Academic Editor: Chang Wook Ahn

学术编辑： Chang Wook Ahn

Received: 18 February 2025

收到： 2025 年 2 月 18 日

Revised: 10 March 2025

修订： 2025 年 3 月 10 日

Accepted:14March 2025

接受： 2025 年 3 月 14 日

Published:18 March 2025

发表于： 2025 年 3 月 18 日

Citation:Zhao, Y.;Yang,M.; Lin,Y.; Zhang,X.; Shi,F; Wang,Z.; Ding,J.; Ning, H. AI-Enabled Text-to-Music Generation: A Comprehensive Review of Methods, Frameworks, and Future Directions. Electronics 2025,14,1197. https://doi.org/10.3390/

引用： 赵，y。；杨，m。；林，y。；张，x。；史，f。；王，z。；丁，j。；宁，h。AI-Enabled Text-to-Music Generation: a Comprehensive Review of Methods，Frameworks，and Future Directions。电子学 2025,14,1197。Https://doi.org/10.3390/

1School of Computer and Communications Engineering, University of Science and Technology Beijing, Beijing 100083, China; zhaoyujia@ustb.edu.cn (Y.Z.); shifeifei@ustb.edu.cn (F.S.) 2Guangxi Tourism Development One-Click Tour Digital Cultural Tourism Industry Co., Ltd., Nanning 530012, China; asanseu@163.com 3Administrative Office, Chunan Academy of Governance, Hangzhou 311700, China; m202120818@xs.ustb.edu.cn

北京科技大学计算机与通信工程学院，北京 100083; zhaoyujia@ustb.edu.cn (y.z。) ; shifeifei@ustb.edu.cn (f.s。) 2 广西旅游发展有限公司，南宁 530012; asanseu@163.com 3 淳安行政学院行政办公室，杭州 311700; m202120818@xs.ustb.edu.cn

45Foreign Department, Jinzhong University, Jinzhong 030606, China; dzxjgz666@163.com

晋中大学外事系，晋中 030606，中国；dzxjgz666@163.com

Department of Computer Science, Blekinge Institute of Technology, 371 79 Karlskrona, Sweden; jianguo.ding@bth.se . Correspondence: wangzj@ustb.edu.cn (Z.W.); ninghuansheng@ustb.edu.cn (H.N.)

布莱金厄理工学院计算机科学系，瑞典卡尔斯克鲁纳 37179;

These authors contributed equally to this work.

这些作者对这项工作做出了同样的贡献。

Abstract: Text-to-music generation integrates natural language processing and music gen-eration, enabling artificial intelligence (AI) to compose music from textual descriptions. While AI-enabled music generation has advanced, challenges in aligning text with mu-sical structures remain underexplored. This paper systematically reviews text-to-music generation across symbolic and audio domains, covering melody composition, polyphony, instrumental synthesis, and singing voice generation. It categorizes existing methods into traditional, hybrid, and end-to-end LLM-centric frameworks according to the usage of large language models (LLMs), highlighting the growing role of LLMs in improving controllabil-ity and expressiveness. Despite progress, challenges such as data scarcity, representation limitations, and long-term coherence persist. Future work should enhance multi-modal in-tegration, improve model generalization, and develop more user-controllable frameworks to advance AI-enabled music composition.

文本到音乐的生成集成了自然语言处理和音乐生成，使得人工智能 (AI) 能够从文本描述中创作音乐。虽然基于人工智能的音乐生成技术已经取得了进步，但是将文本与音乐结构对齐的挑战仍然没有得到充分的探索。本文系统回顾了符号和音频领域的文本 - 音乐生成，包括旋律创作、复调、器乐合成和歌唱声音生成。它根据大型语言模型 (llm) 的使用将现有方法分为传统的、混合的和端到端的以 llm 为中心的框架，突出了 llm 在提高可控性和表达能力方面日益增长的作用。尽管取得了进展，但数据稀缺性、表示限制和长期一致性等挑战依然存在。未来的工作应该加强多模态的集成，提高模型的泛化能力，开发更多用户可控的框架来推进基于人工智能的音乐创作。

Keywords: music generation; text-to-music generation; artificial intelligence; large lan-guage model

关键词： 音乐生成；文本到音乐的生成；人工智能；大型语言模型

1.Introduction

1. 简介

1.1. Background

背景

Music, as a “universal language”[1], bridges different cultures and historical periods, playing a significant role in expressing human emotions and creativity. Traditional music composition often relies on musicians applying their knowledge of music theory to create works using real instruments. In contrast, computers have gradually become tools for music creation, utilizing algorithms and models to replicate the composition process. This evolution has led to the emergence of music generation, a field that originally relied heavily on music theory as prior knowledge to design algorithms. However, recent advance-ments have shifted the focus from knowledge-driven approaches to data-driven methods, leveraging large datasets of musical compositions to enhance generative capabilities.

音乐，作为一种 “通用语言”[1] ，连接着不同的文化和历史时期，在表达人类情感和创造力方面扮演着重要的角色。传统的音乐创作往往依赖于音乐家运用他们的音乐理论知识，用真实的乐器进行创作。相比之下，计算机已逐渐成为音乐创作的工具，利用算法和模型来复制作曲过程。这种进化导致了音乐生成的出现，这是一个最初严重依赖音乐理论

作为先验知识来设计算法的领域。然而，最近的进展已经把重点从知识驱动的方法转移到数据驱动的方法，利用大量的音乐作品数据集来增强生成能力。

Music generation is a typically multi-modal task involving the transformation of symbols, audio, text, images, and other modalities [2]. Multi-modal task refers to the use of multiple types of data, such as text, audio, and images, in a system to enhance the generation process by integrating their complementary features. Among these, text-to-music generation is considered a cross-modal task, as it involves transforming text into music, bridging the gap between two different modalities. This task stands out as a uniquely promising area due to its ability to interpret natural language descriptions and transform them into music. Unlike other modalities, such as images or videos, text provides a more intuitive, user-friendly, and accessible medium for expressing musical intent, allowing users to articulate emotions, styles, or themes with precision and simplicity. This accessibility significantly lowers the barriers to music creation, enabling broader participation from individuals without formal musical training. Furthermore, the potential of text-to-music generation extends beyond user convenienceit offers a transformative tool for diverse applications such as music therapy, dynamic video soundtracks, and immersive experiences in the metaverse. By bridging natural language processing (NLP) with music generation, this field can redefine how music is created and experienced, making it a critical area of study. This review is thus essential for providing a comprehensive understanding of the technological advancements, challenges, and future opportunities in text-to-music generation, setting the stage for continued innovation in this emerging domain.

音乐生成是一个典型的多模态任务，涉及符号，音频，文本，图像和其他形式的转换 [2]。多模态任务是指在一个系统中使用多种类型的数据，如文本、音频和图像，通过整合它们的互补特性来加强生成过程。其中，文本到音乐的生成被认为是一个跨模态的任务，因为它涉及到将文本转换为音乐，弥合两个不同模态之间的差距。这项任务脱颖而出，作为一个独特的有前途的领域，由于其能够解释自然语言的描述，并将其转化为音乐。与图像或视频等其他形式不同，文本为表达音乐意图提供了一种更直观、用户友好和可访问的媒介，允许用户以精确和简单的方式表达情感、风格或主题。这种可访问性显著降低了音乐创作的障碍，使没有接受过正规音乐训练的个人能够更广泛地参与进来。此外，文本到音乐的生成潜力超越了用户的便利性，它为音乐治疗、动态视频配乐和元宇宙中的沉浸式体验等多种应用提供了变革性工具。通过将自然语言处理 (NLP) 与音乐生成结合起来，这个领域可以重新定义音乐是如何创作和体验的，使其成为一个重要的研究领域。因此，本综述对于全面了解文本到音乐生成的技术进步、挑战和未来机遇至关重要，为这一新兴领域的持续创新奠定基础。

1.2.Motiuation

1.2 动机

Text-to-music generation is an emerging research area at the intersection of artificial intelligence, music generation, and natural language processing. While existing reviews have extensively explored general music generation, they have primarily focused on traditional composition tasks, single-modality generation, or deep learning-based music synthesis, often overlooking the unique cross-modal challenges of text-to-music generation. This gap is particularly significant given the increasing integration of large language models (LLMs) in creative AI, which has opened new possibilities for translating textual descriptions into structured, meaningful, and emotionally resonant musical compositions. Despite the transformative potential of LLMs in aligning textual inputs with complex musical outputs, their role in text-to-music generation remains underexplored, highlighting the need for a comprehensive review of this rapidly growing field.

文本到音乐的生成是人工智能、音乐生成和自然语言处理交叉的一个新兴研究领域。虽然现有的综述广泛地探讨了一般的音乐生成，但它们主要集中在传统的作曲任务、单一模态生成或基于深度学习的音乐合成，往往忽略了文本到音乐的生成这一独特的跨模态挑战。鉴于大型语言模型 (llm) 在创造性人工智能中的日益整合，这一差距显得尤为重要，这为将文本描述翻译成结构化的、有意义的、能引起情感共鸣的音乐作品提供了新的可能性。尽管 llm 在将文本输入与复杂的音乐输出结合起来方面具有变革潜力，但它们在文本到音乐的生成方面的作用仍未得到充分探索，突出表明需要对这一快速增长的领域进行全面审查。

Existing reviews have extensively explored the broader landscape of music generation tasks, synthesizing representational levels, compositional processes, and single-modality tasks. Several surveys have comprehensively reviewed the broader field of music gen-eration,offering valuable insights into its methodologies and applications. A significant portion of these studies emphasizes the role of deep learning. For instance, Ji et al. (2020)[1] provide an overview of various compositional tasks at different levels of music generation, while Briot et al. (2020)[3] explore deep learning cases from perspectives such as musical structure, creativity, and interactivity. Another study by Ji et al. (2023)[4]delves into the applications of deep learning in symbolic music generation. Additionally, Hernandez-Olivan and Beltran (2021)[5] examine research advancements by aligning them with the stages and methods involved in the human creative process of composing music.

现有的评论已经广泛地探索了音乐生成任务的更广泛的景观，综合了表现水平，作曲过程和单一模态任务。一些综述全面回顾了音乐生成这一更广泛的领域，为其方法论和应用提供了有价值的见解。这些研究中很大一部分强调了深度学习的作用。例如，Ji et al。(2020)[1] 提供了音乐生成不同层次的各种作曲任务的概述，而 Briot et al。(2020)[3] 从音乐结构、创造性和交互性等角度探索深度学习案例。Ji 等人的另一项研究 (2023)[4] 深入研究了深度学习在符号音乐生成中的应用。此外，Hernandez-Olivan 和 Beltran (2021)[5] 通过将研究进展与人类创作音乐过程中所涉及的阶段和方法相结合来检验研究进展。

Other reviews have examined the field from alternative perspectives. Civit et al. (2022)[6]employed bibliometric methods to analyze the development of artificial intel-ligence in music generation. Herremans et al. (2017) [7] categorized music generation systems based on their functionality. Zhu et al. (2023) [8] introduced various tools for music generation, and Wen and Ting (2023) [9] discussed the evolution of computational intelligence techniques in this domain. Ma et al.(2024) [2] give a comprehensive survey on foundation models for music. These comprehensive reviews highlight the diverse approaches and significant progress in music generation.

