arXiv:2602.16790v1 [cs.SD] 18 Feb 2026

生成式音频延展与变形

Prem Seetharaman⋆、*Oriol Nieto*⋆、*Justin Salamon*

Adobe Research，美国加利福尼亚州旧金山

摘要

在音频相关的创意任务中，声音设计师常常希望延展和变形其素材库中的不同声音。能够以示例作为参考来创作音频的生成式音频模型，为此提供了很有前景的解决方案。(i) 给定一个音频参考，我们可以按指定时长向前和向后延展；(ii) 给定两个音频参考，我们可以在所需时长内将它们无缝变形。此外，我们展示了通过使用不同类型的平稳音频数据对模型进行微调，可以减少潜在的幻觉。客观指标验证了我们方法的有效性：生成音频的 Fr´echet 音频距离（FAD）与训练数据中的真实样本相当。此外，我们还通过主观听众测试验证了结果，受试者对所提出模型生成的音频给予了积极评价。这项技术为更加可控且富有表现力的生成式声音框架铺平了道路，使声音设计师能够减少繁琐重复的工作，更专注于实际的创作过程。

索引词— 音频生成、音频延展、音频修补、音频变形、扩散模型

1. 引言

声音设计是电影、电视、电子游戏和虚拟现实等多个创意领域的重要组成部分。它涉及创造听觉元素的艺术与实践，这些元素能够增强叙事效果、唤起情感并带来沉浸式体验。声音设计师精心制作并处理声音，以实现预期的听觉效果；他们常将自然声音与合成声音变形，创造独特的听觉景观。这一过程要求深入了解声学、心理声学，并具备使用各种音频工具和软件的技术能力 [1]。生成无缝音频延展片段和变形音频对声音设计师至关重要。这些从业者经常面临这样的挑战：延长或变形音频片段，使其适配特定场景或转场，同时避免引入伪影或不自然的元素；因为原始素材往往可能被截短，或比预期稍晚开始。传统方法可能耗时较长，且不一定总能取得令人满意的结果，尤其是在处理复杂声景或非平稳声音时。

基于音频的扩散模型的最新进展已在文本条件 [2, 3, 4] 与无条件 [5, 6] 音频生成方面取得显著进步。此外，已有研究为此类生成模型提出了增强的控制机制，利用音频特征实现基于文本的条件控制 [7, 8]，并优化潜在表征，从而借助时变信号实现细粒度控制 [9, 10, 11, 12]。音频延展也已在通用音频 [13], 中得到探索，尤其是在音乐领域 [14], ；而音频变形技术（通常称为音频修复）已被应用于通用音频场景 [15, 16, 17, 18, 19, 20], 以及音乐 [21, 22]。然而，这些方法并未专门针对声音设计师的需求，因而常常导致次优结果，或产生与他们通常处理的声音类型不兼容的模型，例如特殊效果和环境声音。

我们的工作通过推出能够从一段音频生成高质量、48千赫兹的立体声音频延展片段和变形音频的模型，解决了这些挑战。借助在音频潜变量上运行的扩散Transformer [23]，我们提出了一种新颖的潜变量掩码技术，并将其与无分类器引导 [24], 的一种变体相结合，从而实现无缝且有效的音频延展和变形。此外，我们通过采用创新的微调策略，减少静态生成中可能出现的幻觉。我们使用 Fr´echet Audio Distance [25], 评估结果，结果表明，就音频质量而言，生成的音频往往与各自的音频提示难以区分。此外，我们还通过主观听众测试验证了这些发现，参与者对模型生成的音频给予了积极评价。借助这一新颖模型，我们旨在改进声音设计师的创作流程，并减少其工作中所需的手动、繁琐操作。

我们工作的主要贡献如下：

- 该模型旨在根据一个或两个音频提示生成延展片段并进行变形，从而实现音频片段的无缝续接与融合。
- 一种受无分类器引导启发的新型音频提示引导（APG）技术，可提升生成音频的质量与连贯性。
- 一种解决幻觉问题的新方法：使用合成噪声底数据集对模型进行微调，确保生成的音频忠实于输入提示。

