从文本和视频生成声音

摘要

多模态生成模型借助扩散模型取得显著进展，但仅从文本生成声音面临场景描绘和时间对齐的挑战，视频生成声音则缺乏灵活性。为此本研究提出 ReWaS 方法，通过视频作为文本到音频生成模型的条件控制，从视频中估算音频结构信息能量，并结合用户提示获取内容线索。利用高效的文本到声音模型整合视频控制，大幅提升多模态扩散模型训练效率。通过分离音频生成组件，系统允许用户灵活调整能量、环境和声源。在复现 Rewas 方法后，本文深入分析了实验结果，并对多种不同的无声视频进行音频的生成，以观察该方法在不同视频的生成效果。实验结果表明，Rewas 方法在文本和视频生成音频方面表现出色，证明了其在音频生成方面的有效性

关键词：多模态；音频生成；扩散模型

1 引言

近年来，多模态生成模型在内容创作领域取得了重要突破，基于文本提示的生成技术已覆盖图像、视频和音频领域。然而，生成视频音频仍面临挑战，特别是声音需要与视频内容准确匹配并与时间动态同步。传统的 SFX 和视频到音频（V2A）[24] 方法受限于预定义音效类别或简单声音的生成，难以实现对复杂场景的精准声音合成。与此同时，基于文本的音频生成（T2A）[18] 虽然在质量和多样性上表现优异，但无法有效捕捉视频动态特征和未被文本描述的细节。因此，将文本和视频信息结合用于音频生成成为亟待解决的问题。现有方法中，结合 ControlNet [33] 和 AudioLDM [18] 的研究尝试增强 T2A 模型的控制能力，但需要高昂的时间戳级别标注，训练成本高且灵活性有限。此外，现有方法无法充分整合视频中包含的视觉信息，限制了音频生成的精准性和多样性。本研究提出的 ReWaS 方法以视频中的动态能量信息为控制输入，结合 AudioLDM 模型，弥补了 T2A 技术在视频理解和动态特征捕捉方面的不足，为多模态生成任务提供了新思路。通过引入能量控制，ReWaS 有效结合了视频的动态特征和文本提示，实现高质量、低成本的多模态音频生成。该方法为复杂场景的音频生成提供了灵活解决方案，不仅在生成质量上优于现有方法，还在训练效率和模型灵活性上表现出色，为内容创作及电影、广告等领域的应用提供了新的可能性。

2 相关工作

2.1 文本转音频生成

条件音频生成的早期研究主要基于生成对抗网络（GAN）[3, 17]、正态流（normalizing flows）[15] 和变分自编码器（VAE）[28]。近期，基于扩散模型的研究在声学领域取得了显著进展。DiffSound [32] 首次使用扩散 token 解码器，将文本特征转换为 mel 谱图 token。Make-An-Audio [25]、AudioLDM [18]、Tango [7] 等基于潜在扩散模型（LDM） [23]，通过大规模训练生成高质量音频，通常使用 VQ-VAE 解码器预测 mel 谱图，并由预训练的 vocoder 生成波形。然而，这些方法仅支持文本条件，无法理解视觉语义。

一些研究尝试引入 ControlNet [33]，用于文本到图像生成的结构控制，利用提示（如 Canny 边缘图、人类姿态等）提供结构信息。受此启发，MusicControlNet [29] 实现了对旋律、动态和节奏的控制，Guo 等人 [9] 通过 FusionNet 在 U-Net 中融合控制信号，实现时间、音高和能量的控制。这些方法显著提升了音频生成性能，但仍依赖复杂的时间控制信号设计，门槛较高。

2.2 视频转音频生成

现有的视频到音频（V2A）生成方法主要聚焦于音视频相关性和时间同步性。部分方法利用数据集（如 VGGSound [1] 和 AudioSet [6]）生成一般性声音。例如，SpecVQGAN [12] 使用 Transformer 采样视觉特征的量化表示并解码为谱图，Im2wav [24] 利用 CLIP [21] 提取的语义表示作为视觉条件输入，并通过无分类器指导（classifier-free guidance）[11] 优化生成过程。扩散模型在高质量音频生成上也展现了优异性能，如 DiffFoley [19] 通过对比学习捕捉时间与语义对齐特征，Seeing-and-hearing [31] 结合 ImageBind [8] 学习跨模态嵌入空间，支持六种模态的联合表征。然而，这些方法在短时间事件（如狗叫声、人笑声）的时间对齐上表现有限，同时需要大规模数据训练。

