11https://huggingface.co/sander-wood/text-to-music
11https://huggingface.co/sander-wood/text-to-music
12http://web.mit.edu/music21/
12 http://web.mit.edu/music21/
13https://ai-muzic.github.io/musecoco/
13https://ai-muzic. github.io/musecoco/

one-head BERT by 39.87%, which illustrates that each head can learn their corresponding attribute knowledge, and using multiple heads can improve the overall performance.

单头 BERT 提高了 39.87% ，说明每个头都可以学习到相应的属性知识，使用多个头可以提高整体性能。

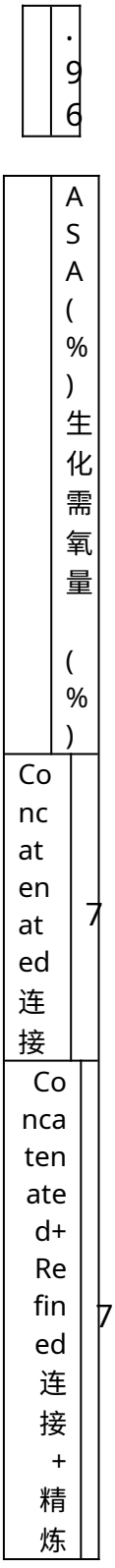
Different text synthetic strategies In order to showcase the effectiveness of the refinement strategy outlined in Section 3.4, we evaluate its impact on enhancing the fluency and diversity of the text within the training set during the text-to-attribute stage. We engage musicians to write 17 text descriptions manually to help evaluate and contrast the effectiveness of synthesis strategies w/ and w/o the refinement step. As shown in Table 6, we observe that fine-tuning the model with 25%partially refined text achieves better performance than with the simply concatenated templates.

不同的文本合成策略为了展示 3.4 节中概述的细化策略的有效性，我们在文本到属性的阶段评估其对提高训练集中文本的流畅性和多样性的影响。我们聘请音乐家手动写 17 个文本描述，以帮助评估和对比合成策略的有效性 w / 和 w/o 的细化步骤。如表 6 所示，我们观察到用 25% 部分精化的文本微调模型比简单连接的模板获得更好的性能。

Table 5: Analysis on multiple classification heads.Table 6: Analysis on text refinement.

表 5: 对多个分类头的分析表 6: 对文本细化的分析。

One Head	60.9999



4.3.2 Analysis on Attribute-to-Music Generation

4.3.2 音乐属性生成分析

Attribute Control Performance To evaluate the controllability of

the attribute-to-music generation model, we report the control accuracy results for each attribute in Appendix B.4. The average attribute control accuracy is 80.42%, demonstrating a strong capability of the model to effectively respond to the specified attribute values during the music generation process.

属性控制性能为了评估属性到音乐生成模型的可控性，我们在附录 B.4 中报告了每个属性的控制精度结果。平均属性控制精度为 80.42% ，表明该模型在音乐生成过程中有效响应指定属性值的强大能力。

Study on Control Methods We compare Prefix Control, which is the default method of our model that uses prefix tokens to control music generation, with two other methods: 1) Embedding:Add attribute input as embedding to token embedding; 2) Conditional LayerNorm: Add attribute input as a condition to the layer norm layer[43, 44]. We utilize Musicality and average attribute control accuracy as evaluation metrics. For more details on this experiment, please refer to the description in Appendix B.4. We report evaluation results in Table 7. It is shown that Prefix Control outperforms other methods in terms of musicality and average attribute control accuracy, with a minimum improvement of 0.04 and 19.94% respectively, highlighting its superior capability to capture the relationship between attributes and music.

控制方法的研究我们比较前缀控制，这是我们的模型的默认方法，使用前缀标记来控制音乐生成，与其他两种方法： 1) 嵌入： 添加属性输入作为嵌入标记嵌入；2) 条件层规范： 添加属性输入作为条件层规范层 [43,44]。我们使用音乐性和平均属性控制准确性作为评估指标。关于这个实验的更多细节，请参考附录 B.4 中的描述。我们在表 7 中报告评估结果。实验结果表明，前缀控制方法在音乐性和平均属性控制准确率方面优于其他方法，最低分别提高了 0.04% 和 19.94% ，体现了其捕捉属性与音乐之间关系的优越能力。

Table 7: Comparison of different control methods. Musicality reflects the quality of the generated music. Average attribute control accuracy represents the control accuracy over all attributes, which can reflect controllability.

表 7: 不同控制方法的比较。音乐性反映了生成音乐的质量。平均属性控制

精度代表对所有属性的控制精度，可以反映音乐的可控性。

Method 方法	Musicality ↑ 音乐性 ↑	Average attribute control accuracy(%) ↑ 平均属性控制精度 (%)↑	
Embedding 嵌入	2.97 ± 0.91 2.97 ± 0.91	36.94	
Conditional LayerNorm 有条件的 LayerNorm Prefix Control 前缀控制	3.11 ± 1.02 3.11 ± 1.02 3.15 ± 1.02 3.15 ± 1.02	47.46 67.40	

Study on Model Size We conduct a comparative analysis between two different model sizes to determine whether increasing the model size would result in improved generated results. The parameter configurations for these model sizes are presented in Table 8. The default model, referred to as large, is the default model for the attribute-to-music generation stage. Additionally, we utilize xlarge model for comparison, which consists of approximately 1.2 billion parameters. The training of xlarge model follows the same settings outlined in Section 4.1. The objective evaluation results are displayed in Table 8, which indicates that increasing the model size enhances controllability. To further evaluate the performance, we will conduct subjective listening tests to compare the subjective evaluation results for these two model sizes.

关于模型大小的研究我们对两种不同的模型大小进行了比较分析，以确定增加模型大小是否会改善生成的结果。这些模型大小的参数配置如表 8 所示。默认模型，被称为 large，是属性到音乐生成阶段的默认模型。此外，我们使用 xlarge 模型进行比较，它由大约 12 亿个参数组成。Xlarge 模型的训练遵循与第 4.1 节相同的设置。客观评估结果如表 8 所示，表明增加模型大小可以增强可控性。为了进一步评估性能，我们将进行主观听力测试来比较这两种模型大小的主观评估结果。

4.3.3 Comments from Musicians

4.3.3 音乐家的意见

We invite professional musicians to give their comments on generated music samples from given texts based on our system. Here are some feedbacks:

我们邀请专业音乐人根据我们的系统对给定文本生成的音乐样本进行评论。下面是一些反馈：

The generated music closely resembles human compositions, displaying a high level of accuracy and creativity. It provides inspiration for creative compositions, showcasing interesting motifs and

生成的音乐与人类作品非常相似，表现出高度的准确性和创造性。它为创意作品提供了灵感，展示了有趣的主题和

Table 8: Comparison of different model sizes in the attribute-to-music generation stage. Average objective attribute control accuracy represents the control accuracy over objective attributes, which can reflect controllability.

