arXiv:2301.12661v1 [cs.SD] 30 Jan 2023

Make-An-Audio：使用提示增强扩散模型进行文本到音频生成

Rongjie Huang [*1] **Jiawei Huang** [*1] **Dongchao Yang** [*2] **Yi Ren** [3] **Luping liu** [1] **Mingze Li** [1] **Zhenhui Ye** [1] **Jinglin Liu** [1] **Xiang Yin** [3] **Zhou Zhao** [1]

摘要

大规模多模态生成建模在文本转图像和文本转视频领域取得了里程碑式进展。然而，音频领域的应用仍然落后，主要有两个原因：缺乏包含高质量文本-音频对的大规模数据集，以及对长时间连续音频数据进行建模的复杂性。在这项工作中，我们提出了Make-An-Audio，这是一种采用提示增强扩散模型的方法，通过以下方式弥补这些差距：1）引入结合先蒸馏后重编程方法的伪提示增强，利用无语言音频进行数量级庞大的概念组合，从而缓解数据稀缺；2）利用频谱图自编码器预测自监督音频表征，而非波形。结合稳健的对比式语言-音频预训练（CLAP）表征，Make-An-Audio 在客观和主观基准评估中均取得了最先进成果。此外，我们展示了其在**“不让任何模态掉队”**理念下对 X到音频 的可控性和泛化能力，首次实现了根据用户定义的模态输入生成高清、高保真音频。音频样本可在此处获取：

https://Text-to-Audio.github.io

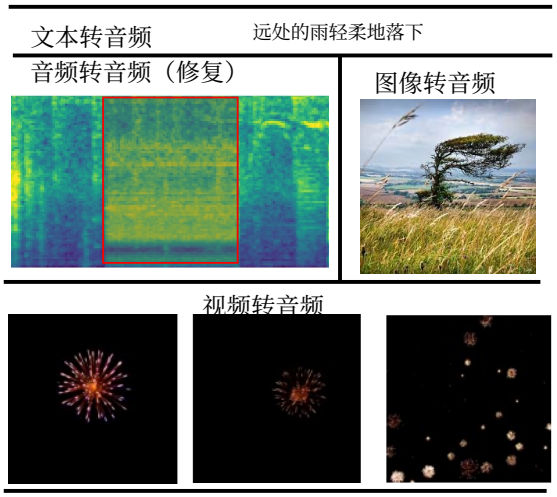


图 *1.***不让任何模态掉队：** Make-An-Audio 能够很好地泛化到具有多种用户自定义输入（文本、音频、图像和视频）的 X 到音频任务，使人们能够创作丰富多样的音频内容，并拓展个性化迁移和细粒度控制等多种应用。

1. 引言

深度生成模型（Goodfellow 等人，2020年；Kingma 和 Dhariwal，2018年；Ho 等人，2020年）近期在各种数据模态中展现出了高质量的样本。凭借大规模训练数据和强大的模型，文本转图像（Saharia 等人，2022年；Ramesh等人，2021年；Nichol 等人，2021年）和文本转视频（Singer 等人，2022年；Hong 等人，2022年）模型如今已能够生动地呈现文本提示词所描述的视觉场景，并使人们能够以前所未有的轻松方式创作丰富多样的视觉内容。然而，由于缺乏大规模、高质量的文本-音频配对数据集，加之对长连续信号数据进行建模极其复杂，这种成功难以复制到音频领域。

在这项工作中，我们提出了 Make-An-Audio，这是一种用于文本转音频（T2A）生成的提示增强扩散模型。为缓解数据稀缺问题，我们引入了一种伪提示增强方法，以构造与音频良好对齐的自然语言，从而能够利用数量级更大的无监督无语言数据。为应对文本转音频生成中对复杂音频信号建模的挑战，我们引入了一种频谱图自编码器，用于预测自监督表征，而非波形，从而确保高效压缩和高层次语义理解。结合对比式语言-音频预训练（CLAP）（Radford 等人，2021年；Elizalde 等人，2022年）和高保真扩散模型（Ho 等人，2020年；Song 等人，2020年；Rombach 等人，

[*]贡献相同 [1]浙江大学 [2]北京大学 [3]字节跳动人工智能实验室语音与音频团队。通讯作者：Zhou Zhao <ZhaoZhou@zju.edu.cn>。

2022年），该模型实现了深层次的语言理解和高保真生成。

尽管概念上简单且易于训练，Make-An-Audio 却取得了出人意料的优异结果。主观和客观评估均表明，Make-An-Audio 在文本转音频领域实现了新的最先进水平，能够进行自然且可控的合成。在基准数据集 AudioCaption 上，Make-An-Audio 展现出更优的音频质量和文本-音频对齐忠实度；甚至还能以零样本方式很好地泛化到无监督的 Clotho 数据集。

我们首次阐明了采用不同输入模态进行音频生成的必要性。除了自然语言之外，Make-An-Audio 还能很好地泛化到多种用户定义的输入模态（音频、图像和视频），使人们能够创作丰富多样的音频内容，并为个性化迁移和细粒度控制开辟了众多应用场景。

本文的主要贡献包括：

- 我们提出了 Make-An-Audio，这是一种有效的方法，利用潜在扩散和频谱图自编码器对长连续波形进行建模。
- 我们采用先蒸馏后重编程方法引入伪提示增强，通过开放无语言音频的使用来纳入大量概念组合，从而缓解数据稀缺问题。
- 我们研究文本表示，并强调对比式语言-音频预训练在高效深入理解自然语言方面的优势。
- 我们对Make-An-Audio进行评估，并展示了最先进的定量结果，以及包含定性结果的全面评估。
- 我们将这一强大模型推广至X到音频生成，首次实现了根据用户定义的模态输入生成高清、高保真音频的能力。

2. 相关工作

2.1. 文本引导的图像/视频合成

随着深度生成模型的快速发展，文本引导的合成已在图像和视频领域得到广泛研究。DALL-E 的开创性工作（Ramesh等人，2021年）使用矢量量化 VAE（Van Den Oord等人，2017年）将图像编码为离散潜在词元，并将文本到图像生成视为序列到序列翻译问题。近年来，借助大规模扩散模型，已取得令人瞩目的视觉效果。GLIDE（Nichol 等人，2021年）训练文本到图像上采样模型，以实现级联生成。Imagen（Saharia 等人，2022年）实现了具有前所未有照片级真实感和深度语言理解能力的文本到图像生成。Stable Diffusion（Rombach 等人，2022年）采用潜空间扩散而非像素空间扩散，以提高计算效率。大量研究也在探索将文本到图像模型用于视频生成。CogVideo（Hong 等人，2022年）基于 CogView2（Ding 等人，2022年）文本到图像模型构建，采用多帧率分层训练策略。Make-A-Video（Singer 等人，2022年）通过时空因子分解扩散模型，将基于扩散的文本到图像模型扩展为文本到视频模型。

我们的研究超越了视觉生成，旨在根据任意自然语言生成高保真音频，而这一方向此前受到的关注相对较少。

2.2. 文本引导的音频合成

尽管文本引导的视觉生成取得了显著进展，文本转音频（T2A）生成的发展仍然滞后，主要有两个原因：缺乏包含高质量文本-音频对的大规模数据集，以及对长连续波形数据进行建模的复杂性。DiffSound（Yang 等人，2022年）率先探索了采用离散扩散过程进行文本转音频生成的方法，该过程作用于从向量量化变分自编码器（VQ-VAE）获取的音频编码，并借助 CLIP表征进行掩码文本生成。AudioLM（Borsos 等人，2022年）引入了在音频上预训练的掩码语言模型的离散化激活，并生成语法上合理的语音或音乐。

最近，同时开展的研究 AudioGen（Kreuk 等人，2022年）提出根据文本输入，以自回归方式生成音频样本；而我们提出的方法与之有以下区别：1）我们引入伪提示增强，并利用对比式语言-音频预训练和扩散模型的优势，实现高保真生成。2）我们预测连续频谱图表征，显著提升计算效率并降低训练成本。