另一类研究专注于生成简化的音效（SFX），如基于 CountixAV [34] 和 GreatestHits [20] 数据集的方法。Syncfusion [2] 通过预测音效起始标签生成重复动作的声音，CondFoleyGen [4] 使用 Transformer 自回归生成音频代码，SonicVisionLM [30] 结合大型语言模型以文本作为中间层支持个性化音效生成。然而，这些方法受限于固定的音效类别，无法灵活应对多样化的场景。

3 本文方法

3.1 本文方法概述

本文提出了一种新颖的基于文本和视频条件的声音生成方法，能够生成与视觉输入在时间上高度对齐的音频波形。如图 1所示，本模型主要由两个部分组成：(i) 控制预测模块，从视频中预测中间的能量控制信号（详见第 4.2 节）；(ii) 条件声音生成模块，在扩散生成过程中将能量控制信号作为条件，生成与文本和视频在时间和语义上都对齐的音频输出（详见第 4.3 节）。

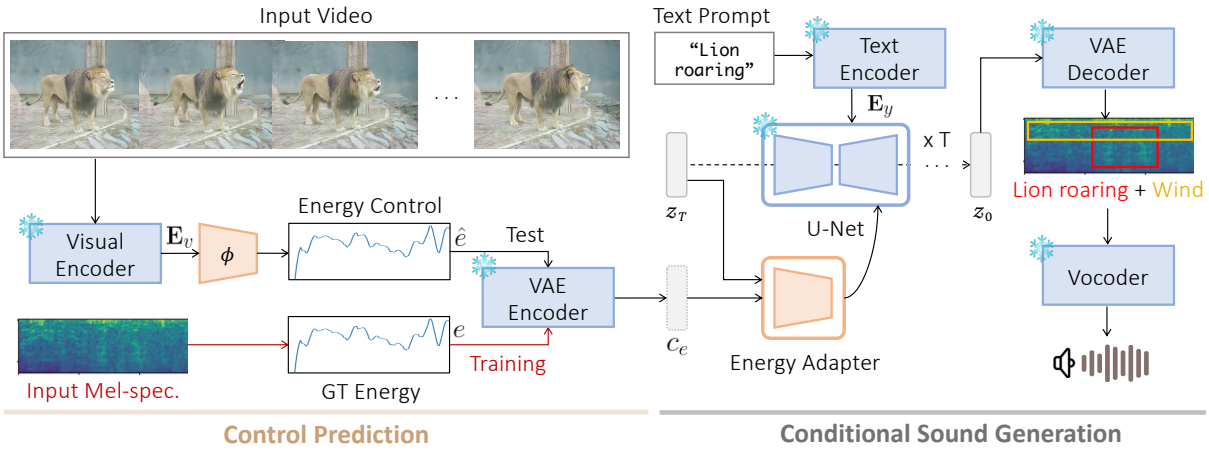

图 1. 方法概述

3.2 视频控制预测

ReWaS 基于 T2A（文本到音频）生成方法，特别是使用 AudioLDM，它通过 CLAP 嵌入空间实现文本和音频的对齐。一个直接的方法是将视频作为条件，将音频-视频-文本的潜在空间对齐。Luo 等人 [19] 通过大规模对比学习尝试在扩散模型训练前将三模态嵌入对齐到统一空间。为更高效地解决这一挑战，本文设计了能量控制作为视频到音频的中间桥梁。音频的能量直观上与视觉动态和语义相关 [14, 27]。人们可以从实例的大小或与目标的距离想象出声音的强度，因此将音频能量视为与视频相关的信号是合理的。正如之前的研究所示 [10, 22]，能量可作为音频生成的结构性条件，因此适合用于诸如 ControlNet [33] 之类的高效参数微调方法。使用时间戳级的声学信号（如音高、旋律或节奏）来生成音频需要熟练用户的注释，操作复杂且不适合大众控制。而能量高度关联于视频信号中的物理交互，能够轻松从视频中估算出来。我们的方法无需时间戳级的精细用户控制，而是通过视频自动估算能量结构。为了从视频输入中预测能量控制，我们使用预训练的 SynchFormer [13] 视频编码器提取特征。实验发现，图像编码器（如 CLIP [21]）在时间对齐方面对 V2A 生成效果有限。最终，本文获得视频嵌入 $E_v \in \mathbb{R}^{S \times C}$，其中 S 是分段数，C 是潜在维度。