其他的评论已经从不同的角度研究了这个领域。Civit 等人。(2022)[6] 采用文献计量学方法分析音乐生成中人工智能的发展。Herremans 等人。(2017)[7] 根据其功能对音乐生成系统进行分类。朱等人。(2023)[8] 介绍了各种音乐生成工具，文和婷 (2023)[9] 讨论了计算智能技术在这个领域的演变。马等人。(2024)[2] 对音乐的基础模型进行了全面的调查。这些综合评论强调了音乐生成的多样化方法和重大进展。

However, the unique challenges and opportunities of cross-modal text-to-music gener-ation remain underexplored. Most existing reviews focus on broader music generation tasks or single-modal approaches, often classifying studies based on network architectures or technical methodologies. This makes it difficult for researchers to gain a precise understand-ing of specific generation tasks, such as generating melodies from lyrics. Furthermore, the transformative potential of LLMs in text-to-music generation has been largely overlooked.

然而，跨模态文本到音乐生成的独特挑战和机遇仍然被低估。大多数现有的综述聚焦于更广泛的音乐生成任务或单模态方法，通常基于网络架构或技术方法对研究进行分类。这使得研究人员很难精确理解具体的生成任务，例如从歌词中生成旋律。此外，llm 在文本到音乐生成中的变革潜力在很大程度上被忽视了。

To the best of our knowledge, this is the first comprehensive review of the text-to-music generation task, offering an introduction to both traditional methods and

LLM-based approaches. Table 1 compares our review with previous reviews on music generation, highlighting the differences in focus, methods, and future directions discussed in each work. This paper aims to fill existing gaps by focusing on the integration of LLMs and the unique challenges of cross-modal generation.

据我们所知，这是对文本到音乐生成任务的第一次全面回顾，介绍了传统方法和基于 llm 的方法。表 1 将我们的评论与以前的音乐生成评论进行了比较，突出了每个工作中讨论的重点，方法和未来方向的差异。本文旨在通过关注 llm 的整合和跨模态生成的独特挑战来填补现有的空白。

Table 1. Comparison of Our Review with Previous Works.

表 1。我们的评论与前人的作品的比较。

Paper 论文	Keywords 关键词	Focus on 关注 Cross-Modal 跨模态 Text-to-Text-to - 文本到 - Music 音乐	Classification 分类 Based on 基于 Symbolic& 象征性的 & Audio 音频	Task-任务 Oriented 面向 Generation 一代	Inclusion of 列入 Kinds of 种类 Methods 方法	General Text-通用文本 to-Music To - 音乐 Framework 框架	Challenges 挑战 and Future 未来 Directions 方向	
Multi-level 多层次	Deep Learning 深度学习							

[1] [1]	Representations 陈述		×			Methods 方法	√	
[2] [2]	Foundation 基金会	×			LLMs Methods LLMs 方法		√	
Model 模型	Deep Learning 深度学习 Methods 方法							
[3] [3]	Deep Learning 深度学习	Deep Learning, 深度学习，	×	×	Only 只是	×		√
[4] [4]	Symbolic Music 象征性音乐		Symbolic 象征性音乐		Deep Learning 深度学习 Methods 方法	X	√	
			Deep	X	X			

			Learning 深度学习					
[5] [5]	Deep Learning 深度学习	Methods 方法						
[6] [6]	Scoping review 范围审查	×	X	×	×	Traditional 传统	X	×
[7] [7]	Functional 功能性 Taxonomy 分类法	Methods and 方法和						
		X	X		Deep Learning 深度学习 Methods 方法	X	√	
Neural Network 神经网								

络								
[8] [8]	Tools and 工具和 Models 模型	×			and Non-Neural 和非神经模型			
					Network 网络 Methods 方法	×		
[9] [9]	Evolutionary 进化 Computation 计算 and Genetic 遗传学 Algorithm 算法		X		EC and GA 欧共体和共同体 Methods 方法		√	
[10] [10]	Deep Learning 深度学				Methods 方法	Neural Network		√

	习				Both Traditional, 传统方法， Neural Network 神经网络 and LLMs 和法学硕士 Methods 方法	神经网络		
Our 我们的 Review 回顾	Text to music 文本到音乐	√	√	√	√	√		

1.3. Objectives

1.3 目标

This paper aims to provide a comprehensive and task-oriented review of advance-ments in text-to-music generation, addressing key gaps in the field and proposing actionable insights for future research. The main objectives of this work are as follows:

本文旨在提供一个全面和面向任务的文本到音乐生成的研究进展综述，指出该领域的关键差距，并为未来的研究提出可行的见解。本工作的主要目标如下：

To systematically classify and analyze text-to-music generation tasks: By categorizing tasks into symbolic and audio domains, this paper examines subtasks such as melody generation, polyphony generation, singing voice synthesis, and complete song composition. This taxonomy offers a clear aspect for understanding the distinct challenges and opportunities within each domain. This framework supports modular method development by providing researchers with a structured reference for locating domain-specific innovations;

为了对文本到音乐的生成任务进行系统的分类和分析，本文通过将任务分为符号域和音频域，分别考察了旋律生成、复调生成、歌声合成和完整歌曲创作等子任务。这种分类为理解每个领域的不同挑战和机遇提供了一个清晰的视角。该框架通过为研究人员提供定位领域特定创新的结构化参考，支持模块化方法开发；

To emphasize the potential of LLMs through framework comparison: This study focuses on traditional methods, hybrid approaches, and end-to-end LLM systems, providing a detailed analysis of their strengths, limitations, and applicability. The analysis highlights the progressive improvements introduced by LLMs, demonstrating their ability to enhance user controllability, generalization capability, etc., offering a clearer perspective on the role of LLMs in advancing AI-enabled music composition;

通过框架比较强调 LLM 的潜力： 本研究聚焦于传统方法、混合方法和端到端 LLM 系统，详细分析了它们的优势、局限性和适用性。分析突出了 llm 引入的渐进改进，展示了其增强用户可控性，泛化能力等的能力，为 llm 在推进人工智能音乐创作中的作用提供了更清晰的视角；

To identify challenges and propose future directions: This objective is crucial because addressing unresolved challenges-such as data scarcity, model generalization, emo-tion modeling, and user interactivityis the foundation for advancing text-to-music generation. By systematically analyzing these barriers, this paper provides a roadmap for overcoming limitations that currently hinder the effectiveness and creativity of such systems. This exploration advances text-to-music generation, establishing it as a key direction for creative industries.

识别挑战并提出未来方向： 这一目标至关重要，因为解决尚未解决的挑战 —— 如数据稀缺性、模型泛化、情感建模和用户交互性 —— 是推进文本到音乐生成的基础。通过系

统地分析这些障碍，本文提供了一个路线图，以克服目前阻碍这类系统的有效性和创造性的局限性。这种探索推进了文本到音乐的生成，使其成为创意产业的一个关键方向。

This paper is organized as follows. Section 2 provides an overview of the evolution of text-to-music generation, tracing its development from rule-based systems to the integration of LLMs. Section 3 discusses the representation forms of text and music, as well as their roles in aligning textual semantics with musical outputs. Section 4 critically reviews text-to-music generation methods, categorizing them into symbolic domain methods and audio domain methods and providing a comparative analysis of existing techniques. Section 5 sorts out three mainstream research frameworks based on the LLMs integration, highlighting the potential of LLMs in enhancing end-to-end generation and multi-modal integration. Section 6 outlines the challenges and future directions and identifies unresolved issues such as data scarcity, emotion modeling, and interactive systems. Finally, Section 7 concludes this paper by summarizing key insights and proposing actionable recommendations for advancing research in this emerging field.

本文组织如下。第二部分概述了文本到音乐生成的发展，从基于规则的系统到 llm 的集成。第三节讨论了文本和音乐的表现形式，以及它们在文本语义与音乐输出对齐中的作用。第 4 节批判性地回顾了文本到音乐的生成方法，将其分为符号域方法和音频域方法，并对现有的技术进行了比较分析。第 5 节梳理了基于 llm 整合的三个主流研究框架，突出了 llm 在加强端到端生成和多模态整合方面的潜力。第六节概述了挑战和未来方向，并指出了尚未解决的问题，如数据稀缺性、情感建模和交互系统。最后，第 7 节总结了本文的主要见解，并提出了推进这一新兴领域研究的可行性建议。

2. Evolution

2、进化

2.1. Early Rule-Based Systems

2.1 早期基于规则的系统

Music generation research dates back to the mid-20th century, initially focusing on using programming languages and mathematical algorithms to simulate the process of music creation. The first computer-generated music was born in 1957 through sound synthesis software developed by Bell Labs [10].

音乐生成的研究可以追溯到 20 世纪中叶，最初集中于使用编程语言和数学算法来模拟音乐创作的过程。第一个计算机生成的音乐诞生于 1957 年，通过贝尔实验室开发的声音合成软件 [10]。

Early works such as Iannis Xenakis' use of probability theory [11], as well as Lejaren Hiller and Leonard Isaacson's work Illiac Suite [12]一 which employed rule-based methodsmarked the birth of automated music generation. These efforts were primarily concerned with generating melodies, harmonic progressions, and rhythmic patterns. In addition to rule-based templates and probabilistic methods, some notable studies also em-ployed other approaches, such as the genetic algorithm-based GenJam [13], which creates jazz compositions by providing a given chord progression, and Cope's "Experiments in Musical Intelligence"(EMI) [14] project, which used search agents to generate numerous pieces in the styles of classical composers.

早期的工作，如 Iannis Xenakis 的概率论的使用 [11] ，以及 Le-jaren Hiller 和 Leonard Isaacson 的工作 Illiac Suite [12] 一采用基于规则的方法标志着自动音乐生成的诞生。这些努力主要涉及生成旋律，和声进程和节奏模式。除了基于规则的模板和概率方法之外，一些值得注意的研究还采用了其他方法，例如基于遗传算法的 GenJam [13] ，它通过提供给定的和弦进程来创建爵士乐作品，以及 Cope 的 "音乐智能实验"(EMI)[14] 项目，使用搜索代理以古典作曲家的风格生成许多作品。

Early approaches to text-to-music generation also relied heavily on predefined rules and templates to create music. These methods included lyric-based melody generation, where algorithms analyzed the content of lyrics to produce melodies that aligned with their emotional and rhythmic structures. For example, systems mapped syllables to notes based on rhythmic patterns and harmonic rules [15,16]. Additionally, textual instruction sequences, often based on music theory, guided melody generation. Such systems translated harmonic progressions (e.g.,I-IV-V-I) [17,18] and other theoretical constructs into melodies by algorithmically processing these instructions.

文本到音乐生成的早期方法也严重依赖于预定义的规则和模板来创建音乐。这些方法包括基于歌词的旋律生成，其中算法分析歌词的内容，以产生符合其情感和节奏结构的旋律。例如，系统根据节奏模式和和声规则将音节映射到音符 [15,16]。此外，通常基于音乐理论的文本指令序列指导旋律生成。这样的系统通过算法处理这些指令，将和声级数(例如，i-iv-v-i)[17,18] 和其他理论结构转换成旋律。

While these methods ensured compliance with musical structures, they were limited by their reliance on rigid rules and templates. The reliance on preprogrammed rules meant that these systems could not adapt to the vast diversity of musical styles and human creativity. They often resulted in outputs that were predictable and repetitive, lacking the emotional depth and originality that human composers can bring to music. This lack of flexibility in generating diverse musical ideas was one of the key limitations that motivated the adoption of machine learning techniques.

