RFM-编辑：矫正流匹配用于文本引导的音频编辑

Liting Gao[1], Yi Yuan[1], Yaru Chen[1], Yuelan Cheng[1], Zhenbo Li[2], Juan Wen[2], Shubin Zhang[3], Wenwu Wang

[1] Centre for Vision, Speech and Signal Processing (CVSSP), University of Surrey, United Kingdom
[2] College of Information and Electrical Engineering, China Agricultural University, China
[3] Fisheries College, Ocean University of China, China

ABSTRACT

扩散模型在文本到音频生成方面取得了显著进展。然而，基于文本引导的音频编辑仍处于起步阶段。这项任务的重点是在保持其他部分不变的情况下修改音频信号中的目标内容，因此需要精确定位并根据文本提示进行忠实编辑。现有的基于训练的方法和依赖于完整标题或昂贵优化的零样本方法往往在处理复杂编辑时遇到困难或缺乏实用性。在这项工作中，我们提出了一种新颖的端到端高效的校正流匹配扩散框架用于音频编辑，并构建了一个具有重叠多事件音频的数据集以支持复杂场景下的训练和基准测试。实验表明，我们的模型能够实现忠实的语义对齐，而无需辅助标题或掩码，同时在各种指标下保持竞争力的编辑质量。

Index Terms— 音频编辑、校正流匹配、扩散模型、CLAP 评分、音频编辑数据集

1. 介绍

近期基于扩散的生成模型在高质量文本到音频（TTA）生成方面取得了显著进展，示例包括基于去噪扩散概率模型（DDPM）的方法 [1]（AudioLDM [2,3]、Make-An-Audio [4,5]）和基于流的方法 [6]（TangoFlux [7]）。文本引导的音频编辑旨在根据自然语言指令或目标描述修改现有音频，同时保留未更改的内容。这使得通过提示进行灵活的音频操作成为可能，并支持声音设计、后期制作和个人化音频生成等应用。然而，关于文本引导音频编辑的研究，包括无训练扩散逆向方法 [8–10] 和基于训练的模型 [11,12]，在性能上仍有限制且处于早期阶段。

无训练的音频编辑通常利用预训练的 TTA 扩散模型 [2,13,14]，通过反转扩散过程从输入音频中恢复潜在噪声，并使用文本提示引导去噪。

诸如 AudioEditor [8] 等方法采用去噪扩散隐式模型 (DDIM) 反转和无文本优化进行高保真编辑，而提示引导的精确音频编辑（PPAE）[9] 和 DDPM 反转零样本 [10] 通过源提示和目标提示之间的语义变化来操纵交叉注意力图以实现局部控制。

它们都引入了 Prompt-to-Prompt 注意力替换机制 [15] 到音频编辑中，以准确对齐目标文本并显著提高 CLAP 分数。WavCraft [16] 进一步扩展这一范式，通过使用大型语言模型 (LLMs) 将提示翻译成专家模块指令来进行灵活的编辑。

相比之下，AUDIT [11] 训练了一个潜在扩散模型（LDM）[17]，使用三元组数据进行指令引导的音频编辑。非刚性提示编辑 [12] 在音频-标题对上微调一个扩散模型，并通过在提示嵌入空间中的插值来进行编辑。尽管基于训练的方法 [11] 通过显式监督实现了指令遵循，但它们的进步受到大规模数据集稀缺性的限制，使得准确定位编辑区域同时保留其余部分变得困难，特别是在有重叠声音的复杂场景中。相比之下，无训练方法在没有标签数据的情况下提供了灵活性，但在推理过程中 [8] 经常需要昂贵的空文本优化。此外，一些方法依赖于完整的标题 [8–10, 12] 或修改后的令牌掩码 [8]，而不是简洁的编辑指令，这既耗时又不切

*Corresponding author. Thanks to XYZ agency for funding.