2.3. 音频表征学习

不同于对信号细粒度细节的建模，高层自监督学习（SSL）（Baevski 等人，2020年；Hsu 等人，2021年；He 等人，2022年）已被证明能有效缩小生成算法的采样空间。受矢量量化（VQ）技术的启发，SoundStream（Zeghidour 等人，2021年）提出了一种用于高层表征的分层架构，可承载语义信息。Data2vec（Baevski 等人，2022年）采用快速卷积解码器，并以自监督方式探索上下文化目标表征。

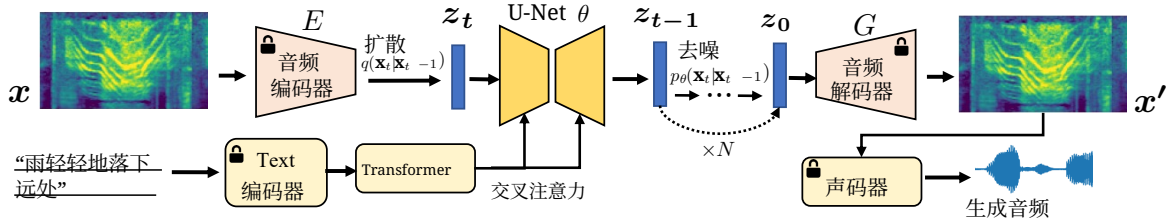


图 2。 Make-An-Audio 的整体概览。请注意，部分模块（以锁标示）在训练文本到音频模型时被冻结。

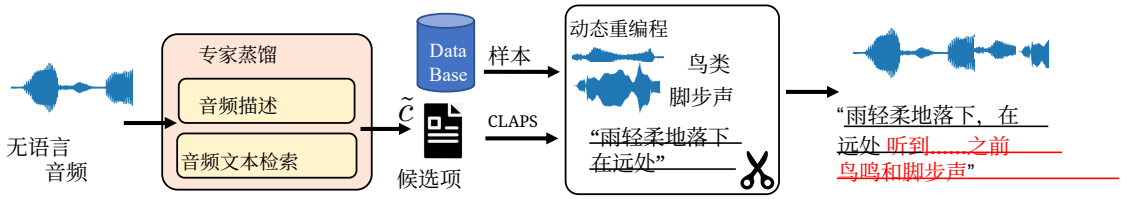


图 3。 伪提示增强的流程。我们的半参数扩散模型由固定的专家蒸馏阶段和动态重编程阶段组成。数据库 D 包含音频示例，并采用采样策略 ξ 来创建未见对象组合。我们用 CLAPS 表示 CLAP选择。

近期，频谱图（类似于单通道二维图像）自编码器（Gong 等人，2022年；He 等人，2022年）以重建目标作为自监督信号，已展现出异构图像到音频迁移的有效性，推动了语音与音频处理领域在多种下游任务上的发展。在这些方法中，Xu等人(2022年)研究了掩码自编码器（MAE）（He 等人，2022年），用于从音频频谱图中进行自监督表示学习。Gong 等人(2022年)采用音频频谱图变换器，将判别式与生成式掩码频谱图建模结合起来。受此启发，我们沿用频域频谱图自监督学习近期取得的成功，从而确保高效压缩和高层语义理解。

3. Make-An-Audio

本节首先概述 Make-An-Audio 框架，并介绍伪提示增强，以更好地对齐文本与音频语义；随后，我们介绍用于多模态学习的文本与音频表示。结合具备无分类器引导能力的扩散模型，Make-An-Audio 实现了高保真合成，并具有出色的泛化能力。

3.1. 概述

深度生成模型在文本引导的视觉合成方面取得了领先的性能。然而，当前文本转音频（T2A）生成的发展受到两大挑战的阻碍：1）模型训练面临数据稀缺的问题，因为人工标注音频的制作成本高昂，而且提供自然语言描述的音频资源很少。2）对长连续波形进行建模（例如，16 kHz 波形每秒通常包含 16,000 个数据点），给所有高质量神经合成器都带来了挑战。

如图所示 2，Make-An-Audio 由以下主要组件构成：1）伪提示词增强，用于缓解数据稀缺问题，从而能够利用数量级更大的无语言音频；2）频谱图自编码器，用于预测自监督表示，而非长连续波形；3）扩散模型，借助对比式语言-音频预训练（CLAP）的能力，将自然语言映射到潜在表示；4）单独训练的神经声码器，用于将梅尔频谱图转换为原始波形。在接下来的章节中，我们将详细介绍这些组件。

3.2. 伪提示增强：先蒸馏后重编程

为缓解数据稀缺问题，我们提出构造与音频良好对齐的提示词，从而能够利用数量级庞大的无监督数据，更好地理解文本—音频动态。如图 3 所示，该方法包含两个阶段：首先采用专家蒸馏方法生成与音频对齐的提示词，然后通过动态重编程过程构造多样的概念组合。

3.2.1. 专家蒸馏

我们将预训练的自动音频描述（Xu等人，2020年）和音频文本检索（Deshmukh等人，2022年；Koepke等人，2022年）系统作为提示词生成的专家。音频描述模型旨在生成多样化的自然语言句子，以描述音频片段的内容。音频文本检索则将自然语言作为查询，从数据库中检索相关音频文件。为此，专家们共同蒸馏知识，构建与音频对齐的描述，随后我们从候选项中选出 CLAP（Elizalde 等人，2022年）得分较高的候选项作为最终描述（我们设置一个阈值，仅纳入忠实度较高的结果

）。这一简单而有效的流程在很大程度上缓解了数据稀缺问题，并显著提升了模型在不同音频领域的泛化能力。有关研究结果的总结，请参阅第 6.3.2 节。详细内容见附录 E.2。

3.2.2. 动态重编程

为防止过拟合并更好地理解概念组合，我们引入了一种动态重编程技术，用于构建多种概念组合。该技术分三个步骤进行，如图 3所示，具体过程如下：1）我们首先准备一个声音事件数据库 D，其中的条目均标注了单一标签。2）每次从数据库中采样 N 个概念 D，其中 $N \in \{0, 1, 2\}$。3）根据模板，将原始文本-音频配对数据与采样事件随机拼接，构造出具有多种概念组合的新训练样本。该过程可以在线进行，从而大幅减少数据准备所耗费的时间。重编程模板见附录F。

3.3. 文本表示

文本引导合成模型需要强大的语义文本编码器来理解任意自然语言输入的含义，这类编码器主要分为两大类：1）对比预训练。与基于图文数据进行预训练的 CLIP（Radford 等人，2021年）类似，近期的对比式语言-音频预训练（CLAP）（Elizalde 等人，2022年）将音频与文本描述映射到联合空间，并展现出在多个下游领域中更优越的零样本泛化能力。2）大规模语言建模（LLM）。Saharia 等人（2022年）和Kreuk 等人（2022年）利用语言模型（例如 BERT（Devlin 等人，2018年）、T5（Raffel 等人，2020年））进行文本引导生成。语言模型在规模远大于配对多模态数据的纯文本语料库上训练，因此能够接触到丰富的文本分布。

按照常见做法（Saharia 等人，2022年；Ramesh等人，2022年），我们冻结这些文本编码器的权重。我们发现，CLAP 和 T5-Large 在基准评估中的表现相近，而 CLAP 无需像 LLM 那样离线计算嵌入，因此效率可能更高。关于研究结果的总结，请参阅第 6.3.1节。

3.4. 音频表示

近年来，以重建目标作为自监督方式的频谱图（类似于单通道二维图像）自编码器（Gong 等人，2022年；He 等人，2022年）已证明异构图像转音频迁移的有效性，推动了语音与音频处理领域在多种下游任务上的发展。

音频信号是一系列梅尔频谱图样本 $\boldsymbol{x} \in [0,1]^{C_{\mathrm{a}} \times T}$，其中$C_{\mathrm{a}}$，$T$分别表示梅尔通道数和帧数。我们的频谱图自编码器由以下部分组成：1）编码器网络 E，它以样本 $\boldsymbol{x}$ 为输入，并输出潜在表示 z；2）解码器网络 G，根据压缩表示 z重建梅尔频谱图信号 $\boldsymbol{x}'$；以及3）多窗口判别器 Dis，它学习在梅尔频谱图的不同多感受野中区分生成样本$G(z)$与真实样本。

