Diff-BGM：用于视频背景音乐生成的扩散模型

杨硕

2024 年 12 月 31 日

摘要

在编辑视频时，一段吸引人的背景音乐是不可或缺的。然而，视频背景音乐生成任务面临如缺乏合适的训练数据集，灵活控制音乐生成过程，顺序对齐视频与音乐方面等挑战。在技术复现中，论文首先提出了一个高质量的音乐-视频数据集 BGM909，提供关于视频和音乐的多模态信息。随后提出评估指标来评估音乐质量，包括音乐多样性和音乐与视频之间的对齐，以及检索精度指标。最后，论文提出了 Diff-BGM 框架，来自动生成给定视频的背景音乐，该框架在生成过程中使用不同的信号来控制音乐的不同方面，即使用动态视频特征来控制音乐节奏，使用语义特征来控制旋律和氛围。

关键词：扩散模型；背景音乐；音乐质量评估；

1 引言

随着多媒体和社交平台的快速发展，视频已经成为一种常见的表达情感和记录生活的方式。在创作视频时，选择合适的背景音乐对于提高视频的吸引力至关重要。但对于缺乏音乐知识或视频编辑经验的人来说，选择或创作合适的背景音乐并不容易。同时，版权保护问题也引起了广泛关注。因此，自动生成视频背景音乐是一个实用的解决方案。

现有的背景音乐生成方法主要使用基于 Transformer 框架建立视频和音乐之间的关系。但这些方法难以灵活地控制音乐生成过程，导致生成结果的可解释性差。论文提出了一种基于扩散模型的 Diff-BGM 框架，可以使用不同的信号来控制音乐生成的不同方面，如使用视频动态特征控制节奏，使用语义特征控制旋律和氛围，同时引入了分段感知的交叉注意力机制，更好地对齐视频和音乐的时间。

同时，文章提出的 Diff-BGM 框架可以自动为给定视频生成高质量的背景音乐，并通过可控的生成过程提高了结果的可解释性。实验结果表明，Diff-BGM 在客观和主观评估中都优于现有的最先进模型，为视频背景音乐生成任务提供了新的解决方案。

2 相关工作

2.1 音乐生成

基于 Transformer 的模型 [1] 在音乐生成任务上取得了不错的结果。但这些模型通常依赖人工生成的音乐标记，因此生成能力受限。而扩散模型在视觉任务和音乐生成任务中都展现

出了出色的生成能力。一些工作提出了基于扩散模型 [4] 的音乐生成方法，通过对音乐的不同方面施加控制来生成音乐。

2.2 背景音乐生成

针对视频背景音乐生成任务，现有的方法主要集中在以人物为核心的视频，如舞蹈或运动视频。这些方法通常依赖画面中人物的动作来控制音乐节奏，但这种方法不适用于自由风格的视频。视频背景音乐生成任务最早由 CMT [2] 提出，之后引起了越来越多的关注。现有的视频背景音乐生成方法大多基于 Transformer 框架。如,CMT 首先编码和弦和音符来表示音乐，然后建立三种节奏关系 (如视频运动速度与音频音符密度) 来缩小视频和背景音乐之间的差距。除了使用基于规则的节奏关系，V-MusProd [5] 和 Video2Music [3] 还关注语义层面的对应关系，提取视频的语义特征来控制生成音乐的风格。但基于 Transformer 的方法仍然存在难以控制端到端生成过程，导致可解释性较差的问题。

3 本文方法

3.1 本文方法概述

Diff-BGM 采用扩散模型作为生成器，在生成过程中灵活地利用视频的不同特征来控制音乐的不同方面。如图 1所示：音乐过程模块将原始 MIDI 文件转换为钢琴谱表示，作为生成器的输入。视频过程模块提取视频的视觉特征和语义特征，用于控制音乐的节奏和旋律。生成模块采用扩散模型，利用视频特征作为条件来生成新的钢琴谱。Diff-BGM 引入了分段感知的交叉注意力机制用于视频和音乐的时间对齐。模型的训练目标是最小化噪声预测误差，同时考虑视频特征作为条件。