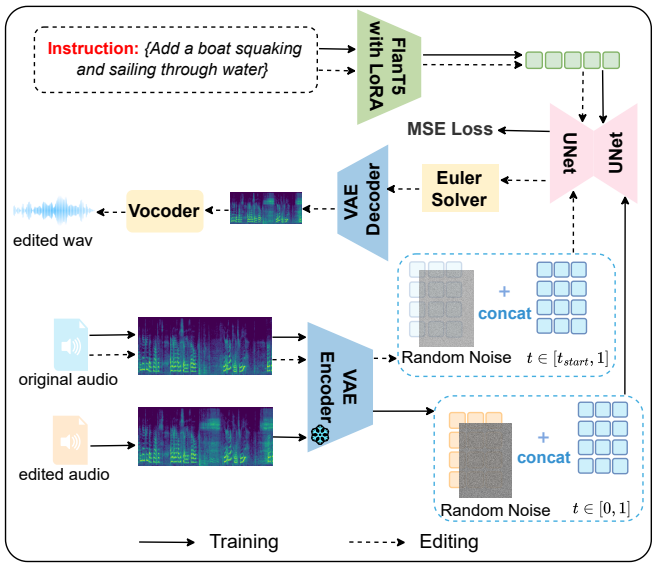


Fig. 1: RFM-编辑的训练和编辑管道。

实际。由于音频通常没有详细的文本描述相伴，我们认为理想的音频编辑系统应从原始音频和编辑指令开始操作，正如在 [11] 中所做的那样。

为了解决这些限制，我们引入了一个基于校正流匹配（RFM）[18] 的有效端到端潜在扩散文本引导音频编辑框架，并将其命名为RFM-Editing。它采用了一种训练范式，直接从指令中学习局部速度场，而不是明确的掩码或标题。为了支持训练，我们构建了一个大规模的音频编辑数据集，其中包含来自 AudioCaps2 [19] 的重叠多事件音频。RFM-Editing 在添加、删除和替换场景下实现了具有竞争力的表现和分布一致性，即使在复杂重叠事件的情况下也无需昂贵的推理时间优化。

2. 提出的方法

图 1 展示了 RFM-Editing 的训练和推理时编辑流水线，这是第一个统一的基于 RFM 的指令引导音频编辑模型，联合训练了三个编辑任务。基于 LDM [17]，RFM-Editing 集成了一个音频特征提取器、一个用于指令理解的低秩适应（LoRA [20]）调谐文本编码器、一个用于文本引导潜在编辑的 UNet，以及一个用于波形重建的 HiFi-GAN 解码器 [21]。RFM-Editing 还将潜在特征与原始特征在通道方向上连接，并重置反向扩散初始化状态以保留未编辑区域。

2.1. 使用修正流匹配进行训练

RFM-Editing 根据文本指令将输入音频转换成目标版本。它通过优化修正流匹配目标，在原始音频片段 x_{input}、它们的编辑版本 x_{edited} 以及相应的指令 $\mathcal{I}$ 上进行训练。x_{input} 和 x_{edited} 均被转换为对数梅尔谱图 $X_{\text{input}} \in \mathbb{R}^{T\times F}$ 和 $X_{\text{edited}} \in \mathbb{R}^{T\times F}$，其中 T 和 F 分别表示时间和频率维度。一个预训练的变分自动编码器（VAE）[22] 被用来将频谱图编码到潜在空间中，其中 x_T 是原始音频的潜在表示，而 x_0 则是编辑后音频的潜在表示。为了更好地捕捉编辑指令，我们将 LoRA [20] 应用于 Flan-T5 文本编码器 [23] 中，冻结预训练权重，并在变换器层中插入可训练的低秩矩阵，这减少了参数数量同时提高了语义理解能力以实现精确编辑。

我们将 RFM [18] 目标应用于学习一个连续向量场，该向量场利用基于 UNet 的扩散模型将噪声分布映射到编辑音频的目标分布。与依赖随机微分方程 (SDEs) 的标准扩散模型相比，RFM 形式化了一个确定性的常微分方程 (ODE) 过程，该过程建模了从噪声 ϵ 到目标 x_0 的直线轨迹，消除了对细粒度时间离散化的需要，从而导致稳定和高效的训练。具体而言，我们在 x_0 上添加随机高斯噪声 ϵ 并在连续时间步骤 $t \in [0, 1]$ 中获得扰动的潜在变量 x_t。受扰动的样本 x_t 沿着从噪声到数据的直线插值路径计算得出：

$$x_t = (1 - (1 - \sigma_{\min}) \cdot t) \cdot \epsilon + t \cdot x_0 \tag{1}$$

其中，$\sigma_{\min}$ 是控制 $t = 0$ 处最小噪声尺度的一个小常数。插值路径的时间导数产生在任何时间步 t 处的真实速度场：

$$\boldsymbol{v}_{\text{target}} = \frac{dx_t}{dt} = x_0 - (1 - \sigma_{\min}) \cdot \epsilon \tag{2}$$