整个系统采用端到端训练，以最小化以下损失：1）重建损失 $\mathcal{L}_{re}$，可提高训练效率和生成频谱图的保真度；2）GAN 损失 $\mathcal{L}_{GAN}$，其中判别器和生成器进行对抗博弈；3）KL 惩罚损失 $\mathcal{L}_{KL}$，它限制频谱图编码器学习标准 z，避免潜在空间出现任意高方差。

为此，Make-An-Audio 利用频谱图自编码器来预测自监督表示，而非波形。这在很大程度上缓解了对长连续数据进行建模的挑战，并确保了高层次的语义理解。

3.5. 生成式潜在扩散

我们在潜在扩散模型（LDM）之上实现了我们的方法（Rombach 等人，2022年）。潜在扩散模型是近期提出的一类去噪扩散概率模型（DDPM）（Ho 等人，2020年），在潜在空间中运行。该模型以文本表示为条件，将生成过程分解为多个条件扩散步骤。训练损失定义为噪声 $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ 空间中的均方误差，并通过随机梯度下降优化t 中的随机项：

$$\mathcal{L}_\theta = \|\boldsymbol{\epsilon}_\theta(\mathbf{z}_t, t, c) - \boldsymbol{\epsilon}\|_2^2, \tag{1}$$

其中，α 表示一个较小的正数常量，$\boldsymbol{\epsilon}_\theta$ 表示去噪网络。总而言之，扩散模型可以通过优化证据下界（ELBO）进行高效训练，而无需对抗性反馈，从而确保生成与真实分布高度一致的极其忠实的重建结果。DDPM 的详细公式已附于附录D中。

3.6. 无分类器引导

对于（Dhariwal 和 Nichol，2021年；Ho 和 Salimans，2022年）中所示的无分类器引导，通过联合训练条件扩散模型和无条件扩散模型，可以结合条件得分和无条件得分，在样本质量与多样性之间取得权衡。潜在扩散模型中的文本条件 $\boldsymbol{\epsilon}_\theta(\mathbf{z}_t,\ t,\ c)$会以固定概率被替换为空提示词 $c_\emptyset$ 进行训练。采样时，模型的输出会

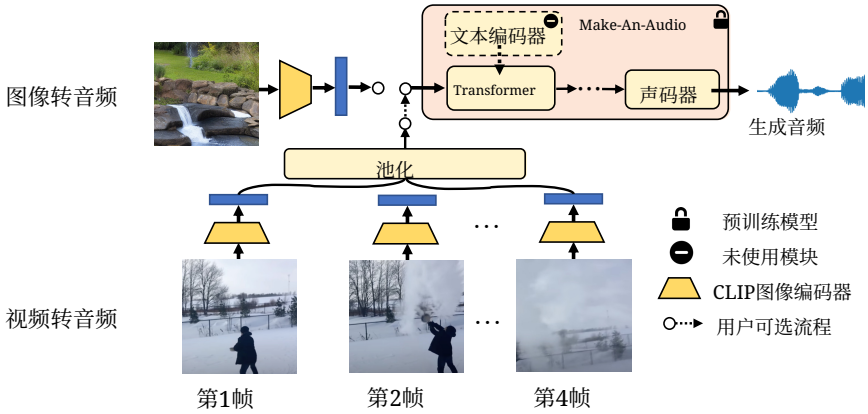


图4. 使用 Make-An-Audio 进行视觉到音频生成（I2A/V2A）的流程概览。

沿着 $\epsilon_\theta(z_t, t, c)$ 的方向进一步外推，并以引导尺度 $s \geq 1$ 远离 $\epsilon_\theta(z_t, t, c_\emptyset)$：$\tilde{\epsilon}_\theta(z_t, t, c) = \epsilon_\theta(z_t, t, c_\emptyset) + s \cdot (\epsilon_\theta(z_t, t, c) - \epsilon_\theta(z_t, t, c_\emptyset))$ (2)

4. X到音频：不让任何模态掉队

在本节中，我们将功能强大的条件扩散模型推广至X到音频生成。我们首次阐释了使用不同条件模态进行音频生成的必要性，包括：1）文本，2）音频（修复），以及3）视觉。Make-An-Audio 让人们能够以前所未有的便捷性创建丰富多样的音频内容，只需给定用户定义的模态输入，即可生成高清、高保真的音频。

4.1. 个性化文本到音频生成

使模型适应（Chen 等人，2020b；Huang 等人，2022年）特定个体或物体，是机器学习研究长期以来的目标。近年来，个性化（Gal 等人，2022年；Benhamdi 等人，2017年）研究也出现在视觉和图形学领域，使人们能够将独特物体注入新的场景，在不同风格之间转换它们，甚至生成新产品。例如，当要求在初始声音“雷声”的基础上生成“婴儿哭泣”时，我们的模型能够生成逼真且忠实描述“雷雨天里一个婴儿在哭泣”的音频。除此之外，它还具有广泛的音频混音和调音用途，例如为现有片段添加背景音，或通过插入说话对象来编辑音频。

我们研究了通过随机微分编辑（Meng 等人，2021年）实现的个性化文本到音频生成。该方法已被证明能够生成真实感强且操控保真度高的样本。给定带有用户指南（提示词）的输入音频，我们选取特定时刻 t_0，总去噪步骤数为 N，并根据公式 $\mathbf{z}_0$，对原始数据 $\mathbf{z}_T$ ($T = t_0 \times N$) 添加噪声。随后，通过由共享参数化的逆向过程对其进行去噪 θ，以根据公式 6提升其真实感。

可以观察到忠实度（文本与描述对齐）和真实感（音频质量）之间的权衡：随着 T 增加，初始音频中会加入大量噪声，生成样本会变得更加真实，但忠实度会降低。有关研究结果的总结，请参阅图 5 。

4.2. 音频修复

修复（Liu 等人，2020年；Nazeri 等人，2019年）是一项填补音频掩蔽区域的任务，即当部分音频遭到损坏或不符合预期时，用新内容填补这些区域。尽管可以通过向初始音频添加噪声并使用 SDEdit 进行采样来实现扩散模型修复，但这可能会产生不理想的边缘伪影，因为采样过程中可能会丢失信息（模型只能看到带噪声的上下文）。为了获得更好的结果，我们对 Make-An-Audio 进行专门微调，使其能够修复音频。

在训练过程中，掩码的生成方式会极大地影响系统的最终性能。因此，我们采用 LaMa（Suvorov 等人，2022年）提出的不规则掩码（厚、中、薄掩码）。该方法均匀地使用经高随机宽度膨胀的多边形链（宽掩码）以及任意宽高比的矩形（框掩码）。此外，我们还研究了语音文献中常用的基于帧的掩蔽策略（Baevski 等人，2020年；Hsu 等人，2021年）。该策略采用 wav2vec 2.0（Baevski 等人，2020年）中的算法实现，其中长度为 p 的片段会以一定概率被掩蔽。

4.3. 视觉到音频生成

深度生成模型的最新进展在视觉诱导音频生成方面取得了令人瞩目的成果（Su 等人，2020年；Gan 等人，2020年），旨在生成能够描述图像或视频内容的逼真音频：Hsu 等人（2020年）表明，视觉基础语音生成模型能够学习口语。Iashin

和 Rahtu（2021年）提出了一种多类别视觉引导声音合成方法，该方法依赖于基于码本先验的变换器。

为了进一步推进这项研究，我们将 Make-An-Audio 扩展到视觉到音频生成。由于图像转音频（I2A）研究缺乏大规模视觉-音频数据集，我们的核心思路是将对比式语言-图像预训练（CLIP）与 CLIP 引导的 T2A 模型结合起来，并利用文本表示来弥合视觉与音频世界之间的模态差距。由于 CLIP 编码器会将图像和文本嵌入联合潜在空间，我们的 T2A 模型提供了一个独特的机会，可以直观呈现 CLIP 图像编码器所看到的内容。考虑到 V2A 生成的复杂性，利用视频的图像先验来简化学习过程是很自然的做法。基于此，我们从视频中均匀选取 4 帧，并对这些 CLIP 图像特征进行池化，以构成“平均”视频表示，随后将其退化为 I2A 生成任务。