2. 方法

图 1 展示了所提出的生成式扩展与变形方法。本节将介绍完整流程：先介绍共用流程（包括预处理、后处理和音频提示引导），再分别介绍这两种应用。

2.1. 输入预处理

我们的主要受众是专门制作静态和非静态特殊效果及环境音效的声音设计师，因此训练过程中不会使用语音数据。因此，我们使用 [26] 移除音频中的语音成分。[1] 随后，将不含语音的音频提示 $\mathbf{x} \in \mathbb{R}^T$ 编码为潜表示 $\mathbf{z} \in \mathbb{R}^{n \times T_z}$，其中 T 是音频样本的数量，n 是潜变量的声道数，T_z 是在潜空间中编码 T 个音频样本所需的帧数（即 $T_z \ll T$）。任何能够生成高度压缩且高质量潜在表征的自编码器架构都适用于此目的，例如变分自编码器（VAEs） [27], 、残差向量量化（RVQ） [28],，或 Descript音频编解码器（DAC）等任何预训练模型 [29]。

⋆贡献相同。

[1]这是一项成熟的技术，其作者分享了 中间表示 代码。

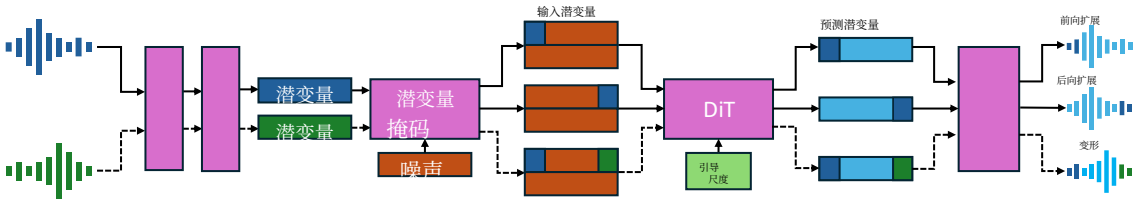


图 1：生成式延展（实线）与变形（虚线）的示意框图。图中展示了两个音频示例：蓝色示例用于生成式延展（向前和向后），而蓝色和绿色示例都用于变形（从蓝色变为绿色）。

假设我们要生成 d_z 帧音频，那么接下来将高斯噪声 $\mathbf{z}_G \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \in \mathbb{R}^{n \times d_z}$ 输入扩散变换器（DiT） [23] f^θ。我们选择 DiT架构，是因为我们有内部实现，不过这种方法也兼容其他扩散架构（例如 U-Net [30]）。请注意，d_z 与 T_z 相互独立，因为生成内容的量 d_z 不取决于音频提示的长度 T_z。这类噪声会通过 $f_M(\mathbf{z}_G, \mathbf{z})$ 使用输入提示潜变量进行掩码处理，其中 f_M：$\mathbb{R} \times \mathbb{R} \to \mathbb{R}$ 是依赖于具体应用的掩码函数，将在以下小节中介绍。DiT 的输入既包括 经过掩码处理的含噪潜变量（$f_M(\mathbf{z}_G, \mathbf{z})$），也包括未经过掩码处理的含噪潜变量（$\mathbf{z}_G$）。图 1 中央的“输入潜变量”示例展示了这一点：输入潜变量的总长度（包含提示的橙色部分）为 d_z，提示的长度（蓝色或绿色的小矩形）为 T_z。DiT不会对 含噪潜变量中经过掩码处理的部分进行去噪，从而强制生成结果与原始潜变量保持一致。请注意，此设置要求 $T_z < d_z$，即音频提示必须短于输入 DiT 的噪声总长度。

2.2. 音频提示引导

为确保生成的音频更忠实地符合音频提示，我们修改了 DiT，使其采用一种变体的无分类器引导（CFG） [24]，并将其称为音频提示引导（APG）。APG 引导生成结果趋向掩码潜变量的分布，并远离未掩码潜变量的分布。这使模型能够生成质量高得多、且更贴合提示的音频。形式化地说：

$$\mathbf{z}' = f^\theta(\mathbf{z}_G) + \gamma\left(f^\theta(f_M(\mathbf{z}_G, \mathbf{z})) - f^\theta(\mathbf{z}_G)\right) \tag{1}$$

其中，$\mathbf{z}' \in \mathbb{R}^{n \times d_z}$ 是 DiT 的 APG 输出，γ 是引导尺度。

这种方法与原始 CFG 公式相似，但其独特之处在于应用于与目标数据相同的模态，即音频。据我们所知，这是首项以这种方式应用 CFG 的工作。

2.3. 输出后处理

随后，对输出$\mathbf{z}'$ 应用输入中使用的相同 f_M 函数进行后处理。因此，最终的后处理输出包含原始提示中的潜在信息（即待延展/变形的音频）：$\hat{\mathbf{z}}' = f_M(\mathbf{z}', \mathbf{z})$。