为了帮助模型区分可编辑和不可编辑的内容，我们通过将原始潜变量 x_T 沿通道维度与噪声潜变量 x_t 进行拼接作为附加条件提供（如图 1 所示）。这使模型在训练和推理过程中都能直接访问未经编辑的输入，帮助保留未更改区域，同时仅在指定位置进行编辑。RFM-Editing 学习一个连续向量场 $\boldsymbol{v}_\theta^*(x_t \oplus x_T, t, E_{\mathcal{I}})$，该向量场根据指令嵌入 $E_{\mathcal{I}}$ 预测从 x_t 到目标潜变量 x_0 的方向。模型通过最小化预测速度场和目标速度场之

间的均方误差（MSE）损失进行训练：

$$\mathcal{L}_{\mathrm{FM}} = \mathbb{E}_{x_0,x_T,t,\epsilon}\left[\|\boldsymbol{v}_\theta^*(x_t \oplus x_T, t, E_{\mathcal{I}}) - \boldsymbol{v}_{\mathrm{target}}\|_2^2\right]. \tag{3}$$

该损失通过对每个采样时间步长鼓励预测的向量场与目标速度对齐来指导模型参数的优化。

2.2. 指令驱动编辑

在推理过程中，我们利用训练好的扩散模型根据原始音频 x_{input} 和文本指令 $\mathcal{I}$ 来执行音频编辑，而不需要完整的目标描述，就像在 [8, 10] 中所做的那样。x_{input} 首先由变分自编码器（VAE）编码为潜表示 x_T，而指令则使用 LoRA 调整的 Flan-T5 编码器嵌入到向量 $E_{\mathcal{I}}$ 中。

我们采用了一种更灵活的初始化策略，灵感来自 DDPM/DDIM 逆向过程，而不是从纯高斯噪声开始采样。由于音频编辑旨在保留大部分原始内容，而非完全重新合成新的音频，初始状态应保留部分原始输入的信息以更好地保持未编辑区域的一致性。具体来说，我们定义了沿着修正插值路径从噪声 ϵ 到原始音频潜变量 x_T 的起点 x_{start}：

$$x_{\mathrm{start}} = (1 - (1 - \sigma_{\min}) \cdot t_{\mathrm{start}}) \cdot \epsilon + t_{\mathrm{start}} \cdot x_T, \tag{4}$$

其中 t_{start} 是一个可调的小参数，在我们的模型中设置为 $t_{\mathrm{start}} = 0.01$。这通过在去噪过程中保留非编辑区域来促进忠实编辑，使编辑后的音频与原始音频之间具有一致性更好。因此，采样间隔变为 $t \in [t_{\mathrm{start}}, 1]$。在每一步 t 中，噪声潜在变量 x_t 沿着通道维度与原始潜在变量 x_T 连接，并且连同当前时间步长 t 和指令嵌入 $E_{\mathcal{I}}$ 一起传递给训练好的 UNet。UNet 预测瞬时速度场 $\boldsymbol{v}_\theta^*(x_t \oplus x_T, t, E_{\mathcal{I}})$，该速度场被连续时间欧拉求解器用于迭代更新潜变量：

$$x_{t+\Delta t} = x_t + \Delta t \cdot \boldsymbol{v}_\theta^*(x_t \oplus x_T, t, E_{\mathcal{I}}). \tag{5}$$

通过迭代此更新直到 $t = 1$，我们得到最终编辑后的潜变量 x_0^*。最后，通过 VAE 解码器对 x_0^* 进行解码以重建编辑后音频的对数梅尔谱图。然后使用 HiFi-GAN 语音编码器 [21] 将谱图转换为波形，生成最终的编辑后音频输出。

