Noise2Music：基于文本条件的扩散模型音乐生成

Qingqing Huang [*1] **Daniel S. Park** [*1] **Tao Wang** [†1] **Timo I. Denk** [†1] **Andy Ly** [†1] **Nanxin Chen** [†1] **Zhengdong Zhang** [1] **Zhishuai Zhang** [1] **Jiahui Yu** [1] **Christian Frank** [1] **Jesse Engel** [1] **Quoc V. Le** [1] **William Chan** [2] **Zhifeng Chen** [2] **Wei Han** [†1]

arXiv:2302.03917v2 [cs.SD] 6 Mar 2023

摘要

我们介绍 Noise2Music：通过训练一系列扩散模型，从文本提示生成高质量的 30 秒音乐片段。我们依次训练并使用两类扩散模型：生成器模型根据文本生成中间表示；级联模型则根据中间表示（也可能结合文本）生成高保真音频，从而生成高保真音乐。我们探索了两种中间表示方案，一种使用频谱图，另一种使用保真度较低的音频。我们发现，生成音频不仅能够忠实地体现文本提示中的音乐流派、速度、乐器、情绪和时代等关键元素，还能进一步准确呈现提示词中细粒度的语义。预训练大语言模型在这一过程中发挥着关键作用——它们用于为训练集中的音频生成配对文本，并提取扩散模型所接收文本提示的嵌入。

生成示例：

https://google-research.github.io/noise2music

1. 引言

已有研究表明，深度神经网络具备出色的生成能力。在本文中，我们探索深度模型在音频生成方面的能力。

我们介绍了 Noise2Music，这是一种基于扩散的方法（Sohl-Dickstein 等人，2015年；Song 和 Ermon，2019年；Ho 等人，2020），可根据文本提示生成音乐；并通过生成时长为 30 秒、采样率为 24kHz 的音乐片段，展示了其能力。

建模： 我们训练了一系列级联扩散模型（Ho 等，2022b），其中第一个模型学习从文本提示生成 30 秒波形压缩表示的生成任务，第二个模型则学习以该压缩表示（以及可选的文本提示）为条件生成 16kHz 波形。我们研究了两种中间表示方案：对数梅尔频谱图或 3.2kHz 波形。我们使用一维 U-Net 学习扩散模型的噪声向量。扩散模型以自由文本形式的用户提示为条件，这些提示由预训练语言模型（LM）编码，再通过交叉注意力输入一维 U-Net 层。最后，使用超分辨率级联器，根据 16kHz 波形生成 24kHz 音频。

数据挖掘： 大量训练数据对于深度生成模型生成高质量样本至关重要。我们采用数据挖掘流程，构建了一个大规模训练数据集，其中包含多样化的音乐音频片段，且每段音频都配有多个描述性文本标签。音频的文本标签由一对预训练深度模型生成：首先，我们使用大语言模型生成大量通用音乐描述句，作为候选说明文字；随后，我们使用预训练的音乐-文本联合嵌入模型，计算每段未标注音乐片段与所有候选说明文字之间的匹配分数，并选取相似度最高的说明文字作为该音频片段的伪标签。通过这种方式，我们为约 15 万小时的音频素材添加了标注，构成训练数据。

MuLaMCap： 作为本研究的副产品，我们推出了 MuLan-LaMDA 音乐描述数据集（MuLaMCap），其中包含约 40 万组音乐-文本配对。这些配对是通过上述流程对 AudioSet 中的音乐内容进行标注而获得的。与原始 AudioSet 本体相比，后者的 632 个标签类别中只有 141 个与音乐相关；MuLaMCap 的说明文字则来自一个由 400 万条音乐描述句和短语构成的大型词汇表，具有更高的多样性和细粒度。我们预计，该数据集可用于声音分类之外的多种应用，例如音乐描述生成、检索或生成。

评估： 我们使用两个指标来衡量文本条件音乐生成模型的质量：Fr´echet 音频距离（FAD）（Kilgour 等人，2018），该指标衡量

[*]贡献相同 [†]核心贡献者 [1]Google Research [2]在 Google 工作期间完成。通讯作者：Qingqing Huang<qqhuang@google.com>。

预印本。

生成音频片段的质量与两个基准数据集（例如 AudioSet 的音乐子集（Gem- meke 等人，2017）和 MagnaTagATune（Law 等人，2009））相比如何；以及 MuLan 相似度分数（Huang 等人，2022），该指标衡量文本提示与相应生成音频片段之间的语义对齐程度。

生成能力： 我们的模型表明，它们能够超越简单的音乐属性条件控制，例如音乐流派、乐器和时代，并能够处理复杂且细粒度的语义，从而体现氛围、感受或活动等软属性。这是通过构建训练数据集实现的：该数据集不仅依赖元数据标签，还利用预训练的音乐-文本联合嵌入模型，将语义与音频特征对应起来。可在 google-research.github.io/noise2music#table-1 和 google-research.github.io/noise2music#table-4 中查看从文本提示生成的精选音乐示例。

2. 相关工作

生成模型： 深度生成模型在广泛的领域中有着悠久且成功的应用历史。近年来，大量工作致力于扩大用于训练模型的数据集规模，以生成质量极高的样本。这里列举了近期相关进展（并不全面），涵盖文本（Brown等，2020年；Thoppilan等，2022年）、语音（Wang等，2018年；Chen等，2021年；Borsos等，2022年）、图像（Ramesh等，2022年；Saharia 等人，2022b；Yu 等，2022年）和音频（Briot，2021年；Dhariwal等，2020年；MubertAI，2022；Kreuk等，2022年）。

扩散模型： 扩散模型由（Sohl-Dickstein等，2015年；Song 和 Ermon，2019年；Ho 等人，2020）提出，已展现出生成高质量图像（Ho 等人，2020；2022b）、音频（Yang等，2022年；Popov等，2021）和视频（Ho 等人，2022c；a）的能力。级联扩散模型（Ho 等，2022b；Saharia 等人，2022b）使用一系列扩散模型生成低保真图像，再依次对图像进行优化以生成高保真图像；本研究将其应用于音频。

音频生成： 人们采用了各种方法来生成以外部输入为条件的音频。文本转音频任务中有一些相关示例，其中，文本条件的频谱图生成和频谱图条件音频已得到深入研究（Popov 等，2021；Chen 等，2021；Kong 等，2021；Wu 和 Shi，2021；Chen 等，2022）。若将关注范围限定在基于描述性文本的音频生成，AudioGen（Kreuk 等，2022）采用自回归方法实现了文本条件的一般声音事件生成，而基于扩散的方法则通过 DiffSound（Yang等，2022年）对离散音频编码进行操作。若将范围进一步缩小至音乐生成，Jukebox（Dhariwal等，2020年）、Mubert（MubertAI，2022）和 MusicLM（Agostinelli等，2023）采用了自回归方法，而 Riffusion（Forsgren与Martiros，2022）则采用扩散进行频谱图生成。

音频生成中的条件信号： 总体而言，针对如何对条件信号进行参数化并将其传递给音频生成模型，已有两种方法。这种信号用于引导模型生成特定风格的音乐。一种方法是将信号投影到预定义且可解释的嵌入空间中：Jukebox（Dhariwal等，2020年）依赖从训练数据中提取的固定艺术家和音乐流派词汇表来对解码器进行条件控制，而 Mubert（MubertAI，2022）则将用户提示与预定义词汇表中的一组标签进行匹配。另一种方法见于 AudioGen（Kreuk等，2022年）和 MusicLM（Agostinelli 等人，2023年）等工作，即使用预训练文本编码器对任意用户提示进行编码。

作者注： 在完成这项工作期间，出现了一项与本工作部分重叠的同期研究（Schneider等，2023年）。

3. 方法

3.1. 扩散模型

扩散模型（Sohl-Dickstein 等人，2015年；Song & Ermon，2019年；Ho 等人，2020）是强大的生成模型，通过反复对随机噪声进行去噪来生成样本。这里我们仅回顾理解本工作所需的最少扩散模型相关信息。更多详情请参阅补充材料。

扩散模型的输入包括条件信号 $\mathbf{c}$、随机采样的时间步 t，以及样本 $\mathbf{x}_t$；该样本是通过具有高斯扩散过程的噪声调度，对原始样本 $\mathbf{x}$ 进行扰动得到的，其中噪声调度由时间 t时噪声的标准差 σ_t 参数化。时间 t 的范围设为 $[0, 1]$, , 训练期间从中均匀采样，扩散过程则被视为沿时间递增的方向推进。高斯扩散的动态过程已得到充分理解——$\mathbf{x}_t$ 的分布完全由一个属于标准正态分布的噪声向量 $\boldsymbol{\epsilon}$ 参数化，因为$\mathbf{x}_t$ 可以写成原始样本、确定性噪声调度和噪声向量 $\boldsymbol{\epsilon}$的函数，即$\mathbf{x}_t(\mathbf{x}\ \sigma,\ \boldsymbol{\epsilon})$，其中应理解为 σ 表示整个噪声调度。模型 $\boldsymbol{\epsilon}_\theta$ 经过训练，能够根据这些输入识别噪声向量。扩散损失可写为

$$\mathbb{E}_{\mathbf{x},\mathbf{c},\boldsymbol{\epsilon},t}\left[w_t\|\boldsymbol{\epsilon}_\theta(\mathbf{x}_t,\mathbf{c},t)-\boldsymbol{\epsilon}\|^2\right], \tag{1}$$