总而言之，视觉到音频推理方案可表示为图4所示。该方案显著减少了对成对视觉数据集的需求，并且结合预训练的 Make-An-Audio，即插即用模块能够让人们从视觉领域创作丰富多样的音频内容。

5. 训练与评估

5.1. 数据集

我们在多个数据集的组合上进行训练：AudioSet、BBC 音效、Audiostock、AudioCaps-train、ESC-50、FSD50K、Free To Use Sounds、Sonniss Game Effects、We-SoundEffects、MACS、Epidemic Sound、UrbanSound8K、WavText5Ks、LibriSpeech 和 Medley-solos-DB。对于没有自然语言标注的音频，我们采用伪提示增强来构建与音频良好对齐的描述文本。总体而言，我们拥有 ∼3千小时音频和 100 万个音频-文本对作为训练数据。在评估文本到音频模型时（Yang 等人，2022年；Kreuk 等人，2022年），我们采用 AudioCaption验证集作为标准基准，其中包含 494 个样本，每个音频片段均配有五条人工标注字幕。针对更具挑战性的零样本场景，我们还在包含多个音频事件的 Clotho（Drossos等人，2020年）验证集上提供结果。更详细的数据设置见附录A。

我们对文本和音频数据进行预处理：1）将音频的采样率转换为 16kHz，并将较短的片段补齐至 10 秒；2）使用 FFT 大小为 1024、跳步大小为 256 的参数提取频谱图，并将其裁剪为大小为 80×624 的梅尔频谱图；3）对非标准词语（例如缩写、数字和货币表达）以及符号类别（Taylor，2009）（表示语义受限的特定实体的文本词元，例如度量短语、地址和日期）进行规范化。

5.2. 模型配置

我们训练一个连续自编码器，将感知空间压缩并下采样为 4 通道的潜在表示，从而兼顾效率与感知上的忠实度。在主要实验中，我们训练了一个基于 U-Net（Ronneberger 等人，2015）的文本条件扩散模型，使用 18 个 NVIDIA V100 GPU 进行优化，直至达到 2M 个优化步骤。基础学习率设为 0.005，并按照 LDM，根据 GPU 数量和批量大小对其进行缩放。我们在所有实验中均使用在 VGGSound 数据集（Chen 等人，2020a）上训练的 HiFi-GAN（Kong 等人，2020）（V1）作为声码器，从生成的梅尔频谱图合成波形。超参数见附录 B。

5.3. 评估指标

我们使用客观和主观指标评估模型，衡量音频质量以及文本-音频对齐忠实度。按照常见做法（Yang 等人，2022年；Iashin 和 Rahtu，2021年），所采用的关键自动化性能指标基于 Melception（Koutini 等人，2021年），包括 FID（Heusel 等人，2017年）和 KL 散度，用于衡量音频保真度。此外，我们引入 CLAP 分数来衡量本研究中的音频-文本对齐情况。CLAP 分数由 CLIP 分数（Hessel 等人，2021年；Radford 等人，2021年）适配至音频领域而来，是一种无需参考的评估指标，与人类感知高度相关。

对于主观指标，我们通过 Amazon Mechanical Turk 众包开展人工评估，请评分者在 20 至 100 的李克特量表上给出平均意见分。我们分别通过 MOS-Q 和 MOS-F 评分来评估音频质量和文本-音频对齐忠实度，并报告 95% 置信区间（CI）。有关评估的更多信息见附录 C。

6. 结果

6.1. 定量结果

自动客观评估 与基线模型 DiffSound（唯一公开可用的文本到音频生成模型）的客观评估对比结果见表1，我们得到以下观察结果：1）在音频质量方面，Make-An-Audio 在 AudioCaption 上实现了最高的感知质量，FID 为 4.61，KL散度为 2.79。在零样本生成方面，其结果也优于基线模型；2）在文本与音频的相似度方面，Make-An-Audio 的 CLAP 得分最高，与真实音频相差 0.037，这表明 Make-An-Audio 能够生成忠实且与描述高度吻合的音频。

主观人工评估 对文本转音频

模型	文本条件	参数	FID	KL	CLAP	MOS-Q	MOS-F	FID-Z	KL-Z
参考	/	/	/	/	0.526	74.7±0.94	80.5±1.84	/	/
Diffsound	CLIP	520M	7.17	3.57	0.420	67.1±1.03	70.9±1.05	24.97	6.53
Make-An-Audio	**CLAP**	**332M**	**4.61**	**2.79**	0.482	**72.5±0.90**	**78.6±1.01**	17.38	**6.98**
	BERT	809M	5.15	2.89	0.480	70.5±0.87	77.2±0.98	18.75	7.01
	T5-Large	563M	4.83	2.81	**0.486**	71.8±0.91	77.2±0.93	**17.23**	7.02
	CLIP	576M	6.45	2.91	0.444	72.1±0.92	75.4±0.96	17.55	7.09

表 *1.* 文本转音频评估 。我们报告的评估指标包括平均意见分（↑）、FID（↓）、KL散度（↓）和 CLAP（↑）。FID-Z 和 KL-Z 表示零样本结果，适用于 Clotho 数据集。

训练掩码	窄掩码			宽掩码		
	FID	KL	MOS-Q	FID	KL	MOS-Q
不规则（细）	1.83	0.46	68.3±1.38	4.01	0.86	66.2±1.20
不规则（中）	1.76	0.31	67.8±1.41	3.93	0.65	66.9±1.22
不规则（粗）	1.73	0.32	69.6±1.36	3.83	0.67	69.3±1.05
帧 (p=30%)	1.64	0.29	66.9±1.60	3.68	0.62	66.1±1.29
帧 (p=50%)	1.77	0.32	68.6±1.42	3.66	0.63	67.4±1.27
帧 (p=70%)	1.59	0.32	71.0±1.12	3.49	0.65	70.8±1.50

表 *2.* 采用多种掩蔽策略的音频修复评估。

方法	MOS-Q	MOS-F
图像到音频生成		
参考	72.0±1.54	76.4±1.83
Make-An-Audio	68.4±1.09	78.0±1.20
视频到音频生成		
参考	69.5±1.22	81.0±1.43
Make-An-Audio	60.0±1.31	69.0±1.08

表 *3*。图像/视频到音频评估。

模型的评估颇具挑战，因为感知质量具有主观性。因此，我们在表中纳入了人工评估 1：Make-An-Audio（ CLAP）的感知质量最高，MOS-Q 达到 72.5，MOS-F 达到 78.6。这表明，就音频自然度和忠实度而言，评分者更喜欢我们模型合成的音频，而非基线模型。

对于音频修复，我们比较了不同的掩蔽设计，包括来自视觉领域的非规则（粗、中、细）策略（Suvorov 等人，2022年），以及语音中常用的基于帧的策略（变化的 p）（Baevski 等人，2020年；Hsu 等人，2021年）。评估过程中，我们随机掩蔽宽或窄区域，并使用 FID 和 KL 指标衡量性能。结果见表 2，我们得出以下观察结果：1）无论采用基于帧还是非规则策略，训练时较大的掩蔽区域都能提升感知质量，这促使网络充分利用连续频谱图的较大感受野。2）在掩蔽区域大小相近时，基于帧的策略始终优于非规则策略，这表明对按时间序列对齐的音频频谱图进行掩蔽可能效果更好。

我们还在表3中展示了视觉到音频生成结果。如图所示，Make-An-Audio 能够泛化到种类繁多的图像和视频。借助对比预训练，该模型能够高层次地理解视觉输入，并生成与其语义含义高度吻合的高保真音频频谱图。

6.2. 定性结果

首先，我们探索文本到音频合成中的无分类器引导。我们遍历不同的引导值，并展示 CLAP 与 FID 分数之间的权衡曲线见图7。与Ho 和 Salimans (2022年）的观察结果一致，选择不同的分类器引导权重可以调整条件合成和无条件合成，在样本对条件文本的忠实度与真实感之间进行权衡。