3. 实验

3.1. 数据集

我们使用 AudioCaps2 [19] 构建了一个基于指令的音频编辑数据集。DeepSeek API 被用来统计每个字幕中的声音事件数量。包含超过三个事件的音频片段被排除，因为它们往往噪声较大且不太适合训练，并将仅含一个事件的片段作为单事件片段用于组合。我们将每个音频 X 与两个随机的单事件片段 A 和 B 混合，以创建重叠且语义上有意义的例子，生成六个指令条件下的三元组：$\langle X, X + A, \text{Add A}\rangle$,$\langle X, X + B, \text{Add B}\rangle$,$\langle X + A, X, \text{Remove A}\rangle$,$\langle X + B, X, \text{Remove B}\rangle$,$\langle X + A, X + B, \text{Replace A with B}\rangle$,$\langle X+B, X+A, \text{Replace B with A}\rangle$，在训练过程中用作模型输入。每个示例都有一个仅用于评估的目标标题。原始标题保留供未来研究使用。

为了确保高质量的监督，我们计算每个音频与其字幕之间的 CLAP 相似度 [24]，并仅保留原始和编辑后的成对样本中 CLAP 相似度均超过 0.35 的样本。最终数据集包含每种任务类型 95,616 个样本，分别为训练、验证和测试提供了 234,639、26,103 和 26,103 个样本，并且每个划分在任务类型上保持平衡。我们还提供了一个相对较小的子集，包含 54 个,123、6 个,021 和 6 个,021 样本用于训练、验证和测试。

3.2. 实验设置和基线

我们在从以 16kHz 采样率采集的 10 秒音频片段中提取出含有 1024 个时间帧和 64 个梅尔频率分频点的日志梅尔谱图上训练模型。该模型使用具有交叉注意力机制的 U-Net 主干，并通过分类器自由引导进行条件设置。在 A100 GPU 上以学习率 5×10^{-5} 进行 100 个周期的训练。我们采用速度参数化，包括 1000 个扩散步骤和线性噪声调度表。在推理过程中，使用 Euler 积分 [25]，并设置 200 个采样步骤。验证基于从每个训练周期中的验证集中随机选取的 1000 个样本的 CLAP 相似度，最佳检查点根据最高的 CLAP 分数选择。

我们将我们的方法与三个强大的基线进行比较：AudioEditor [8]，Zero-Shot [10] 和 AUDIT [11]。相比之下，RFM-Editing 利用基于速度的 RFM 来实现跨

Table 1: 编辑音频的定量评估。

方法	FD ↓	脂肪酸脱氢酶 ↓	KL 下降	是 ↑
AudioEditor [8]	<u>14.24</u>	2.01	4.07	8.40
AUDIT [11]	32.62	7.22	9.99	<u>6.59</u>
Zero-Shot [10]	25.77	3.86	4.09	5.04
RFM-Editing	15.00	2.95	<u>2.90</u>	4.90
RFM-Editing$_{\text{full}}$	13.27	<u>2.50</u>	2.77	5.27

多样任务的指令引导编辑。我们使用预训练的 CLAP [24] 来计算编辑后的音频与目标字幕之间的余弦相似度以进行评估。为了评估整体质量、分布一致性及效率，我们报告了 Frechet 距离（FD）、Frechet 音频距离（FAD）、Kullback-Leibler（KL）散度、Inception 得分 (IS) [2] 和每个音频剪辑的平均编辑时间。FD、KL 和 IS 使用一个基于 PANNs 的模型 [26] 计算得出，该模型提取语义嵌入和类别 logits，而 FAD 则通过 VGGish 模型 [27] 测量得到，后者捕捉音频中的低级感知特征。此外，RFM-Editing 指的是在子集上训练，而 RFM-Editing$_{\text{full}}$ 表示在整个数据集上进行训练。

3.3. 结果

表 3 报告了编辑后音频质量的定量比较，包括 FD、FAD、KL 和 IS。我们观察到使用子集训练的 RFM-Editing 已经达到了具有竞争力的表现，在大多数指标上显著优于 AUDIT [11] 和 Zero-Shot [10]。当在完整数据集上进行训练时，RFM-Editing$_{\text{full}}$ 进一步提升，并实现了最佳的 FD 和 KL 得分，表明其在特征空间中具有更高的分布一致性以及更好的与目标语义分布对齐。这些结果表明基于速度的 RFM 能够更稳定地建模全局分布转变，并且对于由音频编辑指令指定的各种语义变化具有更强的泛化能力。尽管 RFM-Editing 仅获得中等的 IS 得分，这是可以预料到的，因为任务强调的是与编辑指令的真实对齐而非最大化输出多样性。