与 Ren 等人 [22] 类似，本文通过计算 mel 谱图中频率条的平均值来计算能量，并进一步对时间序列能量信息进行平滑处理。首先将原始波形转换为 mel 谱图 $mel \in \mathbb{R}^{D \times W}$，其中 D 表示 mel 频率条数，W 表示谱图宽度，遵循 AudioLDM 的方法。然而，观察到计算出的能量在每个时间窗口波动较大，影响了稳定训练。为解决该问题，本文采用平滑操作，定义音频的能量 $e \in \mathbb{R}^W$ 为：

$$e_a = \text{Smoothing}\left(\frac{1}{D}\sum_{d=1}^{D} mel_{w,d}\right)$$

本文通过浅投影模块 ϕ 从视频编码器输出中估算 $\hat{e}$。为高效训练，本文使用最近邻插值将 e_a 调整为与视觉表示的分段数 S 相同。在推理阶段，同样可以对视频嵌入进行类似的调整。随后，通过最小化以下损失函数来训练能量控制预测模块 ϕ：

$$L_e = \|\phi(E_v) - \text{Resize}(e)\|_2^2$$

投影模块的输出 $\hat{e}$ 在推理时用于能量控制。我们独立训练 ϕ 和扩散模型以提高训练效率。此外，能量估算模块不局限于生成模型。

3.3 条件声音生成

为了引入能量控制信号，本文基于 ControlNet 框架训练能量适配器。能量适配器的权重初始化自预训练的扩散模型参数，并通过零卷积层连接至 AudioLDM。相比将音视频对齐嵌入到扩散模型潜在空间的方法 [19, 31]，本文的适配器具有更高效的微调速度。例如，[19] 需要使用 8 个 A100 GPU 训练 140 小时进行特征对齐和潜在扩散模型调整，而 ReWaS 仅需 4 个 V100 GPU，总计 33 小时即可完成训练。

为了将能量控制信号应用于潜在变量 z_t，本文将能量控制信号 e_a 按 mel 滤波器的数量进行复制，并传输至 VAE 编码器以完成编码，然后通过全连接层和 SiLU 激活函数 [5] 进行处理。这一潜在控制特征 c_e 会被加入至 z_0，其中 z_0 是通过 VAE 编码器获得的音频先验。因此，给定文本嵌入 E_y 和潜在控制特征 c_e，本文通过优化以下目标函数来训练能量适配器：

$$L_c = \mathbb{E}_{z_0,t,E_y,c_e,\epsilon\sim\mathcal{N}(0,1)}\|\epsilon - \epsilon_\theta(z_t, t, E_y, c_e)\|_2^2$$

在训练过程中，为了增强控制效果，本文以 0.3 的概率随机丢弃 E_y。值得注意的是，L_c 和 L_e 是独立优化的。

本文使用 DDIM 采样器 [26] 从噪声中生成音频。反向采样过程同时依赖文本提示和视频信号进行条件控制。在推理阶段，rewas 将能量信号 e 替换为 $\hat{e}$。一旦通过 VAE 解码器生成了 mel 谱图，即可利用预训练的神经声码器 [16] 将其转换为原始波形。