最后，我们将其传入相应的解码器，得到最终输出音频$\mathbf{y} \in \mathbb{R}^d$，其中 d 表示音频样本的总时长。请注意，后处理输出中生成的样本总数，对于生成式延展为 $d-T$，对于变形为 $d-2T$，因为 DiT 的原始 APG 输出将被原始提示信息替换。如图 1 右侧最终生成的音频波形所示，最终完整音频的总长度为d，提示的长度（深蓝色或绿色）为 T，后处理生成音频（浅蓝色）的长度在延展和变形情况下分别为 $d-T$ 和 $d-2T$。

2.4. 生成式延展

生成式延展可以通过两种不同模式实现：向前 和向后。两种情况下都需要一个音频提示$\mathbf{x}$。对于前向扩展（图 1 上方的示例），掩码函数 f_M 会将输入的潜在信息复制到噪声/输出潜变量的开头。这样，提示就会沿时间维度向前延展。另一方面，在后向扩展中（图 1 中间的示例），掩码函数 f_M 会将输入的潜在信息复制到末尾，从而使音频提示向后延展。最终输出的时长为 d 个音频样本，其中 T 个为输入查询样本。

2.5. 生成式变形

我们提出使用这一框架，在特定时间内将音频提示$\mathbf{x}_1$ 变形为$\mathbf{x}_2$，以生成音频信息（示例见图 1 底部）。这可以视为音频修补，且起始音频与结束音频可能大不相同。[2] 为此，我们需要两次应用掩码函数 f_M：第一次用音频查询$\mathbf{x}_1$ 替换开头处的噪声/输出潜变量，第二次用音频查询$\mathbf{x}_2$对噪声/输出潜变量的末尾进行掩码。假设$\mathbf{x}_1$ 和$\mathbf{x}_2$ 都包含 T 个音频样本，则生成的变形音频将长达 $d-2T$ 个音频样本。请注意，可以将两个查询潜变量的掩码位置设得更近，以调整最终所需的时长。最后，只要音频提示的总长度小于固定的生成长度 d，其长度可以任意，亦即 $2T < d$。

3. 训练策略

我们训练了一个立体声音频编码器，并调整标准的 DiT 训练流程，以实现所需的生成式扩展与变形功能。本节将详细介绍这一流程。

3.1. 立体声编码

我们训练了一种变分自编码器，并对其进行了创新性扩展，使其能够编码所有音频领域中的任意立体声音频。其架构采用全卷积结构，改编自 DAC [29]。它将 48千赫兹单声道或立体声音频压缩到 40Hz 下的 256 维潜在空间，并以 KL 正则化连续潜在空间取代残差向量量化。我们的模型能够编码各种类型的立体声音频，并特别注重准确的空间定位。为此，在将音频编码到潜在空间时，我们将音频波形参数化为单声道（左右声道之和）和侧声道（左右声道之差）。训练自编码器时，我们还会在计算重建损失（例如波形、频谱图等之间的差异）时应用这种参数化。最后，我们进行了广泛的数据增强，以确保涵盖所有可能的空间宽度和定位。

[2] 请注意，这与音色变形不同；音色变形是指将两种声音转换成一种新声音，且不包含时间过渡。

3.2. 从头训练 DiT

我们冻结立体声编码器/解码器，并从头开始训练 DiT，将完整生成时长设为 13 秒（即 $d = 13 \cdot 48k$），以确保模型在满足 GPU 内存限制的同时，保持适用于高效训练的实用批量大小。

在标准 DiT 训练过程中，我们进行了四项关键修改，以实现所需的生成能力：(i) 随机设定掩码的时长，使模型能够学会不受音频长度限制地延展/变形音频；(ii)随机掩码嵌入的开头、结尾或两端。这确保模型能在单次训练过程中理解如何延展/变形音频；(iii) 丢弃掩码过程，通过让模型在有掩码和无掩码的情况下接受训练，使其能够在推理期间有效应用 APG 策略；(iv) 对文本条件应用 *CFG*：我们使用成对的音频及其对应的自由文本描述训练模型，并随机省略这些文本条件。因此，模型在训练期间既能生成无条件音频，也能生成有条件音频，且生成时所用的含噪潜变量既可以经过掩码处理，也可以未经掩码处理。