表 8: 音乐创作阶段不同模型尺寸的比较。平均客观属性控制精度代表对客观属性的控制精度，可以反映可控性。

Model Size 型号尺寸	Layers 图层	dmodel Dmodel 模型	Parameters 参数	Average objective attribute control accuracy(%) ↑ 平均目标属性控制精度 (%)↑
large 大型	16	1024	203M 2.03 亿	83.63
xlarge Xlarge 大	24	2048	1.2B 1.2 b 1.2 b 年	87.15

demonstrating skillful organization of musical elements. This work greatly improves composition efficiency, saving approximately one day of time. —- from Musician A

演示音乐元素的熟练组织。这项工作大大提高了作曲效率，节省了大约一天的时间。音乐家 a

The generated music offers arrangement inspiration, exemplified by the idea of combining right-hand arpeggios and left-hand bass melody to facilitate creative expansion in composition. It sparks the concept of blending classical and popular music genres described in texts. —- from Musician B

产生的音乐提供编曲灵感，例如结合右手琶音和左手低音旋律的想法，以促进创作的创造性扩展。它激发了融合文本中描述的古典和流行音乐流派的概念。一一选自

《音乐家 b 》

The generated music significantly reduces my composition time, saving anywhere from 2 days to 2weeks, particularly when it comes to instrumental arrangements that are outside my familiarity. The composition incorporates inspirational elements in some sections of the piece. For example, I very much enjoyed the journey through the conflicts and resolutions. In some areas towards the end it felt like I was embarking on an adventure up a mountain and through grassy fields, very interesting. —-from Musician C

生成的音乐显著地减少了我的创作时间，节省了 2 天到 2 周的时间，特别是当涉及到我不熟悉的乐器编曲时。这首作品在某些部分融入了灵感元素。例如，我非常享受通过冲突和解决方案的旅程。在一些接近终点的地方，我感觉就像是开始了一场登山和穿越草地的冒险，非常有趣。——选自《音乐家 c 》

The generated music presents the various elements of the text well, which provides great convenience for musicians to edit the music. For example, music in the style of Bach is well-created, and multiple instruments in the generated music are arranged harmoniously. Meanwhile, as for music teachers, MuseCoco is very helpful in generating desired musical examples for education. —- from Musician and Music Teacher D

生成的音乐很好地呈现了文本的各种元素，为音乐家编辑音乐提供了极大的方便。例如，巴赫风格的音乐创作精良，生成音乐中的多种乐器排列和谐。同时，对于音乐教师来说，MuseCoco 可以帮助他们创作出符合教育需要的音乐范例。——来自音乐家和音乐教师 d

The feedback from musicians has demonstrated our ability to enhance their workflow efficiency by reducing redundant tasks and providing creative inspiration.

来自音乐家的反馈表明，我们有能力通过减少多余的任务和提供创造性的灵感来提高他们的工作效率。

5 Conclusion

5 结论

In conclusion, this paper makes several significant contributions to the field of music generation and the application of AI in creative tasks. We introduce MuseCoco, a system that seamlessly transforms text descriptions into musically coherent symbolic compositions. This innovative approach empowers musicians and general users from diverse backgrounds and skill levels to create music more efficiently and with greater control.

总之，本文对音乐生成领域和人工智能在创造性任务中的应用做出了重要贡献。我们介绍了 MuseCoco，一个将文本描述无缝转换为音乐连贯符号作品的系统。这种创新的方法使音乐家和来自不同背景和技能水平的普通用户能够更有效地创作音乐，并拥有更大的控制权。

Second, we present a two-stage design that simplifies the learning process and enhances controllability. This design reduces the reliance on large amounts of paired text-music data and improves more explicit control through various aspects of attributes. By leveraging a large amount of symbolic music data, we achieve impressive musicality and establish a coherent connection between text descriptions and the generated music.

其次，我们提出了一个两阶段的设计，简化了学习过程，增强了可控性。这种设计减少了对大量成对文本 - 音乐数据的依赖，并通过属性的各个方面提高了更明确的控制。通过利用大量的符号音乐数据，我们实现了令人印象深刻的音乐性，并在文本描述和生成的音乐之间建立了一致的连接。

Our research demonstrates the potential of AI technologies in facilitating idea generation and streamlining the composition process for music creation tasks. By offering a powerful and adaptable tool like MuseCoco, we aim to inspire and empower artists to overcome the challenges they face in their creative pursuits and unlock new possibilities in music composition. The utilization of generative AI for creative purposes often raises concerns regarding copyright and ownership, which necessitates careful consideration moving forward. Limitations of our work can be seen in Appendix C.

我们的研究证明了人工智能技术在促进创意产生和简化音乐创作任务的作曲过程方面的潜力。通过提供一个像 MuseCoco 这样强大和适应性强的工具，我们旨在激励和授权艺术家克服他们在创作过程中面临的挑战，并开启音乐创作的新可能性。将生成性人工智能用于创作目的通常会引发对版权和所有权的担忧，这需要仔细考虑向前推进。我们工作的局限性可以在附录 c 中看到。

6 Acknowledgment

6 致谢

We would like to express our sincere gratitude to Amy Sorokas for her invaluable help in connecting with musicians. We would also like to thank musician, Grentperez, and musicians from Central Conservatory of Music in China for the collaboration and constructive feedback,

which contributed to the success of this work. Meanwhile, we sincerely appreciate all members of Wenqin Piano Society from Zhejiang University for completing the most questionnaires.

我们衷心感谢艾米·索罗卡斯 (Amy Sorokas)，感谢她在与音乐家交流方面给予的宝贵帮助。我们还要感谢音乐家格伦佩雷斯和中央音乐学院的音乐家们的合作和建设性的反馈，他们为这项工作的成功做出了贡献。同时，我们衷心感谢浙江大学文琴学会所有成员完成的大部分问卷调查。

References

参考文献

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

[1] A.Agostinelli，T.I.Denk，Z.Borsos，J.Engel，M.Verzetti，A.Caillon，q. Huang，A.Jansen，a. Roberts，M.Tagliasacchi et al.，“Musiclm: Generating music from text,”arXiv preprint arXiv: 2301.11325,2023[1] A.Agostinelli，T.I.Denk，Z.Borsos，J.Engel，M.Verzetti，A.Caillon，q. Huang，A.Jansen，a. Roberts，M.Tagliasacchi et al.，“Musiclm: Generating music from text,”arXiv preprint arXiv: 2301.11325,2023。

[2] Q. Huang, D. S. Park, T. Wang, T. I. Denk, A. Ly, N. Chen, Z. Zhang, Z. Zhang, J. Yu, C. Frank et al., “Noise2music: Text-conditioned music generation with diffusion models,” arXiv preprint arXiv:2302.03917, 2023.

[2] 黄庆，帕克，王天，邓俊贤，李安，陈南，张志忠，张志忠，余江，弗兰克等，“Noise2music: Text-conditioned music generation with diffusion models,”arXiv preprint arXiv: 2302.03917,2023，“Noise2music: Text-conditioned music generation with diffusion models,”。

[3] P. Zhu, C. Pang, S. Wang, Y. Chai, Y. Sun, H. Tian, and H. Wu, “Ernie-music: Text-to-waveform music generation with diffusion models,” arXiv preprint arXiv:2302.04456, 2023.