为了更好地比较音频修复效果，我们在图6中展示了不同的掩蔽策略和合成结果。如图所示，给定包含不需要内容的初始音频，我们的模型能够准确填补并重建音频，对不同形状的掩蔽区域均具有稳健性，这表明模型具备高层次理解音频内容的能力。

在个性化文本到音频生成方面，我们探索了不同的 $t_0 \in (0,\ 1)$ 来添加高斯噪声并进行反向采样。如图5所示，忠实度（以 CLAP 分数衡量）与真实感（以 1-MSE 距离衡量）之间存在权衡。我们发现，$t_0 \in [0.2, 0.5]$ 能够在忠实引导与真实生成之间取得良好效果，这表明音频变体（例如速度、音色和能量）可能会随着 t_0 增大而轻易被破坏。

6.3. 分析与消融研究

为验证 Make-An-Audio 中若干设计的有效性，包括伪提示增强、文本表示和音频表示，我们进行了消融研究，并讨论了以下关键发现。关于音频表示的更多分析见附录 E.1。

6.3.1. 文本表示

我们探索了多种预训练文本编码器，包括语言模型 BERT（Devlin 等人，2018年）、T5-Large（Raf-

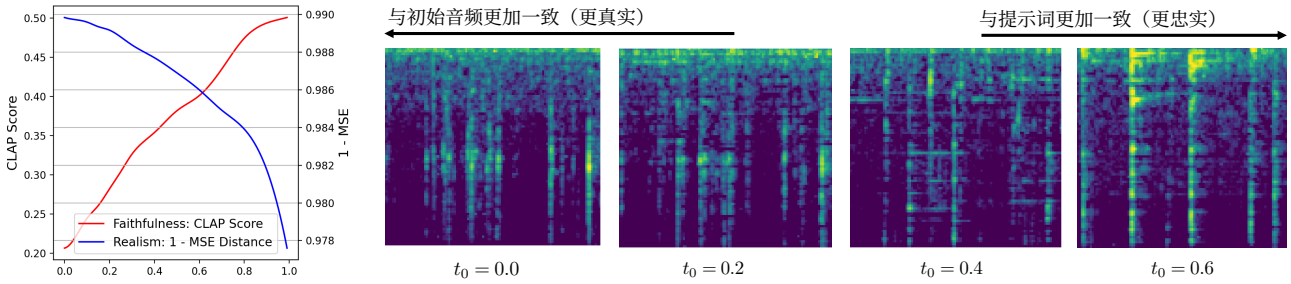


图*5.* 我们展示了采用不同 t_0 初始化的个性化文本到音频结果。 $t_0 = 0$ 表示初始音频本身，而$t_0 = 1$ 表示从头开始进行文本到音频合成。为便于比较，真实感通过生成音频与初始音频之间的 1-MSE 距离衡量，忠实度则通过生成样本之间的 CLAP 分数衡量。提示词：钟表在滴答作响。

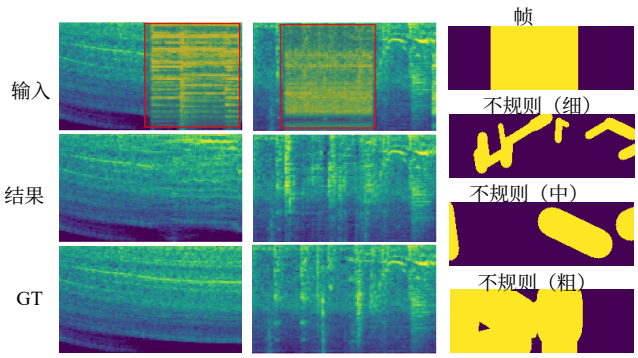


图*6.* 使用我们的修补模型得到的定性结果。

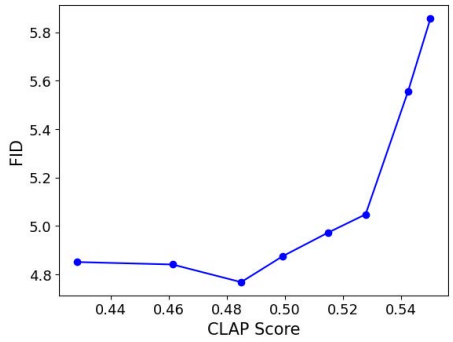


图*7.* 无分类器引导的权衡曲线。

fel 等人，2020年），以及多模态对比预训练编码器 CLIP（Radford 等人，2021年）和 CLAP（Elizalde 等人，2022年）。在文本到音频生成中，我们冻结文本编码器的权重。为便于比较，我们在表1 中展示结果，并得出以下观察：1）由于 CLIP 是一种用于学习文本与图像之间联合表征的可扩展方法，因此与 Yang 等人（2022年）相比，它在为文本到音频生成提取语义表征方面可能效果较差。2）CLAP 和 T5-Large 在基准数据集上的表现相近，而 CLAP 的计算效率可能更高（仅需 %59 个参数），且无需离线计算

大型语言模型中的嵌入 .

6.3.2. 伪提示增强

我们提出的提示增强方法通过先蒸馏、后重编程的两阶段方案，缓解了数据稀缺问题。如表 5 （附录 A）所示，我们计算并比较了各数据集上的平均提示与音频忠实度：联合专家蒸馏生成了高质量且与音频匹配良好的描述文本，并表明该方法能够泛化至多种音频领域。

为凸显所提出的动态重编程策略在创建未见对象组合方面的有效性，我们还在静态训练数据集上训练了 Make-An-Audio，并将结果列于表7（附录E）：1）移除动态重编程方法会导致评估结果略有下降；2）以零样本方式迁移至更具挑战性的 Clotho 场景时，性能显著下降，这证明了该方法在构建多样化对象组合、提升泛化能力方面的有效性。

7. 结论

在本工作中，我们提出了采用提示增强扩散模型的 Make-An-Audio，用于文本转音频生成。借助结合先蒸馏后重编程方法的提示增强，Make-An-Audio 能够利用数量级更大的无监督数据进行多种概念组合。我们研究了文本表示，并强调了对比预训练在以计算效率深入理解自然语言方面的优势。客观评估和主观评估均表明，Make-An-Audio 在文本转音频方面取得了新的业界最佳结果，能够合成逼真且忠实的音频。Make-An-Audio 首次尝试根据用户定义的模态输入生成高清、高保真音频，为个性化迁移和细粒度控制开辟了大量应用场景。我们希望本工作能为未来的音频合成研究奠定基础。

参考

Baevski, A.、Zhou, Y.、Mohamed, A. 和 Auli, M. wav2vec 2.0：语音表征自监督学习框架。神经信息处理系统进展，33，2020年。

Baevski, A., Hsu, W.-N., Xu, Q., Babu, A., Gu, J. 和 Auli, M. Data2vec：一种用于语音、视觉和语言自监督学习的通用框架。*arXiv* 预印本 *arXiv:2202.03555*，2022年。Benhamdi, S., Babouri, A. 和 Chiky, R. 面向电子学习环境的个性化推荐系统。《教育与信息技术》，22(4):1455–1477，2017年。Bittner, R. M., Salamon, J., Tierney, M., Mauch, M., Cannam, C. 和 Bello, J. P. Medleydb：用于注释密集型音乐信息检索研究的多轨数据集。载于*ISMIR*，第14卷，第155–160页，2014年。Borsos, Z., Marinier, R., Vincent, D., Kharitonov, E., Pietquin, O., Sharifi, M., Teboul, O., Grangier, D., Tagliasacchi, M. 和 Zeghidour, N. AudioLM：一种音频生成的语言建模方法。*arXiv* 预印本 *arXiv:2209.03143*，2022年。Chen, H., Xie, W., Vedaldi, A. 和 Zisserman, A. Vggsound：大规模视听数据集。载于*ICASSP 2020-2020 IEEE* 国际声学、语音与信号处理会议（*ICASSP*），第721–725页。IEEE，2020a。Chen, M., Tan, X., Li, B., Liu, Y., Qin, T., Liu, T.-Y. 等。AdaSpeech：面向定制语音的自适应文本转语音。载于国际学习表征会议，2020b。Deshmukh, S., Elizalde, B. 和 Wang, H. 使用 wavtext5k 和 CLAP 训练进行音频检索。*arXiv* 预印本 *arXiv:2209.14275*，2022年。Devlin, J., Chang, M.-W., Lee, K. 和 Toutanova, K. BERT：用于语言理解的深度双向变换器预训练。*arXiv* 预印本 *arXiv:1810.04805*，2018年。Dhariwal, P. 和 Nichol, A. 扩散模型在图像合成方面胜过生成对抗网络。载于*NeurIPS* 论文集，第34卷，2021年。Ding, M., Zheng, W., Hong, W. 和 Tang, J. CogView2：通过层次化变换器实现更快、更优的文本到图像生成。*arXiv* 预印本 *arXiv:2204.14217*，2022年。Drossos, K., Lipping, S. 和 Virtanen, T. Clotho：音频描述数据集。载于*ICASSP 2020-2020 IEEE* 国际声学、语音与信号处理会议（*ICASSP*），第736–740页。IEEE，2020年。

