MusicLM: Generating Music From Text
MusicLM: 从文本生成音乐

Andrea Agostinelli [*] [1] Timo I. Denk [*] [1]
作者: Andrea Agostinelli
Zalan´ Borsos [1] Jesse Engel [1] Mauro Verzetti [1] Antoine Caillon [2] Qingqing Huang [1] Aren Jansen [1]
扎兰 · 博尔索斯 1 杰西 · 恩格尔 1 毛罗 · 韦尔泽蒂 1 安托万 · 卡永 2 黄庆庆 1 阿伦 · 詹森 1
Adam Roberts [1] Marco Tagliasacchi [1] Matt Sharifi [1] Neil Zeghidour [1] Christian Frank [1]
Adam Roberts 1 Marco Tagliasacchi 1 Matt Sharifi 1 Neil Zeghidour 1 Christian Frank 1 亚当 · 罗伯茨 1 马可 · 塔格利亚萨奇 1 马特 · 谢里夫 1 尼尔 · 泽吉多尔 1 克里斯蒂安 · 弗兰克

Abstract
摘要

We introduce MusicLM, a model for generating high-fidelity music from text descriptions such as “a calming violin melody backed by a distorted gui-tar riff ”. MusicLM casts the process of condi-tional music generation as a hierarchical sequence-to-sequence modeling task, and it generates music at 24 kHz that remains consistent over several mi-nutes. Our experiments show that MusicLM out-performs previous systems both in audio quality and adherence to the text descriptions. Moreover, we demonstrate that MusicLM can be conditioned on both text and a melody in that it can transform whistled and hummed melodies according to the style described in a text caption. To support fu-ture research, we publicly release MusicCaps, a dataset composed of 5.5k music-text pairs, with rich text descriptions provided by human experts.

我们介绍了 MusicLM，它是一个模型，可以根据文本描述生成高保真的音乐，比如“一段平静的小提琴旋律，背后是一段扭曲的吉他即兴演奏”。MusicLM 将条件音乐生成过程转换为一个层次化的序列到序列建模任务，它生成的音乐频率为 24khz，在几分钟内保持一致。我们的实验表明，MusicLM 在音频质量和遵守文本描述方面都优于以前的系统。此外，我们还演示了 MusicLM 可以同时受到文本和旋律的制约，因为它可以根据文本标题中描述的风格转换吹口哨和哼唱的旋律。为了支持未来的研究，我们公开发布了 MusicCaps，这是一个由 5.5 k 个音乐-文本对组成的数据集，由人类专家提供丰富的文本描述。

google-research.github.io/seanet/musiclm/examples
Google-research. github.io/seanet/musiclm/examples 谷歌搜索: github.io/seanet/musiclm/example

1. Introduction
1. 引言

Conditional neural audio generation covers a wide range of applications, ranging from text-to-speech (Zen et al., 2013; van den Oord et al., 2016) to lyrics-conditioned music generation (Dhariwal et al., 2020) and audio synthesis from MIDI sequences (Hawthorne et al., 2022b). Such tasks are facilitated by a certain level of temporal alignment between the conditioning signal and the corresponding audio out-put. In contrast, and inspired by progress in text-to-image generation (Ramesh et al., 2021; 2022; Saharia et al., 2022; Yu et al., 2022), recent work has explored generating audio from sequence-wide, high-level captions (Yang et al., 2022; Kreuk et al., 2022) such as “whistling with wind blowing”. While generating audio from such coarse captions repre-sents a breakthrough, these models remain limited to simple acoustic scenes, consisting of few acoustic events over a

有条件的神经音频生成涵盖了广泛的应用，从文本到语音(Zen 等，2013; van den Oord 等，2016)到歌词条件音乐生成(Dhariwal 等，2020)和音频合成 MIDI 序列(honeye 等，2022b)。调节信号和相应的音频输出之间的一定程度的时间对齐促进了这样的任务。相比之下，受到文本到图像生成的进展(Ramesh 等，2021; 2022; Saharia 等，2022; Yu 等，2022)的启发，最近的工作已经探索了从序列范围生成音频，高级标题(Yang 等，2022; Kreuk 等，2022)，如“吹口哨”。虽然从这种粗糙的字幕生成音频代表了一个突破，但是这些模型仍然局限于简单的声学场景，包括一些声学事件

[*]Equal contribution [1]Google Research [2]IRCAM - Sorbonne Universite´ (work done while interning at Google). Correspondence to: Christian Frank <chfrank@google.com>.
1 Google Research 2 ircam-Sorbonne Universite（在 Google 实习期间完成的工作）来信：Christian Frank < chfrank@Google.com > 。

period of seconds. Hence, turning a single text caption into a rich audio sequence with long-term structure and many stems, such as a music clip, remains an open challenge.
秒周期。因此，将单个文本标题转换为具有长期结构和许多词干(如音乐片段)的丰富音频序列，仍然是一个开放的挑战。

AudioLM (Borsos et al., 2022) has recently been proposed as a framework for audio generation. Casting audio synthe-sis as a language modeling task in a discrete representation space, and leveraging a hierarchy of coarse-to-fine audio discrete units (or tokens), AudioLM achieves both high-fidelity and long-term coherence over dozens of seconds. Moreover, by making no assumptions about the content of the audio signal, AudioLM learns to generate realistic audio from audio-only corpora, be it speech or piano music, without any annotation. The ability to model diverse signals suggests that such a system could generate richer outputs if trained on the appropriate data.
AudioLM (Borsos 等，2022)最近被提出作为音频生成的框架。将音频合成作为离散表示空间中的语言建模任务，并利用由粗到细的音频离散单元(或标记)层次结构，AudioLM 在几十秒内实现了高保真和长时间一致性。此外，通过不对音频信号的内容做任何假设，AudioLM 学习从纯音频语料库生成真实的音频，无论是语音还是钢琴音乐，而不需要任何注释。对不同信号进行建模的能力表明，如果对适当的数据进行训练，这样的系统可以产生更丰富的输出。

Besides the inherent difficulty of synthesizing high-quality and coherent audio, another impeding factor is the scarcity of paired audio-text data. This is in stark contrast with the image domain, where the availability of massive datasets contributed significantly to the remarkable image generation quality that has recently been achieved (Ramesh et al., 2021; 2022; Saharia et al., 2022; Yu et al., 2022). Moreover, creat-ing text descriptions of general audio is considerably harder than describing images. First, it is not straightforward to unambiguously capture with just a few words the salient characteristics of either acoustic scenes (e.g., the sounds heard in a train station or in a forest) or music (e.g., the melody, the rhythm, the timbre of vocals and the many instruments used in accompaniment). Second, audio is structured along a temporal dimension which makes sequence-wide captions a much weaker level of annotation than an image caption.
除了固有的难以合成高质量和连贯的音频，另一个阻碍因素是配对音频-文本数据的稀缺性。这与图像领域形成鲜明对比，其中大量数据集的可用性显着促进了最近达到的显着的图像生成质量(Ramesh 等，2021; 2022; Saharia 等，2022; Yu 等，2022)。此外，创建一般音频的文本描述比描述图像要困难得多。首先，仅仅用几个词来清晰地捕捉声学场景(例如，在火车站或森林中听到的声音)或音乐(例如，旋律、节奏、人声的音色和伴奏中使用的许多乐器)的显著特征并不容易。其次，音频是按照时间维度构建的，这使得序列范围的字幕比图像字幕的注释水平要弱得多。

In this work, we introduce MusicLM, a model for genera-ting high-fidelity music from text descriptions. MusicLM leverages AudioLM's multi-stage autoregressive modeling as the generative component, while extending it to incor-porate text conditioning. To address the main challenge of paired data scarcity, we rely on MuLan (Huang et al., 2022), a joint music-text model that is trained to project music and its corresponding text description to representations close to each other in an embedding space. This shared embedding space eliminates the need for captions at training time alto-
在这项工作中，我们介绍了 MusicLM，一个从文本描述生成高保真音乐的模型。MusicLM 利用 AudioLM 的多阶段自回归建模作为生成组件，同时将其扩展到合并文本调节。为了解决配对数据稀缺的主要挑战，我们依赖于 MuLan (Huang et al。，2022)，一种联合音乐-文本模型，该模型被训练用于将音乐及其相应的文本描述投影到嵌入空间中彼此接近的表示。这种共享的嵌入空间消除了在训练时间高音字幕的需要

gether, and allows training on massive audio-only corpora. That is, we use the MuLan embeddings computed from the audio as conditioning during training, while we use MuLan embeddings computed from the text input during inference.
允许在大量只有音频的语料库上进行训练。也就是说，我们在训练期间使用从音频计算的木兰嵌入作为条件，而在推理期间使用从文本输入计算的木兰嵌入。

When trained on a large dataset of unlabeled music, MusicLM learns to generate long and coherent music at 24 kHz, for text descriptions of significant complexity, such as “enchanting jazz song with a memorable saxophone solo and a solo singer” or “Berlin 90s techno with a low bass and strong kick”. To address the lack of evaluation data for this task, we introduce MusicCaps, a new high-quality music caption dataset with 5.5k examples prepared by expert musi-cians, which we publicly release to support future research.
当在大量未标记的音乐数据集上进行训练时，MusicLM 学会以 24khz 生成长而连贯的音乐，用于显著复杂性的文本描述，例如“具有令人难忘的萨克斯独奏和独唱歌手的迷人爵士歌曲”或“具有低音和强劲踢腿的柏林 90 年代电子乐”。为了解决这项任务缺乏评估数据的问题，我们介绍了一个新的高质量音乐字幕数据集 MusicCaps，它包含由专业音乐人准备的 5.5 k 个例子，我们公开发布这个数据集以支持未来的研究。

Our experiments show through quantitative metrics and human evaluations that MusicLM outperforms previous systems such as Mubert (Mubert-Inc, 2022) and Riffu-sion (Forsgren & Martiros, 2022), both in terms of quality and adherence to the caption. Furthermore, since describing some aspects of music with words can be difficult or even impossible, we show how our method supports conditioning signals beyond text. Concretely, we extend MusicLM to accept an additional melody in the form of audio (e.g., whistling, humming) as conditioning to generate a music clip that follows the desired melody, rendered in the style described by the text prompt.
我们的实验表明，通过定量指标和人类评估，MusicLM 优于以前的系统，如 Mubert (Mubert-Inc，2022)和 riffus-sion (Forsgren & Martiros，2022)，在质量和坚持标题方面。此外，由于用文字描述音乐的某些方面可能是困难的，甚至是不可能的，我们展示了我们的方法如何支持文本之外的条件信号。具体来说，我们扩展 MusicLM 以接受音频形式的附加旋律(例如，口哨，哼唱)作为调节，从而生成一个跟随所需旋律的音乐剪辑，以文本提示所描述的风格呈现。

We acknowledge the risks associated with music generation, in particular, the potential misappropriation of creative con-tent. In accordance with responsible model development practices, we conduct a thorough study of memorization by adapting and extending the methodology of Carlini et al. (2022) used for text-based large language models. Our findings show that when feeding MuLan embeddings to MusicLM, the sequences of generated tokens significantly differ from the corresponding sequences in the training set.
我们承认与音乐生成相关的风险，特别是潜在的创造性内容的盗用。根据负责任的模型开发实践，我们通过改编和扩展 Carlini 等人(2022)用于基于文本的大型语言模型的方法，对记忆进行了彻底的研究。我们的研究结果表明，当向 MusicLM 输入花木兰嵌入时，生成的令牌序列与训练集中的相应序列显着不同。

The key contributions of this work are the following:
这项工作的主要贡献如下:

1. We introduce MusicLM, a generative model that pro-duces high-quality music at 24 kHz which is consistent over several minutes while being faithful to a text con-ditioning signal.
我们介绍了 MusicLM，这是一个生成模型，可以以 24khz 的频率生成高质量的音乐，在几分钟内保持一致，同时忠实于文本调节信号。

2. We extend our method to other conditioning signals, such as a melody that is then synthesized according to the text prompt. Furthermore, we demonstrate long and coherent music generation of up to 5-minute long clips.
我们将我们的方法扩展到其他调节信号，比如根据文本提示合成的旋律。此外，我们展示了长达 5 分钟的长片段的长而连贯的音乐生成。

3. We release the first evaluation dataset collected specif-ically for the task of text-to-music generation: Mu-sicCaps is a hand-curated, high-quality dataset of 5.5k music-text pairs prepared by musicians.
我们发布了第一个专门为文本到音乐生成任务收集的评估数据集: Mu-sicCaps 是一个由音乐家准备的 5.5 k 音乐-文本对的手工策划的高质量数据集。

2. Background and Related Work
(二)背景及相关工作

The state-of-the-art in generative modeling for various do-mains is largely dominated either by Transformer-based au-toregressive models (Vaswani et al., 2017) or U-Net-based diffusion models (Ho et al., 2020). In this section, we re-view the related work with an emphasis on autoregressive generative models operating on discrete tokens, which share similarities with MusicLM.

基于 transformer 的自回归模型(Vaswani 等，2017)或基于 u-net 的扩散模型(Ho 等，2020)在很大程度上主导了各种领域的生成建模的最新技术。在本节中，我们回顾了相关的工作，重点介绍了在离散令牌上运行的自回归生成模型，这些模型与 MusicLM 具有相似性。

2.1. Quantization
2.1 量化

Modeling sequences of discrete tokens autoregressively has proven to be a powerful approach in natural language processing (Brown et al., 2020; Cohen et al., 2022) and image or video generation (Esser et al., 2021; Ramesh et al., 2021; Yu et al., 2022; Villegas et al., 2022). Quantization is a key component to the success of autoregressive models for continuous signals, including images, videos, and audio. The goal of quantization is to provide a compact, discrete representation, which at the same time allows for high-fidelity reconstruction. VQ-VAEs (Van Den Oord et al., 2017) demonstrated impressive reconstruction quality at low bitrates in various domains and serve as the underlying quantizer for many approaches.

离散令牌的自回归建模序列已被证明是自然语言处理(Brown 等，2020; Cohen 等，2022)和图像或视频生成(Esser 等，2021; Ramesh 等，2021; Yu 等，2022; Villegas 等，2022)。量化是连续信号(包括图像，视频和音频)的自回归模型成功的关键组成部分。量化的目标是提供一个紧凑的，离散的表示，同时允许高保真重建。Vq-vae (Van Den Oord 等，2017)在各种领域的低比特率下表现出令人印象深刻的重建质量，并作为许多方法的基础量化器。

SoundStream (Zeghidour et al., 2022) is a universal neural audio codec capable of compressing general audio at low bitrates, while maintaining a high reconstruction quality. To achieve this, SoundStream uses residual vector quantization (RVQ), allowing scalability to higher bitrate and quality, without a significant computational cost. More specifically, RVQ is a hierarchical quantization scheme composing a se-ries of vector quantizers, where the target signal is recon-structed as the sum of quantizer outputs. Due to the compo-sition of quantizers, RVQ avoids the exponential blowup in the codebook size as the target bitrate increases. Moreover, the fact that each quantizer is fitted to the residual of coarser quantizers introduces a hierarchical structure to the quan-tizers, where coarser levels are more important for high-fidelity reconstruction. This property is desirable for genera-tion, since the past context can be defined by only attending to the coarse tokens. Recently, SoundStream was extended by EnCodec (Defossez´ et al., 2022) to higher bitrates and stereophonic audio. In this work, we rely on SoundStream as our audio tokenizer, since it can reconstruct 24 kHz mu-sic at 6 kbps with high fidelity.