3.3. 噪声底数据集

在我们的主要数据集上训练的模型表现出幻觉倾向，尤其是在提示中包含环境噪声、房间底噪和低音量白噪声等相对平稳的声音时。发生这种情况时，输出往往会包含音频提示中不存在的突兀、尖锐声音。我们推测，这种现象源于训练数据偏向较短的非平稳声音（例如脚步声、爆炸声）。为解决这一问题，我们提出使用平稳噪声数据对模型进行微调，确保模型遵循输入的平稳特性，并降低在对平稳声音进行延展或变形时产生非平稳声音幻觉的可能性。我们的目标是在提示包含非平稳声音时仍能生成这类声音，同时减少幻觉。为此，我们合成了一个专注于平稳声音的新数据集，称为噪声底数据集，用于微调模型。

噪声底数据集是一个新颖的数据集，包含 130 万小时的噪声底数据。该数据集需要两个关键组成部分：(i) **房间底噪**：这些数据源自 LibriVox 数据集 [31], , 我们将其输入 [26] 进行处理，以去除所有语音，仅保留背景声音，即房间底噪。我们的房间底噪数据由 408 个音频文件组成，总时长为 115 小时；(ii) **白噪声**：可按需生成，生成时指定要创建的音频文件最终长度 n。

获得这两个组成部分后，我们按如下方式创建噪声底数据集：从房间底噪数据中随机选取一个文件，采样其中 d 个连续音频样本。我们将这段音频与按需生成的、长度相同的白噪声进行卷积，得到与房间底噪频率响应相匹配的白噪声，从而合成 d 个全新且独特的噪声底样本。要获得立体声房间底噪，只需对两个声道分别重复上一步。我们重复整个流程，直至生成所需数量的文件。噪声底数据集共包含 10 万个文件。

3.4. 微调：减少幻觉

缓解幻觉的最后一步是使用噪声底数据集对模型进行微调。为此，我们继续训练模型，且仅使用这些新合成的数据。我们通过要求模型生成向前/向后扩展（即不进行变形）来对其进行微调，因为这似乎能带来最佳性能。

我们尝试了不同的微调迭代次数：10k、15k 和 20k。定性结果表明，正如预期，迭代次数越多，模型产生的幻觉就越少。然而，增加迭代次数会显著降低提示词保真度，这可能表明灾难性遗忘会产生轻微影响。因此，我们力求在减少幻觉与保持对原始提示词的忠实度之间取得平衡。定性评估表明，10k 次迭代通常能让单声道模型取得最佳效果，而 20k 次迭代则最适合立体声模型。

4. 实验

本节将介绍我们为客观和主观评估模型而开展的实验。此外，我们强烈建议读者访问配套网站，聆听模型的输出示例。[3]

4.1. 实验设置

我们将专有音效数据集与采用 CC 许可的通用音频数据集合并，共计 110 万个带标签的音频样本。音效数据通常包含单个清晰事件，而通用音频数据噪声较多，也可能包含多个描述不充分的事件。该数据集不包含音乐或语音，因为它们不属于预期应用范围。此外，我们使用 Mixtral 8x7B [32] 为这类数据集补充元数据，生成自由文本描述，从而构建出包含可靠音频-文本对的数据集。整个数据集中的音频采样率为 48千赫兹，且大多为立体声。训练期间，我们会随机将音频下混为单声道，使模型能够生成立体声和单声道内容。我们使用一组经过精心筛选的 98 个立体声音频片段进行评估。这些片段来自专业声音设计测试人员，旨在涵盖该模型所有现实可行的应用场景。上述 100 万条数据与这组评估数据没有重叠。除非另有说明，我们在所有实验中都使用这组评估数据。

我们的 DiT架构采用 8 个头和 24 层，激活函数为 SwiGLU [33], , 且仅在第一层通过交叉注意力处理文本条件。我们在 13 秒（$d=624k$）的音频潜变量上使用 v预测 [34]进行扩散过程，并通过 24 步去噪；除非另有说明，音频提示引导尺度设为 $\gamma=5$。我们使用 AdamW优化器，以潜变量上的标准均方误差损失进行训练，学习率为 1e-4，权重衰减为 1e-2。我们采用 4k 步的线性预热，以及尺度为 0.5 的余弦衰减计划。总训练迭代次数为 400k，并对嵌入层和注意力层应用 10% 的丢弃。每 100 步对模型权重应用指数移动平均，衰减率为 0.99。我们使用 32 张 A100 40GB GPU，在上述 1M 数据集上进行训练，有效批量大小为 256（即每批 3,328 秒）。随后，我们使用 8 张 A100 40GB GPU，在第 3.3 节所述的噪声数据集上微调 15k 次迭代。最后，潜变量会在开头、结尾或两端随机进行掩码处理，并以 50% 的概率丢弃此类掩码。掩码长度从 $\mathcal{U}(0,\ 3。25)$（秒）中采样。文本输入有 20% 的概率被丢弃（即进行无条件训练）。