虽然这些方法确保了与音乐结构的一致性，但它们受到严格规则和模板的限制。对预编程序规则的依赖意味着这些系统不能适应广泛多样的音乐风格和人类的创造力。它们往往导致输出的结果是可预测的和重复的，缺乏人类作曲家所能带给音乐的情感深度和原创性。在产生多样化的音乐创意方面缺乏灵活性是促使人们采用机器学习技术的关键限制因素之一。

2.2. Emergence of Machine Learning

2.2 机器学习的出现

The late 20th and early 21st centuries brought significant shifts with the introduction of machine learning techniques into music generation. In the machine learning era, traditional techniques for music generation focus on learning patterns and structures from large datasets of existing compositions. Early methods often employed Hidden Markov Models(HMMs), which excel at modeling sequential data by capturing probabilistic transitions between states [19]. HMMs were used to generate melodies or harmonies by determining the likelihood of note sequences, though their capacity to handle complex musical structures was limited by their reliance on fixed state-transition probabilities.

20 世纪末 21 世纪初，随着机器学习技术引入音乐生成，带来了重大转变。在机器学习时代，传统的音乐生成技术侧重于从现有作品的大型数据集中学习模式和结构。早期的方法通常采用隐马尔可夫模型 (hmm) ，它擅长通过捕获状态之间的概率转换来建模顺序数据 [19]。Hmm 通过确定音符序列的可能性来生成旋律或和声，尽管它们处理复杂音乐结构的能力受限于它们对固定状态转换概率的依赖。

Building on these early approaches, more advanced models such as Recurrent Neural Networks (RNNs) [20,21] are suited for sequential data like music. These networks generate melodies or chord progressions by predicting the next note based on prior information. An improvement to RNNs, Long Short-Term Memory (LSTM) networks [22] address the challenge of remembering long-term dependencies, allowing for more coherent and extended music sequences.

在这些早期方法的基础上，更先进的模型，如循环神经网络 (RNNs)[20,21] ，适合于连续数据，如音乐。这些网络通过基于先验信息预测下一个音符来产生旋律或和弦进行。对 rnn 的一个改进，长短期记忆 (LSTM) 网络 [22] 解决了记忆长期依赖性的挑战，允许更连贯和扩展的音乐序列。

The same techniques began being applied to text-to-music generation, where systems began linking textual data with musical outputs. In text-to-music generation, this shift enabled the field to move beyond lyric-to-melody mapping. Researchers began exploring models that not only mapped text to melody but also incorporated additional musical elements such as harmony, accompaniment, and vocals.

同样的技术开始应用于文本到音乐的生成，系统开始将文本数据与音乐输出连接起来。在文本到音乐的生成中，这种转变使得这个领域超越了歌词到旋律的映射。研究人员开始探索模型，不仅映射到旋律的文字，而且纳入额外的音乐元素，如和声，伴奏，和声。

The shift to machine learning techniques allowed text-to-music generation to evolve beyond simple lyric-to-melody mapping. Researchers began exploring models that not only mapped text to melody but also incorporated additional musical elements, such as harmony, accompaniment, and vocals. However, these early ML models still had limitations. They were heavily dependent on the patterns present in the training data, and while they generated music with greater flexibility and complexity than rule-based systems, the outputs often lacked true creativity and innovation. This highlighted the need for further advancements in deep learning and cross-modal methods, as these approaches offer the potential to generate more expressive and contextually rich music by bridging multiple modalities.

向机器学习技术的转变使得文本到音乐的生成超越了简单的歌词到旋律的映射。研究人员开始探索不仅将文本映射到旋律，还包括其他音乐元素，如和声、伴奏和人声的模型。然而，这些早期的机器学习模型仍然有其局限性。它们在很大程度上依赖于训练数据中的模式，虽然它们产生的音乐比基于规则的系统具有更大的灵活性和复杂性，但产出往往缺乏真正的创造力和创新。这突出表明需要在深度学习和跨模态方法方面取得进一步进展，因为这些方法有可能通过连接多种模态而产生更具表现力和背景丰富的音乐。

2.3. The Rise of Deep Learning and Cross-Modal Approaches

2.3 深度学习和跨模态方法的兴起

With the advent of deep learning, the capabilities of general music and text-to-music generation greatly expanded. The fundamental difference between machine learning and deep learning lies in their ability to learn hierarchical features, enabling the generation of more sophisticated and creative musical outputs. Deep learning models such as Generative Adversarial Networks (GANs) [23] and Transformers [24] began to offer more realistic and diverse music compositions by capturing complex dependencies in both symbolic music and raw audio data.

随着深度学习的出现，普通音乐和文本到音乐的生成能力大大扩展。机器学习和深度学习的根本区别在于它们学习分层特征的能力，使得生成更加复杂和创造性的音乐输出成为可能。生成对抗网络 (GANs)[23] 和变形金刚 (Transformers)[24] 等深度学习模型开始通过捕获符号音乐和原始音频数据中的复杂依赖关系来提供更真实和多样化的音乐作品。

The emergence of text-to-audio models has opened a new direction for music gen-eration. Models like AudioLM [25] and Suno’s bark https://github.com/suno-ai/bark(accessed on 17 February 2025) combine audio representation with text representation, allowing them to understand textual content and generate corresponding audio. Building on these innovations, researchers began developing more comprehensive text-to-music generation models, which go beyond simple lyric-to-melody mappings. These models now aim to capture emotion, themes, and other non-musical elements from texts to guide the music generation process.

文本到音频模型的出现为音乐生成开辟了新的方向。AudioLM [25] 和 Suno 的 bark https://github.com/Suno-ai/bark (于 2025 年 2 月 17 日访问) 等模型将音频表示与文本表示相结合，使其能够理解文本内容并生成相应的音频。在这些创新的基础上，研究人员开始开发更全面的文本到音乐的生成模型，超越了简单的歌词到旋律的映射。这些模型现在的目标是从文本中捕捉情感、主题和其他非音乐元素来指导音乐生成过程。

The success of diffusion models in image generation tasks [26] has led researchers to apply these models to music generation [27]. This approach has proven effective in creating richer, more expressive outputs and has laid a strong foundation for the continued development of contemporary text-to-music generation techniques.

扩散模型在图像生成任务中的成功 [26] 导致研究人员将这些模型应用于音乐生成 [27]。这种方法已被证明在创造更丰富、更具表现力的输出方面是有效的，并为当代文本 - 音乐生成技术的持续发展奠定了坚实的基础。

While deep learning models significantly improved music generation, they still relied on manually designed architectures and required extensive task-specific training. As the demand for more flexible and generalizable systems grew, researchers shifted toward large-scale pre-trained models(especially LLMs) that leverage vast amounts of data to learn universal representations of music and language.

虽然深度学习模型显着改善了音乐生成，但它们仍然依赖于手动设计的架构，并需要广泛的特定任务培训。随着对更加灵活和普遍化的系统需求的增长，研究人员转向大规模的预训练模型 (尤其是 llm) ，这些模型利用大量的数据来学习音乐和语言的普遍表征。

2.4. The Integration of LLMs

2.4 llm 的整合

Recent breakthroughs in music generation have been driven by multi-modal and cross-modal learning techniques, which integrate various data types such as text, audio, and symbolic representations. These models utilize advanced deep learning frameworks to capture the intricate relationships between these diverse data types, enabling the gener-ation of music that is not only structurally coherent but also emotionally expressive and contextually rich.

多模态和跨模态学习技术整合了文本、音频和符号表示等多种数据类型，推动了音乐生成领域的最新突破。这些模型利用先进的深度学习框架来捕捉这些不同数据类型之间的复杂关系，使得生成的音乐不仅在结构上连贯一致，而且情感表达和上下文丰富。

In parallel, advancements in large-scale models, particularly LLMs, have paved the way for end-to-end text-to-music generation. These models, trained on vast datasets of text and music, are capable of directly mapping textual descriptions to musical outputs. For example, modern LLMs-based systems [28] can interpret detailed textual prompts, including emotional expressions, scene descriptions, or stylistic preferences, and generate highly consistent compositions in style, rhythm, and harmony. This end-to-end paradigm significantly lowers the barrier to music creation, allowing users without formal music training to create complex and expressive musical works. Furthermore, the adaptabil-ity of these models opens new possibilities for personalized music creation, soundtrack generation, and other multi-modal applications.

与此同时，大规模模型的进步，特别是 llm，为端到端的文本到音乐的生成铺平了道路。这些模型在大量的文本和音乐数据集上进行训练，能够直接将文本描述映射到音乐输出。例如，现代的基于 llms 的系统 [28] 可以解释详细的文本提示，包括情感表达，场景描述，或风格偏好，并生成高度一致的风格，节奏，和谐的作品。这种端到端的范式显着降低了音乐创作的障碍，允许没有正式音乐培训的用户创建复杂和表现力强的音乐作品。此外，这些模型的适应性为个性化音乐创作、原声带生成和其他多模态应用开辟了新的可能性。

The data presented in Figure 1 were sourced from Web of Science, where we collected publications related to text-to-music generation until 2024. It is important to note that due to the variability in terminology and the way related research is categorized, some articles may not be captured by the search criteria. However, this dataset still provides valuable insights into the general trends within the field. The significant increase in publications related to neural networks and deep learning since 2019, along with the rise in LLM integration in 2023 and 2024, clearly indicates the growing focus on data-driven approaches in text-to-music generation. This trend highlights the shift toward more complex and contextually nuanced systems that integrate natural language processing, deep learning, and multi-modal techniques to enhance music generation capabilities.

图 1 中显示的数据来源于 Web of Science，我们在那里收集了与文本到音乐生成相关的出版物，直到 2024 年。值得注意的是，由于术语的可变性和相关研究的分类方式，有些文章可能无法通过搜索标准获取。然而，这个数据集仍然为该领域的一般趋势提供了宝贵的见解。自 2019 年以来，与神经网络和深度学习有关的出版物显着增加，以及 2023 年和 2024 年 LLM 集成的增加，清楚地表明在文本到音乐的生成中越来越关注数据驱动的方法。这一趋势突出了向更复杂和上下文细微差别的系统的转变，这些系统整合了自然语言处理、深度学习和多模态技术，以增强音乐生成能力。

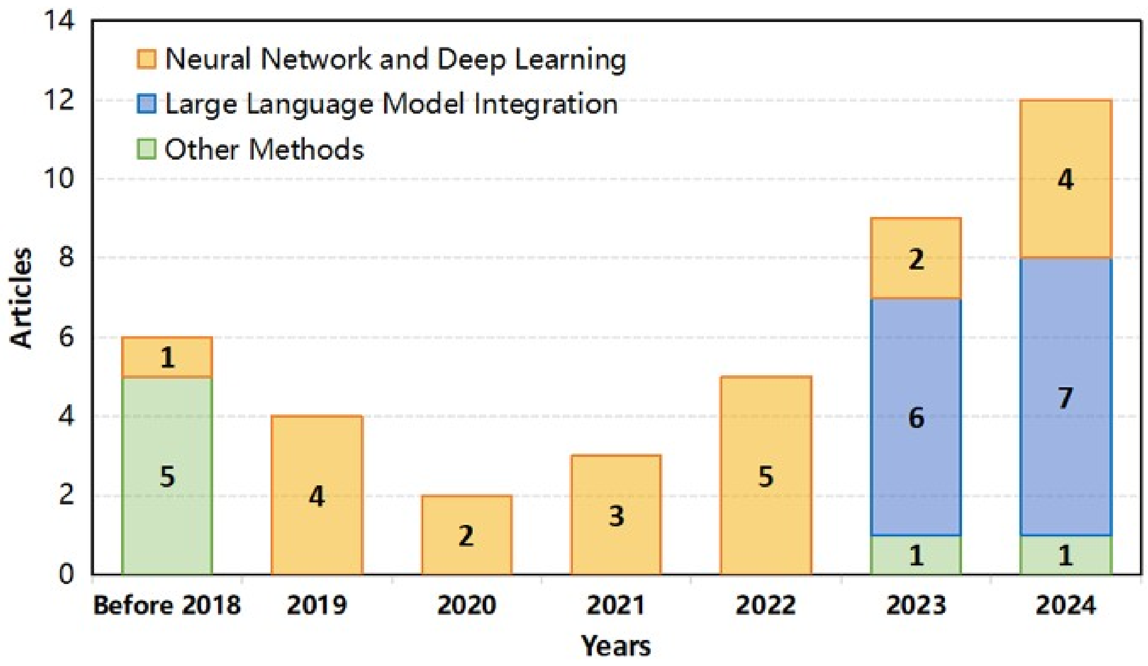


Figure 1. Research Trends in Text-to-Music Generation.