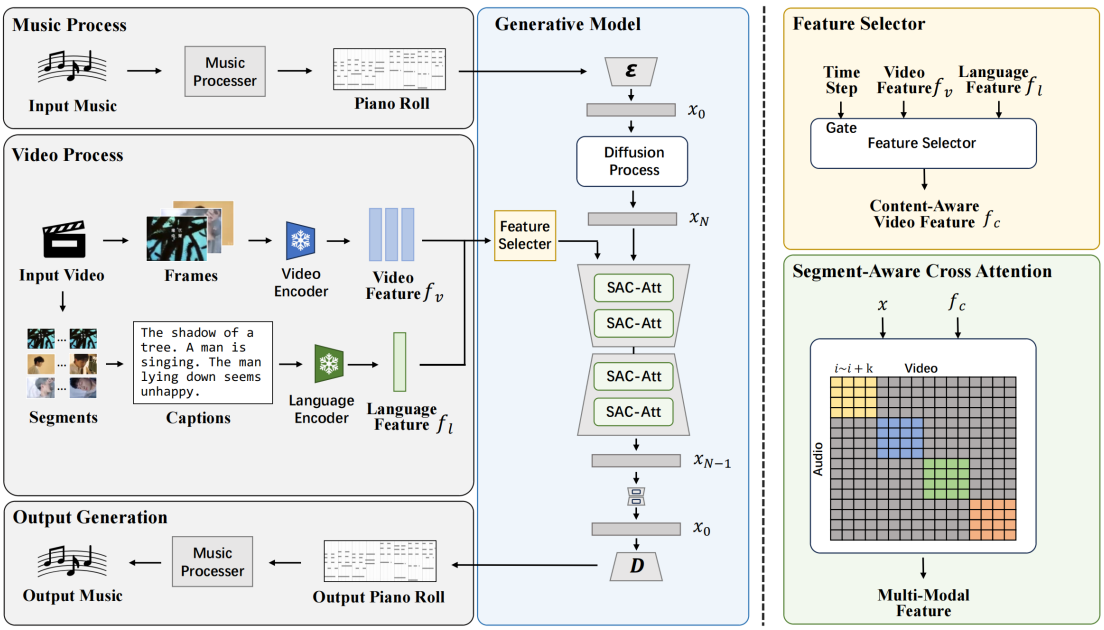

图 1. 方法示意图

3.2 扩散模型生成器

文章采用扩散模型作为生成器，将原始 MIDI 文件转换为钢琴谱表示，作为生成器的输入。扩散模型通过逐步加入噪声并预测噪声，最终生成新的钢琴谱。

3.3 视频特征提取

为了使生成的音乐与视频内容相匹配，文章提取了视频的视觉特征和语义特征，包括视觉特征和语义特征。这些特征将作为条件信息输入到生成模块中，以控制音乐的节奏和旋律。

3.4 分段感知的交叉注意力

为了实现视频和音乐的时间对齐，文章引入了分段感知的交叉注意力机制。具体来说，将相邻的 k 帧划分为短期上下文，只有这些相邻帧的特征才能影响每个时间点的音乐生成。这样可以捕捉视频的短期上下文信息，从而实现精确的时间对齐。

4 复现细节

4.1 与已有开源代码对比

引用的代码库：https://github.com/sizhelee/Diff-BGM。

4.2 实验环境搭建

克隆仓库后，安装 pip 依赖。然后将数据集 POP909 下载后放在./data/下，将数据集 BGM909 下载后放在./data/bgm909/下。下载好训练所需的预训练模型放在./pretrained/下。下载数据集的划分文件来划分数据集。