SoundStream (Zeghidour 等，2022)是一种通用的神经音频编解码器，能够在低比特率下压缩一般音频，同时保持高重建质量。为了实现这一点，SoundStream 使用残差向量量化(RVQ)，允许可伸缩性，以更高的比特率和质量，没有显着的计算成本。具体来说，RVQ 是由一系列矢量量化器组成的分层量化方案，其中目标信号被重构为量化器输出的和。由于量化器的组成，避免了随着目标比特率的增加码本规模的指数爆炸。此外，每个量化器都被拟合到较粗量化器的残差中，这使得量化器具有层次结构，其中较粗的量化器对于高保真重建更为重要。这种特性对于生成来说是可取的，因为过去的上下文可以通过只关注粗糙的 token 来定义。最近，SoundStream 被 EnCodec (Defossez 等，2022)扩展到更高的比特率和立体声音频。在这项工作中，我们依靠 SoundStream 作为我们的音频标记器，因为它可以以 6kbps 的高保真度重建 24khz 的音乐。

2.2. Generative Models for Audio
2.2 音频生成模型

Despite the challenge of generating high-quality audio with long-term consistency, a series of approaches have recently tackled the problem with some success. Jukebox (Dhari-wal et al., 2020), for example, proposes a hierarchy of VQ-VAEs at various time resolutions to achieve high temporal

尽管生成具有长期一致性的高质量音频是一个挑战，但最近一系列方法已经成功地解决了这个问题。例如，Jukebox (Dhari-wal et al。，2020)提出了一个在不同时间分辨率下的 vq-vae 层次结构来实现高时间分辨率

coherence, but the generated music displays noticeable artifacts. PerceiverAR (Hawthorne et al., 2022a), on the other hand, proposes to model a sequence of SoundStream tokens autoregressively, achieving high-quality audio, but compromising the long-term temporal coherence.

连贯性，但生成的音乐显示明显的艺术事实。另一方面,□ 察rar (honeye et al 。 ， 2022a) 提出自動建模一系列 SoundStream 令牌，獲得高品質音效，但是犧牲了長期的時間一致性。

Inspired by these approaches, AudioLM (Borsos et al., 2022) addresses the trade-off between coherence and high-quality synthesis by relying on a hierarchical tokenization and generation scheme. Concretely, the approach distinguishes between two token types: (1) semantic tokens that allow the modeling of long-term structure, extracted from models pretrained on audio data with the objective of masked language modeling; (2) acoustic tokens, provided by a neural audio codec, for capturing fine acoustic details. This allows AudioLM to generate coherent and high-quality speech as well as piano music continuations without relying on tran-scripts or symbolic music representations.

受这些方法的启发，AudioLM (Borsos 等，2022)通过依赖于分层标记和生成方案来解决一致性和高质量合成之间的权衡。具体来说，该方法区分了两种标记类型: (1)语义标记，它们允许对长期结构进行建模，从音频数据上预先训练的模型中提取，目的是对掩蔽语言进行建模; (2)声学标记，它们由神经音频编解码器提供，用于捕捉精细的声学细节。这使得 AudioLM 能够在不依赖转写文本或符号音乐表示的情况下生成连贯和高质量的语音以及钢琴音乐续集。

MusicLM builds on top of AudioLM with three important additional contributions: (1) we condition the generation process on a descriptive text, (2) we show that the conditioning can be extended to other signals such as melody, and

MusicLM 建立在 AudioLM 之上，有三个重要的附加贡献: (1)我们将生成过程限制在一个描述性的文本上，(2)我们表明这种限制可以扩展到其他信号，比如旋律，以及

(3) we model a large variety of long music sequences be-yond piano music (from drum’n’bass over jazz to classical music).

我们模拟了大量的长音乐序列，包括钢琴音乐(从爵士乐的低音鼓到古典音乐)。

2.3. Conditioned Audio Generation
2.3 条件音频生成

Generating audio from a text description (such as “whistling with laughter in the background”) has recently been tack-led by several works. DiffSound (Yang et al., 2022) uses CLIP (Radford et al., 2021) as the text encoder and applies a diffusion model to predict the quantized mel spectrogram features of the target audio based on the text embeddings. AudioGen (Kreuk et al., 2022) uses a T5 (Raffel et al., 2020) encoder for embedding the text, and an autoregressive Transformer decoder for predicting target audio codes produced by EnCodec (Defossez´ et al., 2022). Both approaches rely on a modest amount of paired training data such as AudioSet (Gemmeke et al., 2017) and AudioCaps (Kim et al., 2019) (totalling less than 5k hours after filtering).

根据文本描述生成音频(比如“在背景笑声中吹口哨”)最近已经被一些作品所引导。DiffSound (Yang et al。 ，2022)使用 CLIP (Radford et al。 ，2021)作为文本编码器，并应用扩散模型来预测基于文本嵌入的目标音频的量化 mel 语谱图特征。AudioGen (Kreuk 等，2022)使用 T5(Raffel 等，2020)编码器嵌入文本，并使用自回归变压器解码器预测 EnCodec 产生的目标音频代码(Defossez 等，2022)。这两种方法都依赖于适量的配对训练数据，如 Au-dioSet (Gemmeke 等， 2017) 和 AudioCaps (Kim 等，2019)(过滤后总共少于 5k 小时)。

Closer to MusicLM, there are also works focusing on music generation conditioned on text. In Mubert (Mubert-Inc, 2022), the text prompt is embedded by a Transformer, music tags which are close to the encoded prompt are selected and used to query the song generation API. Based on the selected tags, Mubert generates a combination of sounds, which in turn were generated by musicians and sound designers. This is in contrast to Riffusion (Forsgren & Martiros, 2022), which fine-tunes a Stable Diffusion model (Rombach et al., 2022a) on mel spectrograms of music pieces from a paired music-text dataset. We use both Mubert and Riffusion as baselines for our work, showing that we improve the audio generation quality and adherence to the text description.

接近 MusicLM，还有一些作品关注于以文本为条件的音乐生成。在 Mubert (Mubert-Inc，2022)中，文本提示是由 Transformer 嵌入的，选择靠近编码提示的音乐标签并用于查询歌曲生成 API。根据选中的标签，Mubert 生成一组声音，这些声音又由音乐家和声音设计师生成。这与 Riffusion (Forsgren & Martiros，2022)不同，Riffusion 对来自配对音乐-文本数据集的音乐片段的 mel 声谱图进行了稳定扩散模型(Rombach et al。 ，2022a)的微调。我们使用 Mubert 和 Riffusion 作为我们工作的基线，表明我们提高了音频生成质量和对文本描述的坚持。

Symbolic representations of music (e.g., MIDI) can also be used to drive the generative process as a form of strong conditioning, as demonstrated by Huang et al. (2019); Hawthorne et al. (2019); Engel et al. (2020). MusicLM enables a more natural and intuitive way of providing a con-ditioning signal, for example through a hummed melody, which can also be combined with a text description.
音乐的符号表示(例如 MIDI)也可以用来驱动生成过程作为强制条件的一种形式，如 Huang 等人(2019)；honeye 等人(2019)；Engel 等人(2020)所证明的。MusicLM 能够以更自然和直观的方式提供调节信号，例如通过哼唱的旋律，这也可以与文本描述相结合。

2.4. Text-Conditioned Image Generation
2.4 文本条件图像生成

Precursor to text-conditioned audio synthesis are the text-conditioned image generation models, which made significant progress in quality due to architectural improvements and the availability of massive, high-quality paired train-ing data. Prominent Transformer-based autoregressive approaches include Ramesh et al. (2021); Yu et al. (2022), while Nichol et al. (2022); Rombach et al. (2022b); Saharia et al. (2022) present diffusion-based models. The text-to-image approaches have been extended to generating videos from a text prompt (Wu et al., 2022a; Hong et al., 2022; Vil-legas et al., 2022; Ho et al., 2022).
文本条件音频合成的前身是文本条件图像生成模型，由于架构的改进和海量高质量配对训练数据的可用性，该模型在质量上取得了显著进步。著名的基于变压器的自回归方法包括 Ramesh 等(2021)；Yu 等(2022)，而 Nichol 等(2022)；Rombach 等(2022b)；Saharia 等(2022)目前基于扩散的模型。文本到图像的方法已经扩展到从文本提示生成视频(Wu 等，2022a; Hong 等，2022; Vil-legas 等，2022; Ho 等，2022)。

The closest to our approach among these works is DALL E 2 (Ramesh et al., 2022). In particular, similarly to the way DALL E 2 relies on CLIP (Radford et al., 2021) for text encoding, we also use a joint music-text embed-ding model for the same purpose. In contrast to DALL E 2, which uses a diffusion model as a decoder, our decoder is based on AudioLM. Furthermore, we also omit the prior model mapping text embeddings to music embeddings, such that the AudioLM-based decoder can be trained on an audio-only dataset and the music embedding is simply replaced during inference by the text embedding.
这些工作中最接近我们的方法是 DALL e 2(Ramesh 等，2022)。特别地，类似于 DALL e 2 依赖 CLIP (Radford et al。，2021)进行文本编码的方式，我们也为同样的目的使用了一个联合音乐-文本嵌入模型。与使用扩散模型作为解码器的 DALL e 2 相比，我们的解码器是基于 AudioLM 的。此外，本文还省略了将文本嵌入映射到音乐嵌入的先验模型，使得基于 audiolm 的解码器可以在只有音频的数据集上进行训练，并且在文本嵌入的推理过程中简单地替换音乐嵌入。

2.5. Joint Embedding Models for Music and Text
2.5 音乐和文本的联合嵌入模型

MuLan (Huang et al., 2022) is a music-text joint embedding model consisting of two embedding towers, one for each modality. The towers map the two modalities to a shared embedding space of 128 dimensions using contrastive learning, with a setup similar to (Radford et al., 2021; Wu et al., 2022b). The text embedding network is a BERT (Devlin et al., 2019) pre-trained on a large corpus of text-only data, while we use the ResNet-50 variant of the audio tower.
木兰(Huang et al。，2022)是一个音乐-文本联合嵌入模型，由两个嵌入塔组成，每个嵌入塔对应一个模态。塔使用对比学习将这两种模式映射到 128 维的共享嵌入空间，其设置类似于(Radford 等，2021; Wu 等，2022b)。文本嵌入网络是一个 BERT (Devlin 等，2019)预先训练的大型文本数据库，而我们使用 resnet-50 变体的音频塔。

MuLan is trained on pairs of music clips and their corresponding text annotations. Importantly, MuLan imposes only weak requirements on its training data quality, learn-ing cross-modal correspondences even when the music-text pairs are only weakly associated. The ability to link mu-sic to unconstrained natural language descriptions makes it applicable for retrieval or zero-shot music tagging. In this work, we rely on the pretrained and frozen model of Huang et al. (2022).
花木兰是在成对的音乐剪辑及其相应的文本注释上进行训练的。重要的是，花木兰对其训练数据质量的要求很低，即使在音乐-文本对只是弱关联的情况下也能学习跨模态对应。将音乐与不受约束的自然语言描述相关联的能力使得它适用于检索或零拍摄音乐标注。在这项工作中，我们依赖于 Huang 等人(2022)的预训练和冻结模型。

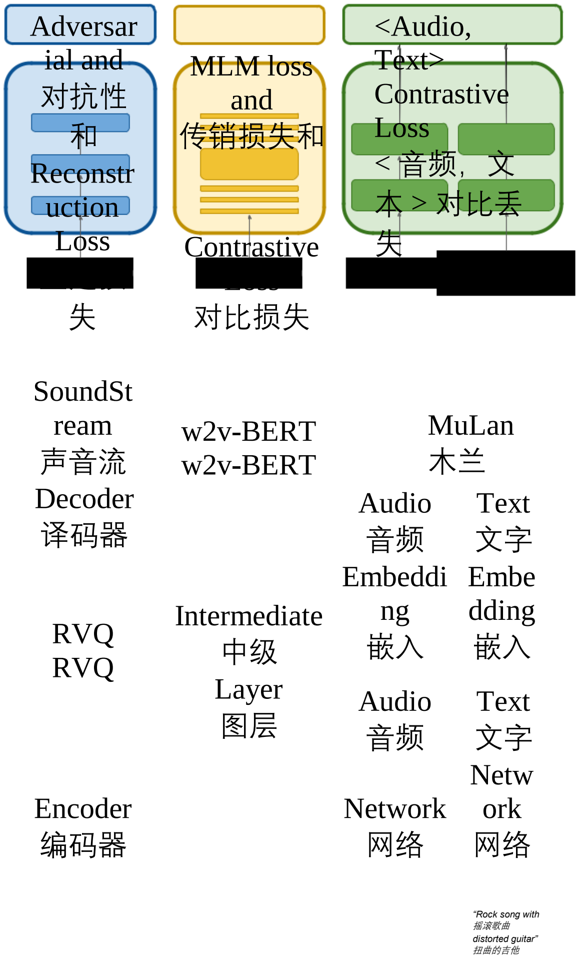


Figure 1. Independent pretraining of the models providing the au-dio and text representations for MusicLM: SoundStream (Zeghi-dour et al., 2022), w2v-BERT (Chung et al., 2021), and MuLan (Huang et al., 2022).

图 1。对为 MusicLM: SoundStream (Zeghi-dour 等，2022)，w2v-BERT (Chung 等，2021)和 milan (Huang 等，2022)提供音频和文本表示的模型进行独立的预训练。

3. Method
3. 方法

In this section, we describe MusicLM and its components. Section 3.1 describes the models that provide the audio representations. Then, we show in Section 3.2 how we use these representations for text-conditioned music generation.

在本节中，我们将介绍 MusicLM 及其组件。第 3.1 节描述了提供音频表示的模型。然后，我们将在 3.2 节展示如何使用这些表示来生成文本条件音乐。

3.1. Representation and Tokenization of Audio and Text
3.1 音频和文本的表示和标记

We use three models for extracting audio representations that will serve for conditional autoregressive music generation, which are illustrated in Figure 1. In particular, by following the approach of AudioLM, we use the self-supervised audio representations of SoundStream (Zeghidour et al., 2022), as acoustic tokens to enable high-fidelity synthesis, and w2v-BERT (Chung et al., 2021), as semantic tokens to facilitate long-term coherent generation. For representing the conditioning, we rely on the MuLan music embedding during training and the MuLan text embedding at inference time. All three of these models are pretrained independently and then frozen, such that they provide the discrete audio and text representations for the sequence-to-sequence modeling.

我们使用三个模型来提取音频表示，这些表示将用于条件自回归音乐生成，如图 1 所示。特别是，通过遵循 AudioLM 的方法，我们使用 SoundStream 的自我监督音频表示(Zeghidour 等，2022)作为声学标记来实现高保真合成，w2v-BERT (Chung 等，2021)作为语义标记以促进长期连贯的生成。为了表示条件，我们依靠在训练期间的花木兰音乐嵌入和在推断时的花木兰文本嵌入。所有这三个模型都是独立的预训练，然后冻结，这样它们就为序列到序列建模提供了离散的音频和文本表示。

SoundStream. We use a SoundStream model for 24 kHz monophonic audio with a striding factor of 480, resulting in 50 Hz embeddings. The quantization of these embeddings is learned during training by an RVQ with 12 quantizers, each with a vocabulary size of 1024. This results in a bitrate of 6 kbps, where one second of audio is represented by 600 to-kens. We refer to these as acoustic tokens, denoted by A.