[3] 朱鹏，庞春，王顺，柴永，孙永，田海，吴海，“Ernie-music: Text-to-waveform music generation with diffusion models,”arXiv preprint arXiv: 2302.04456,2023，“Ernie-music: Text-to-waveform music generation with diffusion models,”arXiv: 2302.04456,2023。

[4] F. Schneider, Z. Jin, and B. Schölkopf, “Moûsai: Text-to-music generation with long-context latent diffusion,” arXiv preprint arXiv:2301.11757, 2023.

[4] f. Schneider，z. Jin，and b. Schölkopf，“ Moûsai: Text-to-music generation with long-context latent diffusion,”arXiv preprint arXiv: 2301.11757,2023。

[5] Y. Zhang, Z. Wang, D. Wang, and G. Xia, “BUTTER: A representation learning framework for bi-directional music-sentence retrieval and generation,” in Proceedings of the 1st Workshop on NLP for Music and Audio(NLP4MusA). Association for Computational Linguistics, 16 Oct.2020, pp. 54–58.

[5] 张勇，王志军，王东，和夏国荣，“BUTTER: 一个双向音乐 - 句子检索和生成的表征学习框架”，载于《第一届音乐和音频自然语言处理研讨会论文集》(NLP4MusA)。计算机语言学协会，2020 年 10 月 16 日，54-58 页。

[6] S. Wu and M. Sun, “Exploring the efficacy of pre-trained checkpoints in text-to-music generation task,” arXiv preprint arXiv:2211.11216, 2022.

[6] S.Wu 和 M.Sun，“探索预训练检查点在文本到音乐生成任务中的功效”，arXiv 预印本 arXiv: 2211.11216,2022。

[7] OpenAI, “Gpt-4 technical report,” arXiv preprint arXiv:2303.08774, 2023.

[7] OpenAI，“gpt-4 技术报告”，arXiv 预印本 arXiv: 2303.08774,2023。

[8] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, “High-resolution image synthesis with latent diffusion models,” in CVPR. IEEE, 2022, pp. 10 674–10 685.

[8] r. Rombach，a. Blattmann，d. Lorenz，p. Esser，and b. Ommer，“高分辨率图像合成与潜在扩散模型，” 在 CVPR。IEEE，2022，pp. 10674-10685.

[9] Q. Huang, A. Jansen, J. Lee, R. Ganti, J. Y. Li, and D. P. W. Ellis, “Mulan: A joint embedding of music audio and natural language,” arXiv preprint arXiv:2208.12415, 2022.

[9] 黄昆，詹森，李建华，甘蒂，李建华，埃利斯，《木兰： 音乐音频与自然语言的联合嵌入》，arXiv 预印本，arXiv: 2208.12415,2022。

[10] S. Bubeck, V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E.

Kamar, P. Lee, Y. T. Lee, Y. Li, S. Lundberg et al., “Sparks of artificial general intelligence: Early experiments with gpt-4,”arXiv preprint arXiv:2303.12712, 2023.

[10] s. Bubeck，v. Chandrasekaran，r. Eldan，j. Gehrke，e. Horvitz，e. Kamar，p. Lee，y. t. Lee，y. Li，s. Lundberg et al.，“ Sparks of artificial general intelligence: Early experiments with gpt-4,”arXiv preprint arXiv: 2303.12712,2023[10] s. Bubeck，v. Chandrasekaran，r. Eldan，j. Gehrke，e. Horvitz，e. Kamar，p. Lee，y. t. Lee，y. Li，s. Lundberg et al。

[11] Z. Wang, D. Wang, Y. Zhang, and G. Xia, “Learning interpretable representation for controllable polyphonic music generation,” arXiv preprint arXiv:2008.07122, 2020.

[11] 王振，王东，张勇，夏国华，“可控复调音乐生成的可解释表征学习”，arXiv 预印本，arXiv: 2008.07122,2020。

[12] H. H. Tan and D. Herremans, “Music fadernets: Controllable music generation based on high-level features via low-level feature modelling,” arXiv preprint arXiv:2007.15474, 2020.

Tan and d. Herremans，“ Music fadernets: Controllable Music generation based on high-level features through the low-level feature modeling,”arXiv preprint arXiv: 2007.15474,2020。

[13] D. von Rütte, L. Biggio, Y. Kilcher, and T. Hofmann, “Figaro: Controllable music generation using learned and expert features,” in The Eleventh International Conference on Learning Representations, 2023.

[13] 冯吕特，l. Biggio，y. Kilcher，和 t. Hofmann，“费加罗： 使用学习和专家特征的可控音乐生成，” 在第十一届国际学习表征会议，2023 年。

[14] P. Neves, J. Fornari, and J. Florindo, “Generating music with sentiment using transformer-gans,”arXiv preprint arXiv:2212.11134, 2022.

[14] p. Neves，j. Fornari，and j. Florindo，“使用变形金刚产生有情感的音乐”，arXiv 预印本 arXiv: 2212.11134,2022。

[15] Y. Zhu, K. Olszewski, Y. Wu, P. Achlioptas, M. Chai, Y. Yan, and S. Tulyakov, “Quantized gan for complex music generation from dance videos,” in Computer Vision–ECCV 2022:17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVII. Springer, 2022, pp. 182–199.

[15] 朱勇，k. Olszewski，y. Wu，p. Achlioptas，m. Chai，y. Yan，and s. Tulyakov，“从舞蹈视频生成复杂音乐的量化 gan”，载于《计算机视觉 - ecv 2022: 第 17 届欧洲会议》，2022 年 10 月 23-27 日，以色列特拉维夫，Proceedings，Part XXXVII。Springer，2022，pp. 182-199.

[16] H.-T. 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,” arXiv preprint arXiv:2108.01374, 2021.

[16] h-t.洪秀柱、程继忠、 s。多赫、 n。金、 j。南和 y。杨，“Emopia: 用于情感识别和基于情感的音乐生成的多模态流行钢琴数据集，”arXiv 预印本 arXiv: 2108.01374,2021。

[17] L. N. Ferreira, L. Mou, J. Whitehead, and L. H. Lelis, “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, vol. 18, no. 1, 2022, pp. 163–170.

[17] L.N.Ferreira，L.Mou，J.Whitehead，和 L.H.Lelis，“用蒙特卡洛树搜索控制符号音乐生成中的感知情绪”，发表在 AAAI 人工智能和交互式数字娱乐会议论文集，第 18 卷，第 1 期，2022 年，第 163-170 页。

[18] C. Bao and Q. Sun, “Generating music with emotions,” IEEE Transactions on Multimedia,2022.

[18] 包春，孙青，“生成有情感的音乐”，《IEEE 多媒体汇刊》，2022。

[19] H. H. Mao, T. Shin, and G. Cottrell, “Deepj: Style-specific music generation,” in 2018 IEEE12th International Conference on Semantic Computing(ICSC). IEEE, 2018, pp. 377–382.

[19] H.h. Mao，t. Shin 和 g. Cottrell，“Deepj: 风格特定的音乐生成”，2018 年 ieee 第 12 届语义计算国际会议 (ICSC)。IEEE，2018，第 377-382 页。

[20] W. Wang, X. Li, C. Jin, D. Lu, Q. Zhou, and Y. Tie, “Cps: Full-song and style-conditioned music generation with linear transformer,” in 2022 IEEE International Conference on Multimedia and Expo Workshops(ICMEW). IEEE, 2022, pp. 1–6.