其中 w_t 是一个选定的固定权重函数。

推理通过在时间 $t = 1$取随机噪声，并利用模型给出的噪声预测对其进行去噪来完成。我们采用祖先采样（或去噪扩散概率模型采样）（Ho 等人，2020），它为推理提供了灵活的框架，其中包含多个可影响生成样本质量的参数。首先，可以通过改变采样器的随机性参数 γ 来控制去噪过程的随机性水平。此外，还可以使用任意的去噪调度，即可以任意划分区间 $0 = t_0 < \cdots < t_n = 1$，以离散化去噪步骤。

因此，在训练扩散模型时会有多种选择。针对以下元素，我们采用了多种方案，更多详情请参阅补充材料：

- 损失权重（w_t）：简化权重 $w_t = 1$（Ho 等人，2020）和西格玛权重 $w_t = \sigma_t^2$
- 方差调度：线性（Ho 等人，2020）和余弦（Nichol&Dhariwal，2021）调度
- 随机性参数：$\gamma = 0$或 1
- 去噪步数调度

无分类器引导（CFG）： CFG（Ho和Salimans，2022）是一种用于提升生成样本与条件输入之间对齐程度的方法。在训练期间，每个训练批次中部分训练样本的条件输入会被网络隐藏，从而使网络学会如何在无条件和有条件的情况下预测噪声向量。在推理时，分别计算有条件输入和无条件输入时的噪声向量，并将最终应用的噪声向量设为 $w\boldsymbol{\epsilon}_\theta(\mathbf{x}_t,\ \mathbf{c}) + (1-w)\boldsymbol{\epsilon}_\theta(\mathbf{x}_t,\ \cdot)$，其中 $w > 1$。采用动态裁剪（Saharia等人，2022b）来避免CFG导致的过饱和。

3.2. 架构

我们采用一维高效U-Net，即 Saharia 等人（2022b）提出的高效U-Net 的一维版本，作为扩散模型。图 1 所示的 U-Net 模型由一系列下采样和上采样块构成，并通过残差连接相连。一个下采样/上采样块由一个下采样/上采样层，以及一系列通过组合一维卷积层、自注意力/交叉注意力层和组合层构成的块组成。组合层使单个向量能够与一串向量交互，其中单个向量用于生成逐通道缩放和偏置。这些块与 Saharia 等人（2022b）构建的高效 U-Net 中的块结构十分相似，只是将二维卷积替换为一维对应层。补充材料中进一步阐述了这些块的具体结构。

模型有四种可能的输入途径。堆叠输入和输出均由长度为 T 的序列组成，而扩散时间 t 则编码为单个时间嵌入向量，并通过前述下采样和上采样块中的组合层与模型交互。鉴于我们希望生成长度为 T 的序列，噪声样本 $\mathbf{x}_t$ 始终位于图中最左侧的堆叠输入中，而输出则被视为噪声预测 $\boldsymbol{\epsilon}$。对于级联模型，可以对模型所依赖的低保真音频进行上采样并堆叠。此外，任意长度的向量序列都可以通过交叉注意力与这些块交互。文本提示正是通过这一途径输入模型的。模型还可以通过在 U-Net 的“U”形结构底部进行相加，以对齐但经过压缩的序列表示作为条件。

3.3. 级联扩散

本文训练了两类扩散模型，以根据文本提示生成高质量的30秒音频。遵循（Ho 等，2022b）的方法，我们训练生成器模型，使其在文本提示的条件下生成最终音频的某种中间表示；同时训练级联器模型，根据该中间表示生成最终音频。对于中间表示，我们同时考虑低保真音频和频谱图。

3.3.1. 波形模型

生成器模型： 生成器模型会生成以文本输入为条件的 3.2 kHz 音频。它会根据文本输入生成一系列向量，并将其作为交叉注意力序列输入网络。

级联模型： 级联模型会生成以文本提示和生成器模型根据该文本提示生成的低保真音频为条件的 16 kHz 音频。文本条件通过交叉注意力实现。同时，低保真音频会经过上采样，并与 $\mathbf{x}_t$ 堆叠后输入模型。上采样的过程是：先对低保真音频序列应用快速傅里叶变换（FFT），再应用逆快速傅里叶变换，从低保真傅里叶系数中获取高保真音频。

3.3.2. 频谱图模型

生成器模型： 该模型根据文本输入生成对数梅尔频谱图。该频谱图包含 80 个通道，每秒有 100 个频率特征。现在，输入和输出序列除了序列维度外，还具有通道维度。对数梅尔频谱图的像素值经过归一化，范围为 $[-1, 1]$。文本条件通过交叉注意力实现。

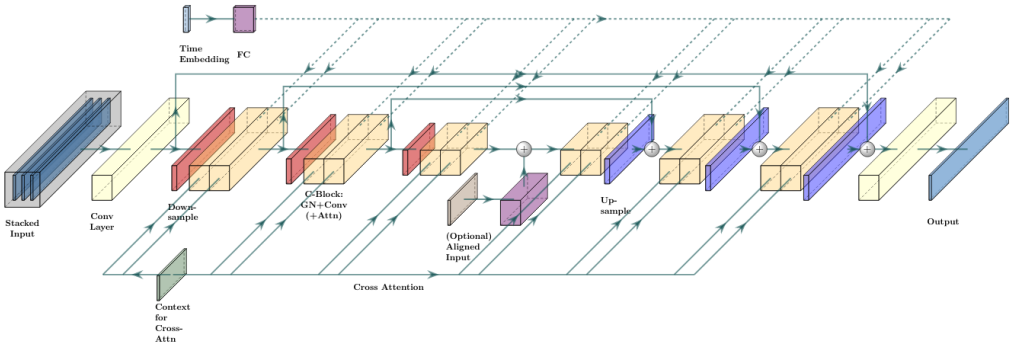


图 *1.* 扩散模型所用的 U-Net 架构。对长度为 T 的主输入序列依次应用一系列下采样块和上采样块，生成长度为 T 的输出序列。下采样块的输出通过残差连接与上采样块的输出相加。模型有四种输入模式。第一种是长度为 T 的主输入（可能经过堆叠），从图示左侧输入。 T 是目标序列长度。第二种是时间嵌入向量。第三种是文本嵌入序列，下采样/上采样块可通过交叉注意力机制关注该序列。最后，还可以输入长度为 T/C 的输入，该输入与长度为 T、压缩率为 C 的序列对齐。

声码器模型： 声码器模型仅根据频谱图生成 16kHz 音频，并将频谱图视为对齐的输入。U-Net 模型的下采样率和上采样率经过调整，以实现频谱图相对于音频的压缩率。

3.3.3. 超分辨率级联器

最后，使用一个轻量级级联器，将任一模型生成的 16kHz 波形转换为 24kHz 音频。将 16kHz 音频上采样后，与 $\mathbf{x}_t$ 堆叠，作为模型输入。该模型不使用文本条件。

3.4. 文本理解

在文本到图像扩散模型的背景下（Saharia 等人，2022b；Rombach 等人，2021），已有研究表明，强大的文本编码器能够捕捉音乐描述性文本提示的复杂性和组合性。我们采用 T5 编码器（Raffel 等人，2020），并使用未池化的词元嵌入序列来为扩散模型提供条件。对其他上下文信号进行全面比较，例如来自不同大语言模型的嵌入，或由在音乐-文本配对上训练的 CLIP 类文本编码器（Radford 等人，2021）生成的单一向量嵌入，不在本研究的讨论范围之内。

3.5. 音乐音频的伪标签

大规模训练数据是确保生成式深度神经网络质量的必要条件。例如，Imagen（Saharia 等人，2022b）使用 O(1B) 图文对进行训练。尽管音乐内容广泛可得，但高质量的音乐文本配对数据仍然稀缺，尤其是自由形式的文本，用来描述音乐属性，而非标题、艺术家姓名、专辑名称和发行年份等高级元数据。

为生成此类音乐-文本配对，我们采用伪标签方法，利用 MuLan（Huang 等人，2022）这一预训练文本与音乐音频联合嵌入模型，以及 LaMDA（Thoppilan 等人，2022）这一预训练大语言模型，为未标注的音乐音频片段赋予细粒度语义伪标签。