SoundStream.我们对 24khz 的单声道音频使用 SoundStream 模型，跨度因子为 480，嵌入频率为 50hz。这些嵌入的量化是在训练期间由具有 12 个量化器的 RVQ 学习的，每个量化器的词汇量为 1024。这导致了 6kbps 的比特率，其中一秒的音频由 600 to-kens 表示。我们称之为声学令牌，用 a 表示。

w2v-BERT. Similarly to AudioLM, we use an intermedi-ate layer of the masked-language-modeling (MLM) mod-ule of a w2v-BERT model with 600M parameters. After pretraining and freezing the model, we extract embeddings from the 7th layer and quantize them using the centroids of a learned k-means over the embeddings. We use 1024 clus-ters and a sampling rate of 25 Hz, resulting in 25 semantic tokens for every second of audio, denoted by S.

w2v-BERT.与 AudioLM 类似，我们使用具有 600m 参数的 w2v-BERT 模型的掩蔽语言建模(MLM)模块的中间层。在预训练和冻结模型之后，我们从第 7 层提取嵌入并使用嵌入上的学习 k 均值的质心对其进行量化。我们使用了 1024 个聚类和 25hz 的采样率，每秒音频产生 25 个语义标记，用 s 表示。

MuLan. To train MusicLM, we extract the representation of the target audio sequence from the audio-embedding network of MuLan. Note that this representation is continuous and could be directly used as a conditioning signal in Transformer-based autoregressive models. However, we opt for quantizing the MuLan embeddings in such a way that both the audio and the conditioning signal have a homogeneous representation based on discrete tokens, aiding further research into autoregressively modeling the conditioning signal as well.

木兰。为了训练 MusicLM，我们从 MusicLM 的音频嵌入网络中提取目标音频序列的表示。注意，这种表示是连续的，可以直接用作基于变压器的自回归模型中的调理信号。然而，我们选择量化木兰嵌入这样一种方式，音频和调理信号都有一个基于离散令牌的同质表示，有助于进一步研究自回归建模的调理信号。

Since MuLan operates on 10-second audio inputs and we need to process longer audio sequences, we calculate the audio embeddings on 10-second windows with 1-second stride and average the resulting embeddings. We then discretize the resulting embedding by applying an RVQ with 12 vector quantizers, each with a vocabulary size of 1024. This process yields 12 MuLan audio tokens M_A for an audio sequence. During inference, we use as conditioning the MuLan text embedding extracted from the text prompt, and quantize it with the same RVQ as the one used for the audio embeddings, to obtain 12 tokens M_T .

由于 MuLan 在 10 秒的音频输入上运行，并且我们需要处理更长的音频序列，所以我们以 1 秒的步长计算 10 秒窗口上的音频嵌入，并对得到的嵌入进行平均。然后，我们通过应用具有 12 个矢量量化器的 RVQ 将所得嵌入离散化，每个矢量量化器的词汇表大小为 1024。这个过程为音频序列产生 12 个 MuLan 音频令牌 MA。在推理过程中，我们使用从文本提示中提取的木兰文本嵌入作为条件，并使用与音频嵌入相同的 RVQ 对其进行量化，得到 12 个标记的机器翻译。

Conditioning on M_A during training has two main advan-tages. First, it allows us to easily scale our training data, since we are not limited by the need of text captions. Sec-ond, by exploiting a model like MuLan, trained using a contrastive loss, we increase the robustness to noisy text descriptions.

在训练期间对 MA 进行调节有两个主要优点。首先，它允许我们轻松地扩展我们的训练数据，因为我们不受文本标题的限制。其次，通过使用一个像花木兰这样的模型，使用对比损失进行训练，我们增加了对噪声文本描述的鲁棒性。

3.2. Hierarchical Modeling of Audio Representations
3.2 音频表示的分层建模

We combine the discrete audio representations presented above with AudioLM to achieve text-conditioned music generation. For this, we propose a hierarchical sequence-to-sequence modeling task, where each stage is modeled autoregressively by a separate decoder-only Transformer. The proposed approach is illustrated in Figure 2.

我们将上面提到的离散音频表示与 AudioLM 相结合，以实现文本条件下的音乐生成。为此，我们提出了一个层次化的序列到序列建模任务，其中每个阶段都由一个单独的只解码器的变换器自回归地建模。所提出的方法如图 2 所示。

The first stage is the semantic modeling stage, which learns the mapping from the MuLan audio tokens to the seman-tic tokens S, by modeling the distribution $p(S_{t}jS_{<t}; M_{A})$, where t is the position in the sequence corresponding to a the time step. The second stage is the acoustic modeling stage, where the acoustic tokens A_q are predicted condi-tioned on both the MuLan audio tokens and the semantic tokens, modeling the distribution $p(A_{t}jA_{<t}; S; M_{A})$.

第一阶段是语义建模阶段，通过建模分布 p (StjS < t; MA)，学习从木兰音频标记到语义标记 s 的映射，其中 t 是时间步对应的序列中的位置。第二阶段是声学建模阶段，声学标记 Aq 的预测条件同时包括木兰语音标记和语义标记，建模分布 p (AtjA < t; s; MA)。

Notably, to avoid long token sequences, AudioLM proposed to further split the acoustic modeling stage into a coarse and fine modeling stage. We rely on the same approach, where the coarse stage models the first four levels from the output of the SoundStream RVQ, and the fine stage models the re-maining eight — we refer to Borsos et al. (2022) for details.

值得注意的是，为了避免长的令牌序列，AudioLM 提出进一步将声学建模阶段划分为粗和细两个建模阶段。我们依赖于同样的方法，其中粗级模拟 SoundStream RVQ 输出的前四个级别，精级模拟剩余的八个级别一一详情请参阅 Borsos 等人(2022)。

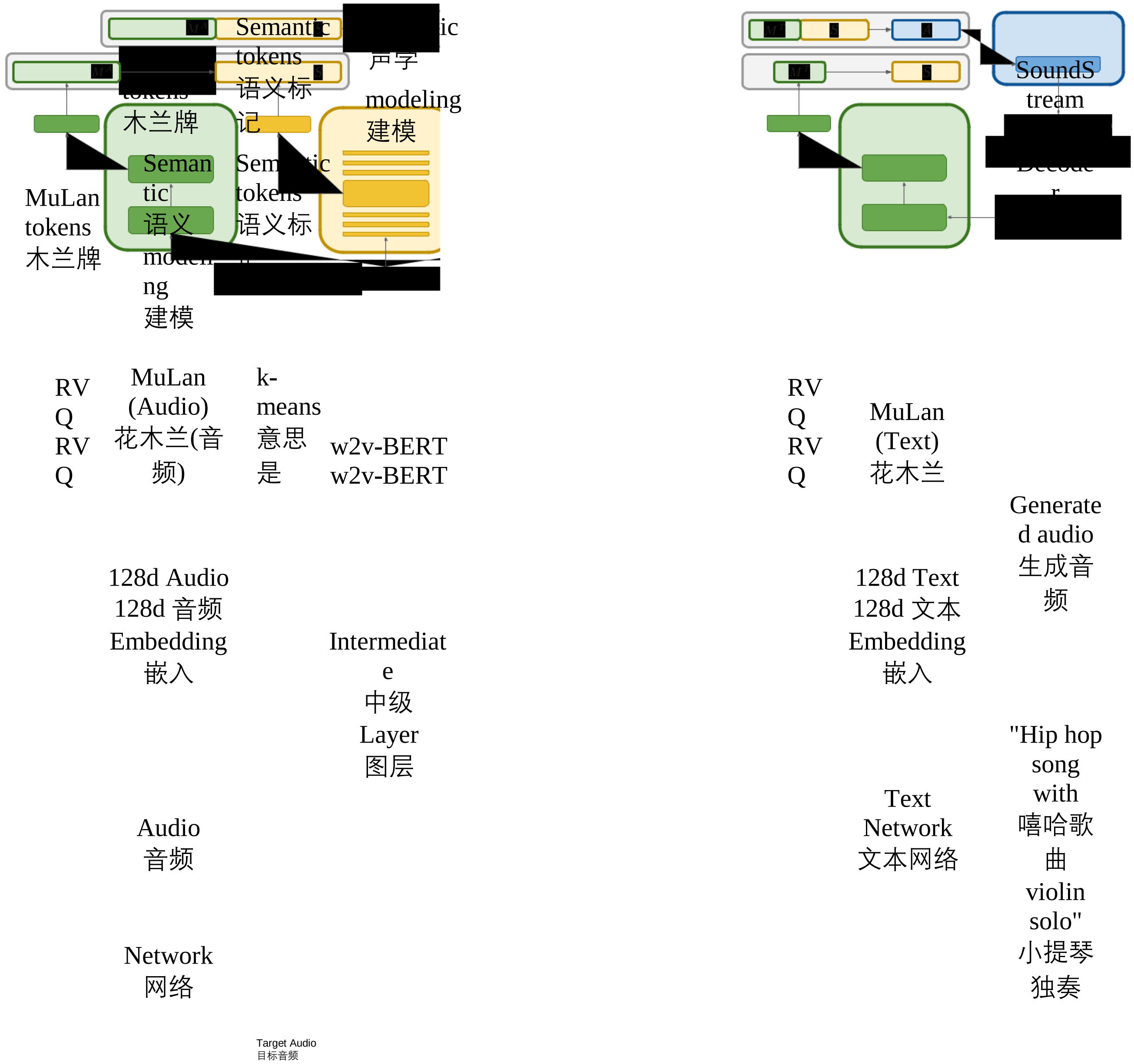


Figure 2. Left: During training we extract the MuLan audio tokens, semantic tokens, and acoustic tokens from the audio-only training set. In the semantic modeling stage, we predict semantic tokens using MuLan audio tokens as conditioning. In the subsequent acoustic modeling stage, we predict acoustic tokens, given both MuLan audio tokens and semantic tokens. Each stage is modeled as a sequence-to-sequence task using decoder-only Transformers. Right: During inference, we use MuLan text tokens computed from the text prompt as conditioning signal and convert the generated audio tokens to waveforms using the SoundStream decoder.
图 2。左图: 在训练过程中，我们从只有音频的训练集中提取木兰音频标记、语义标记和声学标记。在语义建模阶段，我们使用花木兰音频标记作为条件来预测语义标记。在随后的声学建模阶段，我们预测声学标记，给定木兰音频标记和语义标记。每个阶段都被建模为一个序列到序列的任务，使用只有解码器的变换器。右: 在推理过程中，我们使用从文本提示计算出来的 MuLan 文本标记作为调节信号，并使用 SoundStream 解码器将生成的音频标记转换为波形。

4. Experimental Setup
实验装置

4.1. Models
模型

We use decoder-only Transformers for modeling the seman-tic stage and the acoustic stages of AudioLM. The models share the same architecture, composed of 24 layers, 16 atten-tion heads, an embedding dimension of 1024, feed-forward layers of dimensionality 4096, dropout of 0.1, and relative positional embeddings (Raffel et al., 2020), resulting in 430M parameters per stage.

我们使用只有解码器的变压器来建模 AudioLM 的语义阶段和声学阶段。这些模型具有相同的架构，由 24 层，16 个注意头，1024 个嵌入维度，维度 4096 的前馈层，0.1 的丢失和相对位置嵌入(Raffel 等，2020)组成，每个阶段 430m 参数。

4.2. Training and Inference
4.2. 训练和推理

By relying on pretrained and frozen MuLan, we need audio-only data for training the other components of MusicLM. We train SoundStream and w2v-BERT on the Free Music Archive (FMA) dataset (Defferrard et al., 2017), whereas the tokenizers and the autoregressive models for the seman-tic and acoustic modeling stages are trained on a dataset containing five million audio clips, amounting to 280k hours of music at 24 kHz. Each of the stages is trained with multi-ple passes over the training data. We use 30 and 10-second random crops of the target audio for the semantic stage and the acoustic stage, respectively. The AudioLM fine acoustic modeling stage is trained on 3-second crops.

通过依靠预先训练和冻结的花木兰，我们需要只有音频的数据来训练 MusicLM 的其他组件。我们在免费音乐档案馆(FMA)数据集上训练 SoundStream 和 w2v-BERT (Defferrard 等，2017)，而语义和声学建模阶段的标记器和自回归模型在包含 500 万个音频剪辑的数据集上进行训练，总计 28 万小时的音乐在 24khz。每个阶段都通过多次通过训练数据进行训练。我们分别为语义阶段和声学阶段使用 30 秒和 10 秒的目标音频随机作物。声学建模阶段使用 3 秒的随机剪辑进行训练。

During inference, we make use of the joint embedding space between audio and text learned by MuLan, that is, we substitute M_A with M_T . We then follow the stages described above and obtain A given M_T . We use temperature sam-pling for the autoregressive sampling in all stages, with temperature of 1.0 for the semantic modeling stage, 0.95 and 0.4 for the coarse and fine acoustic modeling stages respec-tively. These temperature values were chosen based on sub-jective inspection to provide a good trade-off between diver-sity and temporal consistency of the generated music.

在推理过程中，利用木兰学习到的音频和文本之间的联合嵌入空间，即用机器学习代替机器学习。然后，我们按照上述步骤获得给定的机器翻译。所有阶段的自回归采样都使用温度采样，语义建模阶段的温度为 1.0，粗声学建模阶段的温度为 0.95，精声学建模阶段的温度为 0.4。这些温度值是基于主观检查选择的，以在生成的音乐的多样性和时间一致性之间提供良好的平衡。

4.3. Evaluation Dataset
4.3 评估数据集

To evaluate MusicLM, we prepare MusicCaps, a high-quality music caption dataset, which we make publicly available.[1] This dataset includes 5.5k music clips from Au-dioSet (Gemmeke et al., 2017), each paired with correspond-ing text descriptions in English, written by ten professional musicians. For each 10-second music clip, MusicCaps pro-vides: (1) a free-text caption consisting of four sentences on average, describing the music and (2) a list of music aspects, describing genre, mood, tempo, singer voices, instrumenta-tion, dissonances, rhythm, etc. On average, the dataset in-cludes eleven aspects per clip. See Appendix A for a few caption and aspect list examples.

为了评估 MusicLM，我们准备了 MusicCaps，这是一个高质量的音乐标题数据集，我们对外公开。1 这个数据集包括来自 AudioSet (Gemmeke et al。，2017)的 5.5 k 个音乐片段，每个片段都配有相应的英文文本描述，由十位专业音乐家撰写。对于每个 10 秒的音乐片段，MusicCaps 提供: (1)一个平均包含 4 个句子的自由文本标题，描述音乐；(2)一个音乐方面的列表，描述体裁、情绪、节奏、歌手声音、乐器、不协调音、节奏等等。平均来说，每个音乐片段包含 11 个方面。请参见附录 a 中的一些标题和方面列表示例。

MusicCaps complements AudioCaps (Kim et al., 2019), as they both contain audio clips from AudioSet with corresponding textual descriptions. However, while AudioCaps contains non-music content, MusicCaps focuses exclusively on music and includes highly detailed expert-provided annotations. The examples are extracted from both the train and eval split of AudioSet, covering a diverse distribution of genres, as detailed in Appendix A. MusicCaps also pro-vides a genre-balanced split of the data with 1k examples.