虽然 AudioEditor [8] 通过局部注意力操作达到了最低的 FAD，但其在全局语义一致性（如 KL 所示）上的表现有限。AUDIT [11] 在构建的数据集上训练了 100 个周期后，保真度和分布得分较差，而零样本 [10] 达到了有竞争力的表现但缺乏分布一致性。相比之下，

Table 2: 编辑保真度和效率的比较。

方法	提示	掌声 ↑	编辑时间（秒）↓
AudioEditor [8]	图注 & 修改后的标记词汇	0.4579	101.87
AUDIT [11]	**指令**	0.1113	<u>11.00</u>
Zero-Shot [10]	<u>标题</u>	0.4333	12.52
RFM-Editing	**指令**	0.4250	10.97
RFM-Editing$_{\text{full}}$	**指令**	<u>0.4398</u>	11.27

我们的方法在所有指标上都实现了更加平衡的性能。

在表 2 中，RFM-Editing 显示了编辑后的音频与目标字幕的一致性以及编辑效率方面的明显优势。AudioEditor [8] 由于注意力替换机制 [15]，实现了最佳的 CLAP 分数和与目标文本的对齐，但由于推理时间优化，其编辑速度几乎比我们的慢一个数量级，显著降低了用户体验。此外，它需要完整的字幕和修改后的标记索引，而零样本 [10] 依赖于目标字幕而不是简洁的编辑指令，这突显了基于指令的方法如 [11] 和 RFM-Editing 的优越性。

3.4. 消融和可视化

如表 3 所示，增加 t_{start} 保留了更多的原始音频但减弱了编辑力度，这导致感知质量提高，正如 FAD 最低和 IS 最高时所反映的那样，当 $t_{\text{start}} = 0.1$ 时。然而同时这也导致与目标字幕的语义对齐较差，表现为 CLAP 分数极低。相比之下，设置 $t_{\text{start}} = 0.01$ 提供了最佳折衷方案，在保持竞争性音频质量的同时达到最好的 CLAP 分数。

为了直观地说明 RFM-Editing 的有效性，我们可视化了扩散网络中的交叉注意力权重。成功的编辑需要模型精确地定位要删除或替换的相关声音的时间帧，并关注指令中新添加的事件。我们既可视化特定令牌与音频特征之间的交叉注意力动态，也可视化显示频谱图时间帧上的时序指令序列-音频交叉注意力的热力图。

在图 2 中，特定指令标记与相应音频帧之间的交叉注意力权重揭示了模型如何专注于需要编辑的相关片段。值得注意的是，我们观察到令牌 “beeps” 和 “barking” 在具体时间帧上具有高注意力值，这些时间帧恰好对应于实际音频中的机器蜂鸣声和狗叫声的发生。这一观察表明我们的模型可以自动且准确地定

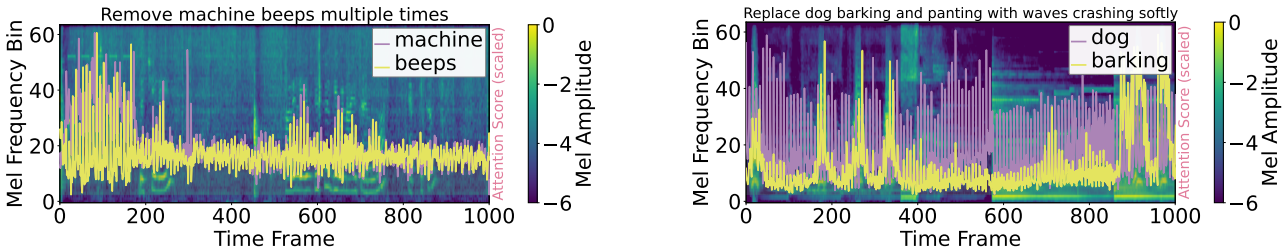
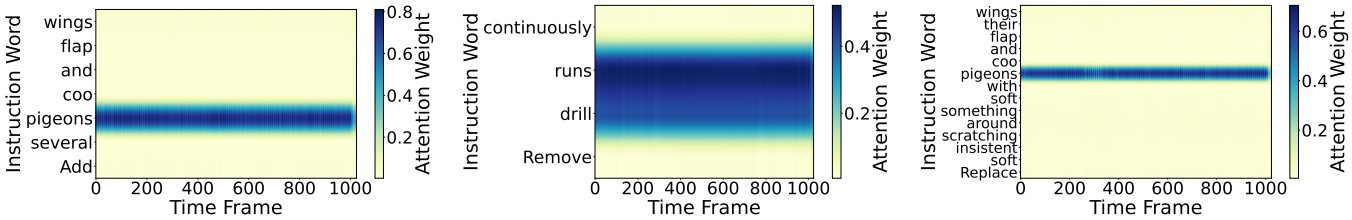