[20] 王伟，李喜，金春，陆东，周庆，铁永，“Cps: 基于线性变压器的全歌曲风格条件化音乐生成”，2022 年 IEEE 多媒体与博览会国际研讨会 (ICMEW)。IEEE，2022，第 1-6 页。

[21] K. Choi, C. Hawthorne, I. Simon, M. Dinculescu, and J. Engel,

“Encoding musical style with transformer autoencoders,” in International Conference on Machine Learning. PMLR, 2020, pp. 1899–1908.

[21] k. Choi，c. Hawthorne，i. Simon，m. Dinculescu，and j. Engel，“用变压器自动编码器编码音乐风格”，在机器学习国际会议上。PMLR，2020，pp. 1899-1908.

[22] B. Yu, P. Lu, R. Wang, W. Hu, X. Tan, W. Ye, S. Zhang, T. Qin, and T.-Y. Liu, “Museformer:Transformer with fine- and coarse-grained attention for music generation,” in Proceedings of Advances in Neural Information Processing Systems(NeurIPS), 2022, pp. 1376–1388.

[22] 余平，吕平，王立军，胡伟，谭向东，叶伟，张松生，秦铁生，杨铁生。刘，“Museformer: 对音乐生成具有细粒度和粗粒度注意力的变压器”，载于《神经信息处理系统进展学报》，2022 年，第 1376-1388 页。

[23] X. Zhang, J. Zhang, Y. Qiu, L. Wang, and J. Zhou, “Structure-enhanced pop music generation via harmony-aware learning,” in Proceedings of the 30th ACM International Conference on Multimedia, 2022, pp. 1204–1213.

[23] 张晓，张建，邱永，王良，周建，“通过和声意识学习的结构增强的流行音乐生成”，载于第 30 届 ACM 国际多媒体会议论文集，2022 年，第 1204-1213 页。

[24] J. Ens and P. Pasquier, “Mmm: Exploring conditional multi-track music generation with the transformer,” arXiv preprint arXiv:2008.06048, 2020.

[24] 恩斯和帕斯奎尔，“嗯： 用变压器探索有条件的多轨音乐生成”，arXiv 预印本 arXiv: 2008.06048,2020。

[25] S. Di, Z. Jiang, S. Liu, Z. Wang, L. Zhu, Z. He, H. Liu, and S. Yan, “Video background music generation with controllable music transformer,” in Proceedings of the 29th ACM International Conference on Multimedia, ser. MM ’21. New York, NY, USA: Association for Computing Machinery, 2021, pp. 2037–2045.

[25] 迪思，江志江，刘思，王志，朱丽，何志和，刘海，严思，“可控音乐变换器的视频背景音乐生成”，载于第 29 届 ACM 国际多媒体会议论文集，ser。MM’21。纽约，纽约，美国： 计算机协会，2021，第 2037-2045 页。

[26] S. Wu, X. Li, and M. Sun, “Chord-conditioned melody choralization with controllable har-monicity and polyphonicity,” arXiv preprint arXiv:2202.08423, 2022.

[26] 吴思，李新，孙明，“和声与复调可控的和弦条件旋律合唱”，arXiv 预印本 arXiv: 2202.08423,2022。

[27] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, “Hierarchical text-conditional image generation with clip latents,” arXiv preprint arXiv:2204.06125, 2022.

[27] a. Ramesh，p. Dhariwal，a. Nichol，c. Chu 和 m. Chen，“使用剪辑潜伏的分层文本条件图像生成”，arXiv 预印本 arXiv: 2204.06125,2022。

[28] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. Gon-tijo Lopes, B. Karagol Ayan, T. Salimans et al., “Photorealistic text-to-image diffusion models with deep language understanding,” Advances in Neural Information Processing Systems, vol. 35, pp. 36 479–36 494, 2022.

[28] C.Saharia，W.Chan，S.Saxena，L.Li，J.Whang，E.L.Denton，K.Ghasemipour，R.Gon-tijo Lopes，B.Karagol Ayan，T.Salimans 等，“具有深度语言理解的逼真文本到图像的扩散模型”，《神经信息处理系统进展》，第 35 卷，第 36479-36494 页，2022 年。

[29] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, “Zero-shot text-to-image generation,” in International Conference on Machine Learning. PMLR,2021, pp. 8821–8831.

[29] a. Ramesh，m. Pavlov，g. Goh，s. Gray，c. Voss，a. Radford，m. Chen，and i. Sutskever，“零样本文本到图像的生成”，在机器学习国际会议上。PMLR，2021，pp. 8821-8831.

[30] P. R. Cook, Music, Cognition, and Computerized Sound: An Introduction to Psychoacoustics. MIT press, 2001.

[30]《音乐、认知与计算机化声音： 心理声学导论》，麻省理工学院出版社，2001 年。

[31] N. S. Keskar, B. McCann, L. R. Varshney, C. Xiong, and R. Socher, “Ctrl: A conditional transformer language model for controllable generation,” arXiv preprint arXiv:1909.05858,2019.

[31] N.s. Keskar，b. McCann，L.r. Varshney，c. Xiong 和 r. Socher，“Ctrl: 可控生成的条件变压器语言模型，” arXiv 预印本 arXiv: 1909.05858,2019。

[32] X. L. Li and P. Liang, “Prefix-tuning: Optimizing continuous prompts for generation,” arXiv preprint arXiv:2101.00190, 2021.

[32] X.L.Li 和 P.Liang，“前缀调优：优化生成的连续提示”，arXiv 预印本 arXiv: 2101.00190,2021。

[33] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, “Pre-train, prompt, and predict:A systematic survey of prompting methods in natural language processing,” ACM Computing Surveys, vol. 55, no. 9, pp. 1–35, 2023.

[33] 刘平，袁伟，傅建华，江泽民，林浩，和 G.Neubig，“预训练，提示和预测：自然语言处理中提示方法的系统调查”，acm 计算概观，第 55 卷，第 9 期，第 1-35 页，2023 年。

[34] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell et al., “Language models are few-shot learners,” Advances in neural information processing systems, vol. 33, pp. 1877–1901, 2020.

[34] t. Brown，b. Mann，n. Ryder，m. Subbiah，j. d. Kaplan，p. Dhariwal，a. Neelakantan，p. Shyam，g. Sastry，a. Askell 等人，“语言模型是少数学习者”，《神经信息处理系统进展》，第 33 卷，第 1877-1901 页，2020 年。

[35] J. Devlin, M. Chang, K. Lee, and K. Toutanova, “BERT: pre-training of deep bidirectional transformers for language understanding,” in NAACL-HLT(1). Association for Computational Linguistics, 2019, pp. 4171–4186.

[35] j. Devlin，m. Chang，k. Lee，和 k. Toutanova，“BERT: 深度双向变压器语言理解的预训练”，NAACL-HLT (1)。计算机语言学协会，2019 年，第 4171-4186 页。

[36] M. Zeng, X. Tan, R. Wang, Z. Ju, T. Qin, and T.-Y. Liu, “MusicBERT: Symbolic music understanding with large-scale pre-training,” in Findings of the Association for Computational Linguistics(ACL Findings), 2021, pp. 791–800.