4.2. 评估指标

我们采用标准的 Fr´echet 音频距离（FAD） [25]，以客观衡量结果的质量。该指标依赖两个关键要素：(i) 用于比较目标音频分布的参考集；(ii) 能将音频嵌入压缩但具有语义意义的空间中的音频模型。作为参考集，我们使用 Audition SFX 数据集 [35], , 其中包含约 1 万个高质量音效。我们选择 CLAP [36] 作为嵌入模型，因为它支持处理 48千赫兹立体声文件。在评估生成的扩展音频和变形音频时，我们只使用实际生成的内容，确保掩码潜变量不会造成信息泄漏并影响 FAD 计算。由于 FAD

[3] https://urinieto.github.io/genextend_ht ml

表 1：客观音频质量评估。

方法	FAD ↓
GenExtend	0.520
GenMorph	**0.432**
卷积噪声匹配 [37]	0.599
白噪声	1.358
噪声底	0.586
原始音频	**0.426**

是一种实际的音频距离指标，分数越低表示质量越好。最后，在本节末尾进行主观评估时，我们采用标准的平均意见分（MOS）。

4.3. 客观音频质量结果

客观音频质量实验结果如表 1 所示。作为基线，我们计算了原始音频（即用于生成的音频提示）、从第 3.3 节所述噪声底数据集中随机抽取的 1k 个文件，以及纯白噪声的 FAD。此外，我们还实现了一种基于 [37], 的经典且稳健的扩展方法，称为“卷积噪声匹配”。该方法将白噪声与输入信号的包络进行卷积，从而重现提示中的房间底噪／噪声底。我们未能找到其他专门解决这一问题的相关开源方法（例如，[38] 只扩展／变形音乐，[39] 图像；[16] 只进行音色变形，而非时间变形；等等）。

不出所料，白噪声和原始音频分别构成了性能下限和上限。卷积噪声匹配的结果与噪声底相近，这符合预期，因为该方法旨在处理平稳声音。有趣的是，生成式扩展和变形取得了不错的分数，接近原始提示。变形音频几乎与音频提示无法区分（0.432 对 0.426），表明其音频质量与实际使用的提示相当。作为合成环境音的噪声底，其 FAD 结果较差，这也在预期之中，因为该数据集中缺少非平稳声音。

4.4. 音频提示引导消融

图 2 展示了针对所提出的音频提示引导的消融研究结果。我们将引导参数 γ 从 0（无引导）调整到 6（强力引导至掩码噪声潜变量）。如图所示，较高的 γ 值能改善 FAD 指标。然而，当$\gamma=6$时，生成式变形的分数会下降，而生成式延展的分数则保持稳定。因此，我们选择 $\gamma=5$ 作为最佳引导参数，从而生成高质量结果，并使其紧密遵循音频提示。

有趣的是，似乎不对掩码潜变量进行引导
这对变形尤其有益，但对扩展则没有。我们推测，尽管生成结果并未忠实遵循经过掩码处理的提示，但其质量相当高，因为它必须保持在同一分布中。另一方面，没有引导时，生成式变形无法生成两个可能不同、可供变形的声音，因此 APG 在推理过程中至关重要。

4.5. 听众研究

我们开展了一项听众研究，对我们提出的任务之一——生成式延展——的质量进行主观评估。研究共涉及 15 名参与者，其中包括 12 名熟悉视频编辑工作流程的专业人士，以及 3 名音频/视频研究人员。参与者需从 1（最差）到 5（最佳）进行评分，从以下三个维度评价延展片段：