Fig. 2: 在 RFM 编辑中移除和替换任务的动态标记级交叉注意力趋势的可视化。



（a）添加　　（b）移除　　（c）替换

Fig. 3: 指令序列和 RFM 编辑中音频特征之间时间交叉注意力热图的可视化。

Table 3: 扩散起始时间初始化的影响。

t_{start}	掌声 ↑	下降 FD	FAD ↓	KL ↓	IS ↑
0	0.4216	17.97	2.45	2.96	4.27
0.001	0.4224	17.94	2.48	2.94	4.27
0.01	0.4249	17.38	2.52	3.06	4.34
0.1	0.3799	16.80	1.49	4.47	5.24

位音频事件，无需时间对齐的掩码，这对于实现精确有效的音频编辑至关重要。

此外，图 3(a)、(b) 和 (c) 中的热图显示，模型在所有任务中始终关注指令的关键部分，确保了准确且符合指令的编辑结果。有趣的是，我们观察到，在替换任务中，如果模型将更多注意力分配给要移除的事件而不是新引入的事件，则编辑质量往往会下降。此外，指令的质量对结果有重大影响，突显了提示在编辑任务中的关键作用。

4. 结论

我们介绍了 RFM-Editing，这是首个无需字幕或蒙版的指令引导音频编辑校正流匹配框架，并附带了一个新数据集。实验表明，RFM-Editing 可以自动定位与指令相关的时间帧，实现与目标语义的真实对齐和精确编辑。结果突显了校正流匹配作为实用范式的地位，并建议未来的工作可以利用语言提示能力。

5. 参考文献

[1] J. Ho, A. Jain, and P. Abbeel, “Denoising diffusion probabilistic models,” Advances in neural information processing systems, vol. 33, pp. 6840–6851, 2020.

[2] H. Liu, Z. Chen, Y. Yuan, X. Mei, X. Liu, D. Mandic, W. Wang, and M. D. Plumbley, “Audioldm: text-to-audio generation with latent diffusion models,” in Proceedings of the 40th International Conference on Machine Learning, 2023, pp. 21 450–21 474.

[3] H. Liu, Y. Yuan, X. Liu, X. Mei, Q. Kong, Q. Tian, Y. Wang, W. Wang, Y. Wang, and M. D. Plumbley, “Audioldm 2: Learning holistic audio generation with self-supervised pretraining,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 32, pp. 2871–2883, 2024.

[4] R. Huang, J. Huang, D. Yang, Y. Ren, L. Liu, M. Li, Z. Ye, J. Liu, X. Yin, and Z. Zhao, “Make-an-audio: Text-to-audio generation with prompt-enhanced diffusion models,” in International Conference on Machine Learning. PMLR, 2023, pp. 13 916–13 932.

[5] J. Huang, Y. Ren, R. Huang, D. Yang, Z. Ye, C. Zhang, J. Liu, X. Yin, Z. Ma, and Z. Zhao, “Make-an-audio 2: Temporal-enhanced text-to-audio generation,” arXiv preprint arXiv:2305.18474, 2023.

[6] Y. Lipman, R. T. Chen, H. Ben-Hamu, M. Nickel, and M. Le, “Flow matching for generative modeling,” arXiv preprint arXiv:2210.02747, 2022.