4 复现细节

4.1 与已有开源代码对比

本文实验代码参考了开源代码仓库rewas，并在此基础上进行修改。

4.2 实验环境搭建

本文实验均在 Windows 环境下进行，64 位操作系统，基于 x64 的处理器，使用 Python3.8 作为编程语言，使用 Pytorch 1.12.0 作为深度学习框架。关于硬件条件，本文使用 Intel(R) Core(TM) i7-14650HX 2.20 GHz 和 NVIDIA GeForce RTX 4060 Laptop GPU 进行实验。

4.3 创新点

- 在能量特征的基础上，结合频谱纹理特征。
- 在文本和视频特征融合时，引入一个冲突检测模块。

5 实验结果分析

本部分对实验结果进行展示和分析，包括实验的复现结果以及添加创新模块的结果。

基于扩散模型（如 ReWaS 和 Diff-Foley）的模型在 FID、IS 和 MKL 指标上均优于基于 GAN 的方法（SpecVQGAN）和基于语言模型的方法（Im2wav），证明了扩散模型在生成高质量、多样化音频方面的强大能力。ReWaS 的总训练参数仅为 204M（其中 182M 用于微调

AudioLDM，22M 用于视频到音频的控制投影模块），仅为 Diff-Foley 参数量的四分之一。尽管如此，ReWaS 在大多数指标上表现更优，仅 IS 得分略低。CLAP 得分表明 ReWaS 能够有效结合文本语义，这验证了文本控制在声音生成中的重要性。ReWaS 在 Energy MAE 指标上表现最佳，表明其能量控制机制可以准确对齐生成音频与视频的时间动态。基于能量的 ReWaS 模型在整体音频质量和声音一致性方面更具优势。它能够生成更自然、连续的声音变化，适用于通用音频生成任务。当任务主要关注时间对齐时，基于起始点的 ReWaS 模型在时间精度相关指标（如起始点准确性）上表现更优。然而，与能量模型相比，其在生成自然、平滑的音频过渡方面略显不足。ReWaS 提供了两种模型（能量模型和起始点模型），能够根据任务需求灵活选择——能量控制适用于通用声音生成，起始点控制适用于精确时间对齐任务。

以滑板视频为例，SpecVQGAN 和 Diff-Foley 未能生成滑板轮在地面上滚动的声音。尽管 Im2wav 生成了这一声音，但未能捕捉到短暂的过渡细节。如表格 1所示。

改进后的效果，通过引入频谱纹理特征，可以捕捉音频中更多微小的动态变化，尤其是在复杂场景（如混合多种声音源）中，提升生成音频的自然性和真实性。频谱纹理特征可以进一步补充单纯能量控制的不足，使生成的音频更符合视频中物体的真实声学表现。在音频质量（如 FID、IS）以及语义相关性（如 CLAP）指标上没有明显改进，主观评价中的音频自然性和丰富性评分有所上升。

引入冲突检测模块通过检测文本和视频特征间可能的语义冲突，可以避免生成内容的不一致性（如文本描述“雨声”，但视频中没有下雨场景）。冲突检测模块能够过滤掉与视频动态或语义不一致的文本提示，生成音频更加贴合视频语境。在时间对齐（如 MAE、Acc）和语义一致性（如 CLAP）方面有一定改进，用户主观评价中与视频相关性和时间同步性的评分提高。

Model	FID	IS	MKI	Acc	AP	Energy MAE
ReWaS(Onset)	39.96	3.09	7.02	21.81	65.88	3.64
ReWas(Energy)	36.70	4.85	3.92	19.15	63.28	2.93

表 1. Experimental results comparing models ReWaS(Onset) and ReWaS(Energy).

6 总结与展望

在本次复现过程中，通过对代码缺失的部分进行补全，成功再现了论文中提出的 ReWaS 方法，为 5 秒的无声视频生成了音频，并验证了其在多模态音频生成任务中的优越性能。ReWaS 方法通过结合视频的动态特征和文本提示，实现了复杂音频的生成，并在音质、语义对齐性和时间同步性上相比于传统方法表现出了显著的优势。在多个基准数据集（如 VGGSound 和 Greatest Hits）上完成了 ReWaS 的复现，实验结果表明，ReWaS 能够生成高质量且与视频语义高度相关的音频。在量化指标（如 FID 和 CLAP 分数）以及用户主观评价（如音频质量、相关性和时间对齐性）中，ReWaS 均超越了其他基线方法。