图 1。文本到音乐生成的研究趋势。

3. Representation Forms of Text and Music

文本和音乐的表现形式

3.1. Text Types

3.1 文本类型

In a text-to-music generation, the common text types are mainly categorized into three types, which serve as input for the generation process and provide guidance for music creation. The common text types and their roles in text-to-music generation are shown in Table 2.

在文本到音乐的生成过程中，常见的文本类型主要分为三类，作为生成过程的输入，为音乐创作提供指导。常见文本类型及其在文本 - 音乐生成中的作用如表 2 所示。

Table 2. Text types and their characteristics.

表 2。文本类型及其特征。

Category 类别	Description 描述	Application 应用	Generation 一代	Challenges 挑战			
		Example 示例	Characteristics 特点				
Category Description 类别简介	tionXXX 项目名称： Example 示例	ticsLyrics-ticsLyrics-tics 抒情诗	Challenges 挑战				
melod	Multi-						

y 梅洛迪	多						
matching 匹配	language 语言						
The singing 唱歌	“Let it be, “随它去吧，	ing 英格	Singing voice 歌声	Cultural 文化			
Lyrics 歌词 Lyrics 歌词	words of 歌词 songs. 歌曲。	let it be…’ 随它去吧..	synthesis 合成 Emotion-and 情感 —— 还有	context 上下文			
		let it be…” 就这样吧… …”		sis 姐姐	Appropriate 合适		
songs. 歌曲。	Rhythm- 节奏 - based 基于		rhythm 节奏				
		Musical 音乐剧 Musical 音乐剧 Attributes 属性		Describes 描述	Music 音乐	Complex 复杂	
theory-based 以理论为基础							
Describes 描述	bpm 每分钟一次		I-IV-V-I,120 bpm I-IV-V-	Using 使用	music theory 音乐理		

musical rules 音乐规则			I，120 bpm		论 Lacking of 缺乏		
utes 几分钟		like chords. 像和弦。	plates 盘子	attribute 属性	creativity 创造力		
Natu- 纳图	templates 模板						
Flexible 灵活性	Abstract 摘要						
Natural 自然	Describes 描述	“Create a 创建一个	description 描述	concepts 概念			
Language 语言	emotion or 情绪或	melody filled 充满旋律	Diverse 多样化	understanding 理解			
Description 描述 X	scene. 场景。	scene. 场景。	hope…” 希望…”	with hope…” 带着希望	tures. 图雷斯。	music 音乐 features. 特点。	Converting 转换 consistency 一致性

Descrip- 描述							

tion

提议

3.2. Musical Representation

3.2 音乐表现

3.2.1. Event Representation: MIDI-like

3.2 事件表现： 类 midi

MIDI (Musical Instrument Digital Interface, https://en.wikipedia.org/wiki/MIDI, accessed on 17 February 2025) is an industry-standard protocol for communicating between electronic musical instruments and exchanging data between an instrument and a computer. MIDI files record information about a player’s actions, such as which key was pressed, how hard it was pressed, and how long it lasted. These messages are called “events”, which are binary data such as Note On, Velocity, Note Of, Aftertouch, Pitch Bend, etc. Table 3 lists the common events used in symbolic music generation research. Note On, Channel 1, Pitch 6

MIDI (乐器数字接口，https://en.wikipedia.org/wiki/MIDI，于 2025 年 2 月 17 日访问) 是用于电子乐器之间通信和乐器与计算机之间交换数据的行业标准协议。MIDI 文件记录了演奏者的动作信息，例如按下了哪个键，按下的力度有多大，持续了多长时间。这些信息被称为 “事件”，是二进制数据，如 Note On，Velocity，Note Of，Aftertouch，Pitch Bend 等。表 3 列出了在符号音乐生成研究中使用的常见事件。注释 On，Channel 1，Pitch 6

Note On Starts a note

开始一个笔记

Table 3. Common events in music generation research.

表 3: 音乐生成研究中的常见事件。

Velocity 100

速度 100

Event Type 事件类型	Ends a note 结束笔记 Description 描述	Example Format 示例格式
Note On 注释	Starts a note 开始一个笔记	Note On, Channel 1, Pitch 60, 注意，第一频道，音高 60, Velocity 100 速度 100
Note Off Note Off 注意	Ends a note 结束笔记	Note Off, Channel 1, Pitch 60, 音符关闭，一频道，音高 60, Velocity 0 速度 0
Program Change 程序变化	Changes in instrument or sound 乐器或声音的变化	Program Change, Channel 1, 第一频道，节目更改， Program 32 节目 32
Control Change Control Change 控制变更	Adjusts control parameters (e.g., 调整控制参数 (例如， volume, sustain pedal) 音量，保持踏板)	Control Change, Channel 1, 控制改变，频道 1, Controller 64,Value 127 控制器 64，值 127
Pitch Bend 俯仰弯曲	Bends pitch slightly or continuously 使沥青轻微或连续弯曲	Pitch Bend, Channel 1, Value 8192 音高弯曲，频道 1，值

		8192
Aftertouch 余音	Pressure applied after pressing a note 按下音符后施加的压力	Aftertouch, Channel 1, Pressure 60 Aftertouch，Channel 1，Pressure 60 后触，第一频道，压力 60
Tempo Change 节奏改变	Sets playback speed in beats per 设置每拍播放速度 minute(BPM) 每分钟 (BPM)	Tempo Change, 120 BPM Tempo Change，120 BPM 节奏变化，每分钟 120 次
Time Signature 时间信号	Defines beat structure (e.g.,4/4, 定义拍结构 (例如，4/4, 3/4 time) 3/4 次)	Time Signature, 4/4 Time Signature，4/4 时间签名，4/4
Key Signature 密钥签名	Sets the song’s key (e.g.,C Major, 设置歌曲的键 (例如，c 大调， GMinor) GMinor)	Key Signature, C Major Key Signature，c Major 键盘签名，c 大调

Channel

通道

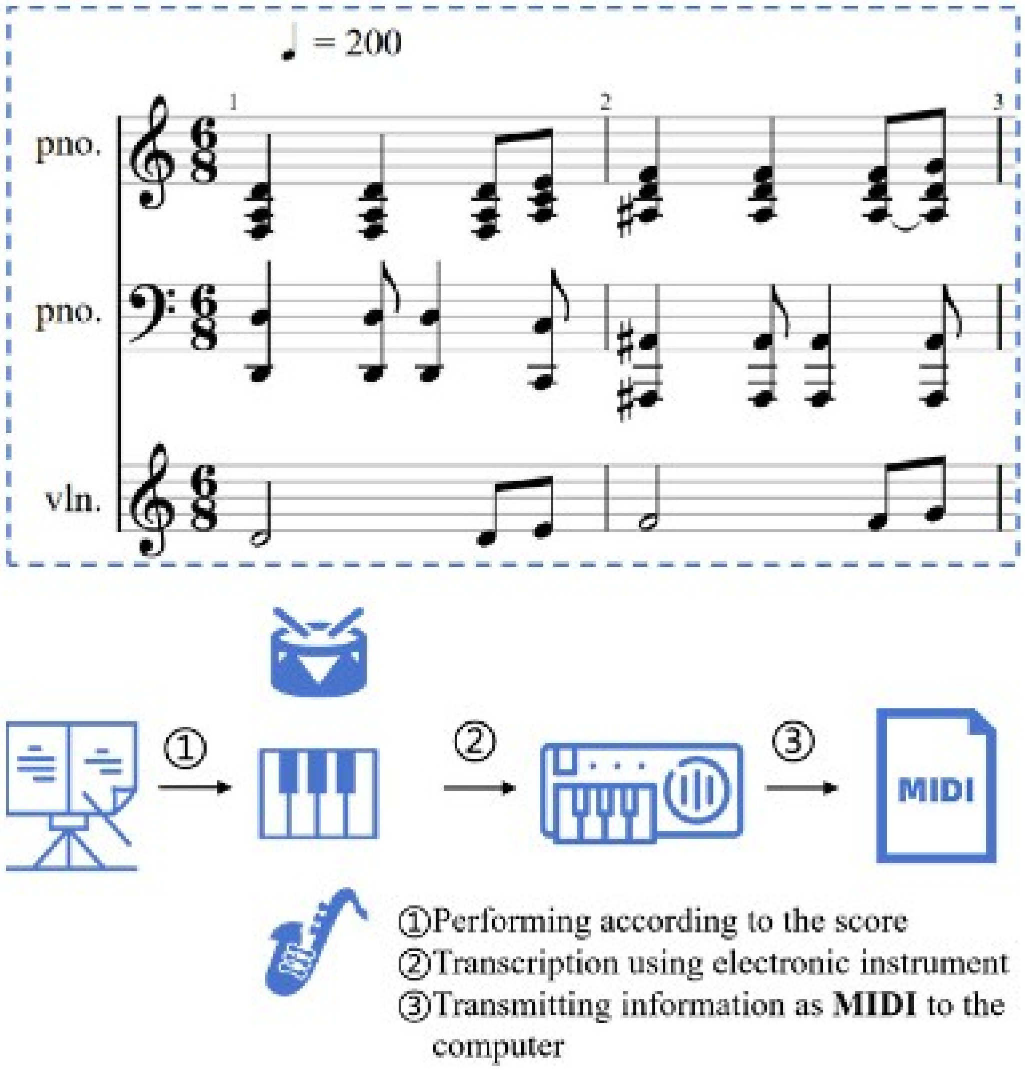
Note Off, Channel 1, Pitch 6

Note Off，Channel 1，Pitch 6 注意，第一频道，第六节

MIDI is a highly compatible and easy-to-edit file format with a small file size that facilitates communication between devices and music creation, as shown in Figure

2. In music generation research, an algorithm or model first slices a melody into sequences of notes and then establishes a mapping relationship between musical elements and numbers through quantization and encoding to obtain a data representation of the music. Native MIDI representations have representational limitations, such as the inability to express the concepts of quarter notes or rest, not being able to represent the musical onset time, etc. Therefore, some studies have improved MIDI representations for music generation by proposing REMI [29] and REMI+ to represent more information.