我们首先整理出若干音乐描述词汇集，每个词汇集都包含大量音乐描述文本。如以下所示，这些文本与标准音乐分类基准（例如 MagnaTagATune、FMA 和 AudioSet）中各标签类别的描述相比，在规模和语义粒度上都大不相同。我们考虑以下三种描述词汇：

LaMDA-LF：我们向大语言模型 LaMDA 提供歌曲标题和艺术家姓名，让其描述一份包含 15 万首流行歌曲的列表。具体的提示模板见补充材料。随后，我们将 LaMDA 回复处理成 400 万条干净的长文本句子，这些句子很可能是在描述音乐。我们选择 LaMDA 作为语言模型，是因为它经过对话应用训练；我们预计它生成的文本会更接近用于生成音乐的用户提示。

Rater-LF：我们从 MusicCaps（Agostinelli 等，2023年）获取了 10,028 条评分者撰写的描述，并将每条描述拆分为独立句子。这一处理产生了 35,333 条描述音乐的长文本句子。

Rater-SF：从上述同一评估数据集中，我们收集了评审者撰写的所有简短音乐属性标签，构成了一个包含 23,906 个词条的词汇表。

描述词汇的示例见表 1。

表 *1.* 描述词汇示例。

词汇表	示例
LaMDA-LF (4M)	“轻盈、空灵的鼓律动营造出热带氛围。”, “轻快的电子舞曲鼓点伴随着贝斯吉他、弦乐, 背景中还有简单的钢琴和打击乐。”
Rater-LF （35k）	“苏格兰次中音鼓演奏着行进节拍。” “音色强劲的贝斯吉他与吉他形成对比。”
Rater-SF （24k）	“50 年代流行乐”、“宽广而富有激情的男声”、“复古氛围”, “爱国情绪”、“活泼的大提琴”、“健身音乐”

我们将 MuLan 模型用作零样本音乐分类器，从词汇表中为未标注的音频片段分配描述。MuLan 由文本编码器和音频编码器组成，二者采用对比学习方案，在大量噪声极高的文本—音乐配对数据上进行训练。与 CLIP（Radford 等人，2021）将图像和文本共同嵌入的方式类似，一段 10 秒长的音乐音频片段和一句描述该音乐的句子，会被置于 MuLan 学得的同一语义嵌入空间中的相近位置。对于每个音频片段，我们先将其分割为互不重叠的 10 秒窗口，再计算每个窗口的 MuLan 音频嵌入的平均值，以此得到该片段的音频嵌入。我们还会计算词汇表中所有候选描述的文本嵌入。对于每个片段，我们选取在嵌入空间中与音频最接近的前 K 个描述。我们统计所有片段中各描述出现的频次，并针对每个片段，按照与频次成反比的概率，进一步从这 K 个描述中抽取 K' 个。最后这一步抽样旨在平衡标签分布并增加描述的多样性。我们使用 $K = 10$和 $K' = 3$。

作为对大型训练集进行伪标签处理的预热，我们将这种伪标签方法应用于 AudioSet 训练集和测试集中，分别筛选出位于 AudioSet 本体音乐子树下且带有标签的 388,262 个和 4,497 个样本，从而构建了音乐描述数据集 MuLaMCap。每段包含音乐内容的 10 秒音频都配有 3 条来自 LaMDA-LF 词汇表的描述、3 条来自 Rater-LF 的描述，以及 6 条来自 Rater-SF 的短描述。

3.6. 训练数据挖掘

为组建大规模音频-文本对集合，我们收集了约 680 万个音乐音频源文件。对于每条配乐，我们提取六个互不重叠的 30 秒片段。这些片段总计接近 34 万小时的音乐。训练超分辨率模型时，音频采样率为 24kHz；训练所有其他模型时，采样率为 16kHz。

对于每首配乐，我们考虑三类噪声文本标签——歌曲标题、与配乐相关的命名实体标签（例如音乐流派、艺术家姓名、乐器）以及伪标签。我们使用三种来自

表 *2*。本文训练的模型。词元长度指训练时文本提示的词元长度。

模型	# 参数量	# 训练步数	词元长度	Loss 权重	噪声调度
波形生成器	724M	1.6M	64	简化	余弦
波形级联器	487M	460k	64	简化	线性
频谱图生成器	745M	1.8M	96	西格玛	线性
频谱图声码器	25.7M	840k	-	简化	线性
超分辨率级联器	81M	270k	-	简化	线性

LaMDA-LF 词汇表的伪标签，以及六种来自 Rater-SF 词汇表的伪标签。LaMDA-LF 和 Rater-SF 中的伪标签为命名实体标签提供了互补信息。与客观且高层次的标签相比，伪标签包含与活动（“适合在高速公路上驾驶时聆听的音乐”）和情绪（“轻松自在的感觉”）相关的主观描述，也包含具有细粒度语义的编曲元素。由于我们在 MusicCaps（Agostinelli等，2023年）上评估模型，而 Rater-LF 的句子正是源自该数据集，因此我们将 Rater-LF 词汇表中的所有伪标签排除在训练数据之外。

我们向大型伪标注训练集中加入了少量高质量音频。这些音频取自内部维护的音乐库中的部分音乐曲目，且无需注明来源。我们将这些音乐曲目切分为互不重叠的 30 秒片段，并将曲目的元数据拼接起来，作为音频的文本提示。这为我们的训练数据贡献了约 300 小时的标注音频。

4. 实验与结果

4.1. 模型训练细节

我们为这项工作训练了四个一维 U-Net 模型：波形生成器和级联器，以及频谱图生成器和声码器。我们在表 2 中总结了这些模型的一些基本信息，而有关模型的更多细节则放在补充材料中。值得注意的是，我们发现西格玛加权损失对于频谱图生成器的收敛至关重要；它会对去噪调度“后端”的损失赋予更高权重。

除声码器外，所有模型均使用音频-文本对进行训练，而声码器仅使用音频进行训练。对于每个音频样本，都会构建一个文本批次。三个较长的提示词构成文本批次中的三个独立元素，而较短的提示词则会先拼接，再切分为一组固定词元长度的片段（长度见表 2），并添加到文本批次中。训练时，会为每个音频片段从相应的文本批次中随机选取一个元素，作为与该音频配对的文本输入模型。

模型采用 Adam优化 进行训练，使用 β_1 =0.9 和 $\beta_2 = 0.999$。采用余弦学习率调度，终点设为 250 万步，峰值学习率为

表3.本文所用模型的推理参数。

模型	去噪步长调度	随机性参数	CFG 引导尺度
波形生成器	前重型	0	10.0
波形级联器	前重型	1	5.0
频谱图生成器	后重型	0	5.0
频谱图声码器	前重型	0	N/A
超分辨率级联器	前重型	0	N/A

1e-4，预热步数为 1 万。对模型参数进行指数移动平均（EMA），衰减率为 0.9999，并在推理时使用。超分辨率级联器的批量大小设为 4096，其他所有模型的批量大小均为 2048。为了在推理时应用无分类器引导（CFG），我们在每个训练批次中对 10% 样本的文本提示进行遮挡。对于这些样本，交叉注意力层的输出设为零。

生成器模型使用自注意力，而级联器和声码器则不使用。因此，生成器模型需要在音频完整的 30 秒表示上进行训练，而级联模型和声码器模型则使用随机采样的 3 到 4 秒片段进行训练。

遵循（Ho 等，2022b），训练级联器/声码器模型时会应用两种数据增强。第一种是通过施加扩散噪声，随机破坏条件低保真音频或频谱图输入。为此，会在 $[0, t_{\text{max}}]$ 内随机选择一个扩散时间，并将其应用于音频的中间表示，即上采样后的低保真音频或频谱图。对于级联器，t_{max} 设为 0.5；对于声码器和超分辨率级联器，则设为 1.0。另一种是模糊增强。对于级联模型，使用大小为 10 的一维模糊核，并采用标准差范围为 0.1 到 5.0 的高斯模糊核。对于声码器模型，则应用二维 5×5 模糊核，其标准差范围为 0.2 到 1.0。

4.2. 模型推理与服务

4.2.1. 模型推理

我们调整了三个推理超参数：去噪调度、随机性参数和 CFG 引导尺度。各模型所使用的参数列于表 3 中。

我们通过时间步长 $[\delta_1, \cdots, \delta_N]$ 对去噪步数调度进行参数化，这些时间步长通过累积转换为第 3.1 节中引入的去噪步骤：$t_n = \sum_{i=1}^{n} \delta_n$。推理成本与时间步的数量成正比。因此，在固定推理成本的情况下优化时间步调度，实质上是分配固定数量的时间步，使其总和为总时间 1。去噪步骤调度的参数空间极其庞大。尽管如此，我们仍对三种不同类型的调度进行了实验，分别称为“前置密集型”“均匀”和“后重型”。前置密集型调度会在“前部”分配较多步数，即

表 4。四个样本在四个 TPU V4 上的推理耗时。

模型	时间/步（毫秒）	步数	时间（秒）
波形生成器	25.0	1000	25.0
波形级联器	75.0	800	60.0
频谱图生成器	8.3	1000	8.3
频谱图声码器	29.9	100	0.3
超分辨率级联器	71.7	800	57.3