[7] C.-Y. Hung, N. Majumder, Z. Kong, A. Mehrish, A. A. Bagherzadeh, C. Li, R. Valle, B. Catanzaro, and S. Poria, "Tangoflux: Super fast and faithful text to audio generation with flow matching and clap-ranked preference optimization," arXiv preprint arXiv:2412.21037, 2024.

[8] Y. Jia, Y. Chen, J. Zhao, S. Zhao, W. Zeng, Y. Chen, and Y. Qin, "Audioeditor: A training-free diffusion-based audio editing framework," in ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2025, pp. 1–5.

[9] M. Xu, C. Li, D. Zhang, D. Su, W. Liang, and D. Yu, "Prompt-guided precise audio editing with diffusion models," in Proceedings of the 41st International Conference on Machine Learning, 2024, pp. 55 126–55 143.

[10] H. Manor and T. Michaeli, "Zero-shot unsupervised and text-based audio editing using ddpm inversion," arXiv preprint arXiv:2402.10009, 2024.

[11] Y. Wang, Z. Ju, X. Tan, L. He, Z. Wu, J. Bian et al., "Audit: Audio editing by following instructions with latent diffusion models," Advances in Neural Information Processing Systems, vol. 36, pp. 71 340–71 357, 2023.

[12] F. Paissan, L. Della Libera, Z. Wang, M. Ravanelli, P. Smaragdis, C. Subakan et al., "Audio editing with non-rigid text prompts," in Proceedings of INTERSPEECH 2024, 2024.

[13] J. Xue, Y. Deng, Y. Gao, and Y. Li, "Auffusion: Leveraging the power of diffusion and large language models for text-to-audio generation," IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2024.

[14] D. Ghosal, N. Majumder, A. Mehrish, and S. Poria, "Text-to-audio generation using instruction guided latent diffusion model," in Proceedings of the 31st ACM International Conference on Multimedia, 2023, pp. 3590–3598.

[15] A. Hertz, R. Mokady, J. Tenenbaum, K. Aberman, Y. Pritch, and D. Cohen-or, "Prompt-to-prompt image editing with cross-attention control," in The Eleventh International Conference on Learning Representations.

[16] J. Liang, H. Zhang, H. Liu, Y. Cao, Q. Kong, X. Liu, W. Wang, M. Plumbley, H. Phan, and E. Benetos, "Wavcraft: Audio editing and generation with natural language prompts." ICLR 2024 Workshop on LLM Agents, 2024.

[17] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 10 684–10 695.

[18] X. Liu, C. Gong, and Q. Liu, "Flow straight and fast: Learning to generate and transfer data with rectified flow," arXiv preprint arXiv:2209.03003, 2022.

[19] C. D. Kim, B. Kim, H. Lee, and G. Kim, "AudioCaps: Generating Captions for Audios in The Wild," in NAACL-HLT, 2019.

[20] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen et al., "Lora: Low-rank adaptation of large language models." ICLR, vol. 1, no. 2, p. 3, 2022.

[21] J. Kong, J. Kim, and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," Advances in neural information processing systems, vol. 33, pp. 17 022–17 033, 2020.

[22] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," arXiv preprint arXiv:1312.6114, 2013.

[23] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, Y. Li, X. Wang, M. Dehghani, S. Brahma et al., "Scaling instruction-finetuned language models," Journal of Machine Learning Research, vol. 25, no. 70, pp. 1–53, 2024.

[24] B. Elizalde, S. Deshmukh, M. Al Ismail, and H. Wang, "Clap learning audio concepts from natural language supervision," in ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2023, pp. 1–5.

[25] C. Lu, Y. Zhou, F. Bao, J. Chen, C. Li, and J. Zhu, "Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps," Advances in neural information processing systems, vol. 35, pp. 5775–5787, 2022.

[26] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "Panns: Large-scale pretrained audio neural networks for audio pattern recognition," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 28, pp. 2880–2894, 2020.

[27] S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A.

Saurous, B. Seybold et al., "Cnn architectures for large-scale audio classification," in 2017 ieee international conference on acoustics, speech and signal processing (icassp). IEEE, 2017, pp. 131–135.