MusicCaps 补充了 AudioCaps (Kim et al。，2019)，因为它们都包含来自 AudioSet 的音频片段和相应的文本描述。然而，AudioCaps 包含非音乐内容，而 MusicCaps 专注于音乐，并包含高度详细的专家提供的注释。这些例子都是从 AudioSet 的 train 和 eval 分类中提取出来的，涵盖了不同流派的分布，详见附录 a。 MusicCaps 也提供了一个 1k 例子的流派平衡数据分类。

4.4. Metrics
4.4 度量

We compute different metrics to evaluate MusicLM, capturing two important aspects of music generation: the audio quality and the adherence to the text description.

我们计算不同的指标来评估 MusicLM，捕获音乐生成的两个重要方面: 音频质量和对文本描述的遵循。

Frechet´ Audio Distance (FAD). The Frechet´ Audio Distance (Kilgour et al., 2019) is a reference-free audio quality metric, which correlates well with human perception. Models producing samples with a low FAD score are expected
Frechet 音频距离(FAD)。Frechet 音频距离(Kilgour et al。 ，2019)是一种无参考的音频质量度量标准，与人类感知相关性很好。预期产生 FAD 评分低的样品的模型

[1] kaggle.com/datasets/googleai/musiccaps
来源: kaggle.com/dataset/googleai/musiccaps

to generate plausible audio. However, the generated sam-ples might not necessarily adhere to the text description pro-vided as conditioning.
然而，所产生的样本未必符合提供作为条件作用的文字说明。

We report the FAD based on two audio embedding models, both of which are publicly available: (1) Trill[2] (Shor et al., 2020), which is trained on speech data, and (2) VGGish[3], (Hershey et al., 2017) which is trained on the YouTube-8M audio event dataset (Abu-El-Haija et al., 2016). Because of the difference in training data, we expect the models to measure different aspects of the audio quality (speech and non-speech, respectively).
我们报告了基于两个音频嵌入模型的 FAD，这两个模型都是公开可用的: (1) Trill2(Shor 等，2020)，它是在语音数据上训练的，和(2) VGGish3，(Hershey 等，2017)在 YouTube-8M 音频事件数据集上训练(Abu-El-Haija 等，2016)。由于训练数据的差异，我们期望模型测量音频质量的不同方面(分别是语音和非语音)。

KL Divergence (KLD). There is a many-to-many relationship between text descriptions and music clips compatible with them. It is therefore not possible to directly compare the generated music with the reference at the level of the audio waveform. To assess the adherence to the input text description, we adopt a proxy method similar to the one proposed in Yang et al. (2022); Kreuk et al. (2022). Specifically, we use a LEAF (Zeghidour et al., 2021) clas-sifier trained for multi-label classification on AudioSet, to compute class predictions for both the generated and the reference music and measure the KL divergence between probability distributions of class predictions. When the KL-divergence is low, the generated music is expected to have similar acoustic characteristics as the reference music, according to the classifier.
KL 发散(KLD)。文本描述和与之相容的音乐片段之间存在多对多的关系。因此，不可能直接比较生成的音乐和音频波形级别的参考。为了评估对输入文本描述的依从性，我们采用类似于 Yang 等人(2022)；Kreuk 等人(2022)提出的代理方法。具体而言，我们使用 LEAF (Zeghidour 等，2021)分类器在 AudioSet 上进行多标签分类训练，以计算生成的和参考音乐的类预测，并测量类预测的概率分布之间的 KL 散度。当 kl 散度低时，根据分类器，生成的音乐预计具有与参考音乐相似的声学特征。

MuLan Cycle Consistency (MCC). As a joint music-text embedding model, MuLan can be used to quantify the similarity between music-text pairs. We compute the MuLan embeddings from the text descriptions in MusicCaps as well as the generated music based on them, and define the MCC metric as the average cosine similarity between these embeddings.
木兰循环一致性(MCC)。作为一种音乐-文本联合嵌入模型，木兰可以用来量化音乐-文本对之间的相似度。我们从 MusicCaps 中的文本描述以及基于它们生成的音乐中计算木兰嵌入，并定义 MCC 度量为这些嵌入之间的平均余弦距离。

Qualitative evaluation. Ultimately, we rely on subjective tests to evaluate the adherence of generated samples to the text description. We set up an A-vs-B human rating task, in which raters are presented with the text description and two samples of music generated by two different models, or one model and the reference music. There are five possible an-swers: strong or weak preference for A or B, and no prefer-ence. The raters are instructed not to take the music quality into account when making their decision, because this as-pect of the evaluation is already covered by the FAD metric.
定性评估。最终，我们依靠主观测试来评估生成的样本是否符合文本描述。我们设置了一个 A-vs-B 的人类评分任务，评分者看到文本描述和两个不同模型生成的音乐样本，或者一个模型和参考音乐。有 5 种可能的答案: 对 a 或 b 的强或弱偏好，以及没有偏好。评分者被指示在做决定时不要考虑音乐质量，因为这方面的评价已经被 FAD 度量所涵盖。

We consider the output of n different models, in addition to the reference music, thus a total of n + 1 conditions and n(n + 1)=2 pairs. To aggregate the results of the pairwise tests and rank conditions, we count the number of “wins”,
我们考虑 n 个不同模型的输出，除了参考音乐，因此总共有 n + 1 个条件和 n (n + 1) = 2 对。为了聚合成对检验和等级条件的结果，我们计算“胜出”的数量，

[2] tfhub.dev/google/nonsemantic-speech-benchmark/trill/3
Dev/google/non-semantic-speech-benchmark/trill/3

[3] tfhub.dev/google/vggish/1
图片来源: tfhub.dev/google/vggish/1

that is, how often a condition is strongly or weakly preferred. The samples are selected from the genre-balanced 1k subset of our evaluation data.

也就是说，一个条件是强优先还是弱优先的频率。样本是从我们的评估数据的类型平衡的 1k 子集中选取的。

Training data memorization. Large language models have the capacity to memorize patterns seen in the training data (Carlini et al., 2020). We adapt the methodology used in Carlini et al. (2022) to study the extent to which MusicLM might memorize music segments. We focus on the first stage, responsible for semantic modeling. We select N examples at random from the training set. For each example, we feed to the model a prompt which includes the MuLan audio tokens M_A followed by a sequence of the first T semantic tokens S, with T 2 f0; : : : ; 250g, corresponding to up to 10 seconds. We use greedy decoding to generate a continuation of 125 se-mantic tokens (5 seconds) and we compare the generated tokens to the target tokens in the dataset. We measure exact matches as the fraction of examples for which generated and target tokens are identical over the whole sampled segment.

训练数据记忆。大型语言模型具有记忆训练数据中的模式的能力(Carlini 等，2020)。我们采用 Carlini 等人(2022)使用的方法来研究 MusicLM 可能记忆音乐片段的程度。我们专注于第一阶段，负责语义建模。我们从训练集中随机选取 n 个样本。对于每个示例，我们向模型提供一个提示，其中包括木兰音频标记 MA，后面是第一个 t 语义标记 s 的序列，其中 t 为 2 f0; : : ; 250g，对应于最多 10 秒。我们使用贪婪解码生成 125 个语义标记的延续(5 秒)，并将生成的标记与数据集中的目标标记进行比较。我们将精确匹配度量为在整个采样段中生成的和目标令牌相同的例子的比例。

In addition, we propose a methodology to detect approx-imate matches, based on the observation that sequences of seemingly different tokens might lead to acoustically sim-ilar audio segments. Namely, we compute the histogram of semantic token counts over the corresponding vocab-ulary f0; : : : ; 1023g from both the generated and target tokens, and define a matching cost measure between his-tograms as follows. First, we compute the distance matrix between pairs of semantic tokens, which is populated by the Euclidean distances between the corresponding k-means centroids used to quantize w2v-BERT to semantic tokens (see Section 3.1). Then, we solve an optimal transport prob-lem to find the matching cost between a pair of histograms using the Sinkhorn algorithm (Cuturi, 2013), considering only the sub-matrix corresponding to non-zero token counts in the two histograms. To calibrate the threshold used to determine whether two sequences might be approximate matches, we construct negative pairs by permuting the examples with target tokens and measure the empirical distribution of matching costs for such negative pairs. We set the match threshold to 0:85, which leads to less than 0.01% false positive approximate matches.

此外，我们提出了一种方法来检测近似匹配，基于观察到的序列似乎不同的标记可能导致声学相似的音频片段。即，我们计算相应词汇表 f0; : : ; 1023g 上生成的和目标标记的语义标记计数直方图，并定义直方图之间的匹配代价度量如下。首先，我们计算语义标记对之间的距离矩阵，该矩阵由用于将 w2v-BERT 量化为语义标记的相应 k-means 质心之间的欧氏距离填充(参见 3.1 节)。然后，我们使用 Sinkhorn 算法(Cuturi，2013)求解一个最优传输问题来找出两个直方图之间的匹配代价，只考虑两个直方图中非零 token 计数对应的子矩阵。为了校准用于判断两个序列是否可能是近似匹配的阈值，我们通过用目标标记置换样本来构造否定对，并测量这种否定对的匹配代价的经验分布。我们将匹配阈值设置为 0:85，这导致低于 0.01% 的假阳性近似匹配。

5. Results

5. 结果

We evaluate MusicLM by comparing it with two recent baselines for music generation from descriptive text, namely Mubert (Mubert-Inc, 2022) and Riffusion (Forsgren & Mar-tiros, 2022). In particular, we generate audio by querying the Mubert API,[4] and by running inference on the Riffusion model.[5] We perform our evaluations on MusicCaps, the eval-uation dataset we publicly release together with this paper.

我们通过将 MusicLM 与来自描述性文本的两个最近的音乐生成基线，即 Mubert (Mubert-Inc，2022)和 Riffusion (Forsgren & Mar-tiros，2022)进行比较来评估 MusicLM。特别地，我们通过查询 Mubert API，4 和在 Riffusion 模型上运行推理来生成音频。5 我们在 MusicCaps 上执行我们的评估，MusicCaps 是我们与本文一起公开发布的评估数据集。

[4] github.com/MubertAI (accessed in Dec 2022 and Jan 2023)

github.com/MubertAI (2022 年 12 月和 2023 年 1 月访问)

[5] github.com/riffusion/riffusion-app (accessed on Dec 27, 2022)

5github.com/riffusion/riffusion-app (2022 年 12 月 27 日访问)

Table 1. Evaluation of generated samples using captions from the MusicCaps dataset. Models are compared in terms of audio quality, by means of Frechet´ Audio Distance (FAD), and faithfulness to the text description, using Kullback–Leibler Divergence (KLD) and MuLan Cycle Consistency (MCC), and counts of wins in pairwise human listening tests (Wins).
表 1。使用 MusicCaps 数据集的标题评估生成的样本。通过 Frechet 音频距离(FAD)和对文本描述的忠实度，使用 Kullback-Leibler 散度(KLD)和花木兰循环一致性(MCC)以及成对人类听力测试(Wins)中的胜利计数，在音频质量方面比较模型。

MODEL 模型	FADTR ILL # FADTR ILL #	FADV GG # #	KLD # 吉隆坡 #	MCC " 管理公司	WIN S " 胜利”
RIFFUSI ON 翻滚	0.76	13.4	1.19	0.34	158
MUBER T 穆伯特	0.45	9.6	1.58	0.32	97
MUSIC LM MUSIC LM 音乐	0.44	4.0	1.01	0.51	312
MUSIC CAPS MUSIC CAPS 音乐大帽	-	-	-	-	472

Comparison to baselines. Table 1 reports the main quan-titative and qualitative results of this paper. In terms of au-dio quality, as captured by the FAD metrics, on FAD$_{\text{VGG}}$ MusicLM achieves better scores than Mubert and Riffusion. On FAD$_{\text{Trill}}$, MusicLM scores similarly to Mubert (0.44 vs. 0.45) and better than Riffusion (0.76). We note that, ac-cording to these metrics, MusicLM is capable of generating high-quality music comparable to Mubert, which relies on pre-recorded sounds prepared by musicians and sound de-signers. In terms of faithfulness to the input text description, as captured by KLD and MCC, MusicLM achieves the best scores, suggesting that it is able to capture more informa-tion from the text descriptions compared to the baselines.
与基线的比较。表 1 报告了本文的主要定量和定性结果。在音频质量方面，如 FAD 指标所捕获的，在 FADVGG MusicLM 上取得了比 Mubert 和 Riffusion 更好的分数。在 FADTrill 上，MusicLM 得分与 Mubert 相似(0.44 vs. 0.45) ，优于 Riffusion (0.76)。我们注意到，根据这些指标，MusicLM 能够生成与 Mubert 相当的高质量音乐，Mubert 依赖于音乐家和声音设计师准备的预先录制的声音。在对输入文本描述的忠实度方面，由 KLD 和 MCC 捕获，MusicLM 取得了最好的分数，表明它能够从文本描述中捕获更多的信息相比，基线。

We further supplement our evaluation of text faithfulness with a human listening test. Participants are presented with two 10-second clips and a text caption, and asked which clip is best described by the text of the caption on a 5-point Likert scale. We collect 1200 ratings, with each source involved in 600 pair-wise comparisons. Table 1 reports the total number of “wins”, that is, counting how often the human raters preferred a model in a side-by-side comparison. MusicLM is clearly preferred over both baselines, while there is still a measurable gap to the ground truth reference music. Full details of the listening study can be found in Appendix B.
我们进一步通过人工听力测试来补充我们对文本忠实度的评估。参与者会看到两个 10 秒的片段和一个文字说明，然后问他们哪个片段最适合用李克特 5 分制来描述。我们收集了 1200 个评分，每个来源涉及 600 个成对比较。表 1 报告了“胜利”的总数，也就是说，计算人类评估者在并排比较中偏好某个模型的频率。MusicLM 显然比这两个基线都更受欢迎，但是与参考音乐相比，还是有一定的差距。听力研究的全部细节可以在附录 b 中找到。

Listening to examples in which the ground truth was pre-ferred over MusicLM reveals the following patterns: (1) cap-tions are extremely detailed, referring to more than five in-struments or describing non musical aspects such as “wind, people talking“; (2) captions describe temporal ordering of the audio being played; (3) negations are used, which are not well captured by MuLan.
听一听这些例子，我们会发现《花木兰》更倾向于基本事实，而不是音乐：(1)标题非常详细，涉及 5 种以上的乐器，或者描述非音乐的方面，比如“风，人们在说话”；(2)标题描述了音频播放的时间顺序；(3)使用了否定，这是《花木兰》没有很好地捕捉到的。

Overall, we conclude that: (1) our approach is able to cap-ture fine-grained information from the rich free-text cap-tions of MusicCaps; (2) the KLD and MCC metrics provide a quantitative measure of the faithfulness to the text descrip-tion, which is in accordance with the human rating study.
实验结果表明：(1)该方法能够从音乐大写字母的自由文本标注中获取细粒度信息；(2)该方法提供了一种量化的文本描述忠实度度量方法，与人工评分研究结果一致。

Importance of semantic tokens. To understand the use-fulness of decoupling semantic modeling from acoustic mod-
语义标记的重要性

eling, we train a Transformer model which directly predicts coarse acoustic tokens from MuLan tokens, by modeling p(AtjA<t; MA). We observe that while the FAD metrics are comparable (0.42 FADTrill and 4.0 FADVGG), KLD and MCC scores worsen when removing the semantic modeling stage. In particular the KLD score increases from 1.01 to 1.05, and the MCC score decreases from 0.51 to 0.49, indi-cating that semantic tokens facilitate the adherence to the text description. We also confirm this qualitatively by listen-ing to the samples. In addition, we observe degradation in long term structure.