在 $t = 0$ 附近，而“后重型”调度则在 $t = 1$附近消耗更多步数。均匀调度使用间隔相等的时间步。具体使用的调度见补充材料。

4.2.2. 模型服务

我们在 Google Cloud TPUV4 上提供模型服务，每个服务请求都会生成四段 30 秒的音乐片段。我们使用 GSPMD（Xu 等人，2021）将模型分布到四个 TPU V4 设备上，从而将服务时间缩短了 50% 以上。表 4 展示了在四个 TPU V4 上提供模型服务以生成四个样本时的推理时间开销。

4.3. 评估

4.3.1. 模型参数选择

模型参数包括架构、训练超参数、检查点和推理参数，均以启发式方式进行选择。作者设计了一小组开发提示词，这些提示词独立于文中所述各评估集中的提示词，并用于从训练好的模型生成音频。模型参数的选择依据生成结果的质量（由作者评判），同时也会考虑计算资源和时间是否充足等实际限制。评估均基于 16kHz 波形进行——生成评估指标所需的波形时，不使用超分辨率级联器。

4.3.2. 评估指标

我们使用两类指标来衡量文本条件音乐生成模型的质量：Fr´echet 音频距离（FAD）（Kilgour 等人，2018）和 MuLan 相似度得分（Huang 等人，2022）。

FAD 衡量生成音频示例的音频质量与一组参考音频片段的音频质量之间的差异。具体而言，我们使用音频编码器，分别计算生成音频示例集合和评估数据集中的背景音频片段的音频嵌入。假设来自每个集合的嵌入均服从高斯分布，并根据两组的平均嵌入向量和相关矩阵计算两个分布之间的 Fr´echet 距离。

计算 FAD 指标时会使用三种音频编码器

：一种是在 YouTube-8M（Abu-El-Haija 等人，2016）上训练的 VGG[1] 音频事件嵌入模型（Hershey 等人，2017）；一种是 Trill（Shor 等人，2020）模型[2]，，即在包含 AudioSet 片段的语音数据上训练的卷积式语音表征学习模型；还有 MuLan 音频编码器。VGG 和 Trill 生成逐帧嵌入，而 MuLan 的嵌入则针对整个片段。由于这些音频编码器是在不同的数据集和任务上训练的，使用这些音频表征计算出的 FAD 关注音频的不同方面。我们推测，FAD_{VGG} 评估总体音频质量，FAD_{Trill} 更能反映人声质量，而 FAD_{Mulan} 则捕捉全局音乐语义。

对比模型 MuLan 为我们提供了一种量化音频-文本对以及音频-音频对之间相似度的方法。对于给定的文本-音频对或音频-音频对，我们将 MuLan 相似度定义为两个实体的 MuLan 嵌入之间的余弦相似度。对于给定的音乐-文本对评估集，我们计算由数据集文本提示生成的音频与对应文本或真实音频之间的平均相似度。作为参考，我们还计算评估集与真实音频之间的平均 MuLan 相似度，以及通过打乱真实音频获得的“随机”音频配对的平均 MuLan 相似度。

4.3.3. 评估数据集

我们针对以下三个由文本-音乐对组成的数据集报告 FAD 和 MuLan 相似度。

首先，我们重新利用音频标签基准 MagnaTa-gATune（MTAT）（Law 等人，2009年），以评估我们的模型生成的 29 秒音乐片段。MTAT 包含 25,863 个音乐片段，其中 21,638 个片段与词汇表中的 188 个音乐标签中的多个标签相关联。我们仅使用这 21,638 个示例进行评估，并将每个示例的音乐标签连接成一个字符串，作为对应的文本提示。在评估过程中，我们针对这 21,638 个示例中每个示例对应的提示词生成一个 29 秒的音频片段。

其次，我们使用 AudioSet-Music-Eval，即 AudioSet 中的音乐部分（Gemmeke 等，2017）。AudioSet 的评估划分中有 1,482 个与音乐相关的样本，其中每段 10 秒的片段都标注了 AudioSet 本体中音乐子树非琐碎部分的标签。对于该数据集中的每个样本，我们将拼接后的标签用作文本提示，生成一段 30 秒的片段，并取其中间的 10 秒部分进行评估。

最后，我们在 MusicCaps（Agostinelli等，2023）上进行评估。该数据集包含 5.5K 段来自 Au- dioSet、与评分者撰写的说明配对的 10 秒音频片段。我们使用评分者撰写的说明作为文本提示，并使用生成音频样本中间的 10 秒片段报告各项指标。

[1]tfhub.dev/google/vggish/1

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

表*5*。参考数据集音频与使用参考数据集提示词生成的音频之间的 FAD。我们使用了三种音频编码器：VGG、Trill 和 MuLan，来测量 FAD。数值越低，表示质量越接近。

数据集/模型	FAD_{VGG}	FAD_{Trill}	FAD_{MuLan}
MusicCaps（Agostinelli等，2023）			
Riffusion（Forsgren与Martiros，2022）	13.371	0.763	0.487
Mubert（MubertAI，2022）	9.620	0.449	0.366
MusicLM（Agostinelli等，2023）	4.0	0.44	-
Noise2Music 波形	**2.134**	**0.405**	**0.110**
Noise2Music 频谱图	3.840	0.474	0.180
AudioSet-Music-Eval			
Noise2Music 波形	2.240	0.252	0.193
Noise2Music 频谱图	3.498	0.323	0.276
MagnaTagATune			
Noise2Music 波形	3.554	0.352	0.235
Noise2Music 频谱图	5.553	0.419	0.346

4.4. 评估结果

在表5中，我们报告了模型在三个评估数据集上的 FAD，并将其与来自 Riffusion[3] 和 Mubert[4] 的基线模型进行比较。在表6中，我们报告了生成音频与评估数据集之间的平均音频-文本和音频-音频 MuLan 相似度分数。我们还列出了针对真实音频以及经过打乱的真实音频计算的指标。

评估指标应谨慎解读，因为与所呈现的基线相比，我们的结果可能具有潜在优势。首先，与基线相比，我们的训练数据分布可能更接近评估数据集。此外，人们可能会怀疑，基于 MuLan 的指标是否会偏向我们的模型，因为 MuLan 模型曾被用于为我们的数据生成伪标签。因此，读者不应仅依据这些指标，就对本文所使用方法相较于基线方法的有效性得出结论。不过，这些指标确实能够代表训练模型自身在 AudioSet 领域中的性能，并为最终模型性能提供定量衡量。如果我们假设 MuLan 模型与 CLIP（Radford 等人，2021）类似，已经学习到了对文本和音频数据无偏且忠实的表示，那么这一点理想情况下也应适用于基于 MuLan 的指标。

我们还进行了人工听音测试，以衡量语义对齐程度，其设置与

[3]截至2022年12月24日，我们通过 github.com/MubertAI 调用 MubertAPI，根据评估数据集中的文本提示生成10秒音频片段。[4]截至2022年12月24日，我们使用 github.com/riffusion/riffusion-app 提供的 riffusion-model-v1 进行推理，生成10秒音频片段。

表 *6*。对于每个评估集，生成音频与文本提示或真实音频之间的 MuLan 平均相似度。数值越高，表示语义对齐效果越好。

数据集/模型	音频 ↔ 真实文本	音频 ↔ 真实音频
MusicCaps（Agostinelli等，2023年）		
真实音频	0.452	(1.000)
随机打乱的音频	0.248	0.278
Riffusion（Forsgren与Martiros，2022年）	0.342	0.312
Mubert（MubertAI，2022年）	0.323	0.280
MusicLM（Agostinelli等，2023年）	**0.51**	-
Noise2Music 波形	0.478	**0.489**
Noise2Music 频谱图	0.434	0.464
AudioSet-Music-Eval		
真实音频	0.470	(1.000)
随机打乱的音频	0.274	0.265
Noise2Music 波形	**0.563**	**0.429**
Noise2Music 频谱图	0.490	0.389
MagnaTagATune		
真实音频	0.498	(1.000)
随机打乱的音频	0.277	0.315
Noise2Music 波形	**0.518**	**0.479**
Noise2Music 频谱图	0.459	0.444

表*7*。人类听力研究中成对比较的获胜次数。数值越高表示语义对齐效果越好。

MusicCaps	Noise2Music 波形	MusicLM	Mubert	Riffusion
959	718	692	254	308

（Agostinelli等，2023）中使用的设置相同，但现在包含五个来源，详见表7。测试参与者会看到一条来自MusicCaps评估集的文本描述，以及来自两个不同来源的10秒音频片段；随后，他们需要在5点李克特量表上判断哪个片段更符合该文本描述。我们共收集了3,000份评分，其中每个来源参与了1,200次成对比较。每一对来源之间的正面交锋比较均匀分布。在表7中，我们报告了每个模型在其参与的1,200次比较中获得的“胜出”总数。我们的波形模型表现与MusicLM相当，但落后于真实音频。