[36] 曾明，谭 x，R.Wang，Z.Ju，T.Qin，and T.-Y。刘，“MusicBERT: 大规模预训练的符号音乐理解”，载于《计算机语言学协会的发现》，2021 年，第 791-800 页。

[37] J. Ens and P. Pasquier, “Building the metamidi dataset: Linking symbolic and audio musical data.” in ISMIR, 2021, pp. 182–188.

[37] j。 Ens 和 p。 Pasquier，“建立 metamidi 数据集：连接符号和音频

音乐数据。”在 ISMIR，2021，第 182-188 页。

[38] 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 Proceedings of International Society for Music Information Retrieval Conference(ISMIR), 2020, pp. 38–45.

[38] 王振，陈凯，江建，张永，徐明，戴世华，夏国华，“POP909: 用于音乐编排生成的流行歌曲数据集”，载于《国际音乐信息检索会议学报》，2020 年，第 38-45 页。

[39] J. Liu, Y. Dong, Z. Cheng, X. Zhang, X. Li, F. Yu, and M. Sun, “Symphony generation with permutation invariant language model,” arXiv preprint arXiv:2205.05448, 2022.

[39] 刘建军，董永东，郑志成，张晓军，李晓军，余文中，孙文中，“置换不变语言模型的交响乐生成”，arXiv 预印本，arXiv: 2205.05448,2022。

[40] Y.-S. Huang and Y.-H. Yang, “Pop music transformer: Beat-based modeling and generation of expressive pop piano compositions,” in Proceedings of ACM International Conference on Multimedia(MM), 2020, pp. 1180–1188.

[40] y-s.黄和 y-h。杨，“流行音乐转换器： 基于节拍的建模和一代表达流行钢琴作品，” 在美国计算机协会国际会议多媒体 (MM) ，2020 年，第 1180-1188 页。

[41] A. Katharopoulos, A. Vyas, N. Pappas, and F. Fleuret, “Transformers are rnns: Fast autore-gressive transformers with linear attention,” in International Conference on Machine Learning. PMLR, 2020, pp. 5156–5165.

[41] a. Katharopoulos，a. Vyas，n. Pappas，和 f. Fleuret，“变压器是 rnns: 具有线性注意力的快速自回归变压器，” 在机器学习国际会议上。PMLR，2020，pp. 5156-5165.

[42] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” arXiv preprint arXiv:1412.6980, 2014.

[42] D.P.Kingma 和 J.Ba，“Adam: a method for stochastic optimization,”arXiv preprint arXiv: 1412.6980,2014。

[43] E. Perez, F. Strub, H. De Vries, V. Dumoulin, and A. Courville, “Film: Visual reasoning with a general conditioning layer,” in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32, no. 1, 2018.

[43] e. Perez，f. Strub，h. De Vries，v. Dumoulin 和 a. Courville，“电

影： 具有一般条件层的视觉推理”，在 AAAI 人工智能会议论文集，第 32 卷，第 1 期，2018 年。

[44] M. Chen, X. Tan, B. Li, Y. Liu, T. Qin, S. Zhao, and T.-Y. Liu, “Adaspeech: Adaptive text to speech for custom voice,” arXiv preprint arXiv:2103.00993, 2021.

[44] m. Chen，x. Tan，b. Li，y. Liu，t. Qin，s. Zhao 和 T.-Y。Liu，“ Adaspeech: Adaptive text to speech for custom voice,”arXiv preprint arXiv: 2103.00993,2021。

A Attribute Information

属性信息

Table 9 shows the detailed pre-defined musical attribute values. The value NA of each attribute refers to that this attribute is not mentioned in the text. Objective attributes can be extracted from MIDI files with heuristic algorithms and subjective attributes are collected from existing datasets, as shown in Table 10.

表 9 显示了详细的预定义音乐属性值。每个属性的值 NA 表示这个属性在文本中没有提到。客观属性可以用启发式算法从 MIDI 文件中提取，主观属性从现有的数据集中收集，如表 10 所示。

Table 9: Detailed attribute values.

表 9: 详细的属性值。

Attributes 属性	v
28 instrument	

s: piano, keyboard, percussion, organ, guitar, bass, violi

n，viola，28种乐器：钢琴，键盘，打击乐器，风琴，吉他，贝司，小提琴，中提琴，

c
b
E
0

-11:octaves,12:NA 0-11:octaves,12:NA 0-11:八度音阶

R

0

:danceable,1:notdanceable,2:NA 0:可跳，1:不可跳，2:不可跳	0

:serene,1:moderate,2:intense,3:NA

0:宁静,1:适中,2:激烈

，3：不安

I

0：1-4bars,1：5-8bars,2：9-12bars,3：13-16bars,4：N

A0:1-4格，1:5-8格，2:9-12格，3:13-16格，4:NA
0:4/4，1:2/4，

2:3/4,3:1/4,4:6/8,5:3/8,6:othertempos,7:NA0:4/4,

1:2/4,2:3/4,3:1/4,4:6/8,5:3/8,6:其他节奏,7:NAS

0:major

r,1:minor,2:NA 0:大调 1:小调 2:无调
0:slow(<=76BPM),1:mode

rato(76-120BPM),2:fast(>=120BPM),0:慢(<=76bpm)，1:中等

	(76-120bpm)，2：快（>=120bpm），3
	0：0-15s，1：15-30s，2：

30-45s,3:45-60s,4:>60s,5:NA0:0-15s1:15-30s2:30-4

5s3:45-60s4:>60s5:NA
0-16artists:Beethoven,Mozart,C

hopin, Schubert, Schumann, J. S. Bach, Haydn, Brahms, 0-16

艺术家：贝多芬，莫扎特，肖邦，舒伯特，舒曼，巴赫，海顿，勃拉姆斯，
Handel,Tchaik

ovsky, Mendelssohn, Dvorak, Liszt, Stravinsky, Mahler, Prokofiev

fiev,Shostakovich，汉德尔，柴可夫斯基，门德尔松，德沃夏克，李斯特，斯特拉文斯基，

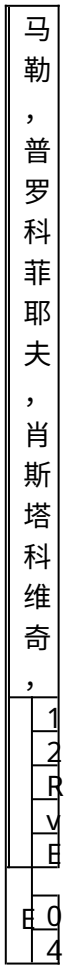


Table 10: Extraction methods and sources of each attributes.