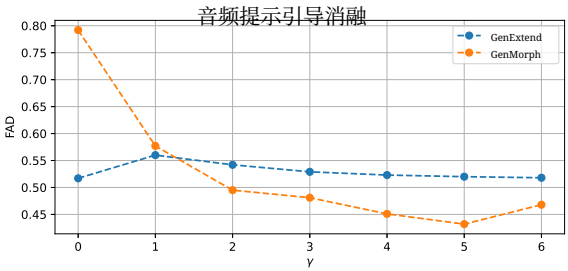


图 2：音频提示引导技术的消融实验。

表 2：听众研究：各音频维度中不同评分（1-5）所占的受试者百分比及平均分。

评分 ↑	平滑度	一致性	质量
1	5.9%	2.1%	3.8%
2	9.1%	2.8%	6.6%
3	27.5%	13.6%	26.5%
4	**36.9%**	**54.0%**	**41.1%**
5	20.6%	27.5%	22.0%
平均值	3.5	3.8	3.5

平滑度——从音频提示到延展片段的过渡是否平滑；一致性——延展片段在语义上与提示是否一致；以及质量——延展片段整体听起来有多好。平均而言，每位参与者对从包含 298 个样本的数据集抽取的 100 个音频扩展进行了评分。总计收集了 1,435 份评分。

表 2 展示的结果包含向前和向后扩展。Consistency 的平均得分最高，平均意见分（MOS）为 3.8。不过，平滑度和质量的 MOS 仅略低，为 3.5，这表明受试者对这些特征的感知同样积极。三项音频特征的中位数均为 4，分别对应平滑度、一致性和质量的描述“相当连贯”“相当相关”和“良好”。这些令人鼓舞的结果彰显了我们的模型在实际应用中的巨大潜力。

5. 结论与未来工作

我们提出了一种新颖的方法，利用扩散Transformer和一种无分类器引导变体生成高质量的音频延展片段与变形音频，我们将其称为音频提示引导。我们的方法有效应对了声音设计师面临的挑战，能够生成连贯流畅、紧密遵循输入提示的音频。客观与主观评估结果证明了该方法的有效性：生成音频的 Fr´echet 音频距离与原始音频相当，且获得了声音设计专家的积极评价。未来，我们计划探索以下方向，进一步提升模型性能：(i) 研究将本方法应用于其他类型的音频数据，例如语音和音乐；(ii) 整合更多控制机制，使声音设计师能够指定生成音频更为详细的特征，例如特定音色或动态范围；(iii) 探索使用本方法微调预训练生成模型，而非从头训练；(iv) 应用蒸馏或其他近期技术，例如流匹配 [40] 或归纳矩匹配 [41]，以减少推理时所需的扩散步数。