4.5. 推理参数消融

我们改变模型的推理参数，并观察其影响。需要指出的是，我们进行消融实验时使用的模型检查点训练程度略低于上一小节中用于生成评估数值的检查点。消融实验以表3所列的基础参数为基准进行。

在图2中，我们展示了推理过程中改变去噪步数调度和 CFG 引导尺度时，使用 VGG 测得的 FAD 以及 MuLan 相似度分数如何变化。每次只改变一个参数，其余参数均固定为基准值。

我们发现，FAD 指标与相似度分数总体上呈相关关系，但级联器的情况除外：此时 FAD 可能变差，而相似度分数却有所提升。我们

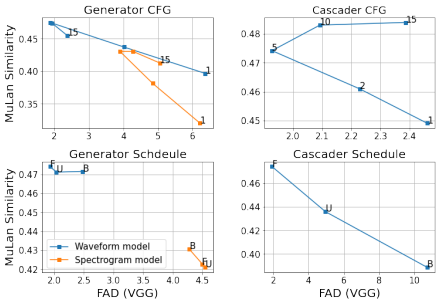


图*2*。我们绘制了调整推理参数时 FAD_{VGG} 和 MuLan 相似度分数的变化情况。CFG 参数取值范围为 [1, 2, 5, 10, 15], , 同时采用了“B”后置密集型、“U”均匀型和“F”前置密集型去噪步数调度。

还发现，CFG 引导尺度存在一个最优值，过大的 CFG 引导尺度会损害生成质量。还可以看出，生成器的 CFG 引导尺度比生成器的去噪调度影响更大，而级联器的去噪调度影响则极其显著。

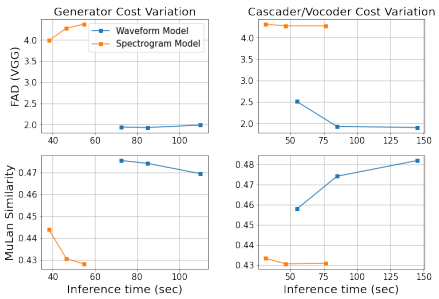


图 *3*。 生成示例的质量指标与以推理时为参数的计算成本之间的关系。

4.6. 推理成本与性能

在图 3 中，我们绘制了质量指标与通过推理时测量的推理成本之间的关系。我们减少或增加生成器或级联器/声码器的推理步数，并按反比例缩放推理调度中的步长。我们发现，增加生成器的推理成本所产生的影响并不一致，而随着级联器/声码器推理步数的增加，生成质量通常会有所提升。

5. 定性分析

内容表征： 我们在 google-research.github.io/noise2music#table-2 展示了生成示例，以说明模型能够将文本提示中所表征的音乐要素具象化。具体而言，我们发现，文本所暗示的音乐流派、乐器、情绪、人声特征和音乐时代，都体现在生成的音乐中。

创意提示词： 尽管我们的模型常常难以根据分布外提示词生成高质量音频，但仍能生成一些有趣的示例。在 google-research.github.io/noise2music#table-3 中，我们收集了一些创意提示词示例，模型能够根据这些提示词生成优质音乐。

6. 讨论

频谱图与波形方法： 频谱图方法和波形方法各有优势。与波形模型相比，本研究采用的频谱图模型训练和部署成本低得多，并且更容易扩展到更长的时间长度。这是因为频谱图的序列长度远短于低保真波形。此外，频谱图包含低保真音频中缺失的高频信息。与此同时，波形模型在生成过程的每一步都会产生可解释的表示，因此便于调试和调整模型。这也是我们能够更轻松地训练波形模型的部分原因。

未来方向： 尽管我们展示了基于文本提示的音乐生成的潜力，但除本研究之外仍有很大的改进空间。提高模型的可解释性、进一步改善文本-音频对齐、降低训练和推理成本，以及扩展音频生成长度，只是本研究有待改进的几个方向。另一个有趣的方向是，将本研究训练的模型微调用于多种音频任务，包括音乐补全和修改，就像 Saharia 等人（2022a）在图像领域所做的那样。

7. 更广泛的影响

我们相信，我们的工作有望发展成为一款实用工具，帮助艺术家和内容创作者进一步丰富他们的创作实践。为了兑现这一承诺，我们还需要与音乐家及其他利益相关者开展更多合作，将模型开发成有意义的共创工具。

我们承认所提出模型的局限性。尤其是，大型生成模型会学习模仿训练集中固有的模式和偏置；就我们的情况而言，模型可能会传播用于训练模型的文本和音乐语料中潜藏的偏置。这些偏置往往以微妙且难以预测的方式显现，因此很难被发现，也无法被我们目前的评估基准充分捕捉。由于模型学到的关联或偶然因素，模型输出可能会生成贬损性或其他有害语言。

除此之外，我们认识到，音乐流派十分复杂，关键的音乐属性具有语境依赖性，并且会随时间而变化。鉴于来自全球音乐文化的样本在录音和数字化方面并不均衡，训练数据只能反映有限的音乐样本和音乐流派语料。音乐的分类和标注可能会将音乐流派本质化；而且，这些标签可能在未有相关社群参与的情况下被构建和应用。当读者查看随附网站上发布的生成示例时，我们提醒读者不要想当然地认为每个样本都能代表整个音乐流派，也不要认为单一标签能够涵盖某一地区所产生的音乐流派的多样性（即“拉丁音乐”包含广泛的文化和风格）。此外，对于这些社群之外的人来说，音乐样本听起来可能很“真实”，因为要识别音乐传统中的细微差异，需要经过训练的听觉和文化知识。在生成人声时，文本提示在请求具有文化或宗教属性的音乐流派，或在某些社群政治斗争中形成的音乐流派时，可能产生刻板 caricature、11模仿口音”、戏仿或其他贬损性的语言伤害（例如，在请求“灵魂乐人声”时生成“模仿黑人歌唱”，或在拉丁音乐请求中生成“模仿西班牙语”）。

与任何其他技术一样，我们的研究成果也可能遭到误用或滥用。我们承认，当生成内容与训练数据中的示例完全一致时，存在被不当挪用的风险。按照负责任的模型开发实践，重复检查已纳入我们当前生成和发布示例的流程中，未来的任何工作也将继续如此。

识别潜在安全问题并加以解决，是改进这类生成模型的重要环节。在更清楚地了解其局限性和风险之前，我们无意发布该模型。

致谢

我们感谢 Aren Jansen 构建了 MuLan，它是本项目不可或缺的组成部分。感谢 Austin Tarango、Fernando Diaz、Kathy Meier-Hellstern、Molly FitzMorris 和 Renee Shelby 帮助我们将重要的负责任的人工智能实践纳入本项目。感谢 Blake Cunningham 和 Cara Adams 在项目过程中提供建议，并协助我们完成出版流程。我们感谢 Alex Ku、Andrea Agostinelli、Ankur Bapna、Chen Liang、Ed Chi、Ekin Dogus Cubuk、Er- ica Moreira、Esteban Real、Heiga Zen、Jaehoon Lee、James Qin、Nathan Park、Stephen Kelly、Thang Luoung、Weizhe Hua、Ye Jia、Yifeng Lu、Yonghui Wu、Yu Zhang 和 Yuma Koizumi 提供宝贵的反馈与支持。特别感谢 MusicLM 的作者进行了富有帮助的讨论与合作，尤其感谢他们在发表前分享评估集和手稿。

参考文献

Abu-El-Haija, S., Kothari, N., Lee, J., Natsev, P., Toderici, G., Varadarajan, B. 和 Vijayanarasimhan, S.《Youtube-8m：大规模视频分类基准》，2016 年。网址：https://arxiv.org/abs/1609.08675。

Agostinelli, A., Denk, T. I., Borsos, Z., Engel, J., Verzetti, M., Caillon, A., Huang, Q., Jansen, A., Roberts, A., Tagliasacchi, M., Sharifi, M., Zeghidour, N., and Frank, C. Musiclm：从文本生成音乐。*arxiv* 预印本 *arxiv:2301.11325*，2023年。

Borsos, Z., Marinier, R., Vincent, D., Kharitonov, E., Pietquin, O., Sharifi, M., Teboul, O., Grangier, D., Tagliasacchi, M., and Zeghidour, N. Audiolm：一种用于音频生成的语言建模方法，2022年。网址 https://arxiv.org/abs/2209.03143。

Briot, J.-P. 从人工神经网络到用于音乐生成的深度学习：历史、概念与趋势。神经计算与应用，33(1):39–65，2021年。

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shya m, P., Sastry, G., Askell, A., 等。语言模型是少样本学习者。神经信息处理系统进展，33：1877–1901，2020年。