MIDI 是一种高度兼容且易于编辑的文件格式，其文件大小很小，便于设备之间的通信和音乐创作，如图 2 所示。在音乐生成研究中，算法或模型首先将旋律切分成音符序列，然后通过量化和编码建立音乐元素与数字之间的映射关系，得到音乐的数据表示。原生 MIDI 表征具有表征的局限性，例如不能表达四分音符或休止符的概念，不能表征音乐的起始时间等。因此，一些研究通过提出 REMI [29] 和 REMI + 来代表更多的信息，改进了音乐生成的 MIDI 表示。


♩ = 200
1
2
3
pno.
pno.
vln.
①
②
③
MIDI
①Performing according to the score
②Transcription using electronic instrument
③Transmitting information as **MIDI** to the computer

Track o: He’s a Pirate

他是个海盗

MetaMessage(‘track_name’, name=“He’s a Pirate”, time=o) Track 1: Right Hand note_on channel=o note=62 velocity=114 time=o control change channel=o control=101 value=o time=o note_off channel=o note=62 velocity=64 time=48o

音轨 1: 右手 note _ on channel = o note = 62 velocity = 114 time = o control change channel = o control = 101 value = o time = o note _ off channel = o note = 62 velocity = 64 time = 48o

+.++.+

+.++.+

Track 2: Left Hand MetaMessage(‘track_name’,name=‘Left Hand’, time=o) program_change channel=2 program=o time=o note_on channel=2 note=38 velocity=114 time=o control change channel=2 control=101 value=o time=o note off channel=2 note=38 velocity=64 time=48o+.++.+Track 3: Staff-1

Track 2: Left Hand MetaMessage (‘ Track _ name’，name = ‘ Left Hand’，time = o-rr) program _ change channel = 2 program = o time = o note _ on channel = 2 note = 38 velocity = 114 time = o control change channel = 2 control = 101 value = o time = o note off channel = 2 note = 38 velocity = 64 time = 48o + 。++.+ Track 3: staff-1 第三轨： Staff-1

MetaMessage(‘track_name’,name=‘vloin’, time=o) program_change channel=4 program=48 time=o note_on channel=4 note=62 velocity=76 time=o control change channel=4 control=101 value=o time=o note_off channel=4 note=62 velocity=64 time=96o+.+..+

译者注) program _ change channel = 4 program = 4 program = 48 time = o note _ on channel = 4 note = 62 velocity = 76 time = o control change channel = 4 control = 101 value = o time = o note _ off channel = 4 note = 62 velocity = 64 time = 96o + 。 +..+

3.2.2. Audio Representation: Waveform and Spectrogram

音频表示： 波形和声谱图

Audio representation is a continuous form, typically categorized into one-dimensional and two-dimensional forms. One-dimensional representations, usually in the time domain, are the simplest type. In this form, the audio signal is represented as a time series, often visualized as a waveform. Each point in the waveform corresponds to the amplitude value at a specific time, and the entire sequence shows how the audio signal changes over time.

音频表征是一种连续的形式，通常分为一维和二维形式。一维表征，通常在时间域，是最简单的类型。在这种形式中，音频信号被表示为一个时间序列，通常可视化为一个波形。波形中的每个点对应于特定时刻的振幅值，整个序列显示了音频信号随时间的变化。

Electronics 2025, 14, x FOR PEER REVIEW

电子学 2025,14，x 供同行评审

In contrast, two-dimensional representations, such as spectrograms, transform the audio signal from the time domain to the frequency domain. These representations break down the audio signal into various frequency components using methods like the Short-Time Fourier Transform (STFT), as shown in Figure 3.

相比之下，二维表示，如声谱图，将音频信号从时域转换到频域。这些表示使用短时距傅里叶变换 (STFT) 等方法将音频信号分解为各种频率成分，如图 3 所示。

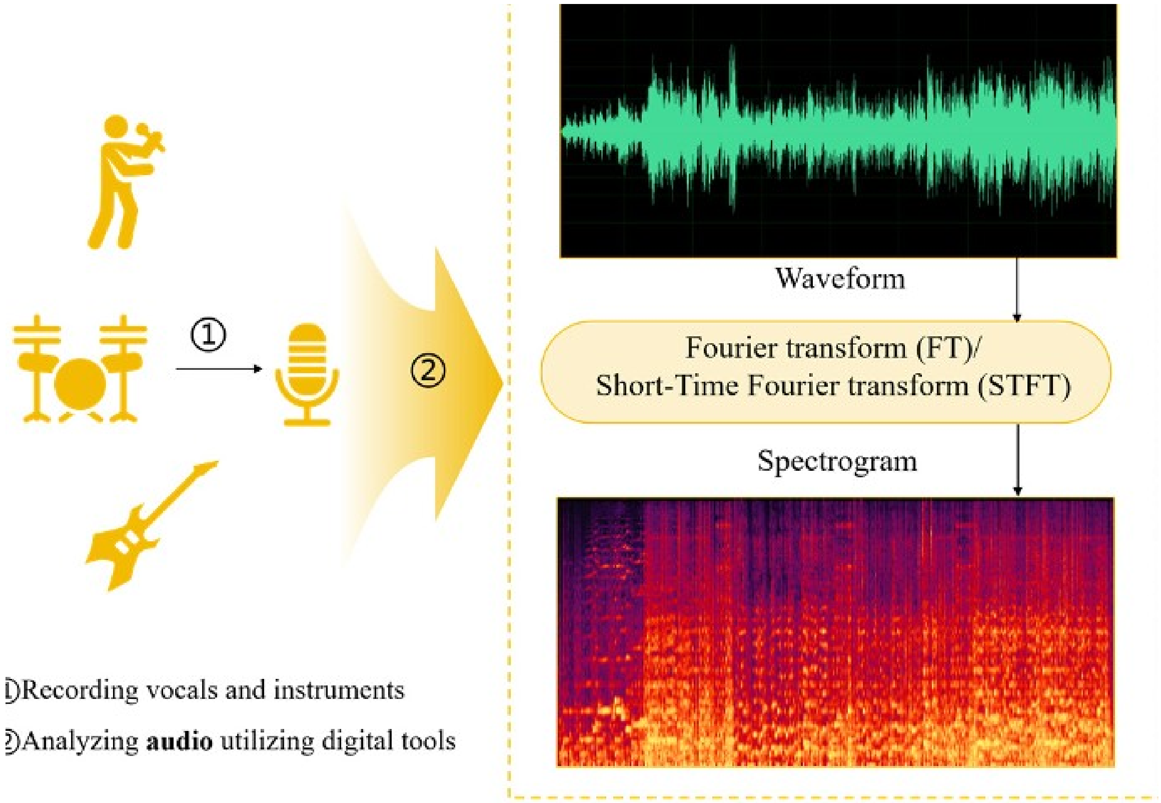

①
②
Waveform
Fourier transform (FT)/
Short-Time Fourier transform (STFT)
Spectrogram
Recording vocals and instruments
Analyzing audio utilizing digital tools

Figure 3. Waveform and Spectrogram Representation.

图 3. 波形和谱图表示。

Compared to MIDI representation, audio waveform and spectrogram retain more de-tails, enabling the style, timbre, emotion, and vocal performance to be modeled. As a result, they offer a greater advantage in creating natural and expressive musical compositions.

与 MIDI 表示法相比，音频波形和声谱图保留了更多的细节，使得风格、音色、情感和声音表演能够被建模。因此，它们在创作自然而富有表现力的音乐作品方面具有更大的优势。

tions.3.2.3. Text Representation: ABC Notation

文本表示法： *ABC* 符号

ABC notation is an ASCII-based text format for representing musichttps://abcnotation.com(accessedon17February2025),usingsimplelettersandsymbolstoencodeinformationsuchasnotes,rhythms,andkeysignatures.Itconsistsoftwoparts:theheaderfieldsandthetunebody.Theheaderfieldstypicallyincludetracknumber(X),title(T),meter(M),notelengthunit(L),tempo(Q),key(K),andothers.Thetuningbodyrepresentsthesequenceofnotes,inwhichthelettersAtoGcorrespondtomusicalnotes,andthenumbers1to8indicatepitchvariations.Forexample,“C”representstheCnote,and”C2”representsthesecondoctaveofC.Thesymbol”|“indicatesbardivisions,whilenumbersspecifythedurationofnotes.Forinstance,”C/2”indicates that the C note lasts for two eighth notes. Additional symbols are used to represent note lifts, sustains, rests, and other musical elements.

ABC 符号是一种基于 ascii 的文本格式，用于表示音乐 http://abcnotation.com (accessedon17 february2025) ，使用简单的字母和符号来编码信息，如音符、节奏和键签名。它由两部分组成： headerfields 和 tunebody。标题字段通常包括曲目编号 (x)、标题 (t)、米 (m)、注释长度单位 (l)、节奏 (q)、键 (k) 等。调音体代表音符序列，其中字母 tog 对应于音符，数字 1 到 8 表示音高变化。例如，“c” 表示 cnote，“C2” 表示 c.thesymbol”| “表示 bardivisions，而 numbers 指定 durationofnotes。例如，“c/2” 表示 c 音符持续两个第八音符。附加的符号用来表示音符的升降、停顿、休息和其他音乐元素。

After being encoded, the text files of ABC notation can extract information such as notes, rhythms, and chords. Based on the extracted information, such as note start, note end, and note strength, they can eventually be interconverted with MIDI files, as shown in Figure 4. Therefore, in this paper, we also categorize the research that generates the form of ABC notation into the symbolic domain.

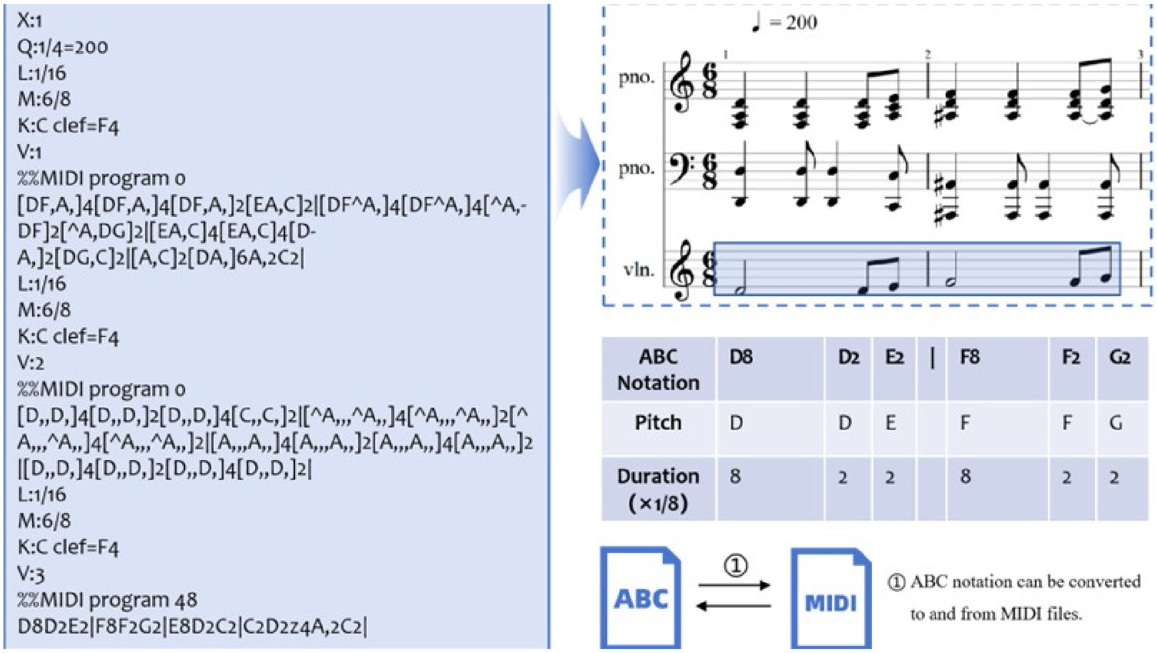
经过编码后，ABC 符号的文本文件可以提取音符、节奏和和弦等信息。基于提取的信息，例如 note start、 note end 和 note strength，它们最终可以与 MIDI 文件相互转换，如图 4 所示。因此，在本文中，我们也将生成 ABC 符号形式的研究归类到符号领域。

After being encoded, the text files of ABC notation can extract information notes, rhythms, and chords. Based on the extracted information, such as note st end, and note strength, they can eventually be interconverted with MIDI files, a in Figure 4. Therefore, in this paper, we also categorize the research that gener

经过编码后，ABC 符号的文本 files 可以提取信息音符、节奏和和弦根据提取的信息，如 note st end 和 note strength，它们最终可以与 MIDI files 相互转换，如图 4 中的 a 所示因此，在这篇文章中，我们也将这些研究归类

form of ABC notation into the symbolic domain.