Elizalde, B., Deshmukh, S., Ismail, M. A., 和 Wang, H. CLAP：通过自然语言监督学习音频概念。*arXiv* 预印本 *arXiv:2206.04769*，2022年。Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A. H., Chechik, G., 和 Cohen-Or, D. 一张图像胜过千言：使用文本反演实现文本到图像生成的个性化。*arXiv* 预印本 *arXiv:2208.01618*，2022年。Gan, C., Huang, D., Chen, P., Tenenbaum, J. B., 和 Torralba, A. 拟音音乐：从视频中学习生成音乐。载于欧洲计算机视觉会议，第758–775页。施普林格，2020年。Gemmeke, J. F., Ellis, D. P., Freedman, D., Jansen, A., Lawrence, W., Moore, R. C., Plakal, M., 和 Ritter, M. 音频集：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。Gong, Y., Lai, C.-I., Chung, Y.-A., 和 Glass, J. SSAST：通过掩码预测进行自监督音频频谱图变换器学习。载于*AAAI*人工智能会议论文集，第36卷，第10699–10709页，2022年。Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., 和 Bengio, Y. 生成对抗网络。*ACM*通讯，63(11):139–144，2020年。He, K., Chen, X., Xie, S., Li, Y., Doll´ar, P., 和 Girshick, R. 掩码自编码器是可扩展的视觉学习器。载于*IEEE/CVF* 计算机视觉与模式识别会议论文集，第16000–16009页，2022年。Hessel, J., Holtzman, A., Forbes, M., Bras, R. L., 和 Choi, Y. CLIPScore：一种无参考图像字幕生成评估指标。*arXiv* 预印本 *arXiv:2104.08718*，2021年。Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., 和 Hochreiter, S. 通过双时间尺度更新规则训练的生成对抗网络收敛至局部纳什均衡。神经信息处理系统进展，第30卷，2017年。Ho, J. 和 Salimans, T. 无分类器扩散引导。*arXiv* 预印本 *arXiv:2207.12598*，2022年。Ho, J., Jain, A., 和 Abbeel, P. 去噪扩散概率模型。载于*NeurIPS* 论文集，2020年。Hong, W., Ding, M., Zheng, W., Liu, X., 和 Tang, J. CogVideo：通过变换器实现文本转视频生成的大规模预训练。*arXiv* 预印本 *arXiv:2205.15868*，2022年。Hsu, W.-N., Harwath, D., Song, C., 和 Glass, J. 使用学习得到的片段单元进行无文本图像转语音合成。*arXiv* 预印本 *arXiv:2012.15454*，2020年。

Hsu, W.-N., Bolte, B., Tsai, Y.-H. H., Lakhotia, K., Salakhutdinov, R., 和 Mohamed, A. Hubert：通过对隐单元进行掩码预测实现自监督语音表征学习。*IEEE/ACM* 音频、语音与语言处理汇刊，29:3451–3460，2021年。

Huang, R., Ren, Y., Liu, J., Cui, C., and Zhao, Z. Gen- erspeech：面向可泛化的域外文本转语音合成的风格迁移。*arXiv* 预印本 *arXiv:2205.07211*，2022年。Iashin, V. and Rahtu, E. 驯服视觉引导的声音生成。*arXiv* 预印本 *arXiv:2110.08791*，2021年。Kim, C. D., Kim, B., Lee, H., and Kim, G. AudioCaps：为真实场景中的音频生成描述文本。载于计算语言学协会北美分会 *2019* 年会议论文集：人类语言技术，第 *1* 卷（长篇与短篇论文），第 119–132 页，2019年。Kingma, D. P. and Dhariwal, P. Glow：基于可逆 1×1 卷积的生成流。神经信息处理系统进展，31:10215–10224，2018年。Koepke, A. S., Oncescu, A.-M., Henriques, J., Akata, Z., and Albanie, S. 使用自然语言查询进行音频检索：一项基准研究。*IEEE* 多媒体汇刊，2022年。Kong, J., Kim, J., and Bae, J. HiFi-GAN：用于高效、高保真语音合成的生成对抗网络。*arXiv* 预印本 *arXiv:2010.05646*，2020年。Koutini, K., Schl¨ uter, J., Eghbal-zadeh, H., and Widmer, G. 使用补丁丢弃高效训练音频 Transformer。*arXiv* 预印本 *arXiv:2110.05069*，2021年。Kreuk, F., Sy nnaeve, G., Polyak, A., Singer, U., D´efossez, A., Copet, J., Parikh, D., Taigman, Y., and Adi, Y. Audio-gen：文本引导的音频生成。*arXiv* 预印本 *arXiv:2209.15352*，2022年。Liu, H., Jiang, B., Song, Y., Huang, W., and Yang, C. 通过具有特征均衡的互编码器-解码器重新思考图像修复。载于欧洲计算机视觉会议，第 725–741 页。施普林格，2020年。Mart´ın-Morat´o, I. and Mesaros, A. 什么是真实标注？多标注者音频标注数据的可靠性。载于*2021*年 第 *29* 届欧洲信号处理会议（*EUSIPCO*），第 76–80 页。IEEE，2021年。Meng, C., He, Y., Song, Y., Song, J., Wu, J., Zhu, J.-Y., and Ermon, S. SDEdit：使用随机微分方程进行引导式图像合成与编辑。载于国际学习表征会议，2021年。

Nazeri, K., Ng, E., Joseph, T., Qureshi, F. Z., and Ebrahimi, M. Edgeconnect：通过对抗式边缘学习进行生成式图像修复。*arXiv* 预印本 *arXiv:1901.00212*，2019年。Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., and Chen, M. Glide：利用文本引导扩散模型实现逼真图像生成与编辑。*arXiv* 预印本 *arXiv:2112.10741*，2021年。Piczak, K. J. ESC：用于环境声音分类的数据集。发表于第*23* 届 *ACM* 国际多媒体会议论文集，第1015–1018页，2015年。Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. 从自然语言监督中学习可迁移视觉模型。发表于国际机器学习大会，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., et al. 探索统一文本到文本转换器的迁移学习极限。机器学习研究期刊，21(140):1–67，2020年。Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., and Sutskever, I. 零样本文本转图像生成。发表于国际机器学习大会，第8821–8831页。PMLR，2021年。Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. 利用 CLIP 潜变量进行分层文本条件图像生成。*arXiv* 预印本 *arXiv:2204.06125*，2022年。Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and O mmer, B. 利用潜在扩散模型进行高分辨率图像合成。发表于*IEEE/CVF* 计算机视觉与模式识别会议论文集，第10684–10695页，2022年。Ronneberger, O., Fischer, P., and Brox, T. U-Net：用于生物医学图像分割的卷积网络。发表于国际医学图像计算与计算机辅助干预会议，第234–241页。施普林格，2015年。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., et al. 具备深度语言理解能力的逼真文本到图像扩散模型。*arXiv* 预印本 *arXiv:2205.11487*，2022年。Salamon, J., Jacoby, C., and Bello, J. P. 城市声音研究的数据集与分类体系。发表于第*22*届 *ACM* 国际多媒体会议（*ACM-MM'14*），第1041–1044页，美国佛罗里达州奥兰多，2014年11月。