Chen, N., Zhang, Y., Zen, H., Weiss, R. J., Norouzi, M., 和 Chan, W. Wavegrad：估计波形生成的梯度。载于国际学习表征会议，2021年。

Chen, Z., Tan, X., Wang, K., Pan, S., Mandic, D., He, L., 和 Zhao, S. Infergrad：通过在训练中考虑推理来改进声码器的扩散模型。载于*ICASSP 2022-2022 IEEE* 国际声学、语音与信号处理会议 *(ICASSP)*，第 8432–8436 页。IEEE，2022年。

Dhariwal, P., Jun, H., Payne, C., Kim, J. W., Radford, A. 和 Sutskever, I. Jukebox：一种用于音乐生成的生成模型。*arXiv* 预印本 *arXiv:2005.00341*，2020年。

Forsgren, S. 和 Martiros, H. Riffusion——用于实时音乐生成的稳定扩散。2022年。网址：https: //riffusion.com/about。

Gemmeke, J. F., Ellis, D. P., Freedman, D., Jansen, A., Lawrence, W., Moore, R. C., Plakal, M. 和 Ritter, M. Audio Set：一个用于音频事件的本体和人工标注数据集。载于*2017*年*IEEE*国际声学、语音与信号处理会议（*ICASSP*），第776–780页。IEEE，2017年。

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, K. 用于大规模音频分类的卷积神经网络架构。载于国际声学、语音与信号处理会议 *(ICASSP)*。2017年。网址：https:// arxiv.org/abs/1609.09430。Ho, J. and Salimans, T. 无分类器扩散引导。*arXiv* 预印本 *arXiv:2207.12598*，2022年。Ho, J., Jain, A., and Abbeel, P. 去噪扩散概率模型。神经信息处理系统进展，33:6840–6851，2020年。

Ho, J., Chan, W., Saharia, C., Whang, J., Gao, R., Gritsenko, A., Kingma, D. P., Poole, B., Norouzi, M., Fleet, D. J. 等。Imagen Video：使用扩散模型生成高清视频。*arXiv* 预印本 *arXiv:2210.02303*，2022a。

Ho, J., Saharia, C., Chan, W., Fleet, D. J., Norouzi, M., 和 Salimans, T. 用于高保真图像生成的级联扩散模型。*J. Mach.Learn. Res.*，23:47–1，2022b。

Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., 和 Fleet, D. J. 视频扩散模型。*arXiv* 预印本 *arXiv:2204.03458*，2022c。

Huang, Q., Jansen, A., Lee, J., Ganti, R., Li, J. Y., 和 Ellis, D. P. W. MuLan：音乐音频与自然语言的联合嵌入。载于国际音乐信息检索学会第 *23* 届会议（*ISMIR*）论文集，2022年。

Kilgour, K., Zuluaga, M., Roblek, D., 和 Sharifi, M. Fr\'echet 音频距离：一种用于评估音乐增强算法的指标。*arXiv* 预印本 *arXiv:1812.08466*，2018年。

Kong、Z.、Ping、W.、Huang、J.、Zhao、K. 和 Catanzaro、B. Diffwave：用于音频合成的通用扩散模型。载于国际学习表征会议，2021年。

Kreuk、F.、Synnaeve、G.、Polyak、A.、Singer、U.、D´efossez、A.、Copet、J.、Parikh、D.、Taigman、Y. 和 Adi、Y. Audio-gen：文本引导的音频生成。*arXiv* 预印本 *arXiv:2209.15352*，2022年。

Law、E.、West、K.、Mandel、M. I.、Bay、M. 和 Downie、J. S. 使用游戏评估算法：音乐标注案例。载于 *ISMIR*，第 387–392 页，2009年。

Manco、I.、Benetos、E.、Quinton、E. 和 Fazekas、G. 面向音乐的对比式音频-语言学习，2022年。URL https://arxiv.org/abs/2208.12208。

MubertAI。https://github.com/mubertai/mubert-text-to-music。2022年。

Nichol, A. Q. 和 Dhariwal, P. 改进的去噪扩散概率模型。载于国际机器学习会议，第 8162–8171 页。PMLR，2021年。

Popov, V., Vovk, I., Gogoryan, V., Sadekova, T., 和 Kudinov, M. Grad-tts：一种用于文本转语音的扩散概率模型。载于国际机器学习会议，第 8599–8608 页。PMLR，2021年。

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., 等。通过自然语言监督学习可迁移视觉模型。载于国际机器学习会议，第 8748–8763 页。PMLR，2021年。

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P. J. 等。利用统一的文本到文本转换器探索迁移学习的极限。机器学习研究期刊，21(140):1–67，2020年。Ramesh, A., Dhariwal, P., Nichol, A., Chu, C. 和 Chen, M. 利用 CLIP 潜变量进行层次化文本条件图像生成。*arXiv* 预印本 *arXiv:2204.06125*，2022年。Rombach, R., Blattmann, A., Lorenz, D., Esser, P. 和 Ommer, B. 利用潜扩散模型进行高分辨率图像合成，2021年。Saharia, C., Chan, W., Chang, H., Lee, C., Ho, J., Salimans, T., Fleet, D. 和 Norouzi, M. Palette：图像到图像扩散模型。载于*ACM SIGGRAPH 2022* 会议论文集，第 1–10 页，2022a。Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., Ghasemipour, S. K. S., Ayan, B. K., Mahdavi, S. S., Lopes, R. G. 等。借助深度语言理解实现照片级真实感文本到图像扩散模型。*arXiv* 预印本 *arXiv:2205.11487*，2022b。Schneider, F., Jin, Z. 和 Sch¨olkopf, B. Moˆusai：利用长上下文潜扩散进行文本到音乐生成，2023年。Shor, J., Jansen, A., Maor, R., Lang, O., Tuval, O., de Chau-mont Quitry, F., Tagliasacchi, M., Shavitt, I., Emanuel, D. 和 Haviv, Y. 迈向学习语音的通用非语义表征。载于*Interspeech 2020*论文集，第 140–144 页，2020年。doi: 10.21437/Interspeech.2020-1242。Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N. 和 Ganguli, S. 利用非平衡热力学进行深度无监督学习。载于国际机器学习会议，第 2256–2265 页。PMLR，2015年。Song, Y. 和 Ermon, S. 通过估计数据分布的梯度进行生成建模。神经信息处理系统进展，32，2019年。

Thoppilan, R., De Freitas, D., Hall, J., Shazeer, N., Kul-shreshtha, A., Cheng, H.-T., Jin, A., Bos, T., Baker, L., Du, Y. 等。LaMDA：用于对话应用的语言模型。*arXiv* 预印本 *arXiv:2201.08239*，2022年。

Wang, Y., Stanton, D., Zhang, Y., Ryan, R.-S., Batten- berg, E., Shor, J., Xiao, Y., Jia, Y., Ren, F., 和 Saurous, R. A. 风格标记：端到端语音合成中的无监督风格建模、控制与迁移。载于国际机器学习会议，第 5180–5189 页。PMLR，2018年。

Wu, S. and Shi, Z. It\ˆottsand it\ˆ owave：音频生成只需线性随机微分方程。*arXiv*预印本 *arXiv:2105.07583*，2021年。

Xu, Y., Lee, H., Chen, D., Hechtman, B. A., Huang, Y., Joshi, R., Krikun, M., Lepikhin, D., Ly, A., Maggioni, M., Pang, R., Shazeer, N., Wang, S., Wang, T., Wu, Y., 和 Chen, Z. GSPMD：面向机器学习计算图的通用且可扩展并行化。*CoRR*, abs/2105.04663, 2021年。网址：https://arxiv.org/abs/2105.04663。Yang, D., Yu, J., Wang, H., Wang, W., Weng, C., Zou, Y., 和 Yu, D. Diffsound：用于文本到声音生成的离散扩散模型。*arXiv* 预印本 *arXiv:2207.09983*, 2022年。Yu, J., Xu, Y., Koh, J. Y., Luong, T., Baid, G., Wang, Z., Vasudevan, V., Ku, A., Yang, Y., Ayan, B. K. 等。扩展自回归模型，实现丰富内容的文本到图像生成。*arXiv*预印本 *arXiv:2206.10789*，2022年。

A. 扩散模型

在本节中，我们回顾了 扩散模型的相关信息，并说明论文正文中使用的符号
本文遵循（Saharia等人，2022b）的表述。

扩散模型假设这样一种设置：分布中的样本 $\mathbf{x}$ 经过高斯扩散过程和噪声调度的扰动，其中噪声调度由时间 t处单调递增的标准差 σ_t 表示。更准确地说，在给定 $\mathbf{x}$ 或对于 $s < t$的 $\mathbf{x}_s$ 的条件下，时间 t 处的受扰样本 $\mathbf{x}_t$ 的分布为:

$$q(\mathbf{x}_t|\mathbf{x}) = \mathcal{N}(\alpha_t \mathbf{x}, \sigma_t^2 \mathbf{I})\,, \qquad q(\mathbf{x}_t|\mathbf{x}_s) = \mathcal{N}((\alpha_t/\alpha_s)\mathbf{x}_s, \sigma_{t|s}^2 \mathbf{I})\,. \tag{2}$$