ABC 符号形式的研究归类到符号领域。



```
X:1
Q:1/4=200
L:1/16
M:6/8
K:C clef=F4
V:1
%%MIDI program 0
[DF,A,]4[DF,A,]4[DF,A,]2[EA,C]2|[DF^A,]4[DF^A,]4[^A,-
DF]2[^A,DG]2|[EA,C]4[EA,C]4[D-
A,]2[DG,C]2|[A,C]2[DA,]6A,2C2|
L:1/16
M:6/8
K:C clef=F4
V:2
%%MIDI program 0
[D,,D,]4[D,,D,]2[D,,D,]4[C,,C,]2|[^A,,,^A,,]4[^A,,,^A,,]2[^
A,,,^A,,]4[^A,,,^A,,]2|[A,,,A,,]4[A,,,A,,]2[A,,,A,,]4[A,,,A,,]2
|[D,,D,]4[D,,D,]2[D,,D,]4[D,,D,]2|
L:1/16
M:6/8
K:C clef=F4
V:3
%%MIDI program 48
D8D2E2|F8F2G2|E8D2C2|C2D2z4A,2C2|
```

ABC Notation	D8	D2	E2	|	F8	F2	G2
Pitch	D	D	E		F	F	G
Duration (×1/8)	8	2	2		8	2	2

① ABC notation can be converted to and from MIDI files.

Figure 4. ABC Notation. Figure 4. ABC Notation.

图 4。 ABC 符号。图 4。 ABC 符号。

4. Methods

方法

In a text-to-music generation, methods are broadly categorized into symbolic domain and audio domain based on data representation formats, as shown in Figure 5. The textual inputs—comprising lyrics, musical instructions, and natural language descriptions—serve as semantic drivers for generation tasks. The symbolic domain, anchored in structured rep-resentations such as MIDI and ABC notation, facilitates melody generation and polyphony generation. In contrast, the audio domain operates on raw waveform and spectrogram data to achieve instrumental music synthesis, singing voice generation, and complete song com-position. The evolution of text-to-music systems reflects a clear trajectory: advancing from single-track to multi-track generation, from simplistic structures to intricate compositions, pieces. and from localized musical fragments to holistic, contextually coherent pieces.

在文本到音乐的生成中，基于数据表示格式，方法大致分为符号域和音频域，如图 5 所示。文本输入一包括歌词、音乐指令和自然语言描述一作为生成任务的语义驱动器。符号域锚定在 MIDI 和 ABC 符号等结构化表示中，有助于旋律生成和复调生成。相比之下，音频域对原始波形和声谱图数据进行操作，实现器乐合成、歌唱声音生成和完成歌曲创作。文本到音乐系统的演变反映了一个清晰的轨迹： 从单音轨到多音轨的生成，从简单的结构到复杂的作曲、作品。从局部的音乐片段到整体的，上下文连贯的片段。

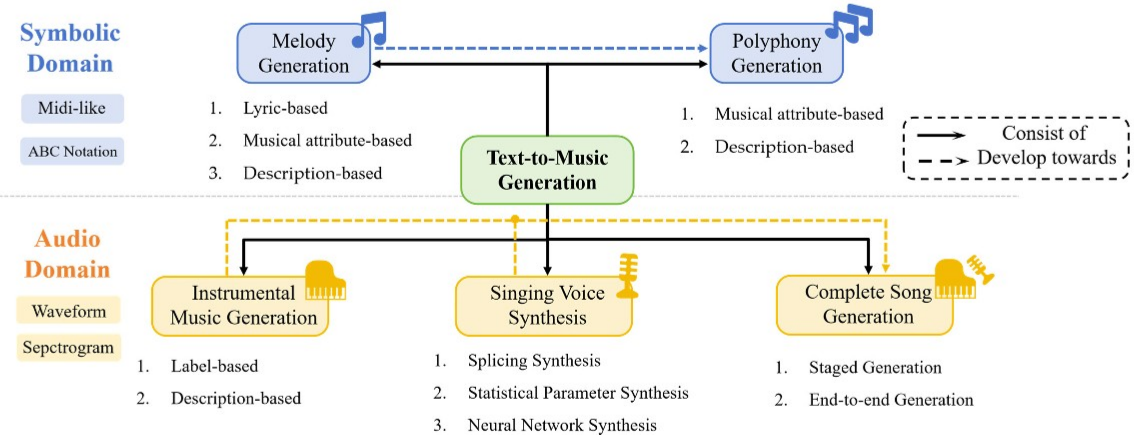


Figure 5. Overview of Text-to-Music Generation Based on Representation and Task Domains.

图 5。基于表征和任务域的文本 - 音乐生成概述。

This section categorizes existing methods based on task types (e.g. melody genera-tion, polyphony generation). To help readers understand the technical evolution of these methods, we annotate each model in the tables with its corresponding framework category(traditional learning-based, hybrid LLM-augmented, or end-to-end LLM). Section 5 will provide a detailed analysis of the commonalities and differences among these frameworks.

本节根据任务类型 (如旋律生成、复调生成) 对现有方法进行分类。为了帮助读者理解这些方法的技术演变，我们在表中用相应的框架类别 (传统的基于学习的、混合 LLM 增强的或端到端 LLM) 注释了每个模型。第五部分将详细分析这些框架之间的共性和差异。

gory(traditional learning-based, hybrid LLM-augmented, or end-to-end LLM). Section 5 works.

Gory (传统的基于学习的，混合的 LLM-augmented，或者端到端的 LLM) 第五部分是有效的。

4.1. Symbolic Domain Methods

4.1 符号域方法

Symbolic-domain music generation is the task of automatically generating symbolic music representations by using computational models. In this process, algorithms create new musical sequences with coherent and creative characteristics based on previously learned patterns or rules. Symbolic music representations usually refer to discrete musical information structures, such as MIDI files or digitized sheet music (e.g., ABC notation), which decompose music into a series of discrete time and frequency units, such as notes, rhythms, pitches, and intensities.

符号域音乐生成是利用计算模型自动生成符号域音乐表示的任务。在此过程中，算法根据先前学习到的模式或规则创建具有连贯性和创造性特征的新音乐序列。符号化音乐表征通常指离散的音乐信息结构，如 MIDI 文件或数字化乐谱 (如 ABC 符号) ，它们将音乐分解成一系列离散的时间和频率单位，如音符、节奏、音高和强度。

Figure 6 illustrates the general workflow of text-to-music generation in the symbolic domain. First, textual input (e.g., lyrics or descriptive prompts) is provided to a sequence generation model, which translates the text into discrete symbolic music notation (such as MIDI or ABC notation). These symbolic sequences, representing musical events (e.g., pitches, durations, chords), are then converted into an audio signal using synthesizers or virtual instruments, ultimately producing the audible music output.

图 6 展示了符号域中文本到音乐生成的一般工作流程。首先，文本输入 (例如，歌词或描述性提示) 提供给序列生成模型，该模型将文本转换为离散的符号音乐符号 (例如 MIDI 或 ABC 符号)。这些符号序列，代表音乐事件 (例如，音高，持续时间，和弦) ，然后转换成音频信号使用合成器或虚拟乐器，最终产生可听的音乐输出。

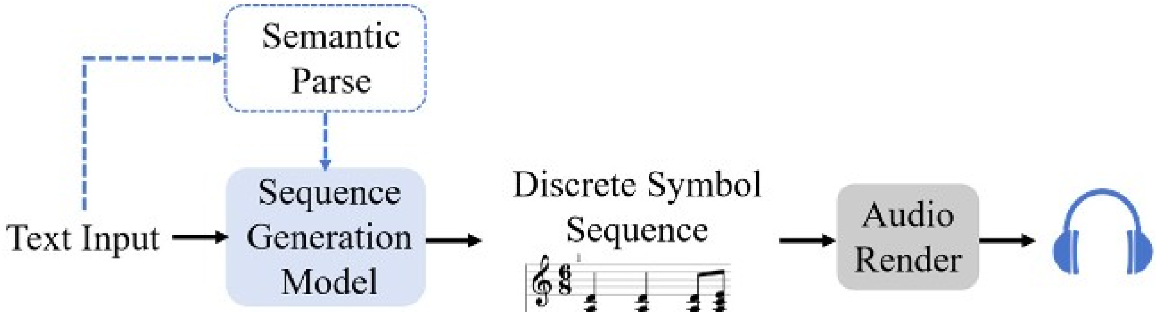


Figure 6. Workflow of Symbolic Domain Text-to-Music Generation.

图 6. 符号域文本到音乐生成的工作流程。

4.1.1. Melody Generation
旋律生成

Melody is one of the fundamental elements of music, which consists of a series of notes arranged in a specific rhythm and pitch. The melody generation task is the process of automatically generating new melodic lines through algorithms or models. This task is a core component of music generation. The melody generation task aims to create a new melody that conforms to the rules of music theory and is artistically pleasing. The text-based melody generation task is mainly divided into lyrics-based melody generation and text description-based melody generation.

旋律是音乐的基本元素之一，由一系列按照特定节奏和音高排列的音符组成。旋律生成任务是通过算法或模型自动生成新旋律线的过程。这个任务是音乐生成的核心组成部分。旋律生成任务的目的是创造出符合音乐理论规律的、艺术上令人愉悦的新旋律。基于文本的旋律生成任务主要分为基于歌词的旋律生成和基于文本描述的旋律生成。

1. Lyric-based Melody Generation

1. 基于歌词的旋律生成

The earliest attempt at lyric-based melody generation was based on rule-based systems. Fukayama et al. (2010) [15] developed an algorithm for generating melodies when specific Japanese lyrics, rhythmic patterns, and harmonic sequences were provided. The algorithm treats composition as an optimal solution search problem under the constraints of lyrics’rhymes and searches for the optimal composition through dynamic programming. In addition, the algorithm innovatively integrates text with melody, which is also considered to be the beginning of the task of generating melodies from lyrics.