Singer, U., Polyak, A., Hayes, T., Yin, X., An, J., Zhang, S., Hu, Q., Yang, H., Ashual, O., Gafni, O., et al. Make-a-video：无需文本-视频数据的文本转视频生成。*arXiv* 预印本 *arXiv:2209.14792*，2022年。

Song, J., Meng, C. 和 Ermon, S. 去噪扩散隐式模型。发表于国际学习表征会议论文集，2020年。Su, K., Liu, X. 和 Shlizerman, E. Audeo：为无声表演视频生成音频。发表于神经信息处理系统进展，33:3325–3337，2020年。Suvorov, R., Logacheva, E., Mashikhin, A., Remizova, A., Ashukha, A., Silvestrov, A., Kong, N., Goka, H., Park, K. 和 Lempitsky, V. 基于傅里叶卷积的分辨率鲁棒型大型掩码图像修复。发表于*IEEE/CVF* 冬季计算机视觉应用会议论文集，第 2149–2159 页，2022年。Taylor, P. 文本转语音合成。剑桥大学出版社，2009年。Van Den Oord, A., Vinyals, O. 等。神经离散表征学习。发表于神经信息处理系统进展，30，2017年。Xu, H., Li, J., Baevski, A., Auli, M., Galuba, W., Metze, F., Feichtenhofer, C. 等。能够聆听的掩码自编码器。*arXiv* 预印本 *arXiv:2207.06405*，2022年。Xu, X., Dinkel, H., Wu, M. 和 Yu, K. 一种基于 CRNN-GRU 的音频描述强化学习方法。发表于 *DCASE*，第 225–229 页，2020年。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。Zeghidour, N., Luebs, A., Omran, A., Skoglund, J. 和 Tagliasacchi, M. SoundStream：端到端神经音频编解码器。发表于*IEEE/ACM* 音频、语音与语言处理汇刊，30:495–507，2021年。Zen, H., Dang, V., Clark, R., Zhang, Y., Weiss, R. J., Jia, Y., Chen, Z. 和 Wu, Y. LibriTTS：一个源自 LibriSpeech 的文本转语音语料库。*arXiv* 预印本 *arXiv:1904.02882*，2019年。

附录

Make-An-Audio：使用提示增强扩散模型的文本到音频生成

A. 详细实验设置

数据集	小时数	Type	来源
Clotho	152	字幕	Drossos等人(2020年)
AudioCaps	109	字幕	Kim等人（2022年）
MACS	100	图注	Mart´ın-Morat´o & Mesaros (2021年)
WavText5Ks	25	图注	Deshmukh等人（2022年）
BBC 音效	481	说明	https://sound-effects.bbcrewind.co.uk/
Audiostock	43	说明	https://audiostock.net/se
筛选 AudioSet	2084	标签	Gemmeke 等人（2017年）
ESC-50	3	标签	Piczak (2015)
FSD50K	108	标签	https://annotator.freesound.org/fsd/
Sonniss Game Effects	20	标签	https://sonniss.com/gameaudiogdc/
WeSoundEffects	11	标签	https://wesoundeffects.com/
Epidemic Sound	220	标签	https://www.epidemicsound.com/
UrbanSound8K	8	标签	Salamon 等(2014年)
LibriTTS	300	无语言	Zen 等 (2019年)
Medley-solos-DB	7	无语言	Bittner 等人 (2014年)

表*4*。 多个数据集组合的统计信息。

如表5所示，我们收集了一个大规模音频文本数据集，其中包含 100 万个音频样本，总时长为~3千小时。该数据集包含人类活动、自然声音和音效的音频，数据源来自多个公开网站。对于带有文本描述的音频，我们下载其平行音频文本数据。对于没有自然语言标注（或仅带有标签）的音频，我们会丢弃相应的类别标签（如果有），并采用伪提示增强，构建与音频良好对齐的自然语言描述。

由于语音和音乐是 AudioSet 中占主导地位的类别，我们筛选出这些样本，以构建一个更加均衡的数据集。最终，我们得到 3 千小时的数据，其中包含 100 万个音频-文本对，用作训练数据。为了评估文本到音频模型（Yang 等人，2022年；Kreuk 等人，2022年），AudioCaption 验证集是标准基准，其中包含 494 个样本，每个音频片段有五条人工标注字幕。在训练和推理过程中，我们将较短的片段补齐至 10 秒，并从 10 秒、16 kHz 音频中随机裁剪一段 624×80 梅尔频谱图。

方法	FSD50K	ESC-50	UrbanSound8K
原始	0.40	0.43	0.33
字幕生成	0.35	0.46	0.37
检索	0.31	0.44	0.38
两者 + CLAP Select	0.54	0.62	0.55

表*5*。 在单标签数据集上取平均的文本-音频对齐 CLAP 分数。

B. 模型配置

我们在表格中列出了 Make-An-Audio 的模型超参数6。

超参数		Make-An-Audio
频谱图自编码器	输入/输出通道	1
	隐藏通道	4
	残差块	2
	频谱图尺寸	80×624
	通道倍数	$[1, 2, 2, 4]$
去噪 U-Net	输入/输出通道	4
	模型通道	320
	注意力头	8
	条件通道	1024
	潜变量维度	10×78
	通道倍数	$[1, 2]$
CLAP 文本编码器	Transformer 嵌入通道	768
	输出投影通道	1024
	词元长度	77
参数总数		332M

表*6*。 Make-An-Audio 模型的超参数。

C. 评估

为评估音频质量，我们进行了平均意见分（平均意见分）测试，并明确要求评分者“重点考察音频质量和自然度。”。测试人员展示并评定样本，每位测试人员需在 20–100 的李克特量表上评估主观自然度。

为探究文本-音频对齐情况，研究人员向人类评分者展示一段音频和一个提示词，并询问“自然语言描述是否与音频忠实地对齐？”。他们必须在 20–100 的李克特量表上选择“完全”“大部分”或“某种程度上”。

我们的主观评估测试通过众包方式开展，并借助亚马逊土耳其机器人进行。我们分别对模型样本和参考音频进行独立评分，并报告两者的结果。测试说明的截图见图8。我们按小时向参与者支付 8 美元，总计约花费 750 美元作为参与者报酬。测试中使用的一小部分语音样本可在以下网址获取：https://Text-to-Audio.github.io/。

D. 去噪扩散概率模型的详细公式

我们将数据分布定义为 $q(\mathbf{x}_0)$。扩散过程定义为从数据 $\mathbf{x}_0$ 到潜变量 $\mathbf{x}_T$ 的固定马尔可夫链：

$$q(\mathbf{x}_1, \cdots, \mathbf{x}_T|\mathbf{x}_0) = \prod_{t=1}^{T} q(\mathbf{x}_t|\mathbf{x}_{t-1}), \tag{3}$$

对于一个较小的正数常量 β_t，在函数 $q(\mathbf{x}_t|\mathbf{x}_{t-1})$ 的作用下，会向$\mathbf{x}_{t-1}$ 添加少量高斯噪声，使其转变为$\mathbf{x}_t$ 的分布。

整个过程按照固定的噪声调度 $\beta_1, \cdots, \beta_T$，逐步将数据 $\mathbf{x}_0$ 转换为白化潜变量 $\mathbf{x}_T$，其中 $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$：

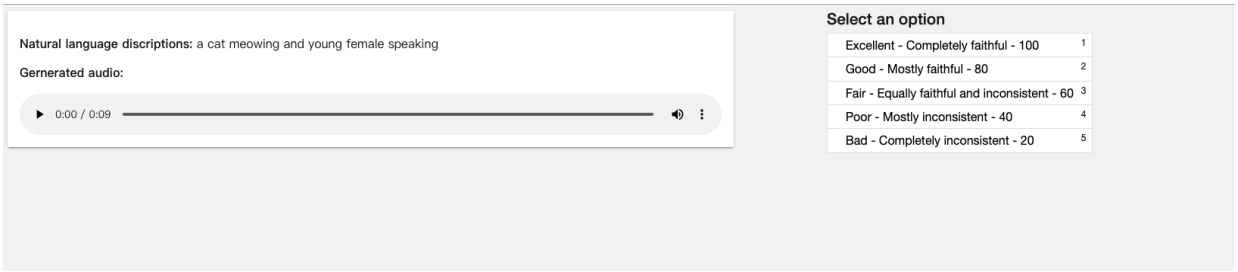
$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) := \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t \boldsymbol{I}) \tag{4}$$