我们可以定义一组变量，让这些方程更加简洁:

$$\alpha_t = \sqrt{1-\sigma_t^2}\,, \quad \lambda_t = \ln(\alpha_t^2/\sigma_t^2)\,, \quad \sigma_{t|s}^2 = (1-e^{\lambda_t-\lambda_s})\sigma_t^2\,, \quad \tilde{\sigma}_{s|t}^2 = (1-e^{\lambda_t-\lambda_s})\sigma_s^2\,. \tag{3}$$

假设时间变量 s 和 t 的取值范围为 $[0,1]$。

如第一个方程所示，受损样本 $\mathbf{x}_t$ 的随机性由单个噪声向量 $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$编码。扩散模型的目标是根据受损样本、时间 t 以及上下文 $\mathbf{c}$:$\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, \mathbf{c}, t)$对该噪声向量进行建模。

模型训练完成后，我们在时间 $t = 1$采样噪声，即 $x_1 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$，并反转扩散过程，从原始分布中生成一个“干净”的样本 $\mathbf{x}_0$。为此，我们采用祖先采样（或去噪扩散概率模型采样）（Ho 等人，2020）。在这种采样方法中，我们选取一些时间步 $0 = t_0 < \cdots < t_N = 1$，并通过应用更新规则反转扩散过程，从 $\mathbf{x}_t$得到$\mathbf{x}_s$：

$$\mathbf{x}_s = \frac{\alpha_s}{\alpha_t}\mathbf{x}_t - (1-e^{\lambda_t-\lambda_s}) \cdot \frac{\alpha_s}{\alpha_t} \cdot \sigma_t \cdot \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, \mathbf{c}, t) + \tilde{\sigma}_{s|t}^{1-\gamma} \cdot \sigma_{t|s}^{\gamma} \cdot \tilde{\boldsymbol{\epsilon}}\,, \tag{4}$$

其中 $s = t_{N-n}$ 和 $t = t_{N-n+1}$ 位于第 n次更新步骤。这里， $\tilde{\boldsymbol{\epsilon}}$ 是在每个推理步骤采样的随机标准正态向量。 γ 是控制扩散过程随机性的超参数。随着 γ 增加，在给定推理步骤中引入的方差会更大，因为对于 $t > s$， $\sigma_{t|s} > \tilde{\sigma}_{s|t}$。

A.1. 噪声调度

我们使用两种 噪声调度——线性调度（Ho 等人，2020）和余弦调度（Nichol 和 Dhariwal，2021） s.

线性调度 （Ho 等人，2020）定义为 σ_t^2 随时间线性增加 :

$$\sigma_t^2 = (\sigma_1^2 - \sigma_0^2)t + \sigma_0^2\,. \tag{5}$$

我们在本文中使用初始值 $\sigma_0^2 = 0.0001$ 和最终值 $\sigma_1^2 = 0.02$。

我们采用余弦调度（Nichol 和 Dhariwal，2021年）的一种轻微变体:

$$\alpha_t = \cos(at+b) \tag{6}$$

其中， $a = \arctan\ e^{10} - \arctan\ e^{-10}$和 $b = \arctan e^{-10}$的取值使得 $\lambda_{t=0} = 20$和 $\lambda_{t=1} = -20$。

A.2. 去噪步数调度

这里列出了生成器以及级联器/声码器模型在推理时使用的实际去噪步数调度。这些调度按照第 4.3.1 节中的说明，通过试错进行手动调优。表 8 以 Python 表示法给出了每个模型及用于消融的调度的时间步长数值。用于最终评估的调度以斜体显示。请注意，对于频谱图生成器，后置密集型调度已针对评估进一步修改。同时，图 4 绘制了每种模型调度在给定推理步骤所达到的去噪时间。

B. 模型架构

我们介绍本文所用模型架构的一些细节。U-Net 的整体结构如图 1 所示。输入经过入口卷积层处理后，会依次通过一系列

表 8。去噪 调度的时间步值。每个模型选用的去噪调度以斜体标示 d.

模型	调度	时间步
波形生成器	前重型 后重型 均匀	前重型 = [0.01/200] * 200 + [0.04/400] * 400 + [0.15/200] * 200 + [0.3/150] * 150 + [0.5/50] * 50 后重型 = 前重型[::-1] 均匀 = [1.0/1000] * 1000
波形级联器	前重型 后重型 均匀	前重型 = [0.05/400] * 400 + [0.15/200] * 200 + [0.3/150] * 150 + [0.5/50] * 50 后重型 = 前重型[::-1] 均匀 = [1.0/800] * 800
频谱图生成器	前重型 后重型 均匀	前重型 = [0.01/400] * 400 + [0.04/800] * 800 + [0.15/400] * 400 + [0.3/300] * 300 + [0.5/100] * 100 后重型 = 前重型[::-1] 均匀 = [1.0/1000] * 1000
	后重型	后重型评估 = [0.3/50] * 50 + [0.3/150] * 150 + [0.2/300] * 300 + [0.2/500] * 500
频谱图声码器	前重型 后重型 均匀	前重型 = [0.05/50] * 50 + [0.15/30] * 30 + [0.3/15] * 15 + [0.5/5] * 5 后重型 = 前重型[::-1] 均匀 = [1.0/100] * 100
超分辨率级联器	前重型	前重型 = [0.05/400] * 400 + [0.15/200] * 200 + [0.3/150] * 150 + [0.5/50] * 50

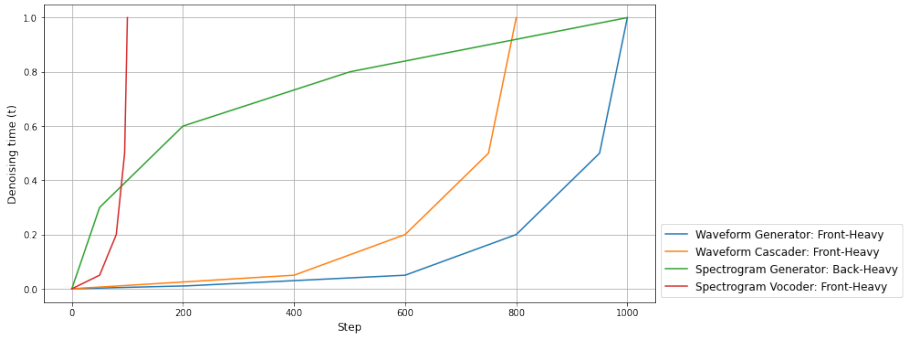


图 4。去噪时间 在论文所使用的四种去噪时间步调度中，于给定推理步骤达到的去噪时间。前重型 y
调度将其大部分 步数花费在接近 $t = 0$ 的位置，而后重型调度则将其大部分步数花费在接近 $t = 1$ 的位置。

下采样层和上采样层，并在其间穿插卷积块。下面我们将详细说明 U-Net 下采样/上采样部分中卷积块的结构，并解释生成流水线中所用各个网络的配置方式。

B.1. 下采样块和上采样块

模型架构与高效 U-Net（Saharia 等人，2022b）高度相似，只是将二维卷积层替换为一维卷积层。两者之间还存在一些细微差异，下面将对此进行说明。

如图 1 所示，U-Net 模型是自身的镜像。给定模型深度 D，模型采用 D 个下采样层和 D 个上采样层，我们可以将它们分别标记为 $I = 1$、$\cdots$、D。每个下采样层都是步幅为 S_I 的一维卷积层。因此，设深度为 I 处前向传播输入的特征长度和特征维度分别为 T_I 和 C_I，则下采样层接受长度和通道维度为 $(T_{I-1}$、$C_{I-1})$ 的输入，并将其映射为维度为 $(T_I$、$C_I)$ 的输出，$T_I = T_{I-1}/S_I$。上采样层执行镜像操作：维度为 $(T_I$、$C_I)$ 的输入被映射为维度为 $(T_{I-1}$、$C_{I-1})$ 的输出，上采样步幅为 S_I。

在下采样层和上采样层之间，使用维度均匀的“卷积块”。U-Net 的下采样部分和上采样部分使用完全相同的块，并且在相同深度使用相同数量的块。块的结构如图 5 所示。所有块都会通过“组合嵌入”层与时间嵌入向量交互；该向量通过位置嵌入将浮点数转换为向量，再经由线性层得到。而自注意力层和交叉注意力层只会在选定的深度启用，有些情况下甚至完全不使用。“组合嵌入”层对时间嵌入应用一个全连接层，以计算逐通道缩放和偏置向量，并将其应用于输入序列。自注意力层和交叉注意力层包含标准的注意力后残差层，其隐藏层的维度是 C_I 的两倍。

与此同时，入口卷积层和出口卷积层的核大小均为 3。第一个下采样层的输入和最后一个上采样层的输出的“第零”通道宽度均设为“基础模型维度”。时间嵌入维度也设为这一数值。

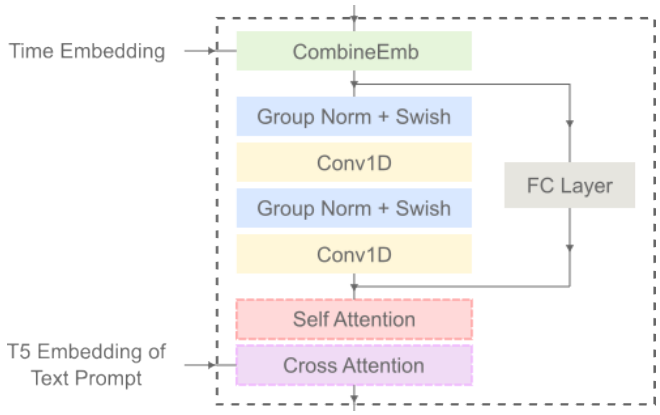


图 *5*。构成一维 U-Net 基本操作单元的卷积块结构。自注意力层和交叉注意力层为可选项。

现在让我们总结决定网络架构的关键超参数 k.