基于歌词的旋律生成的最早尝试是基于规则的系统。Fukayama 等人。(2010)[15] 开发了一种算法，用于在提供特定的日本歌词，节奏模式和和声序列时生成旋律。该算法将作曲视为歌词韵律约束下的最优解搜索问题，并通过动态规划搜索最优作曲。此外，该算法创新性地将文本与旋律融为一体，也被认为是由歌词生成旋律任务的开端。

As statistical methods began to gain traction, researchers like Monteith et al. (2012) [30] moved away from rule-based systems by applying probabilistic models, such as n-gram models, to generate melodies. The system produced hundreds of rhythm and pitch com-binations for given lyrics, and the best result was selected using metrics to evaluate the generated melody. This approach shifted from strict rule-based generation to probabilistic modeling, allowing for more variety in melody generation, though it still depended heavily on predefined patterns and lacked the complexity needed for capturing the full depth of melody generation. Similarly, Scirea et al. (2015) [16]expanded this idea by constructing Markov chains over note sequences using lyric syllables, showcasing how statistical models could link lyrics with melody generation through probabilistic transitions.

随着统计方法开始获得牵引力，像 Monteith 等人 (2012)[30] 这样的研究人员通过应用概率模型 (例如 n-gram 模型) 来生成旋律，从而远离了基于规则的系统。该系统为给定的歌词产生了数百个节奏和音高组合，并且使用度量来评估生成的旋律选择最佳结果。这种方法从严格的基于规则的生成转变为概率建模，允许生成更多样化的旋律，但仍然严重依赖于预定义的模式，缺乏捕获旋律生成的全部深度所需的复杂性。同样，Scirea 等人。(2015)[16] 通过使用抒情音节在音符序列上构建马尔可夫链来扩展这个想法，展示了统计模型如何通过概率转换将歌词与旋律生成联系起来。

The next major shift came with machine learning and neural network algorithms. Ackerman et al. (2017)[31]applied random forests to predict note durations and scales, marking an early attempt at using machine learning to model melodic structures. While this approach improved the generation of rhythmic and melodic patterns, it still required handcrafted features and could not fully capture the complex relationships between lyrics and melodies. Using neural networks further explores the intrinsic connection between lyrics and melody. Bao et al. (2019) [32] developed SongWriter, a sequence-to-sequence(seq2seq) model built on RNNs, which generates melodies from lyrics while precisely aligning them. This model used two encoders: one to encode the lyrics and the other to encode the melody context. The hierarchical decoder generated the musical notes and their corresponding alignments with the lyrics. The use of seq2seq models marked a significant improvement, as they could learn complex mappings between lyrics and melodies, resulting in more cohesive and flexible melodies. This approach outperformed earlier machine learning methods, allowing for better alignment between the generated melodies and the input lyrics.

下一个重大转变来自机器学习和神经网络算法。Ackerman 等人 (2017)[31] 应用随机森林来预测音符的持续时间和音阶，标志着早期尝试使用机器学习来建模旋律结构。虽然这种方法改善了节奏和旋律模式的生成，但它仍然需要手工制作的特征，并且不能完全捕捉歌词和旋律之间的复杂关系。使用神经网络进一步探索歌词和旋律之间的内在联系。Bao 等人。(2019)[32] 开发了 SongWriter，这是一种建立在 rnn 上的序列对序列 (seq2seq) 模型，它从歌词生成旋律，同时精确地对齐它们。该模型使用两个编码器： 一个编码歌词，另一个编码旋律上下文。分级解码器生成音符及其与歌词的对应关系。Seq2seq 模型的使用标志着一个显着的改进，因为他们可以学习歌词和旋律之间的复杂映射，从而产生更有凝聚力和灵活的旋律。这种方法优于早期的机器学习方法，允许生成的旋律和输入歌词之间更好的对齐。

Building on the success of RNN-based models like SongWriter, Long Short-Term Memory (LSTM) networks have been used in music generation due to their ability to capture long-term dependencies in sequential data. Unlike standard RNNs, LSTMs address the vanishing gradient problem, making them particularly effective for modeling the complex temporal structures inherent in music. In parallel, GANs have emerged as a powerful framework for music generation, particularly in creating high-quality and diverse musical outputs. The research combining the above two structures has become a hot topic. Yu et al. (2021) [33] used a conditional LSTM-GAN for the lyrics of generation-based melodies. They combined syllable embedding vectors converted from text lyrics with noise vectors and input them into the generator. A deep discriminator was also trained to distinguish the generated MIDI note sequences from the real ones (Figure 7). This approach demonstrates both LSTM’s ability to capture long-term dependencies of melodies and GAN’s advantage in enhancing the realism and naturalness of melodies through an adversarial learning mechanism. To further improve generation quality, a three-branch conditional LSTM-GAN network is used by Srivastava et al. (2022) [34]. Research utilizing a single structure independently also investigates the potential of LSTMs and GANs. Yu et al.(2022) [35] also proposed a three-branch structure for modeling three independent melodic attributes. The difference is that they did not use LSTM but used a conditional hybrid GAN. Zhang et al. (2023) [36] introduced inter-branch memory fusion (Memofu), which facilitates information flow between multi-branch stacked LSTM networks. This allows for better modeling of dependencies across multiple musical attributes and sequences, improving the overall coherence of the generated melodies.

基于 rnn 模型 (如 SongWriter) 的成功，长短期记忆 (LSTM) 网络由于能够捕捉连续数据中的长期依赖关系而被用于音乐生成。与标准的 rnn 不同，lstm 解决了渐变消失的问题，使得它们在建模音乐中固有的复杂时间结构时特别有效。与此同时，gan 已经成为音乐生成的强大框架，特别是在创建高质量和多样化的音乐输出方面。将上述两种结构结合起来的研究已经成为一个热门话题。Yu 等人 (2021)[33] 使用条件 LSTM-GAN 作为基于生成的旋律的歌词。他们将从文本歌词转换的音节嵌入向量与噪声向量相结合，并将其输入到发生器中。一个深层鉴别器也被训练来区分生成的 MIDI 音符序列和真实的音符序列 (图 7)。这种方法既显示了 LSTM 捕捉旋律长期依赖关系的能力，也显示了 GAN 通过对抗学习机制增强旋律真实性和自然性的优势。为了进一步提高生成质量，Srivastava 等人使用了三分支条件 LSTM-GAN 网络。(2022)[34]。独立使用单一结构的研究也调查了 lstm 和 gan 的潜力。Yu 等人。(2022)[35] 也提出了一个三分支结构来建模三个独立的旋律属性。区别在于他们没有使用 LSTM，而是使用了条件混合 GAN。Zhang 等人 (2023)[36] 引入了分支间存储融合 (Memofu) ，它促进了多分支堆叠 LSTM 网络之间的信息流动。这允许对多个音乐属性和序列的依赖性进行更好的建模，提高生成旋律的整体一致性。

Transformer-based models have been widely applied in music generation. Howevert limited by the quality of the melody-lyrics pairing dataset and the diversity of variations in real melodies, a large number of studies are still highly dependent on the dataset and are not highly transferable. SongMASS, proposed by Sheng et al. (2021) [37], effectively alleviates data dependency by using separate lyrics and melody encoder-decoder structures within a Transformer-based framework. It also enhances the matching accuracy of lyrics and melodies by introducing sentence-level and word-level alignment constraints. However, the complex alignment mechanism may increase the difficulty of training and reduce the controllability of melody generation. To overcome data scarcity and improve generation controllability, Ju et al. (2022) [38] proposed TeleMelody, a two-stage generation pipeline based on music templates. The system first converts lyrics to templates and then generates melodies based on the templates. The music templates include tonality, chord progressions,rhythmic patterns, and terminations, and they use a self-supervised approach to realize13 of 53template-generated melodies, which solves data dependency.

基于 transformer 的模型在音乐生成中得到了广泛应用。然而，受限于旋律 - 歌词配对数据集的质量和真实旋律变化的多样性，大量的研究仍然高度依赖于数据集，并且不具有高度的可移植性。SongMASS，由 Sheng 等人 (2021)[37] 提出，通过在基于 transformer 的框架中使用独立的歌词和旋律编码器 - 解码器结构，有效地减轻了数据依赖。它还通过引入句子级和词级对齐约束来提高歌词和旋律的匹配精度。然而，复杂的对齐机制可能会增加训练难度，降低旋律生成的可控性。为了克服数据稀缺性和提高生成可

控性，Ju 等人 (2022)[38] 提出了 TeleMelody，一个基于音乐模板的两阶段生成管道。系统首先将歌词转换为模板，然后基于模板生成旋律。这些音乐模板包括调性、和弦进行、节奏模式和结尾，它们使用自我监督的方法来实现模板生成的 53 个旋律中的 13 个，从而解决了数据依赖问题。

Electronics 2025, 14, x FOR PEER REVIEW

电子学 2025,14，x 供同行评审

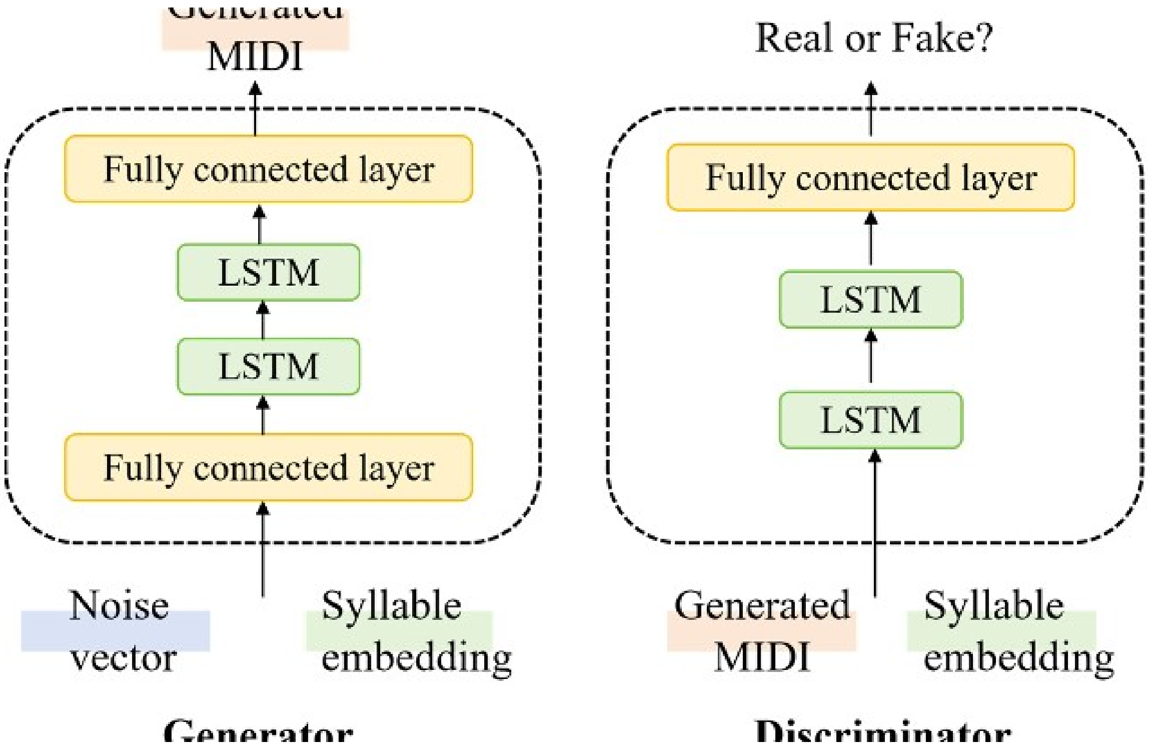


Figure 7. Frameworks of Conditional LSTM-GAN. Reproduced from[33]

图 7。有条件 LSTM-GAN 的框架。转载自 [33]

In a recent study, thanks to the development of LLMs, Ding et al. (2024) [39] proposed SongComposer, an LLM for lyrics and melody composition. It employs a single language model architecture instead of the separate encoders and decoders of traditional approaches and uses the next-token prediction technique. This approach allows the model to predict subsequent notes based on the current lyrics and a portion of the melody until the entire song is generated, outperforming SongMASS and TeleMelody. In contrast to traditional methods, SongComposer does not require complex rules or preset musical templates but rather learns patterns from large amounts of data to make predictions that can generate high-quality melodies without the guidance of explicit music theory. In addition to generating melodies from lyrics, this model can also perform the tasks of generating lyrics from melodies, song continuation, and songs from text. The text-to-song task in this context refers to a task pipeline consisting of textual cues to generate lyrics, lyrics to generate melodies, and artificially produced vocals and accompaniment (as demonstrated by the authors in the project demo https://pjlab-songcomposer.github.io/,accessed on 17 February 2025) and is, therefore, distinct from the text-to-song task mentioned later.