表 10: 每个属性的提取方法和来源。

Type 类型	Attribute 属性	Extraction Method 提取方法
Objective 目的 Subjective 主观	Instrument 乐器	directly extracted from MIDI 直接从 MIDI 中提取
	Pitch range 音高范围	calculated based on the pitch range 基于音高范围计算
	Rhythm danceability 节奏舞蹈性	judged with the ratio of downbeat 用强拍的比率来判断
	Rhythm intensity	judged with the average note

	节奏强度	density 用平均音符密度判断
	Bar 酒吧 Time signature 时间签名	directly extracted from MIDI 直接从 MIDI 中提取 directly extracted from MIDI 直接从 MIDI 中提取
	Key 钥匙	judged with the note pitches based on musical rules 根据音乐规则判断音高
	Tempo 节奏	directly extracted from MIDI 直接从 MIDI 中提取
	Time 时间： Artist 艺术家 Genre 流派 Emotion 情绪	derived from the time signature and the number of bars 从时间标记和小节数得出 provided by a classical music dataset in MMD[36] 由 MMD 中的古典音乐数据集提供 [36] provided by MAGD14, a classical music dataset in MMD[36] and Symphony[39] 由 MMD [36] 和 Symphony [39] 中的古典音乐数据集 magd14 提

		供 provided by EMOPIA[16] and the emotion-gen dataset 由 EMOPIA [16] 和情绪数据集提供

B Experiments

B 实验

B.1 User study with baselines

B. 1 基线用户研究

In the user study, participants were provided with generated music samples along with their corre-sponding textual prompts. For each text description, each model(i.e., BART-base, GPT-4, MuseCoco) generated three different music clips. In each questionnaire, three samples generated with the same text conditions were randomly picked from samples generated by BART-base, GPT-4, and MuseCoco respectively as a group. Each participant was asked to evaluate 7 groups for comparison. Three subjective metrics, musicality, controllability, and an overall score, are rated on a scale of 1(lowest) to 5(highest). The participants were first requested to evaluate their music profession level, as depicted in Table 11. To ensure the reliability of the assessment, only individuals with at least music profession level 3 were selected, resulting in a total of 19 participants. Secondly, they were instructed to independently evaluate two separate metrics: musicality and controllability, ensuring that scoring for one metric did not influence the other. They are also asked to give an overall score to evaluate the generated music comprehensively. For the collected results, we computed the mean and variance for each metric. The results can be found in Table 4.

在用户研究中，参与者被提供生成的音乐样本以及相应的文本提示。对于每个文本描述，每个模型 (即 BART-base，GPT-4，MuseCoco) 生成三个不同的音乐片段。在每份问卷中，分别从 BART-base、 gpt-4 和 MuseCoco 生成的样本中随机抽取三个具有相同文本条件的样本作为一个组。每个参与者

被要求评估 7 组进行比较。三个主观指标，音乐性，可控性，和一个总分，被评为 1 (最低) 到 5 (最高) 的等级。参与者首先被要求评估他们的音乐专业水平，如表 11 所示。为了确保评估的可靠性，只有至少具有音乐专业水平 3 的个人被选中，共有 19 名参与者。其次，他们被指示独立评估两个独立的指标： 音乐性和可控性，确保一个指标的评分不会影响另一个。他们还被要求给出一个综合评分，以全面评估生成的音乐。对于收集的结果，我们计算每个度量的均值和方差。结果见表 4。

Table 11: Music Profession Level

表 11: 音乐专业水平

LevelDescription LevelDescription 水平描述
1I rarely listen to music. 我很少听音乐。
2I haven't received formal training in playing or music theory, but I often listen to music and have my 2 我没有接受过正规的演奏或音乐理论训练，但我经常听音乐 preferred styles, musicians, and genres. 喜欢的风格、音乐家和音乐流派。
3I have some basic knowledge of playing an instrument or music theory, but I haven't received formal 我有一些乐器演奏或者音乐理论的基本知识，但是我还没有收到正式的 training. 训练。
4I haven't received formal training, but I have self-taught myself some aspects such as music theory or 我没有接受过正式的培训，但是我自学了一些方面，比如音乐理论或者 playing an instrument. I am at an amateur level(e.g., CCOM piano level 6 or above). 演奏乐器。我处于业余水平 (例如，CCOM 钢琴 6 级或以上)。
5I have received professional training in a systematic manner. 我接受过系统的专业训练。

B.2 Objective Comparison with baselines

B. 2 与基线的客观比较

In this section, we introduce how to calculate the objective metric, the average sample-wise accuracy(ASA), in Table 4. As for MuseCoco, ten music clips are generated per prompt and we report ASA of them

among the overall standard test set. Since it is labor-intensive to leverage GPT-4 with the official web page, we only guide GPT-4 to produce five music clips per prompt and calculate the ASA of 21 prompts randomly sampled from the standard test set. Besides, we utilize the released text-tune BART-base checkpoint15 to generate five music clips per prompt and report the ASA of 44prompts randomly chosen from the standard test set.

在本节中，我们将介绍如何计算表 4 中的客观指标，即平均样本精度 (ASA)。至于 MuseCoco，每个提示符生成十个音乐片段，我们在整个标准测试集中报告它们的 ASA。由于在官方网页上利用 gpt-4 是劳动密集型的，我们只引导 gpt-4 为每个提示符生成 5 个音乐片段，并计算从标准测试集中随机抽取的 21 个提示符的 ASA。此外，我们利用发布的文本调整 BART-base 检查点 15 为每个提示符生成 5 个音乐片段，并报告从标准测试集中随机选择的 44 个提示符的 ASA。

B.3 Text-to-attribute understanding

B. 3 文本到属性的理解

As shown in Table 12, all attribute control accuracy is close or equal to 100%, which indicates our model with multiple classification heads in the text-to-attribute understanding stage performs quite well.

如表 12 所示，所有属性控制准确率接近或等于 100% ，这表明我们的模型在文本到属性的理解阶段具有多个分类头，表现相当好。

B.4 Details of Analysis on Attribute-to-music Generation

B. 4 音乐属性生成分析的细节

Attribute Control Accuracy We report the control accuracy for each attribute on the test dataset, as shown in Table 13. The average attribute control accuracy of 80.42%, which provides substantial evidence for the model’s proficiency in effectively controlling music generation using music attributes.

属性控制精度我们报告测试数据集中每个属性的控制精度，如表 13 所示。平均属性控制精度为 80.42% ，这为模型在使用音乐属性有效控制音乐生成方面的熟练程度提供了实质性的证据。

Study on Control Methods To verify the effectiveness of the control method

in the attribute-to-music generations stage, we compare Prefix Control with two methods: Embedding and Conditional LayerNorm. For efficiency, we conducted this study on reduced-size models as follows: The backbone model of this experiment is a 6-layer Linear Transformer with causal attention. The hidden size is512 and the FFN hidden size is 2048. The other experiment configuration is the same as Section 4.1. Since the control accuracy of objective attributes can be easily calculated, we only need to measure the controllability of each subjective attribute in listening tests. The control accuracy of each attribute is shown in Table 13. Finally, the average attribute control accuracy can be calculated based on the accuracy results from both types of attributes. To measure the controllability of subjective attributes

控制方法的研究为了验证控制方法在属性 - 音乐生成阶段的有效性，我们比较了前缀控制和两种方法： 嵌入和条件层规范。为了提高效率，我们在缩小尺寸的模型上进行了如下研究： 本实验的骨干模型是一个具有因果注意力的 6 层线性变压器。隐藏大小为 512，FFN 隐藏大小为 2048。另一个实验配置与第 4.1 节相同。由于客观属性的控制准确性很容易计算，我们只需要测量听力测试中每个主观属性的可控性。每个属性的控制精度如表 13 所示。最后，可以根据两类属性的精度结果计算平均属性控制精度。衡量主观属性的可控性

15https://huggingface.co/sander-wood/text-to-music

15https://huggingface.co/sander-wood/text-to-music
15https://huggingface.co/sander-wood/text-to-music

Table 12: Attribute control accuracy(%) of the text-to-attribute understanding model. I: Instrument.