通过建模 p (AtjA < t; MA)，我们训练了一个 Transformer 模型，该模型可以直接预测木兰令牌的粗声令牌。我们观察到，虽然 FAD 指标是可比较的(0.42 FADTrill 和 4.0 FADVGG)，但是当去除语义建模阶段时，KLD 和 MCC 分数恶化。特别是 KLD 得分从 1.01 增加到 1.05，MCC 得分从 0.51 降低到 0.49，表明语义标记有助于遵守文本描述。我们还通过听取样本来定性地证实这一点。此外，我们观察到长期结构的退化。

Information represented by audio tokens. We conduct additional experiments to study the information captured by the semantic and the acoustic tokens. In the first study, we fix the MuLan text tokens as well as the semantic tokens, running the acoustic modeling stage multiple times to gen-erate several samples. In this case, by listening to the gen-erated music, it is possible to observe that the samples are diverse, yet they tend to share the same genre, rhythmical properties (e.g., drums), and part of the main melody. They differ in terms of specific acoustic properties (e.g., level of reverb, distortion) and, in some cases, different instruments with a similar pitch range can be synthesized in different examples. In the second study, we fix only the MuLan text tokens and generate both the semantic and acoustic tokens. In this case, we observe a much higher level of diversity in terms of melodies and rhythmic properties, still coher-ent with the text description. We provide samples from this study in the accompanying material.

由音频标记表示的信息。我们进行了额外的实验来研究语义标记和声学标记所捕获的信息。在第一个研究中，我们固定木兰文本标记和语义标记，多次运行声学建模阶段，生成多个样本。在这种情况下，通过聆听生成的音乐，可以观察到样本是多样化的，但它们往往具有相同的体裁、节奏特性(如鼓)和主旋律的一部分。它们在具体的声学特性(如混响程度、失真)方面有所不同，在某些情况下，具有相似音高范围的不同乐器可以在不同的例子中合成。在第二个研究中，我们只固定木兰文本标记，并生成语义和声学标记。在这种情况下，我们观察到在旋律和节奏属性方面的更高水平的多样性，仍然与文本描述相一致。我们在附带的材料中提供了这项研究的样本。

Memorization analysis. Figure 3 reports both exact and approximate matches when the length of the semantic token prompt is varied between 0 and 10 seconds. We observe that the fraction of exact matches always remains very small (< 0:2%), even when using a 10 second prompt to generate a continuation of 5 seconds. Figure 3 also in-cludes results for approximate matches, using = 0:85. We can see a higher number of matches detected with this methodology, also when using only MuLan tokens as input (prompt length T = 0) and the fraction of matching exam-ples increases as the length of the prompt increases. We inspect these matches more closely and observe that those with the lowest matching score correspond to sequences characterized by a low level of token diversity. Namely, the average empirical entropy of a sample of 125 semantic to-kens is 4.6 bits, while it drops to 1.0 bits when considering sequences detected as approximate matches with matching score less than 0.5. We include a sample of approximate matches obtained with T = 0 in the accompanying material. Note that acoustic modeling carried out by the second stage introduces further diversity in the generated samples, also when the semantic tokens match exactly.

记忆分析。图 3 报告了当语义标记提示符的长度在 0 到 10 秒之间变化时的精确匹配和近似匹配。我们观察到，即使使用 10 秒的提示生成 5 秒的延续，精确匹配的比例也总是非常小(< 0:2%)。图 3 也包含了近似匹配的结果，使用 = 0:85。我们可以看到使用这种方法检测到更多的匹配，也可以使用仅仅木兰令牌作为输入(提示长度 t = 0)，并且随着提示长度的增加，匹配示例的分数增加。我们更仔细地检查这些匹配，并观察到那些匹配得分最低的序列对应于拥有属性多样性水平较低的序列。即对于 125 个语义标记样本，其平均经验熵为 4.6 bit，而当检测序列为匹配得分小于 0.5 的近似匹配时，其平均经验熵下降到 1.0 bits。我们在附带的材料中包含了一个在 t = 0 时获得的近似匹配的样本。注意，第二阶段进行的声学建模在生成的样本中引入了进一步的多样性，当语义标记完全匹配时也是如此。

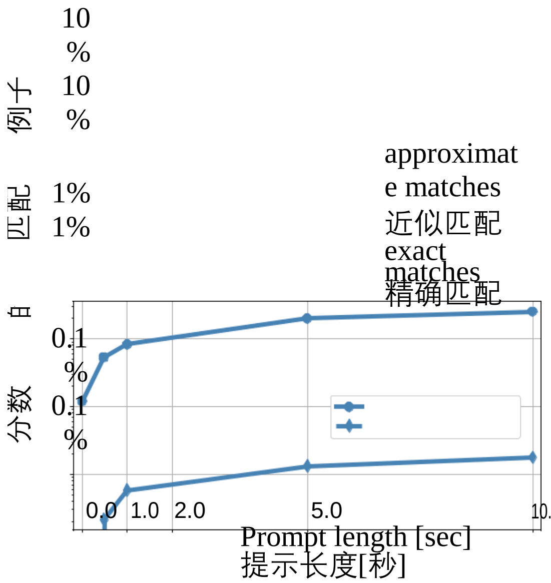


Figure 3. Memorization results for the semantic modeling stage. We compare the semantic tokens generated for 5 seconds of audio to corresponding tokens in the training set, considering exact and approximate matches.

图 3。语义建模阶段的记忆结果。我们将 5 秒钟的音频生成的语义标记与训练集中相应的标记进行比较，考虑精确匹配和近似匹配。

6. Extensions
扩展

Melody conditioning. We extend MusicLM in such a way that it can generate music based on both a text descrip-tion and a melody, which is provided in the form of hum-ming, singing, whistling, or playing an instrument. This requires extending the conditioning signal in a way that cap-tures the target melody. To this end, we create a synthetic dataset composed of audio pairs with matching melodies but different acoustics. To create such pairs, we use differ-ent versions of the same music clip, such as covers, instru-mentals, or vocals. Additionally, we acquire data pairs of people humming and singing. We then train a joint embed-ding model such that when two audio clips contain the same melody, the corresponding embeddings are close to each other. For implementation details we refer to Appendix C.

旋律调节。我们对 MusicLM 进行了扩展，使其能够基于文本描述和旋律生成音乐，这些旋律以哼唱、唱歌、吹口哨或演奏乐器的形式提供。这就需要扩展调节信号，捕捉目标旋律。为此，我们创建了一个由音频对组成的合成数据集，这些音频对具有匹配的旋律，但是不同的声学特性。为了创建这样的音频对，我们使用相同音乐片段的不同版本，比如翻唱、乐器或人声。此外，我们还获取了哼唱的人的数据对。然后训练一个联合嵌入模型，使得当两个音频片段包含相同的旋律时，对应的嵌入是相近的。关于实现细节，我们参考附录 c。

To extract the melody conditioning for MusicLM, we quan-tize the melody embeddings with RVQ, and concatenate the resulting token sequences with the MuLan audio tokens M_A. During inference, we compute melody tokens from the input audio clip and concatenate them with the MuLan text tokens M_T . Based on this conditioning, MusicLM can success-fully generate music which follows the melody contained in the input audio clip, while adhering to the text description.

为了提取 MusicLM 的旋律条件，我们使用 RVQ 对旋律嵌入进行量化，并将得到的标记序列与 MuLan 音频标记 MA 连接。在推断过程中，我们从输入音频剪辑中计算旋律标记，并将它们与 MuLan 文本标记 MT 连接起来。基于这个条件，MusicLM 可以成功地生成跟随输入音频片段中包含的旋律的音乐，同时遵循文本描述。

Long generation and story mode. In MusicLM, gene-ration is autoregressive in the temporal dimension which makes it possible to generate sequences longer than those used during training. In practice, the semantic modeling stage is trained on sequences of 30 seconds. To generate longer sequences, we advance with a stride of 15 seconds, using 15 seconds as prefix to generate an additional 15 sec-onds, always conditioning on the same text description. With this approach we can generate long audio sequences which are coherent over several minutes.

长代和故事模式。在 MusicLM 中，生成在时间维度上是自回归的，这使得生成比训练时更长的序列成为可能。在实际应用中，语义建模阶段是在 30 秒的序列上进行训练的。为了生成更长的序列，我们以 15 秒为步长前进，用 15 秒作为前缀再生成 15 秒，总是以相同的文本描述为条件。使用这种方法，我们可以生成长的音频序列，这些音频序列在几分钟内是连贯的。

With a small modification, we can generate long audio se-quences while changing the text description over time. Bor-rowing from Villegas et al. (2022) in the context of video generation, we refer to this approach as story mode. Con-

稍加修改，我们就可以生成长的音频序列，同时随着时间的推移改变文本描述。Villegas 等(2022)在视频生成的背景下，我们将这种方法称为故事模式。反对

cretely, we compute MT from multiple text descriptions and change the conditioning signal every 15 seconds. The model generates smooth transitions which are tempo consistent and semantically plausible, while changing music context according to the text description.
具体来说，我们从多个文本描述中计算机器翻译，并且每 15 秒改变一次调节信号。该模型在根据文本描述改变音乐上下文的同时，生成节奏一致且语义合理的平滑过渡。

7. Conclusions
结论

We introduce MusicLM, a text-conditioned generative model that produces high-quality music at 24 kHz, consis-tent over several minutes, while being faithful to the text conditioning signal. We demonstrate that our method outper-forms baselines on MusicCaps, a hand-curated, high-quality dataset of 5.5k music-text pairs prepared by musicians.
我们介绍 MusicLM，这是一个文本调节的生成模型，它以 24khz 生成高质量的音乐，持续几分钟，同时忠实于文本调节信号。我们证明了我们的方法超过了 MusicCaps 上的基线，MusicCaps 是一个由音乐家准备的 5.5 k 音乐-文本对的手工策划的高质量数据集。

Some limitations of our method are inherited from MuLan, in that our model misunderstands negations and does not adhere to precise temporal ordering described in the text. Moreover, further improvements of our quantitative evaluations are needed. Specifically, since MCC also relies on MuLan, the MCC scores are favorable to our method.
我们的方法的一些局限性是继承自花木兰，因为我们的模型误解了否定，并且没有遵守文本中描述的精确的时间顺序。此外，我们的定量评估还需要进一步改进。具体而言，由于 MCC 也依赖于花木兰，所以 MCC 得分对我们的方法是有利的。

Future work may focus on lyrics generation, along with improvement of text conditioning and vocal quality. Another aspect is the modeling of high-level song structure like introduction, verse, and chorus. Modeling the music at a higher sample rate is an additional goal.
未来的工作可能会集中在歌词生成，以及改善文本条件和声乐质量。另一个方面是高层次歌曲结构的建模，如序言、韵文和合唱。更高采样率的音乐建模是另一个目标。

8. Broader Impact
更广泛的影响

MusicLM generates high-quality music based on a text description, and thus it further extends the set of tools that as-sist humans with creative music tasks. However, there are several risks associated with our model and the use-case it tackles. The generated samples will reflect the biases present in the training data, raising the question about ap-propriateness for music generation for cultures underrepre-sented in the training data, while at the same time also rais-ing concerns about cultural appropriation.
MusicLM 基于文本描述生成高质量的音乐，因此它进一步扩展了辅助人类完成创造性音乐任务的工具集。然而，我们的模型和它处理的用例存在一些风险。生成的样本将反映出训练数据中存在的偏差，提出了对训练数据中代表性不足的文化的音乐生成是否适当的问题，同时也提出了对文化挪用的关注。

We acknowledge the risk of potential misappropriation of creative content associated to the use-case. In accordance with responsible model development practices, we con-ducted a thorough study of memorization, adapting and extending a methodology used in the context of text-based LLMs, focusing on the semantic modeling stage. We found that only a tiny fraction of examples was memorized ex-actly, while for 1% of the examples we could identify an ap-proximate match. We strongly emphasize the need for more future work in tackling these risks associated to music gene-ration — we have no plans to release models at this point.
我们承认与用例相关的创造性内容存在潜在的挪用风险。根据负责任的模型开发实践，我们对基于文本的语义模型的记忆、调整和扩展方法进行了深入的研究，重点是语义建模阶段。我们发现只有一小部分的例子被准确地记住，而对于 1% 的例子，我们可以确定一个近似匹配。我们强烈强调，未来需要开展更多工作，解决与音乐产生相关的风险——目前我们没有发布模型的计划。

References
参考文献

Abu-El-Haija, S., Kothari, N., Lee, J., Natsev, P., Toderici, G., Varadarajan, B., and Vijayanarasimhan, S. Youtube-8m: A large-scale video classification benchmark. arXiv:1609.08675, 2016.

Abu-El-Haija，s.，Kothari，n.，Lee，j.，Natsev，p.，Toderici，g.，Varadarajan，b.，and Vijayanarasimhan，s. Youtube-8m: 大规模视频分类基准。arXiv: 1609.08675,2016.

Borsos, Z., Marinier, R., Vincent, D., Kharitonov, E., Pietquin, O., Sharifi, M., Teboul, O., Grangier, D., Tagliasacchi, M., and Zeghidour, N. Audiolm: a language modeling approach to audio generation. arXiv:2209.03143, 2022.

Borsos，z。，Marinier，r。，Vincent，d。，Kharitonov，e。，Pietquin，o。，Sharifi，m。，Teboul，o。，Grangier，d。，Tagliasacchi，m。，和 Zeghidour，n。 Audiolm: 语言建模方法的音频生成。arXiv: 2209.03143,2022.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. Language models are few-shot learners. In Advances in Neural Information Processing Systems (NeurIPS), 2020.

Brown，T.b.，Mann，b.，Ryder，n.，Subbiah，m.，Kaplan，j.，Dhariwal，p.，Neelakantan，a.，Shyam，p.，Sastry，g.，Askell，a.，Agarwal，s.，Herbert-Voss，a.，Krueger,g.，Henighan，t.，Child，r.，Ramesh，a.，Ziegler，D.m.，Wu，j.，Winter，c.，Hesse，c.，Chen，m.，Sigler，e.，Litwin，m.，Gray,语言模型是少量的学习者。神经信息处理系统进展(NeurIPS)，2020。

Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, U., Oprea, A., and Raffel, C. Extracting training data from large language models, 2020. URL https://arxiv.org/abs/2012.07805.