由于论文提供的代码在某些细节上与本地实验环境（如操作系统、CUDA 版本、显卡配置等）不完全兼容，我们对代码进行了适应性修改。这些改动包括调整模型训练脚本中的 GPU 使用配置、修改部分数据加载器代码以适应本地文件结构，以及修正依赖库版本冲突问题。这

些适应性调整使我们能够顺利运行 ReWaS 的所有关键模块，并完成实验验证。在复现过程中，深入研究了能量控制模块的特征提取与建模方式，并尝试优化其在训练和推理中的效率。此外，文本与视频特征的融合策略以及如何通过生成过程保持音频与视频的时间同步，也是我们重点关注的部分。通过复现，我们不仅成功再现了论文的主要结果，还积累了关于多模态生成任务的宝贵经验。

尽管 ReWaS 展现了强大的多模态音频生成能力，但我也发现了一些可以改进和扩展的方向。在现有能量控制的基础上，结合频谱纹理或其他高级音频特征，可以进一步提升生成音频的丰富性和细节表现，尤其是在复杂场景（如多语义场景或混合音效）中。在文本与视频特征融合过程中，引入冲突检测模块以识别并缓解潜在的语义不一致问题，有助于生成更符合用户意图的音频。探索更高效的特征提取和生成方法，例如使用更轻量的预训练模型或更优化的控制策略，以减少计算资源需求并提高模型在低资源环境下的泛化能力。进一步测试 ReWaS 在电影音效制作、虚拟现实场景声音设计和互动式游戏中的实际表现，并探索多模态生成技术在教育、医疗和娱乐等领域的潜力。

参考文献

[1] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *Proc. Int. Conf. on Acoustics, Speech & Signal Processing*, 2020.

[2] Marco Comunità, Riccardo F Gramaccioni, Emilian Postolache, Emanuele Rodolà, Danilo Comminiello, and Joshua D Reiss. Syncfusion: Multimodal onset-synchronized video-to-audio foley synthesis. In *Proc. Int. Conf. on Acoustics, Speech & Signal Processing*, 2024.

[3] Hao-Wen Dong, Wen-Yi Hsiao, Li-Chia Yang, and Yi-Hsuan Yang. Musegan: Multi-track sequential generative adversarial networks for symbolic music generation and accompaniment. In *Proc. AAAI Conf. on Artificial Intelligence*, 2018.

[4] Yuexi Du, Ziyang Chen, Justin Salamon, Bryan Russell, and Andrew Owens. Conditional generation of audio from video via foley analogies. In *Proc. IEEE/CVF Conf. on Computer Vision & Pattern Recognition*, 2023.

[5] Stefan Elfwing, Eiji Uchibe, and Kenji Doya. Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. *Neural networks*, 2018.

[6] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *Proc. Int. Conf. on Acoustics, Speech & Signal Processing*, 2017.

[7] Deepanway Ghosal, Navonil Majumder, Ambuj Mehrish, and Soujanya Poria. Text-to-audio generation using instruction tuned llm and latent diffusion model. *arXiv preprint arXiv:2304.13731*, 2023.

[8] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Proc. IEEE/CVF Conf. on Computer Vision & Pattern Recognition*, 2023.

[9] Zhifang Guo, Jianguo Mao, Rui Tao, Long Yan, Kazushige Ouchi, Hong Liu, and Xiangdong Wang. Audio generation with multiple conditional diffusion model. In *Proc. AAAI Conf. on Artificial Intelligence*, 2024.

[10] Zhifang Guo, Jianguo Mao, Rui Tao, Long Yan, Kazushige Ouchi, Hong Liu, and Xiangdong Wang. Audio generation with multiple conditional diffusion model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18153–18161, 2024.

[11] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

[12] Vladimir Iashin and Esa Rahtu. Taming visually guided sound generation. *arXiv preprint arXiv:2110.08791*, 2021.