- 基础模型维度
- 卷积核大小
- 网络深度 D
- 下采样因子列表的长度-D
- 长度为 D，表示各深度所使用的块数列表
- 长度为 D，表示各深度是否使用自注意力/交叉注意力的布尔值列表

注意力所使用的头数

架构的其余部分保持固定。

B.2. 模型规格

本文所用模型的架构参数可由表9概括。

表*9.* 本研究所用模型的架构参数。

参数	波形生成器	波形级联器	频谱图生成器	频谱图声码器	超分辨率级联器
基维度	256	256	256	128	256
核大小	7	7	9	5	7
深度	6	5	5	4	4
步幅	[4, 4, 4, 4, 4, 4]	[4, 4, 4, 4, 4]	[4, 3, 2, 2, 2]	[2, 4, 4, 5]	[4, 4, 4, 4]
# 通道	256 * [1, 1, 2, 3, 4, 4]	256 * [1, 1, 2, 3, 4]	256 * [1, 2, 2, 3, 4]	128 * [1, 1, 2, 4]	256 * [1, 1, 2, 2]
# 块	[3, 3, 3, 4, 4, 4]	[3, 3, 4, 6, 6]	[6, 6, 6, 6, 6]	[4, 4, 4, 4]	[4, 4, 4, 4]
自注意力	[假、假、假、真、真、真]	[F, F, F, F, F]	[F、F、F、T、T]	[F、F、F、F]	[F、F、F、F]
交叉注意力	[假、假、假、真、真、真]	[F、F、F、T、T]	[F、F、F、T、T]	[F、F、F、F]	[F、F、F、F]
注意力头	8	8	8	-	-

C. 提示模板：引导 LaMDA 模型生成音乐描述性文本

Katrina & The Waves 的《Walking on Sunshine》：这首歌是一首流行/摇滚歌曲。它具有欢快、明朗的情绪，配以强劲的贝斯线和简单、重复的鼓点。这首歌由合成器和吉他伴奏。

Lady Gaga 的《Born This Way》：这首电音流行歌曲以轰鸣的合成器音效、嗡鸣的贝斯和额外的副歌打击乐为伴奏，临近结尾时则加入了独奏管风琴。女歌手的嗓音自信而有力。

Emine 的《Till I Collapse》m：这是一首嘻哈歌曲，情绪强烈而富有攻击性，节拍强劲且充满活力 t.

席琳·迪翁的《My Heart Will Go On》：这首流行歌曲的情绪悲伤而心碎。歌曲特别注重器乐编排，锡哨的运用十分突出，并辅以弦乐和节奏吉他的旋律。这首歌融合了原声和电子乐器。女歌手的演唱情感充沛，极具感染力。

梅根·特瑞娜的《Me Too》：这是一首中速流行歌曲，情绪欢快、积极，伴有强劲的贝斯线和简单、重复的鼓点。歌曲由合成器和吉他伴奏。

埃inaudi 的《Nuvole Bianche》：这是一首钢琴古典乐曲，情绪温柔而忧郁，节奏舒缓缓慢。

皇后乐队的《Don’t Stop Me Now》：这首充满活力的摇滚歌曲以钢琴、贝斯吉他和鼓为基础。歌手们兴奋不已，蓄势待发，歌声振奋人心。

酷玩乐队的《Strawberry Swing》：这首另类/独立歌曲融入了非洲流行乐和海莱夫音乐的元素，以指弹的无失真吉他、厚重的贝斯线和迷幻合成器为基础。这是一首中速歌曲，包含回响吉他、受钢琴抒情曲启发的旋律，以及苦乐交织、充满颂歌气息的假声演唱。

盖瑞·朱尔斯的《Mad World》：这首新浪潮/合成器流行歌曲仅以一组钢琴和弦、一架模仿大提琴的梅洛特朗电子琴、轻柔点缀的电钢琴，以及副歌中适度使用的声码器，为男歌手伴奏。

萨姆·库克的《A Change is Gonna Come》：这首灵魂乐/节奏布鲁斯歌曲情绪沉思而怀旧。男歌手的人声清晰地位于前景，伴奏由管乐器、弦乐以及贯穿过渡段的定音鼓构成。圆号传达出一种忧郁感。

涅槃乐队的《Smells Like Teen Spirit》：这首另类摇滚/硬摇滚歌曲以安静的主歌开场，伴有摇曳的合唱效果吉他，随后转入宏大、响亮的受硬核音乐启发的副歌。整体情绪叛逆而振奋。吉他和弦采用双轨录音，营造出更具力量感的声音。

披头士乐队的《Strawberry Fields Forever》：这首迷幻摇滚歌曲采用了反向录制的乐器声、梅洛特朗长笛音色、印度斯瓦尔曼达尔琴、磁带循环和渐弱/渐强尾声，并加入了大提琴和铜管编配。人声略带不协和感，营造出苦乐交织且阴郁的氛围。

{标题} 作者 {艺术家}：

D. AudioSet 音乐标签

AudioSet 标签采用知识共享署名-相同方式共享 4.0 国际版（CC BY-SA 4.0）许可协议。/m/0z9c,/m/0mkg,/m/042v gx,/m/0fd3y,/t/dd00036,/m/025td0t,/m/0192l,/m/018j2,/m/0bm02,/m/018vs,/m/02cz 7,/m/0395lw, /m/0gg8l,/m/0155w,/m/0l14 3,/m/01kcd,/m/015vgc,/m/01xqw,/m/02bk07,/m/0l14jd,/m/02mscn,/m/0140xf,/m/01wy6,/m/0ggq0m, / m/01lyv,/m/0239kh,/m/01qbl,/m/0ggx5q,/m/02bxd,/m/026z9,/m/02fsn,/m/0283d,/m/02hnl,/m/02k mr,/m/026t6,/m/07s72n, / m/02sgy,/m/08cyft,/m/02lkt,/m/03xq f,/m/0m0jc,/t/dd00035,/m/0326g,/m/0l14j ,/m/02w4v,/m/0319l,/m/02x8m,/t/dd00032, / m/0dwtp,/m/0mbct,/m/0dls3,/m/0342h,/m/03gvt,/t/dd00031,/m/03qjg,/m/03m5k,/m/03q5t,/m/03lty,/m/0glt670,/m/03mb9, / m/05rwpb,/m/03 d0,/m/03r5q ,/m/05148p4,/m/07pkxdp,/m/0j45pbj,/m/04rzd,/m/0dwsp,/m/06j64v,/m/05fw6t,/m/0164x2, / m/028sqc,/m/0dq0md,/m/0g293,/m/02v2lh,/m/05pd6,/m/013y1f,/m/0l14md,/m/05r5c,/m/0fx80y,/m/064t9,/m/0dl5d,/m/05w3f, / m/05r6t,/m/05r5wn,/m/06cqb,/m/06j6l,/m/03t3fj,/m/07sbbz2,/m/06by7,/t/dd00033,/m/0ln16,/m/06ncr,/t/dd00037, / m/01hgjl,/m/0l14l2,/m/0l14t7,/m/0jtg0,/m/06rqw,/m/06rvn,/m/0gywn,/m/0l14gg,/m/06w87,/m/0l156b,/m/02qmj0d, / m/07s0s5r,/m/015y n,/m/0l14qv,/m/01p970,/m/07brj,/m/01glhc,/m/07gxw,/t/dd00034,/m/02cjck,/m/07kc ,/m/011k j, / m/02p0sh1,/m/07lnk,/m/07c6l,/m/07gql,/m/016622,/m/07xzm,/m/0dwt5,/m/01z7dr,/m/07y 7,/m/0y4f8,/m/04wptg,/m/085jw, / m/01sm1g,/m/01bns