Carlini，n.，Tramer，f.，Wallace，e.，Jagielski，m.，Herbert-Voss，a.，Lee，k.，Roberts，a.，Brown，t.，Song，d.，Erlingsson，u.，Oprea，a. 和 Raffel，c. 从大型语言模型中提取训练数据，2020 年。网址: https://arxiv.org/abs/2012.07805。

Carlini, N., Ippolito, D., Jagielski, M., Lee, K., Tramer, F., and Zhang, C. Quantifying memorization across neural language models, 2022. URL https://arxiv.org/ abs/2202.07646.

Carlini，n.，Ippolito，d.，Jagielski，m.，Lee，k.，Tramer，f. 和 Zhang，c. 神经语言模型的量化记忆，2022。网址: https://arxiv.org/abs/2202.07646。

Chung, Y., Zhang, Y., Han, W., Chiu, C., Qin, J., Pang, R., and Wu, Y. W2v-bert: Combining contrastive learn-ing and masked language modeling for self-supervised speech pre-training. arXiv:2108.06209, 2021.

钟，y.，张，y.，韩，w.，邱，c.，秦，j.，庞，r.，和吴，Y.W2v-bert: 结合对比学习和掩蔽语言建模的自监督语音预训练。arXiv: 2108.06209,2021.

Cohen, A. D., Roberts, A., Molina, A., Butryna, A., Jin, A., Kulshreshtha, A., Hutchinson, B., Zevenbergen, B., Aguera-Arcas, B. H., ching Chang, C., Cui, C., Du, C., Adiwardana, D. D. F., Chen, D., Lepikhin, D. D., Chi, E. H., Hoffman-John, E., Cheng, H.-T., Lee, H., Kri-vokon, I., Qin, J., Hall, J., Fenton, J., Soraker, J., Meier-Hellstern, K., Olson, K., Aroyo, L. M., Bosma, M. P., Pickett, M. J., Menegali, M. A., Croak, M., D´ıaz, M., Lamm, M., Krikun, M., Morris, M. R., Shazeer, N., Le, Q. V., Bernstein, R., Rajakumar, R., Kurzweil, R., Thop-pilan, R., Zheng, S., Bos, T., Duke, T., Doshi, T., Zhao, V. Y., Prabhakaran, V., Rusch, W., Li, Y., Huang, Y., Zhou, Y., Xu, Y., and Chen, Z. Lamda: Language models for dialog applications. arXiv:2201.08239, 2022.

Cohen，A.d.，Roberts，a.，Molina，a.，Butryna，a.，Jin，a.，Kulshreshtha，a.，Hutchinson，b.，Zevenbergen，b.，Aguera-Arcas,b.h.，ching Chang，c.，Cui，c.，D.，c.，Adiwardana，D.d. f.，Chen，d.，Lepikhin，D.d.，Chi，E.h.，Hoffman-John，e.，Cheng，H.-T.，Lee，h.，Kri-vokon，i.，Qin，j.，Hall，j.，Fenton，j.，Soraker，j.，Meier-Hellstern，k.，Olson，k.，Aroyo，L.m.，Bosma，M.p.，Pickett，M.j.，Menegali，M.a.，Croak，m.，d A.，m.，Lamm，m.，Krikun，m.，Morris，M.r.，Shazeer，n.，L.，Q.v.，Bernstein，r.，Rajakumar，r.，Kurzweil，r.，Thop-pilan,r.，Zheng，s.，Bos，t.，Duke，t.，Doshi，t.，Zhao，V.y.，Prabhakaran，v.，Rusch，w.，Li，y.，Huang，y.，Zhou，y.，Xu，y. 和 Chen，Z.Lamda: 对话应用程序的语言模型。arXiv: 2201.08239,2022.

Cuturi, M. Sinkhorn distances: Lightspeed computation of optimal transport. In Advances in Neural Information Processing Systems (NeurIPS), 2013.

Sinkhorn 距离: 最佳传输的光速计算，《神经信息处理系统进展》，2013。

Defferrard, M., Benzi, K., Vandergheynst, P., and Bresson,
Defferrard，m。，Benzi，k。，Vandergheynst，p，

X. FMA: A dataset for music analysis. In International Society for Music Information Retrieval Conference (IS-MIR), 2017.
fMA: 音乐分析数据集。国际音乐信息检索学会会议，2017。

Devlin, J., Chang, M., Lee, K., and Toutanova, K. BERT: pre-training of deep bidirectional transformers for language understanding. In NAACL-HLT, 2019.
Devlin，j.，Chang，m.，Lee，k.，and Toutanova，k. BERT: 深度双向变换器的语言理解预训练。在 NAACL-HLT，2019 年。

Dhariwal, P., Jun, H., Payne, C., Kim, J. W., Radford, A., and Sutskever, I. Jukebox: A generative model for music. arXiv:2005.00341, 2020.
Dhariwal，p。，Jun，h。，Payne，c。，Kim，j。，Radford，a。，和 Sutskever，i。点唱机: 音乐的生成模型。arXiv: 2005.00341,2020.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby,
Dosovitskiy，a。，Beyer，l。，Kolesnikov，a。，魏森博恩，d。，翟，x。，underthiner，t。，Dehghani，m。，Minderer，m。，Heigold，g。，Gelly，s。，Uszkoreit，j。和 Houlsby,
N. An image is worth 16x16 words: Transformers for image recognition at scale. In International Conference on Learning Representations (ICLR), 2021.
一张图片胜过 16x16 个单词: 用于大规模图像识别的变形金刚。《学习表征国际会议》(ICLR)，2021 年。

Defossez,´ A., Copet, J., Synnaeve, G., and Adi, Y. High fidelity neural audio compression. arXiv:2210.13438, 2022.
高保真神经音讯压缩。 arXiv: 2210.13438,2022。

Engel, J. H., Hantrakul, L., Gu, C., and Roberts, A. DDSP: differentiable digital signal processing. In International Conference on Learning Representations (ICLR), 2020.
Engel，J.h。，Hantrakul，l。，Gu，c。和 Roberts，a。DDSP: 可微数字信号处理。学习表征国际会议(ICLR)，2020。

Esser, P., Rombach, R., and Ommer, B. Taming transformers for high-resolution image synthesis. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2021.
埃塞尔，p。，Rombach，r。和 Ommer，b。驯服高分辨率图像合成变压器。在 IEEE 计算机视觉和模式识别会议(CVPR)，2021 年。

Forsgren, S. and Martiros, H. Riffusion - Stable diffusion for real-time music generation, 2022. URL https:// riffusion.com/about.
Riffusion-实时音乐生成的稳定扩散，2022。

Gemmeke, J. F., Ellis, D. P., Freedman, D., Jansen, A., Lawrence, W., Moore, R. C., Plakal, M., and Ritter, M. Audio set: An ontology and human-labeled dataset for audio events. In IEEE international conference on acous-tics, speech and signal processing (ICASSP). IEEE, 2017.
Gemmeke，j. f。，Ellis，D.p。，Freedman，d。，Jansen，a。，Lawrence，w。，Moore，R.c。，Plakal，m。，和 Ritter，m。音频集: 音频事件的本体和人类标记的数据集。在 IEEE 国际会议上关于声学，语音和信号处理(ICASSP)。IEEE，2017.

Hawthorne, C., Stasyuk, A., Roberts, A., Simon, I., Huang, C. A., Dieleman, S., Elsen, E., Engel, J. H., and Eck, D. Enabling factorized piano music modeling and generation with the MAESTRO dataset. In International Conference on Learning Representations (ICLR), 2019.
H.，Stasyuk，a.，Roberts，a.，Simon，i.，Huang，C.a.，Dieleman，s.，Elsen，e.，Engel，J.h.，and Eck，d. 使用 MAESTRO 数据集实现分解钢琴音乐建模和生成。2019 学习表征国际会议。

Hawthorne, C., Jaegle, A., Cangea, C., Borgeaud, S., Nash, C., Malinowski, M., Dieleman, S., Vinyals, O., Botvinick, M. M., Simon, I., Sheahan, H., Zeghidour, N., Alayrac, J., Carreira, J., and Engel, J. H. General-purpose, long-context autoregressive modeling with perceiver AR. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvari,´ C., Niu, G., and Sabato, S. (eds.), International Conference on Machine Learning (ICML), 2022a.
Hosane，c.，Jaegle，a.，Cangea，c.，Borgeaud，s.，Nash，c.，Malinowski，m.，Dieleman，s.，Vinyals，o.，Botvinick，M.m.，Simon，i.，Sheahan，h.，Zeghidour，n.，Alayrac，j.，Carreira，j. 和 Engel，J.h. 用感知者 A. 进行通用，长上下文自回归建模。在 Chaudhuri，k.，Jegelka，s.，Song，l.，Szepesvari，c. ,Niu，g. 和 Sabato，s. (编辑)，国际机器学习会议(ICML)，2022a。

Hawthorne, C., Simon, I., Roberts, A., Zeghidour, N., Gardner, J., Manilow, E., and Engel, J. H. Multi-instrument music synthesis with spectrogram diffusion. arXiv:2206.05408, 2022b.
霍桑，c。，西蒙，i。，罗伯茨，a。，Zeghidour，n。，加德纳，j。，Manilow，e。和恩格尔，j。arXiv: 2206.05408,2022b.

Hershey, S., Chaudhuri, S., Ellis, D. P. W., Gemmeke, J. F., Jansen, A., Moore, C., Plakal, M., Platt, D., Saurous, R. A., Seybold, B., Slaney, M., Weiss, R., and Wilson,
Hershey，s.，Chaudhuri，s.，Ellis，D.P.w.，Gemmeke，j. f.，Jansen，a.，Moore，c.，Plakal，m.，Platt，d.，Saurous，R.a.，Seybold，b,
K. Cnn architectures for large-scale audio classification. In International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2017.
用于大规模音频分类的 Cnn 架构。国际声学、语音和信号处理会议(ICASSP)，2017。

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion proba-bilistic models. Advances in Neural Information Process-ing Systems (NeurIPS), 2020.
《神经信息处理系统进展》，2020。

Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., and Fleet, D. J. Video diffusion models. arXiv:2204.03458, 2022.
Ho，j。，Salimans，t。，grissenko，a。，Chan，w。，Norouzi，m。，and Fleet，d。《视频扩散模型》 arXiv: 2204.03458,2022。

Hong, W., Ding, M., Zheng, W., Liu, X., and Tang, J. Cogvideo: Large-scale pretraining for text-to-video gene-ration via transformers. arXiv:2205.15868, 2022.
Hong，w.，Ding，m.，Zheng，w.，Liu，x.，and Tang，j. Cogvideo: 通过变形金刚进行文本到视频生成的大规模预训练。arXiv: 2205.15868,2022.

Huang, C. A., Vaswani, A., Uszkoreit, J., Simon, I., Hawthorne, C., Shazeer, N., Dai, A. M., Hoffman, M. D., Dinculescu, M., and Eck, D. Music transformer: Gene-rating music with long-term structure. In International Conference on Learning Representations (ICLR), 2019.
黄，C.a。，Vaswani，a。，Uszkoreit，j。，西蒙，i。，霍桑，c。，Shazeer，n。，Dai，A.m。，Hoffman，M.d。，Dinculescu，M.和 Eck，d。音乐转换器: 具有长期结构的基因评级音乐。2019 年学习表征国际会议。

Huang, Q., Jansen, A., Lee, J., Ganti, R., Li, J. Y., and Ellis, D. P. W. Mulan: A joint embedding of music audio and natural language. In International Society for Music Information Retrieval Conference (ISMIR), 2022.
Huang，q.，Jansen，a.，Lee，j.，Ganti，r.，Li，J.y.，and Ellis，D.P.w.在国际音乐信息检索会议(ISMIR)，2022 年。

Kilgour, K., Zuluaga, M., Roblek, D., and Sharifi, M. Frechet´ audio distance: A reference-free metric for evalu-ating music enhancement algorithms. In INTERSPEECH, 2019.
Kilgour，k。，Zuluaga，m。，Roblek，d。，and Sharifi，m。 Frechet 音频距离: 评估音乐增强算法的无参考指标。In INTERSPEECH，20192019 年 INTERSPEECH。

Kim, C. D., Kim, B., Lee, H., and Kim, G. Audiocaps: Generating captions for audios in the wild. In NAACL-HLT, 2019.
Kim，C.d。 Audiocaps: 生成野外音频的字幕。在 NAACL-HLT，2019 年。

Kreuk, F., Synnaeve, G., Polyak, A., Singer, U., Defossez,´ A., Copet, J., Parikh, D., Taigman, Y., and Adi, Y. Audio-gen: Textually guided audio generation, 2022.
Kreuk，f.，Synnaeve，g.，Polyak，a.，Singer，u.，Defossez，a.，Copet，j.，Parikh，d.，Taigman，y.，and Adi，y. Audio-gen: 文本引导的音频生成，2022。

Mubert-Inc. Mubert. https://mubert.
Mubert-Inc。Mubert。https://Mubert。
com/, https://github.com/MubertAI/ Mubert-Text-to-Music, 2022.
Com/，https://github.com/muberttai/Mubert-Text-to-Music，2022.

Nichol, A. Q., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., and Chen, M. GLIDE: towards photorealistic image generation and edit-ing with text-guided diffusion models. In International Conference on Machine Learning (ICML), 2022.
Nichol，A.q。，Dhariwal，p。，Ramesh，a。，Shyam，p。，Mishkin，p。，McGrew，b。，Sutskever，i。，和 Chen，m。国际机器学习会议(ICML)，2022。

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning transferable visual models from natural language supervision. In International Conference on Machine Learning (ICML), 2021.

Radford，a.，Kim，J.w.，Hallacy，c.，Ramesh，a.，Goh，g.，Agarwal，s.，Sastry，g.，Askell，a.，Mishkin，p.，Clark，j.，Krueger，g. 和 Sutskever，i. 从自然语言监督中学习可迁移的视觉模型。在国际机器学习会议(ICML)，2021 年。

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P. J., et al. Explor-ing the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research (JMLR), 2020.

Raffel，c.，Shazeer，n.，Roberts，a.，Lee，k.，Narang，s.，Matena，m.，Zhou，y.，Li，w.，Liu，P.j. 等。用一个统一的文本-文本转换器探索迁移学习的极限。机器学习研究杂志，2020。

Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Rad-ford, A., Chen, M., and Sutskever, I. Zero-shot text-to-image generation. In Meila, M. and Zhang, T. (eds.), Inter-national Conference on Machine Learning (ICML), 2021.

Ramesh，a。，Pavlov，m。，Goh，g。，Gray，s。，Voss，c。，Rad-ford，a。，Chen，m。，和 Sutskever，i。在 Meila，m。和 Zhang，t 国际机器学习会议(ICML)，2021。

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen,
Ramesh，a。，Dhariwal，p。，Nichol，a。，Chu，c，
M. Hierarchical text-conditional image generation with clip latents. arXiv:2204.06125, 2022.
基于文本条件的层次化图像生成。

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Om-mer, B. High-resolution image synthesis with latent diffu-sion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10684–10695, June 2022a.

Rombach，r.，Blattmann，a.，Lorenz，d.，Esser，p. 和 Om-mer，b. 用潜在扩散模型进行高分辨率图像合成。在 IEEE/cv f 计算机视觉和模式识别会议论文集(CVPR)，10684-10695 页，2022 年 6 月。

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffu-sion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022b.