通过探索这些方向，我们希望为声音设计师打造更强大、更多用途的工具，使他们能够专注于创作过程，并更轻松、高效地制作高质量音频内容。

6. 参考文献

[1] Vanessa Theme Ament，《拟音圣杯：电影、游戏与动画声音表演艺术》，Routledge，纽约，第3版，2021年9月。[2] Zach Evans、CJ Carr、Josiah Taylor、Scott H. Hawley 和 Jordi Pons，“快速时间条件潜音频扩散”，载于《国际机器学习会议论文集》（ICML），2024年，第235卷，第12652–12665页。[3] Haohe Liu、Zehua Chen、Yi Yuan、Xinhao Mei、Xubo Liu、Danilo Mandic、Wenwu Wang 和 Mark D Plumbley，“AudioLDM：基于潜扩散模型的文本到音频生成”，载于《国际机器学习会议论文集》（ICML），2023年，第202卷，第21450–21474页。[4] Dongchao Yang、Jianwei Yu、Helin Wang、Wen Wang、Chao Weng、Yuexian Zou 和 Dong Yu，“Diffsound：用于文本到声音生成的离散扩散模型”，IEEE《音频、语音与语言处理汇刊》，第31卷，第1720–1733页，2023年。[5] Zhifeng Kong、Wei Ping、Jiaji Huang、Kexin Zhao 和 Bryan Catanzaro，“DiffWave：一种多用途扩散模型，用于音频合成”，载于《国际学习表征会议论文集》（ICLR），2021年。[6] Haohe Liu、Yi Yuan、Xubo Liu、Xinhao Mei、Qiuqiang Kong、Qiao Tian、Yuping Wang、Wenwu Wang、Yuxuan Wang 和 Mark D. Plumbley，“AudioLDM 2：通过自监督预训练学习整体音频生成”，IEEE《音频、语音与语言处理汇刊》，第32卷，第2871–2883页，2024年。[7] Zhifang Guo、Jianguo Mao、Rui Tao、Long Yan、Kazushige Ouchi、Hong Liu 和 Xiangdong Wang，“基于多条件扩散模型的音频生成”，载于《AAAI人工智能会议论文集》，2024年。[8] Sonal Kumar、Prem Seetharaman、Justin Salamon、Dinesh Manocha 和 Oriol Nieto，“SILA：用于增强文本到音频生成控制的信号到语言增强”，载于WASPAA论文集，2025年。[9] Hugo Flores Garc´ıa、Oriol Nieto、Justin Salamon、Bryan Pardo 和 Prem Seetharaman，“Sketch2Sound：通过时变信号和声音模仿实现可控音频生成”，载于《国际声学、语音与信号处理会议论文集》（ICASSP），2025年。[10] Marco Furio Colombo、Francesca Ronchini、Luca Comanducci 和 Fabio Antonacci，“MambaFoley：使用选择性状态空间模型生成拟音”，载于《国际声学、语音与信号处理会议论文集》（ICASSP），2025年。[11] Yoonjin Chung、Junwon Lee 和 Juhan Nam，“T-FOLEY：一种用于时间事件引导拟音合成的可控波形域扩散模型”，载于《国际声学、语音与信号处理会议论文集》（ICASSP），2024年。[12] Marco Comunit`a、Riccardo F Gramaccioni、Emilian Postolache、Emanuele Rodol`a、Danilo Comminiello 和 J oshua D Reiss，“Syncfusion：多模态起始点同步视频到音频拟音合成”，载于《国际声学、语音与信号处理会议论文集》（ICASSP）。IEEE，2024年，第936–940页。[13] Gal Greshler、Tamar Shaham 和 Tomer Michaeli，“Catch-a-Waveform：学习从单个短示例生成音频”，载于《神经信息处理系统大会论文集》（NeurIPS），2021年，第34卷。[14] Mark Levy、Bruno Di Giorgi、Floris Weers、Angelos Katharopoulos 和 Tom Nickson，“利用扩散模型与引导梯度实现可控音乐制作”，CoRR，第abs/2311.00613卷，2023年。[15] Zeyue Tian、Yizhu Jin、Zhaoyang Liu、Ruibin Yuan、Xu Tan、Qifeng Chen、Wei Xue 和 Yike Guo，“AudioX：用于任意内容到音频生成的扩散变换器”，arXiv预印本 arXiv:2503.10522，2025年。[16] Purnima Kamath、Chitralekha Gupta 和 Suranga Nanayakkara，“Mor- phFader：利用文本到音频模型实现细粒度可控变形”，载于《国际声学、语音与信号处理会议论文集》（ICASSP），2025年。[17] Marco Comunit`a、Zhi Zhong、Akira Takahashi、Shiqi Yang、Mengjie Zhao、Koichi Saito、Yukara Ikemiya、Takashi Shibuya、Shusuke Takahashi 和 Yuki Mitsufuji，“SpecMaskGIT：音频频谱图的掩码生成建模，用于高效音频合成及其他应用”，载于《国际音乐信息检索会议论文集》（ISMIR），2024年。[18] Xinlei Niu、Jing Zhang 和 Charles Patrick Martin，“Sound Morpher：基于扩散模型的感知均匀声音变形”，arXiv预印本 arXiv:2410.02144，2024年。[19] Federico Miotello、Mirco Pezzoli、Luca Comanducci、Fabio Antonacci 和 Augusto Sarti，“使用多分辨率谐波卷积神经网络进行基于深度先验的音频修复”，IEEE/ACM《音频、语音汇刊