表 12: 文本到属性理解模型的属性控制精度 (%)。

Attribute 属性	Accuracy(%) 精确度 (%)	Attribute 属性	Accuracy(%) 精确度 (%)	Attribute 属性	Accuracy(%) 精确度 (%)
I_piano 我会弹钢琴	100.00	I_clarinet 单簧管	99.92	Genre_comedy_spoken 类型喜剧口语	100.00
I_keyboard 键盘	99.92	I_piccolo 短笛	99.94	Genre_pop_rock 流行摇滚	100.00
I_percussion 打击乐	100.00	I_flute 我吹长笛	99.62	Genre_reggae 雷鬼音乐	100.00

I_organ 译者： 王士杰	100.00	I_pipe I _ pipe i _ pipe	100.00	Genre_stage Genre _ stage 流派阶段	100.00
I_guitar 吉他	99.92	I_synthesizer I _ synthesizer i _ synthesizer i _ synthesizer	100.00	Genre_folk 流派民谣	100.00
I_bass 低音提琴	99.84	I_ethnic_instruments I _ ethnic _ instruments 民族乐器	99.98	Genre_blues 流派布鲁斯	100.00
I_violin I _ violin 小提琴	99.92	I_sound_effects I _ sound _ effects 声音效果	99.98	Genre_vocal 流派 _ 声乐	100.00
I_viola 中提琴	99.96	I_drum 我打鼓	100.00	Genre_holiday Genre _ holiday 类型假期	100.00
I_cello 大提琴	99.92	Genre_new_age 流派新时代	99.98	Genre_country 类型国家	100.00
I_harp 竖琴	100.00	Genre_electronic 电子类型	100.00	Genre_symphony 体裁交响曲	100.00
I_strings I _ strings	99.96	Genre_rap 说唱类型	100.00	Bar 酒吧	100.00
I_voice 我的声音	99.70	Genre_religious 宗教类型	100.00	Time Signature 时间签名	100.00
I_trumpet 小号 i _ trumpet	99.96	Genre_international 国际类型	100.00	Key 钥匙	100.00
I_trombone I _ trombone 长号	99.94	Genre_easy_listening 音乐类型： 轻松聆听	100.00	Tempo 节奏	99.84
I_tuba 大号	100.00	Genre_avant_garde 先锋派风格	100.00	Octave 八度音阶	100.00
I_horn 我是喇叭	99.94	Genre_rnb 类型 _ rnb	100.00	Emotion 情绪	99.80
I_brass I _ brass 黄铜	100.00	Genre_latin 类型拉丁语	100.00	Time 时间：	100.00
I_sax 萨克斯	99.84	Genre_children 儿童类型	100.00	Rhythm Danceability 节奏舞蹈能力	100.00
I_oboe I _ oboe	99.94	Genre_jazz Genre _ jazz 类	100.00	Rhythm Intensity 节奏强度	99.88

双簧管		型爵士乐			
I_bassoon 巴松管	99.96	Genre_classical Genre _ classical 类型：经典	100.00	Artist 艺术家	100.00

Table 13: Accuracy(%) of each attribute for attribute-to-music generation. I: Instrument.

表 13: 属性到音乐生成的每个属性的准确率 (%)。 i: 乐器。

Attribute 属性	Accuracy (%) 精确度 (%)	Attribute 属性	Accuracy (%) 精确度 (%)
I_piano	96.20	I_clarinet	90.63

o			
我会弹钢琴		单簧管	
I_keyboard键盘	7	I_piccolo短笛	
I_percussion打击乐	65.19	I_flute 我吹长笛	86.73
I_organ译者：王士杰	80.55	I_pipe I_pipe_i_	70.73

		pipe	
I_guitar 吉他	9	I_synthesizer I_synthesizer i_synthesizer i_synthesizer	7
I	9	I_ethnic_instruments I_ethnic_instruments 民族乐器	7
I_vio	8	I_sound_effe	5

lin I _violin 小提琴		cts I _sound _effects 声音效果			
I _viola 中提琴	92.03	I _drum 我打鼓	95.96		
I_cello 大提琴	86.50		B	71.80	
I _harp 竖琴	7	Time Signature 时间签名		9	
I_st	86		K	57	

rings I_strings	.08		.42
I_voice 我的声音	75.82	Tempo 节奏	92.71
I_trumpet 小号 i_trumpet	84.86	Octave 八度音阶	61.56
I_trombone I_trombone 长号	84.64	Time 时间：	65.82
I	9	Rhyt	8

		hm Danceability 节奏舞蹈能力	
I_horn 我是喇叭	8	Rhythm Intensity 节奏强度	8
I_brassI_brass 黄铜	77.27	Genre 流派	73.08
I_sax 萨克斯	81.74	Emotion 情绪	69.45
I_o	85.	Art	50.

boeI_oboe双簧管	23	ist艺术家	03
I_bassoon 巴松管	90.72		

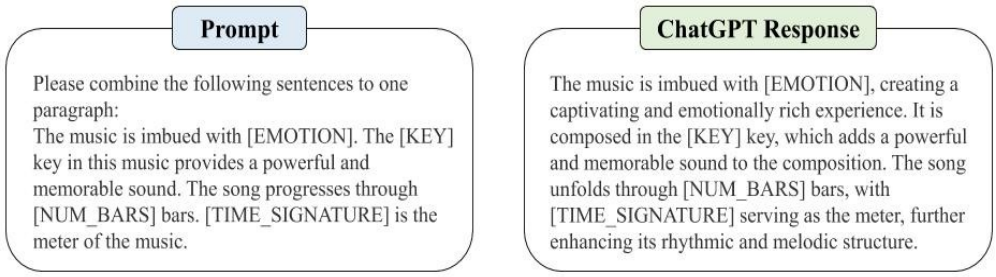


Figure 3: ChatGPT refines concatenated templates in the prompt.