高效训练是通过随机梯度下降优化 t 中的一个随机项：

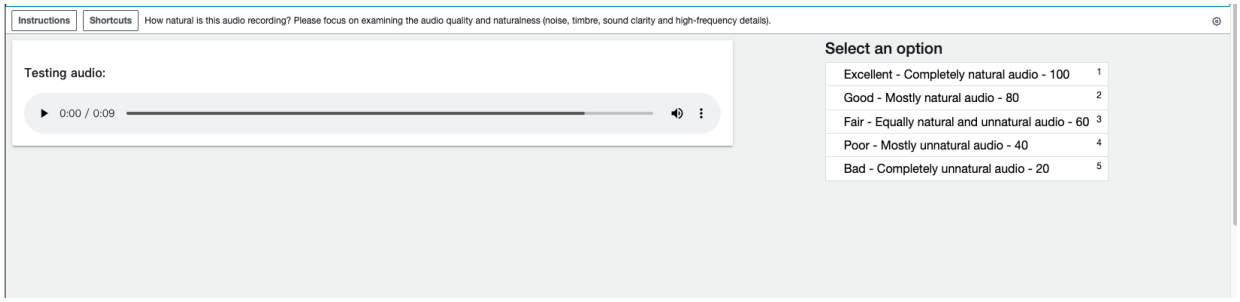
$$\mathcal{L}_\theta = \left\| \boldsymbol{\epsilon}_\theta \left(\alpha_t \mathbf{x}_0 + \sqrt{1-\alpha_t^2}\boldsymbol{\epsilon} \right) - \boldsymbol{\epsilon} \right\|_2^2 \tag{5}$$

与扩散过程不同，逆向过程旨在从高斯噪声中恢复样本。逆向过程是一个从 x_T到 x_0、由共享的 θ参数化的马尔可夫链：

$$p_\theta(\mathbf{x}_0, \cdots, \mathbf{x}_{T-1}|\mathbf{x}_T) = \prod_{t=1}^{T} p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t), \tag{6}$$



(a) MOS-F 测试截图。



(b) MOS-Q 测试截图。

图*8.* 主观评估截图。

其中，每次迭代都会消除扩散过程中添加的高斯噪声：

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) := \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t), \sigma_\theta(\mathbf{x}_t, t)^2 \boldsymbol{I}) \tag{7}$$

E. 实现细节

E.1. 频谱图自编码器

我们还在表7中研究了几种音频自编码器变体的有效性，并发现，更深的表示（即 32 或 128）能够带来相对更高的压缩率，但信息退化可能会增加生成建模中 U-Net 模型的负担。

方法	通道	FID	KL
AudioCaps 数据集上的有监督评估			
Base	4	5.15	2.89
	32	9.22	3.54
	128	10.92	3.68
无伪提示增强	4	5.37	3.05
Clotho 数据集上的零样本评估			
Base	4	18.75	7.01
不使用伪提示增强	4	22.31	7.19

表 *7.* 音频质量比较　　针对 Make-An-Audio BERT 的消融研究比较。我们用伪提示增强表示伪提示增强　　t.

E.2. 文本转音频

我们首先将文本编码为由 K 个词元组成的序列，并利用交叉注意力机制，在一个强大的模型中学习语言与梅尔频谱图表示之间的映射。完成初始训练后，我们对基础模型进行了微调，使其支持无条件生成，其中 20% 的文本词元序列被替换为空序列。这样

一来，模型既保留了生成文本条件输出的能力，也能无条件生成频谱图表示。

我们采用预训练的自动音频描述（Xu等人，2020年）和音频文本检索（Deshmukh等人，2022年；Koepke等人，2022年）系统作为提示词生成专家。自动音频描述模型由一个10层卷积神经网络（CNN）编码器和一个时间注意力单层门控循环单元（GRU）解码器组成。CNN编码器在大规模 AudioSet 数据集上进行了预训练。至于音频文本检索，该模型采用 BERT 和多模态 Transformer 编码器进行表征学习，并在 AudioCaps 和 Clotho 数据集上训练。

E.3. 视觉到音频

对于视觉到音频（图像/视频）合成，我们采用 CLIP 引导的 T2A 模型，并利用全局文本表征来弥合视觉与音频领域之间的模态差距。然而，我们通过实验发现，全局 CLIP 条件在控制忠实合成并实现较高文本与音频相似度方面能力有限。因此，我们使用带有类别标签标注的 FSD50K 中 110 小时音频进行训练；这种简化避免了复杂分布下的多模态预测（一个条件向量可能指代不同概念）。

我们开展了消融研究，以比较不同的训练设置，包括数据集和全局条件。结果已呈现在表 8中，以下是我们的观察结果：1）将 FSD50K 数据集替换为 AudioCaps（Kim 等人，2019年）后，忠实度显著下降。动态概念组合会使全局条件模型产生混淆，而多模态分布则限制了其进行可控合成的能力；2）移除条件向量中的归一化会导致以 FID 衡量的真实感下降，这表明归一化在减少潜在空间中的方差方面是有效的。

训练/测试数据集	条件	FID	KL	CLAP
AudioCaption	全局	/	/	0.12
FSD50k	全局	40.7	8.2	0.40
FSD50K	归一化全局	31.1	8.0	0.42

表 *8.* 使用全局条件训练 Make-An-Audio 的消融研究。

F. 动态重编程模板

下面列出了提供动态重编程时使用的文本模板：

- 在 $v\ q\ a\ n$之前，&、X
- X 在 $v\ q\ a\ n$之前，&、
- 在 $v\ q\ a\ n$前面，&、X
- 第一个是 X，第二个是 $q\ a\ n$和 &
- 在 X 之后，$v\ q\ a\ n$和 &
- 在 $v\ q\ a\ n$之后，&、X
- 位于 $v\ q\ a\ n$的 & 后方，X
- $v\ q\ a\ n$的 &，然后是 X
- $v\ q\ a\ n$的 &，随后是 X
- $v\ q\ a\ n$的 &，之后是 X
- X 在 $v\ q\ a\ n$的 & 之后
- 在 X 之前，$v\ q\ a\ n$的 &

具体来说，我们分别将 X 和 & 替换为数据库中采样数据的自然语言描述和采样事件的类别标签。

对于动词（记为 v），我们有 {‘听见’、‘注意到’、‘聆听’、‘出现’}；对于形容词（记为 a），我们有 {‘清晰’、‘嘈杂’、‘特写’、‘奇怪’、‘干净’}；对于名词（记为 n），我们有 {‘音频’、‘声音’、‘语音’}；对于数词/量词（记为 q），我们有 {‘一个’、‘这个’、‘一些’}；

G. 潜在的负面社会影响

本文旨在推动开放域文本到音频生成的发展，从而降低短视频和数字艺术创作的难度。高效的训练方法还能将文本到音频模型中的知识迁移到 X 到音频生成中，帮助避免从头开始训练，从而缓解数据稀缺问题。负面影响之一是存在传播错误信息的风险。为减轻这一风险，我们可以训练一个额外的分类器来辨别伪造内容。我们认为，收益大于弊端。

Make-An-Audio 降低了实现高质量文本到音频合成的门槛，这可能导致相关职业从业者失业，例如音响工程师和电台主持人。此外，未经同意进行语音克隆或生成虚假媒体可能造成危害，而且录音中的声音可能会被过度使用，超出录音者的预期。

H. 局限性

Make-An-Audio 采用生成式扩散模型进行高质量合成，因此本质上需要经过多轮迭代优化才能获得更好的结果。此外，潜在扩散模型通常需要更多计算资源，而且训练数据减少时，生成质量可能会下降。我们未来的研究方向之一是开发轻量快速扩散模型，以加速采样。