’语言处理 ，第32卷，第113–123页，2024年。

[20] Nathana¨el Perraudin、Nicki Holighaus、Piotr Majdak 和 Peter Balazs，《使用相似性图修复长音频片段》，IEEE/ACM 音频、语音与语言处理汇刊，第 26 卷，第 6 期，第 1083–1094 页，2018 年。[21] Andr´es Marafioti、Piotr Majdak、Nicki Holighaus 和 Nathana¨el Perraudin，《GACELA：用于音乐长音频修复的生成对抗上下文编码器》，IEEE 信号处理精选主题期刊，第 15 卷，第 1 期，第 120–131 页，2021 年。[22] Marcelo Caetano，《感知、表征、图像、声音、音乐》，章节《使用声音变形工具箱中的正弦模型对乐器声音进行变形》，第 481–503 页，施普林格，2021 年。[23] William Peebles 和 Saining Xie，《基于 Transformer 的可扩展扩散模型》，载于 IEEE 计算机视觉国际会议论文集，2023 年，第 4172–4182 页。[24] Jonathan Ho 和 Tim Salimans，《无分类器扩散引导》，载于神经信息处理系统大会深度生成模型与下游应用研讨会，2021 年。[25] Kevin Kilgour、Mauricio Zuluaga、Dominik Roblek 和 Matthew Sharifi，《Fr´echet 音频距离：一种用于评估音乐增强算法的无参考指标》，载于国际语音通信协会会议论文集，2019 年，第 2350–2354 页。[26] Ke Chen、Jiaqi Su 和 Zeyu Jin，《MDX-GAN：通过对抗训练提升多类别源分离的感知质量》，载于国际声学、语音与信号处理会议论文集，2024 年，第 741–745 页。[27] Diederik P. Kingma 和 Max Welling，《自编码变分贝叶斯》，载于国际学习表征会议论文集，2014 年。[28] Aaron Van Den Oord、Oriol Vinyals 和 Koray Kavukcuoglu，《神经离散表征学习》，载于神经信息处理系统大会论文集，2017 年，第 30 卷，第 6307–6316 页。[29] Rithesh Kumar、Prem Seetharaman、Alejandro Luebs、Ishaan Kumar 和 Kundan Kumar，《采用改进 RVQGAN 的高保真音频压缩》，载于神经信息处理系统大会论文集，2023 年，第 36 卷，第 27980–27993 页。[30] Olaf Ronneberger、Philipp Fischer 和 Thomas Brox，《医学图像计算与计算机辅助干预——MICCAI 2015》，章节《U-Net：用于生物医学图像分割的卷积网络》，第 234–241 页，施普林格国际出版，尚，2015 年。[31] LibriVox，《LibriVox 免费有声读物合集》，https: //librivox.org/，2005 年至今，访问日期：2026-01-28。[32] Albert Q. Jiang、Alexandre Sablayrolles、Antoine Roux、Arthur Mensch、Chris Bamford、Devendra Singh Chaplot、Diego de las Casas、Emma Bou Hanna、Guillaume Lample、Lucile Saulnier、Marie-Anne Lachaux、Pierre Stock、Sandeep Subramanian、Sophia Yang、Szymon Antoniak 和 Teven Le Scao，《专家混合模型》，arXiv 预印本 arXiv:2401.04088，2024 年。[33] Noam Shazeer，《GLU 变体改进 Transformer》，arXiv 预印本 aXiv:2002.05202，2020 年。[34] Tim Salimans 和 Jonathan Ho，《用于快速采样扩散模型的渐进式蒸馏》，载于国际学习表征会议论文集，2022 年。[35] Adobe 公司，《Audition 音效库》，https://www.adobe.com/products/audition/offers/adobeauditiondlcsfx. html，访问日期：2026-01-28。[36] Yusong Wu*、Ke Chen*、Tianyu Zhang*、Yuchen Hui*、Taylor Berg- Kirkpatrick 和 Shlomo Dubnov，《采用特征融合与关键词到字幕增强的大规模对比式语言-音频预训练》，载于国际声学、语音与信号处理会议论文集，2023 年。[37] Josh H. McDermott 和 Eero P. Simoncelli，《通过听觉外周统计特征感知声音纹理：来自声音合成的证据》，《神经元》，第 71 卷，第 5 期，第 926–940 页，2011 年。[38] Hugo Flores Garcia、Prem Seetharaman、Rithesh Kumar 和 Bryan Pardo，《VampNet：通过掩码声学令牌建模生成音乐》，载于国际音乐信息检索会议论文集，2023 年。[39] Andreas Lugmayr、Martin Danelljan、Andr´es Romero、Fisher Yu、Radu Timofte 和 Luc Van Gool，《RePaint：使用去噪扩散概率模型进行修复》，载于 IEEE 计算机视觉与模式识别会议论文集，2022 年，第 11451–11461 页。[40] Yaron Lipman、Ricky T. Q. Chen、Heli Ben-Hamu、Maximilian Nickel 和 Matthew Le，《用于生成式建模的流匹配》，载于国际学习表征会议论文集，2023 年。[41] Linqi Zhou、Stefano Ermon 和 Jiaming Song，《归纳矩匹配》，载于国际机器学习会议论文集，2025 年。