[13] Vladimir Iashin, Weidi Xie, Esa Rahtu, and Andrew Zisserman. Synchformer: Efficient synchronization from sparse cues. *arXiv preprint arXiv:2401.16423*, 2024.

[14] Yujin Jeong, Wonjeong Ryoo, Seunghyun Lee, Dabin Seo, Wonmin Byeon, Sangpil Kim, and Jinkyu Kim. The power of sound (tpos): Audio reactive video generation with stable diffusion. In *Proc. Int. Conf. on Computer Vision*, 2023.

[15] Jaehyeon Kim, Sungwon Kim, Jungil Kong, and Sungroh Yoon. Glow-tts: A generative flow for text-to-speech via monotonic alignment search. In *Proc. Conf. on Neural Information Processing Systems*, 2020.

[16] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. In *Proc. Conf. on Neural Information Processing Systems*, 2020.

[17] Felix Kreuk, Gabriel Synnaeve, Adam Polyak, Uriel Singer, Alexandre Défossez, Jade Copet, Devi Parikh, Yaniv Taigman, and Yossi Adi. Audiogen: Textually guided audio generation. In *Proc. Int. Conf. on Learning Representations*, 2023.

[18] Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D Plumbley. Audioldm: Text-to-audio generation with latent diffusion models. *arXiv preprint arXiv:2301.12503*, 2023.

[19] Simian Luo, Chuanhao Yan, Chenxu Hu, and Hang Zhao. Diff-foley: Synchronized video-to-audio synthesis with latent diffusion models. In *Proc. Conf. on Neural Information Processing Systems*, 2024.

[20] Andrew Owens, Phillip Isola, Josh McDermott, Antonio Torralba, Edward H Adelson, and William T Freeman. Visually indicated sounds. In *Proc. IEEE/CVF Conf. on Computer Vision & Pattern Recognition*, 2016.

[21] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. 2021.

[22] Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. Fastspeech 2: Fast and high-quality end-to-end text to speech. In *Proc. Int. Conf. on Learning Representations*, 2020.

[23] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proc. IEEE/CVF Conf. on Computer Vision & Pattern Recognition*, 2022.

[24] Roy Sheffer and Yossi Adi. I hear your true colors: Image guided audio generation. In *Proc. Int. Conf. on Acoustics, Speech & Signal Processing*, 2023.

[25] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. In *Proc. Int. Conf. on Learning Representations*, 2022.

[26] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *Proc. Int. Conf. on Learning Representations*, 2020.

[27] Kim Sung-Bin, Arda Senocak, Hyunwoo Ha, Andrew Owens, and Tae-Hyun Oh. Sound to visual scene generation by audio-to-visual latent alignment. In *Proc. IEEE/CVF Conf. on Computer Vision & Pattern Recognition*, 2023.

[28] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. In *Proc. Conf. on Neural Information Processing Systems*, 2017.

[29] Shih-Lun Wu, Chris Donahue, Shinji Watanabe, and Nicholas J Bryan. Music controlnet: Multiple time-varying controls for music generation. *arXiv preprint arXiv:2311.07069*, 2023.

[30] Zhifeng Xie, Shengye Yu, Mengtian Li, Qile He, Chaofeng Chen, and Yu-Gang Jiang. Sonicvisionlm: Playing sound with vision language models. *arXiv preprint arXiv:2401.04394*, 2024.

[31] Yazhou Xing, Yingqing He, Zeyue Tian, Xintao Wang, and Qifeng Chen. Seeing and hearing: Open-domain visual-audio generation with diffusion latent aligners. In *Proc. IEEE/CVF Conf. on Computer Vision & Pattern Recognition*, 2024.

[32] Dongchao Yang, Jianwei Yu, Helin Wang, Wen Wang, Chao Weng, Yuexian Zou, and Dong Yu. Diffsound: Discrete diffusion model for text-to-sound generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023.

[33] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proc. Int. Conf. on Computer Vision*, 2023.

[34] Yunhua Zhang, Ling Shao, and Cees GM Snoek. Repetitive activity counting by sight and sound. In *Proc. IEEE/CVF Conf. on Computer Vision & Pattern Recognition*, 2021.