Rombach，r.，Blattmann，a.，Lorenz，d.，Esser，p. 和 Om-mer，b. 用潜在扩散模型进行高分辨率图像合成。在 IEEE/cv f 计算机视觉和模式识别会议论文集，2022b。

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Den-ton, E. L., Ghasemipour, S. K. S., Ayan, B. K., Mahdavi, S. S., Lopes, R. G., Salimans, T., Ho, J., Fleet, D. J., and Norouzi, M. Photorealistic text-to-image diffusion mod-els with deep language understanding. arXiv:2205.11487, 2022.

Saharia，c.，Chan，w.，Saxena，s.，Li，l.，Whang，j.，Den-ton，E.l.，gassemipour，S.K.s.，Ayan，B.k.，Mahdavi，S.s.，Lopes，R.g.，Salimans，t.，Ho，j.，Fleet，D.j. 和 Norouzi，m. 具有深度语言理解的照片现实主义文本到图像扩散模型。arXiv: 2205.11487,2022.

Schroff, F., Kalenichenko, D., and Philbin, J. Facenet: A unified embedding for face recognition and clustering. In Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), 2015.
人脸识别和聚类的统一嵌入。2015 年 IEEE 计算机视觉和模式识别会议论文集。

Shor, J., Jansen, A., Maor, R., Lang, O., Tuval, O., de Chau-mont Quitry, F., Tagliasacchi, M., Shavitt, I., Emanuel, D., and Haviv, Y. Towards Learning a Universal Non-Semantic Representation of Speech. In INTERSPEECH, 2020.

Shor，j.，Jansen，a.，Maor，r.，Lang，o.，Tuval，o.，D. Chau-mont Quitry，f.，Tagliasacchi，m.，Shavitt，i.，Emanuel，d.，and Haviv，y.In INTERSPEECH，2020 In INTERSPEECH，2020 年。

van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A. W., and Kavukcuoglu, K. Wavenet: A generative model for raw audio. In ISCA, 2016.

Van den Oord，a。，Dieleman，s。，Zen，h。，Simonyan，k。，Vinyals，o。，Graves，a。，Kalchbrenner，n。，Senior，a。，和 Kavukcuoglu，k。 Wavenet: 原始音频的生成模型。在 ISCA，2016 年。

Van Den Oord, A., Vinyals, O., et al. Neural discrete repre-sentation learning. Advances in neural information pro-cessing systems (NeurIPS), 2017.
范登・奥德，维尼亚斯，等。神经离散表征学习。神经信息处理系统进展，2017。

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones,
Vaswani，a。，Shazeer，n。，Parmar，n。，Uszkoreit，j,
L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Atten-
L.，戈麦斯，A.n.，凯撒，和波洛苏金，i. 注意-
tion is all you need. Advances in neural information pro-
你所需要的就是信息，神经信息的进步
cessing systems (NeurIPS), 2017.
终止系统(NeurIPS)，2017 年。

Villegas, R., Babaeizadeh, M., Kindermans, P.-J., Moraldo,
Villegas，r.，Babaeizadeh，m.，Kindermans，p,
H., Zhang, H., Saffar, M. T., Castro, S., Kunze, J., and
H.，Zhang，h.，Saffar，M.t.，Castro，s.，Kunze，j.，and
Erhan, D. Phenaki: Variable length video generation
Erhan，D.Phenaki: 可变长度视频生成
from open domain textual description. arXiv:2210.02399,
来自开放域文本描述. arXiv: 2210.02399,
2022.
2022.

Wu, C., Liang, J., Ji, L., Yang, F., Fang, Y., Jiang, D.,
吴，c.，梁，j.，季，l.，杨，f.，方，y.，江，d.,
and Duan, N. Nuwa:¨ Visual synthesis pre-training for
女娲: 视觉合成预训练
neural visual world creation. In European Conference on
神经视觉世界的创造
Computer Vision (ECCV), 2022a.
计算机视觉(ECCV)，2022a。

Wu, H., Seetharaman, P., Kumar, K., and Bello, J. P.
吴，h.，Seetharaman，p.，Kumar，k. 和 Bello，J.p。
Wav2clip: Learning robust audio representations from
Wav2clip: 从
CLIP. In International Conference on Acoustics, Speech
国际声学会议演讲
and Signal Processing (ICASSP), 2022b.
和信号处理(ICASSP)，2022b。

Yang, D., Yu, J., Wang, H., Wang, W., Weng, C., Zou, Y.,
杨，d.，于，j.，王，h.，王，w.，翁，c.，邹，y.,
and Yu, D. Diffsound: Discrete diffusion model for text-
diffsound: 文本的离散扩散模型
to-sound generation. arXiv:2207.09983, 2022.
To-sound generation. arXiv: 2207.09983,2022.

Yu, J., Xu, Y., Koh, J. Y., Luong, T., Baid, G., Wang, Z.,
Yu，j.，Xu，y.，Koh，J.y.，Luong，t.，Baid，g.，Wang，z.,
Vasudevan, V., Ku, A., Yang, Y., Ayan, B. K., Hutchinson,
Vasudevan，v.，Ku，a.，Yang，y.，Ayan，B.k.，Hutchinson,
B., Han, W., Parekh, Z., Li, X., Zhang, H., Baldridge, J.,
b.，Han，w.，Parekh，z.，Li，x.，Zhang，h.，Baldridge，j.,
and Wu, Y. Scaling autoregressive models for content-

标度自回归模型的内容-
rich text-to-image generation, 2022.
富文本图像生成，2022。

Zeghidour, N., Teboul, O., de Chaumont Quitry, F., and
Zeghidour，n。，Teboul，o。，de Chaumont Quitry，f
Tagliasacchi, M. LEAF: A learnable frontend for audio
Tagliasacchi，m。LEAF: 一个可学习的音频前端
classification. In 9th International Conference on Learn-
分类。在第九届国际学习会议-
ing Representations, ICLR 2021, Virtual Event, Austria,
国际图联 2021 年，奥地利，虚拟事件,
May 3-7, 2021. OpenReview.net, 2021. URL https:
2021 年 5 月 3 日至 7 日，OpenReview.net，2021，网址 https:
//openreview.net/forum?id=jM76BCb6F9m.
//openreview.net/forum? id = jm76bcb6 f9m.

Zeghidour, N., Luebs, A., Omran, A., Skoglund, J., and
Zeghidour，n。，Luebs，a。，Omran，a。，Skoglund，j
Tagliasacchi, M. Soundstream: An end-to-end neural
Tagliasacchi，m. Soundstream: 一个端到端的神经元
audio codec. IEEE ACM Trans. Audio Speech Lang.
音频语音编解码器。
Process., 30, 2022.
进程，30,2022。

Zen, H., Senior, A., and Schuster, M. Statistical paramet-
Zen，h.，Senior，a. 和 Schuster，m. 统计参数-
ric speech synthesis using deep neural networks. In In-
Ric 语音合成使用深度神经网络
ternational Conference on Acoustics, Speech and Signal
国际声学、语音和信号会议
Processing (ICASSP), 2013.
处理(ICASSP)，2013。

A. MusicCaps Dataset
MusicCaps 数据集

Together with this paper, we release MusicCaps, a high-quality music caption dataset.[6] This dataset includes music clips from AudioSet (Gemmeke et al., 2017), paired with corresponding text descriptions in English. It contains a total of 5,521 examples, out of which 2,858 are from the AudioSet eval and 2,663 from the AudioSet train split. We further tag 1,000 examples as a balanced subset of our dataset, which is balanced with respect to the genres of the music contained. All examples in the balanced subset are from the AudioSet eval split.
与本文一起，我们发布了 MusicCaps，一个高质量的音乐字幕数据集。这个数据集包括 AudioSet (Gemmeke et al。，2017)的音乐片段，与相应的英文文本描述配对。它总共包含 5,521 个示例，其中 2,858 个来自 AudioSet eval，2,663 个来自 AudioSet train 拆分。我们进一步标记了 1000 个例子作为我们数据集的一个平衡子集，这个子集与包含的音乐类型是平衡的。平衡子集中的所有例子都来自 AudioSet eval 分割。

Examples of free text captions:
自由文本标题的例子:

- “This folk song features a male voice singing the main melody in an emotional mood. This is accompanied by an accordion playing fills in the background. A violin plays a droning melody. There is no percussion in this song. This song can be played at a Central Asian classical concert.”
“这首民歌以男声唱主旋律，情绪激昂。伴随着手风琴演奏的是背景音乐。小提琴演奏嗡嗡作响的旋律。这首歌里没有打击乐。这首歌可以在中亚古典音乐会上演奏。”
- “This is a live recording of a keyboardist playing a twelve bar blues progression on an electric keyboard. The player adds embellishments between chord changes and the piece sounds groovy, bluesy and soulful.”
“这是一个键盘手在电子键盘上演奏十二小节布鲁斯进行曲的现场录音。演奏者在和弦变化之间添加了一些修饰，这首曲子听起来很棒，布鲁斯和深情。”
- “A synth is playing an arpeggio pluck with a lot of reverb rising and falling in velocity. Another synth sound is playing pads and a sub bassline. This song is full of synth sounds creating a soothing and adventurous atmosphere. This song may be playing at a festival during two songs for a buildup.”
“一个合成器在演奏一段琶音，混响在速度上起伏不定。另一个合成器声音是播放垫子和低音贝斯线。这首歌充满了合成器的声音，创造了一种舒缓和冒险的氛围。这首歌可能会在一个音乐节上播放，在两首歌曲的过程中逐渐积累。”
- “A low sounding male voice is rapping over a fast paced drums playing a reggaeton beat along with a bass. Something like a guitar is playing the melody along. This recording is of poor audio-quality. In the background a laughter can be noticed. This song may be playing in a bar.”
“一个低沉的男声在快节奏的鼓上敲打，伴随着雷鬼音乐的节拍和低音。像吉他一样的东西伴随着旋律。这段录音的音质很差。在背景中可以看到笑声。这首歌可能是在酒吧里演奏的。”
- “The electronic music features a section that repeats roughly every two seconds. It consists of a beat that’s made of a kick drum and claps. A buzzing synth sets the pulsation of the music by playing once every two beats. The whole music sounds like a loop being played over and over. Towards the end of the excerpt a crescendo-like buzzing sound can be heard, increasing the tension.”
“电子音乐的特点是大约每两秒钟重复一次。它包括由鼓点和拍子组成的节拍。一个嗡嗡作响的合成器通过每两个节拍播放一次来设定音乐的节奏。整个音乐听起来就像一个循环一遍又一遍地播放。在节选的最后，可以听到一种渐强的嗡嗡声，增加了紧张感。”

Examples of aspect lists:
方面列表的例子:

- “pop, tinny wide hi hats, mellow piano melody, high pitched female vocal melody, sustained pulsating synth lead, soft female vocal, punchy kick, sustained synth bass, claps, emotional, sad, passionate”
流行音乐，轻巧宽大的帽子，圆润的钢琴曲，高亢的女声曲，持续跳动的合成器领唱，柔和的女声，有力的踢腿，持续的合成器低音，鼓掌，情绪化，悲伤，激情
- “amateur recording, finger snipping, male mid range voice singing, reverb”

业余录音，剪指，男中音歌唱，混响

- “backing track, jazzy, digital drums, piano, e-bass, trumpet, acoustic guitar, digital keyboard song, medium tempo”

伴奏曲，爵士乐，数码鼓，钢琴，低音提琴，小号，原声吉他，数码键盘歌，中速

- “rubab instrument, repetitive melody on different octaves, no other instruments, plucked string instrument, no voice, instrumental, fast tempo”

摩擦乐器，不同八度重复旋律，无其他乐器，拨弦乐器，无人声，器乐，快节奏

- “instrumental, white noise, female vocalisation, three unrelated tracks, electric guitar harmony, bass guitar, keyboard harmony, female lead vocalisation, keyboard harmony, slick drumming, boomy bass drops, male voice backup vocalisation”

器乐、白噪音、女声、三首不相关的曲目、电吉他和声、低音吉他和声、键盘和声、女主唱和声、键盘和声、圆滑的鼓声、低音大提琴和男声伴唱

[6] kaggle.com/datasets/googleai/musiccaps
来源: kaggle.com/dataset/googleai/musiccaps

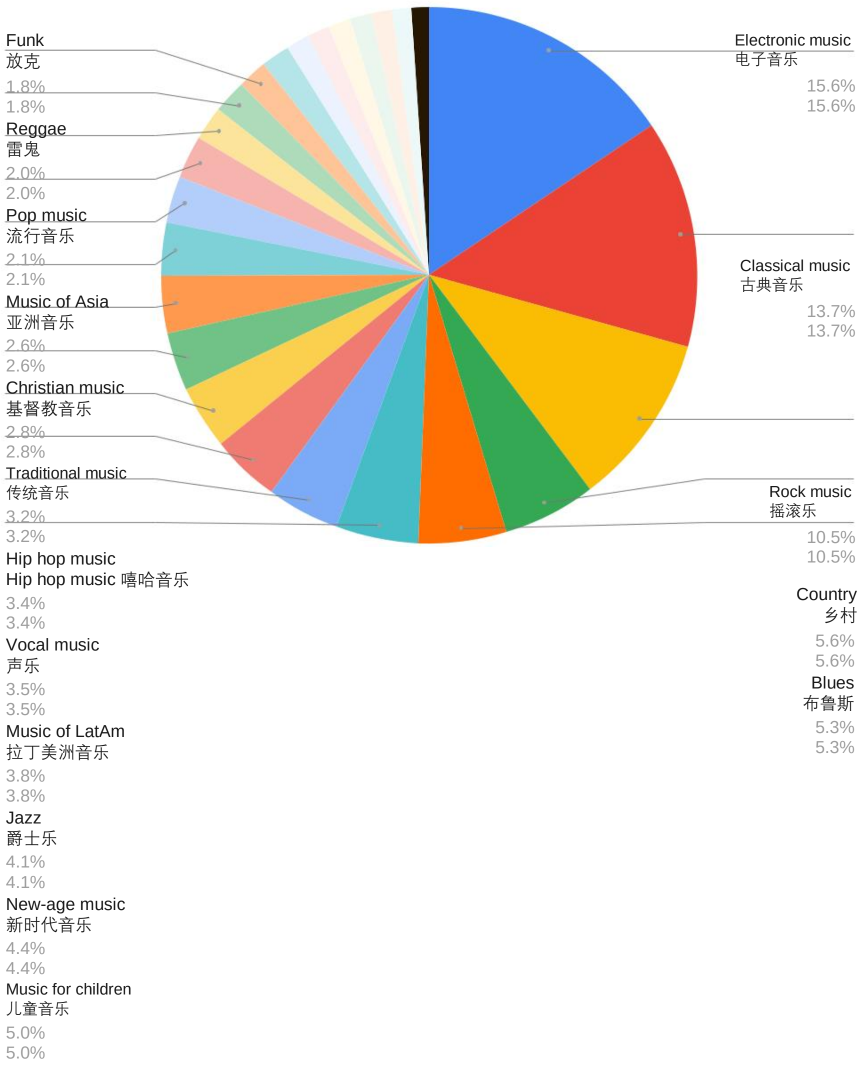
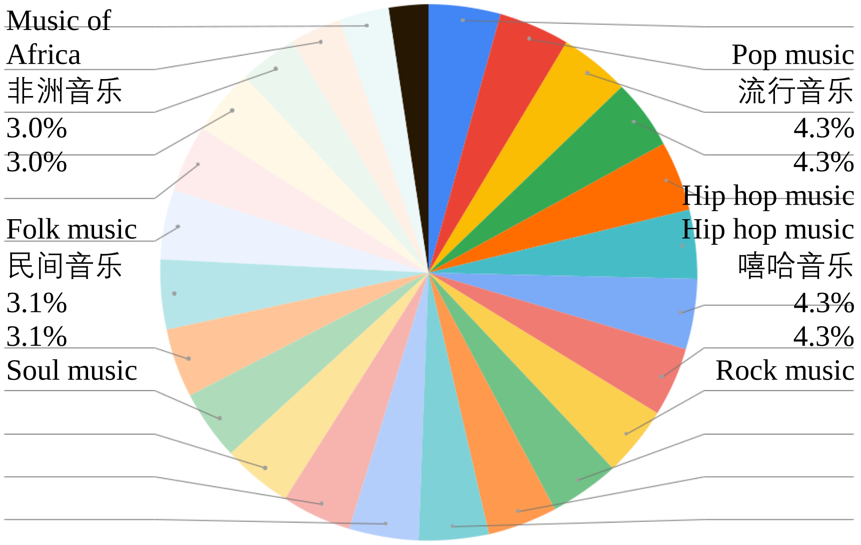