5 实验结果分析

实验结果可见视频 https://github.com/sizhelee/Diff-BGM/blob/master/video.mp4。

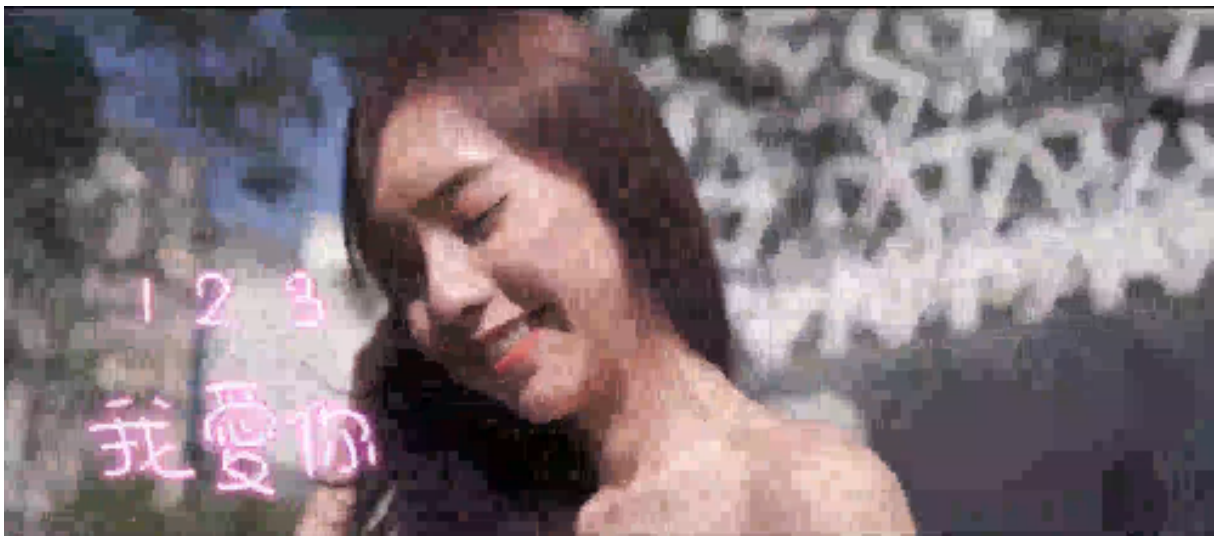

图 2. 实验结果示意

6 总结与展望

复现文章提出了一个基于扩散模型的视频背景音乐生成框架 Diff-BGM。Diff-BGM 通过利用视频的动态特征和语义特征，分别控制生成音乐的节奏和旋律，实现了视频和音乐的时间对齐。实验结果表明,Diff-BGM 生成的背景音乐质量更高, 与视频内容的对应性也更好, 优于现有的方法。

对于未来展望, 文章提出可以进一步提高生成音乐的质量和多样性, 例如通过引入更丰富的音乐理论知识。还可以探索更复杂的视频-音乐对应关系，如情感、节奏等更高层次的特征。将 Diff-BGM 应用于更广泛的场景, 如游戏、广告等。

参考文献

[1] Ke Chen, Yusong Wu, Haohe Liu, Marianna Nezhurina, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Musicldm: Enhancing novelty in text-to-music generation using beat-synchronous mixup strategies. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1206–1210, 2024.

[2] Shangzhe Di, Zeren Jiang, Si Liu, Zhaokai Wang, Leyan Zhu, Zexin He, Hongming Liu, and Shuicheng Yan. Video background music generation with controllable music transformer. In *Proceedings of the 29th ACM International Conference on Multimedia*, MM '21, page 2037–2045, New York, NY, USA, 2021. Association for Computing Machinery.

[3] Jaeyong Kang, Soujanya Poria, and Dorien Herremans. Video2music: Suitable music generation from videos using an affective multimodal transformer model. *Expert Systems with Applications*, 249:123640, 2024.

[4] Ye Zhu, Yu Wu, Kyle Olszewski, Jian Ren, Sergey Tulyakov, and Yan Yan. Discrete contrastive diffusion for cross-modal music and image generation. In *The Eleventh International Conference on Learning Representations*, 2023.

[5] Le Zhuo, Zhaokai Wang, Baisen Wang, Yue Liao, Chenxi Bao, Stanley Peng, Songhao Han, Aixi Zhang, Fei Fang, and Si Liu. Video background music generation: Dataset, method and evaluation. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15591–15601, 2023.