图 3: ChatGPT 在提示符中细化连接的模板。

Prompt I want a music in the ABC notation format and assign proper MIDI instrument IDs as the following:X:1T:Forest Stroll M:3/4L:1/4Q:1/4=80K:C V:1% %MIDI program 90% Synthesizer E/2F/2|G2A|G2F|E2D|C2E|F2G|A3-|A2E/2F/2|G2A|G2F|E2D|C2E|F2G|A2F|E2D|C3||V:2%%MIDI program 68% Oboe G/2A/2|B2C|B2A|G2F|E2G|A2B|c3-|c2G/2A/2|B2C|B2A|G2F|E2G|A2B|c2A|G2F|E3||V:3%%MIDI program 73% Flute z/2A/2|B2C|B2A|G2F|E2G|A2B|c3-|c2z/2A/2|B2C|B2A|G2F|E2G|A2B|c2A|G2F|E3||V:4%%MIDI program 42% Cello C/2D/2|E2F|E2D|C2B,|A,2C|D2E|F3-|F2C/2D/2|E2F|E2D|C2B,|A,2C|D2E|F2D|C2B,|A,3||Compose

ABC notation music as the format above according to the music description as follow: The music should feature the cello and flute. The music is in 4/4meter. The song is 31~ 45 seconds in length.
提示我想要一首 ABC 符号格式的音乐，并分配正确的 MIDI 乐器 id 如下： x: 1T: Forest Stroll m: 3/4L: 1/4Q: 1/4 = 80K: c v: 1% MIDI 程序 90% Synthesizer e/2 f/2 |G2A | G2A | G2 f | g2 f | E2D | E2D | C2E | F2G | F2G | A3-| A2E/2 f/2 | G2A | G2 f | g2 f | E2D | E2D | C2E | F2G | F2G | A2 f | a2 fE2D | C3 | | V: 2% MIDI program 68% Oboe g/2A/2 | B2C | B2C | B2A | B2A | G2 f | g2 f | E2G | A2B | A2B | C3-| c2G/2A/2 | B2C | B2C | B2A | B2A | G2 f | g2 f | E2G | A2B | A2B | c2A | c2A | G2 f | g2 f | E3 | | V: 3% MIDI program 73% Flute z/2A/2 | B2C | B2A | B2A | G2 f | g2 f | E2G | A2B | A2B | C3-| 图片来源：c2z/2A/2B2C | B2C | B2A | B2A | G2 f | g2 f | E2G | A2B | A2B | c2A | c2A | G2 f | g2 f | E3 | | V: 4% MIDI program 42% Cello c/2D/2 | v: 4% MIDI 程序 42% Cello c/2D/2 |E2 f | E2D | C2B，| C2BA，2 c | D2E | F3-| F3-| F2C/2D/2 | 图片来源： e2 fE2D | C2B，| C2BA，2 c | D2E | F2D | C2B，| A，3 | 根据下面的音乐描述，按照上面的格式创作 ABC 符号音乐： 音乐应以大提琴和长笛为特色。音乐以 4/4 米为单位。歌曲长度为 31 ~ 45 秒。

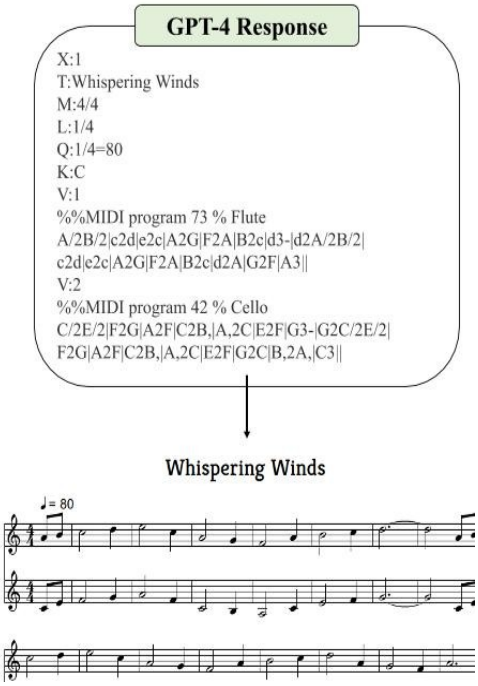


Figure 4: GPT-4 generates ABC notation tunes based on the prompt.

图 4: gpt-4 根据提示符生成 ABC 符号曲调。

(such as emotion and genre), we invite 12 participants to conduct a listening test. Each participant was provided with 18 music pieces(6 pieces per control method) with corresponding subjective attributes. We asked each participant to answer: 1) Musicality(five-point scale): How similar it sounds to the music composed by a human. 2) Controllability: Does it align with the given attributes. Then we report the musicality and average attribute accuracy score in Table Table 7. The experimental results clearly demonstrate that Prefix Control outperforms the other two methods in terms of musicality and controllability.

我们邀请了 12 个参与者进行听力测试。每个参与者被提供 18 个音乐片段 (每个控制方法 6 个片段) 和相应的主观属性。我们要求每个参与者回答： 1) 音乐性 (五分制) : 它听起来与人类创作的音乐有多相似。2) 可控性： 它是否与给定的属性一致。然后，我们在表 7 中报告音乐性和平均属性准确性得分。实验结果清楚地表明，前缀控制在音乐性和可控性方面优于其他两种方法。

B.5 Usage of GPT models

B.5 GPT 模型的使用

Refine texts with ChatGPT As shown in Figure 3, in order to make text descriptions more coherent and fluent, we feed concatenated templates into ChatGPT with a prompt Please combine the following sentences to one paragraph and then ChatGPT will give a response containing all templates within a compact paragraph.

使用 ChatGPT 完善文本如图 3 所示，为了使文本描述更加连贯和流畅，我们通过一个提示将连接的模板输入到 ChatGPT 中。请将下面的句子合并到一个段落中，然后 ChatGPT 将给出一个包含一个紧凑段落中所有模板的响应。

Generate ABC notation music with GPT-4 To use GPT-4 as the baseline method for comparison, we design the instruction to guide GPT-4 as shown in Figure 4. GPT-4 can only generate symbolic music in ABC notation, so we need to explicitly point out the format. Besides, since GPT-4 can generate various ABC notation formats, some of which cannot be processed by music21, we provide an ABC notation example, teaching GPT-4 to follow its format. Meanwhile, we use the prompt, Compose ABC notation music as the format above according to the music description as follows:[text descriptions] to let GPT-4 generate music according to the text description. And we finally convert the ABC notations into MIDI for a fair comparison.

使用 gpt-4 生成 ABC 符号音乐为了使用 gpt-4 作为比较的基线方法，我们设计了指导 gpt-4 的指令，如图 4 所示。Gpt-4 只能以 ABC 记谱法生成符号音乐，所以我们需要明确指出格式。此外，由于 gpt-4 可以生成各种 ABC 符号格式，其中一些格式不能被 music21 处理，所以我们提供了一个 ABC 符号示例，教 gpt-4 遵循其格式。同时，我们使用提示符，根据音乐描述如下： [文本描述] ，让 gpt-4 根据文本描述生成音乐。最后我们将 ABC 符号转换为 MIDI 进行公平比较。

C Limitation

C 限制

This work is mainly about generating symbolic music from text descriptions, which does not consider long sequence modeling especially. To address this, we can employ Museformer[22] as the backbone model, which proposes fine- and coarse-grained attention for handling long sequences.

这项工作主要是关于从文本描述生成符号音乐，而没有特别考虑长序列建模。为了解决这个问题，我们可以使用 Museformer [22] 作为骨干模型，它提出了细粒度和粗粒度的注意力来处理长序列。

The attribute set provided in this work represents only a subset of all music attributes. We aim to further explore additional attributes to offer a broader range of control options for music generation, ensuring greater diversity in the creative process.

本工作中提供的属性集仅代表所有音乐属性的子集。我们的目标是进一步探索更多的属性，为音乐生成提供更广泛的控制选项，确保在创作过程中有更大的多样性。

The possibility of regenerating music based on additional text descriptions to assist users in refining their compositions is an aspect that is worth exploring.

基于额外的文本描述重新生成音乐以帮助用户完善他们的作品的可能性是一个值得探索的方面。