Figure 4. Genre distribution of all 5.5k examples of MusicCaps, according to an AudioSet classifier.
图 4。根据 AudioSet 分类器，所有 5.5 k 个 musiccap 例子的类型分布。



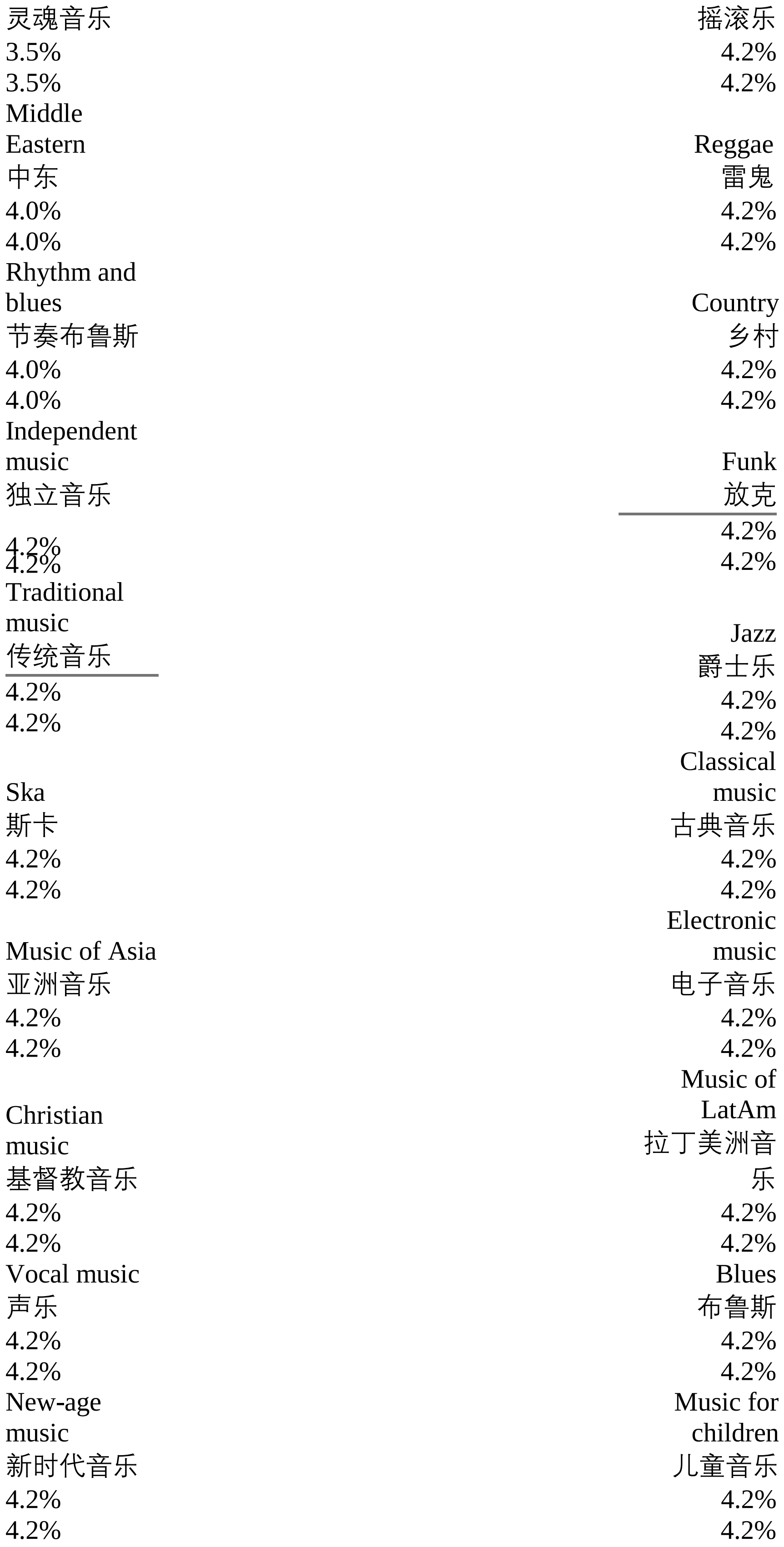


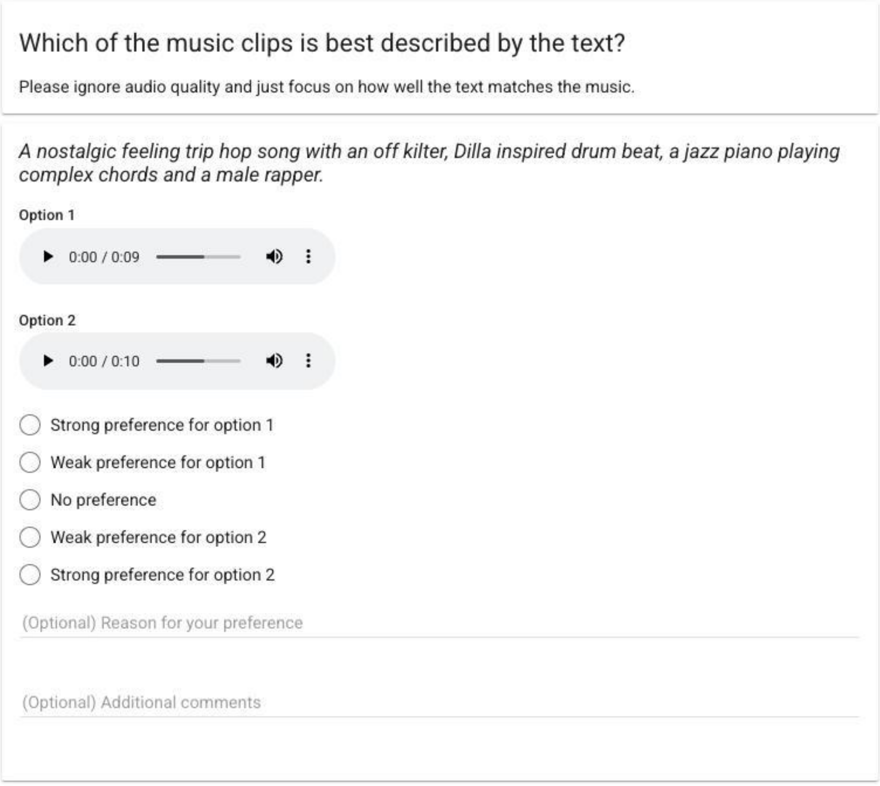
Figure 5. Genre distribution of a balanced 1k example subset of MusicCaps, according to an AudioSet classifier.

图 5。根据 AudioSet 分类器，均衡的 1k musiccap 示例子集的类型分布。

B. Qualitative Evaluation
B. 定性评估

Participants in the listening test were presented with two 10-second clips and a text caption, and asked which clip is best described the text of the caption on a 5-point Likert scale. They were also instructed to ignore audio quality and focus just on how well the text matches the music (similar to MuLan score). Figure 6 shows the user interface presented to raters.
听力测试的参与者被呈现两个 10 秒钟的片段和一个文字说明，并被问及在李克特 5 分量表中哪个片段最能描述说明的文字。他们还被要求忽略音频质量，只关注文字与音乐的匹配程度(类似于《花木兰》的乐谱)。图 6 展示了给评分者的用户界面。

We collected 1200 ratings, with each source involved in 600 pair-wise comparisons. Figures 7 and 8 show the granular results of pairwise comparisons between the models. According to a post-hoc analysis using the Wilcoxon signed-rank test with Bonferroni correction (with p < 0:01=15), the orderings shown in Figure 8 from raters are all statistically significant.
我们收集了 1200 个评分，每个来源涉及 600 个成对比较。图 7 和 8 显示了模型之间成对比较的粒度结果。根据使用具有邦弗朗尼校正的威尔克科逊检验的事后比较分析(p < 0:01 = 15)，评估者的图 8 所示的排序都是统计学显着的。



Which of the music clips is best described by the text?

Please ignore audio quality and just focus on how well the text matches the music.

A nostalgic feeling trip hop song with an off kilter, Dilla inspired drum beat, a jazz piano playing complex chords and a male rapper.

Option 1

0:00 / 0:09

Option 2

0:00 / 0:10

○ Strong preference for option 1
○ Weak preference for option 1
○ No preference
○ Weak preference for option 2
○ Strong preference for option 2

(Optional) Reason for your preference

(Optional) Additional comments

Figure 6. User interface for the human listener study.
图 6。人类听众研究的用户界面。

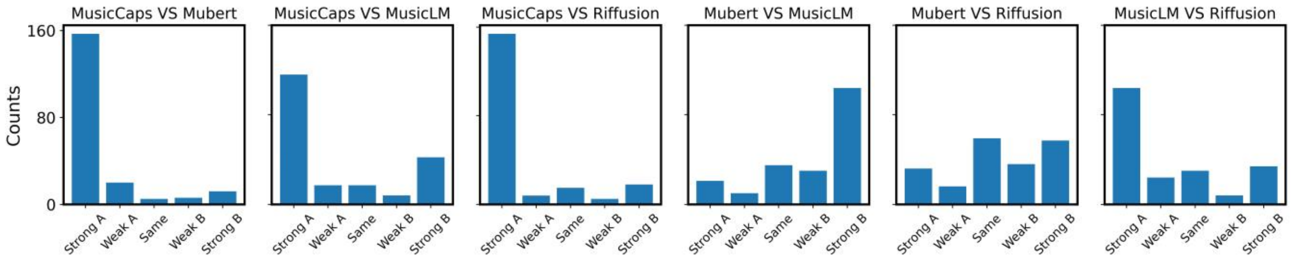


Figure 7. Pairwise comparisons from the human listener study. Each pair is compared on a 5-point Likert scale. Raters had a decisive model preference in all cases except Mubert vs. Riffusion.

图7。来自人类听众研究的成对比较。每一对都用5分的李克特量表进行比较。评分者在所有情况下都有决定性的模型偏好，除了Mubert vs. Riffusion。

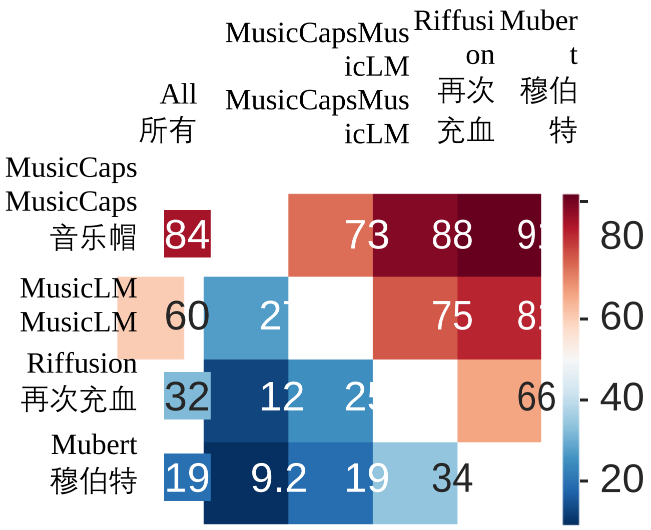


Figure 8. Win percentage from the human listener study. Each row indicates the % of times listeners found the music to better match the caption from that system to those from any other system (first column, N = 1200) and each system individually (other columns,
图 8。从人类听众研究中获胜的百分比。每行表示监听器找到该系统的音乐与其他任何系统(第一列，n = 1200)的标题更匹配的次数的百分比，以及每个系统单独匹配的次数的百分比(其他列，n = 1200),
N = 600). The ground truth data (MusicCaps) clearly is the best match to the captions, but followed closely by MusicLM, which even beats the ground truth in 27% of comparisons.
= 600).基本事实数据(MusicCaps)显然是与标题最匹配的，但紧随其后的是 MusicLM，它甚至在 27% 的比较中击败了基本事实。

C. Melody Conditioning
旋律调节

We provide here implementation details of the model used for conditioning the music generation on melody. The model is based on a small ViT (Dosovitskiy et al., 2021) composed of 12 layers, 6 attention heads, embedding dimension of 512 and feed-forward layer of dimension 1024. The input to the model are the temporal frames of the mel spectrogram of the audio. We use semi-hard triplet loss (Schroff et al., 2015) to train the melody embedding model to generate 192 dimensional embeddings for each 4 seconds of audio. The model learns to generate embeddings which are representative of a melody while being invariant to acoustic properties related to the instruments being played. This is particularly advantageous, since this representation is complementary to the representation learned by the MuLan embeddings. Hence, our melody embeddings and the MuLan can be jointly and complementarily used for conditioning the music generation process. During training, we consider input audio with a duration of 10 seconds. We extract three melody embeddings, with a hop length of 3 seconds, discretize each of them to tokens with residual vector quantization (RVQ) and concatenate the resulting token sequences with the MuLan audio tokens M_A. We use an RVQ composed of 24 quantizers, each with a vocabulary size of 512.
我们在这里提供了用于调节旋律的音乐生成的模型的实现细节。该模型基于一个小的 ViT (Dosovitskiy et al。，2021)，由 12 层、6 个注意头、512 维嵌入层和 1024 维前馈层组成。模型的输入是音频 mel 谱图的时间帧。我们使用半硬三重损失(Schroff 等，2015)来训练旋律嵌入模型，以便为每 4 秒的音频生成 192 维嵌入。该模型学习生成代表旋律的嵌入，同时不变于与正在演奏的乐器相关的声学属性。这是特别有利的，因为这种表示是对木兰嵌入学习到的表示的补充。因此，我们的旋律嵌入和木兰可以共同和互补地用于调节音乐生成过程。在训练过程中，我们考虑输入持续时间为 10 秒的音频。首先提取 3 个跳长为 3 秒的旋律嵌入，然后利用残差向量量化(RVQ)将每个旋律嵌入离散化为标记，最后将得到的标记序列与木兰音频标记 MA 连接。我们使用由 24 个量化器组成的 RVQ，每个量化器的词汇表大小为 512。