在最近的一项研究中，由于 LLM 的发展，Ding 等人 (2024)[39] 提出了 SongComposer，一个歌词和旋律创作的 LLM。它采用单一语言模型架构，而不是传统方法的独立编码器和解码器，并使用下一个令牌预测技术。这种方法允许模型根据当前歌词和旋律的一部分预测后续音符，直到生成整首歌曲，性能优于 SongMASS 和 TeleMelody。与传统的方法相比，SongComposer 不需要复杂的规则或预设的音乐模板，而是从大量的数据中学习模式，在没有明确的音乐理论指导的情况下做出预测，生成高质量的旋律。除了从歌词中生成旋律，这个模型还可以执行从旋律中生成歌词、歌曲延续和从文本中生成歌曲的任务。在这种情况下，文本到歌曲的任务是指任务管道，包括生成歌词的文本线索，生成旋律的歌词，以及人工产生的声音和伴奏 (如项目演示 https://pjlab-songcomposer.github.io/ 中的作者所示，2025 年 2 月 17 日访问) ，因此不同于后面提到的文本到歌曲的任务。

The field of lyric-based melody generation has evolved from conditional constraints to deep learning and has reached new heights driven by LLMs. Despite its theoretical appeal, this approach still faces several challenges in practice. First, the mapping relationship between lyrics and melodies is not one-to-one. The same lyrics can correspond to multiple plausible melodic configurations, making it more difficult for the model to learn the correct mapping relationship. Second, high-quality, diverse, and representative datasets of lyrics-melody pairings are relatively scarce, which further limits the learning effectiveness and generalization ability of the model.

基于歌词的旋律生成领域已经从条件限制发展到深度学习，并且在 LLMs 的驱动下达到了新的高度。尽管这种方法具有理论上的吸引力，但在实践中仍然面临着一些挑战。首先，歌词和旋律之间的映射关系不是一对一的。同一个歌词可以对应多个合理的旋律配置，这使得模型更难学习到正确的映射关系。其次，高质量、多样化、具有代表性的歌词 - 旋律配对数据集相对匮乏，进一步限制了模型的学习效果和泛化能力。

2. Musical attribute-based Melody Generation

2. 基于音乐属性的旋律生成

Earlier studies, limited by the mapping of textual semantics to music, have very limited generative capabilities. TransProse, proposed by Davis et al. (2014) [40],contains several mapping rules for sentiment labels to musical elements. TransProse generates music based on the density of sentiment words in a given text. However, TransProse does not reflect non-emotional information in the text, and its creativity is limited by manually formulated mapping rules. Rangarajan et al. (2015)[41] devised three strategies for mapping text to music: using all letters, using only vowels, and using vowels in mulated mapping rules. Rangarajan et al. (2015)[41] devised three strategies for mapping conjunction with the part of speech (POS) of words. However, since it is based on character-level mapping, the generated music is very random and does not reflect the semantic information in the text. In order to optimize the mapping between text and music, Zhang et al. (2020)[42]proposed a framework called Butter, which is a multimodal representation learning system for bidirectional music and text retrieval and generation. The system learns music, keyword descriptions, and their cross-modal representations based on a Variational Auto Encoder (VAE) (Figure 8). Butter can generate three types of potential representations:music representations, keyword embeddings, and cross-modal representations, and it can generate ABC notation representations of music from text containing three musical keywords (e.g.,key, beat, and style). However, this method is limited by the fact that the three keywords must be specified precisely, and the generated music is restricted to the Chinese folk song dataset.

早期的研究受限于文本语义到音乐的映射，生成能力非常有限。由 Davis 等人 (2014) [40] 提出的 TransProse 包含几个情感标签到音乐元素的映射规则。TransProse 根据给定文本中情感词的密度生成音乐。然而，trans prose 并不反映文本中的非情感信息，其创造力受限于人工制定的映射规则。Rangarajan 等人 (2015)[41] 设计了三种将文本映射到音乐的策略： 使用所有字母，只使用元音，以及在模拟映射规则中使用元音。Rangarajan 等人。(2015)[41] 设计了三种与词性 (POS) 映射连接的策略。然而，由于它是基于字符级映射，生成的音乐是非常随机的，并没有反映文本中的语义信息。为了优化文本和音乐之间的映射，Zhang 等人。(2020)[42] 提出了一个名为 Butter 的框架，这是一个用于双向音乐和文本检索和生成的多模态表示学习系统。该系统基于变分自动编码器 (VAE) 学习音乐，关键词描述及其跨模态表示 (图 8)。Butter 可以生成三种潜在的表征： 音乐表征、关键词嵌入和跨模态表征，并且它可以从包含三个音乐关键词 (例如，

键、节拍和风格) 的文本中生成音乐的 ABC 符号表征。然而，这种方法受到三个关键字必须精确指定的限制，并且生成的音乐仅限于中国民歌数据集。

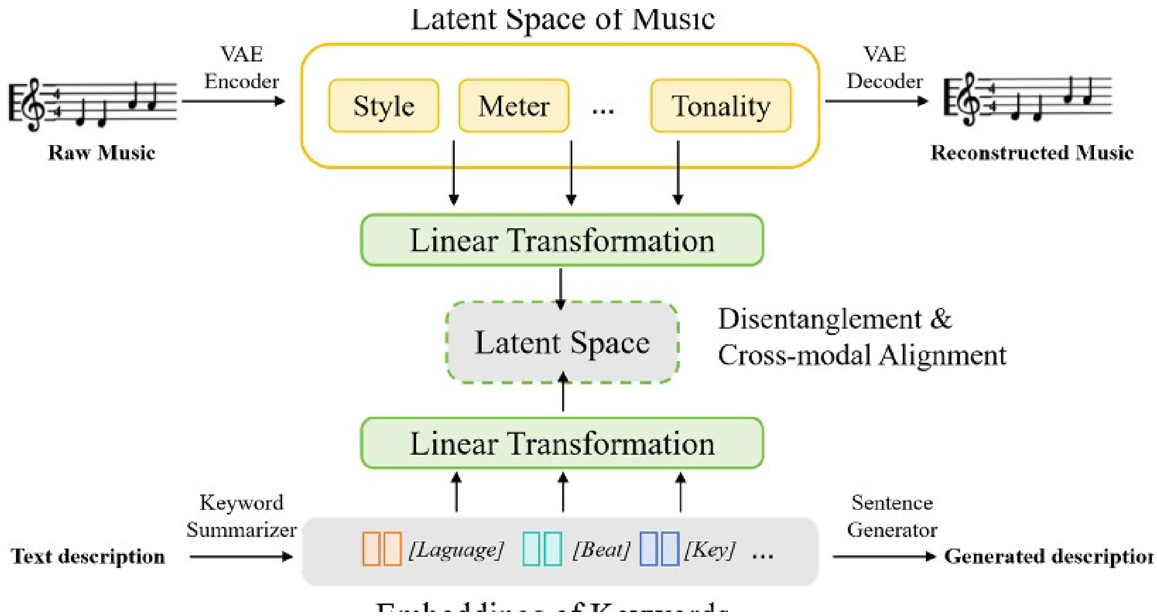


Figure 8. Frameworks of BUTTER. Based on VAE. Reproduced from [42].

图 8。 BUTTER 的框架。基于 VAE。转载自 [42]。

3. I Description-based Melody Generation

3. 基于描述的旋律生成

To escape the limitations of manually formulated rules, Wu et al. (2023)[43]developed a Transformer-based model. The model achieved, for the first time, the generation of complete and semantically consistent musical scores directly from text-based natural language descriptions and also demonstrated the effectiveness of using publicly available pre-trained BERT, GPT-2, and BART checkpoints on music generation tasks.

为了摆脱人工制定规则的局限性，Wu 等人 (2023)[43] 开发了一个基于 transformer 的模型。该模型首次实现了直接从基于文本的自然语言描述生成完整的和语义一致的音乐乐谱，并且还证明了使用公开可用的预先训练的 BERT，gpt-2 和 BART 检查点对音乐生成任务的有效性。

With the development of LLMs, models not specifically designed for music, such as GPT-4 [28] and LLaMA-2[44], have shown some level of music comprehension, but they still perform poorly in music generation. However, ChatMusician, introduced by Yuan et al. (2024)[45], represents a significant progress.

ChatMusician uses music as a second language for LLMs. This new approach, based on the continuously pre-trained and fine-tuned LLaMA2 model, is trained on a 4B dataset and utilizes the ABC notation to seamlessly fuse music and text. In so doing, ChatMusician enabled in-house music composition and analysis without relying on external multi-modal frameworks. Compared to traditional LLMs, ChatMusician can understand music, generate structured, full-length musical compositions, and condition text, chords, melodies, motifs, musical forms, etc., beyond the GPT-4 baseline.

随着 llm 的发展，不是专门为音乐设计的模型，如 GPT-4 [28] 和 LLaMA-2 [44] ，已经显示出一定程度的音乐理解，但是它们在音乐生成中仍然表现不佳。然而，Yuan 等人引入的 ChatMusician (2024)[45] 代表了一个重大的进展。ChatMusician 使用音乐作为 llm 的第二语言。这种新方法，基于不断预先训练和微调的 llama2 模型，在 4b 数据集上进行训练，并利用 ABC 符号无缝融合音乐和文本。通过这样做，ChatMusician 支持内部音乐创作和分析，而不依赖于外部多模态框架。与传统的 llm 相比，ChatMusician 可以理解音乐，生成结构化的全长音乐作品，以及超越 gpt-4 基线的条件文本，和弦，旋律，主题，音乐形式等。

Early research on description-based melody generation was limited by the ability to generate effective mappings from textual semantics to melodies. In recent years, with technological advances, researchers have been able to generate semantically consistent melodies from natural language descriptions. In the latest progress, models such as ChatMusician are not only able to understand music but also generate structurally complete and moderate-length musical compositions, which significantly improves the ability of text-generated music. The generation of independent melodic lines lays a foundation for polyphony generation. Relevant studies are summarized in Table 4.

基于描述的旋律生成的早期研究受限于从文本语义到旋律生成有效映射的能力。近年来，随着技术的进步，研究人员已经能够从自然语言描述中生成语义一致的旋律。在最新进展中，ChatMusician 等模型不仅能够理解音乐，还能生成结构完整、篇幅适中的音乐作品，这大大提高了文本生成音乐的能力。独立旋律线的生成为复调音乐的生成奠定了基础。表 4 总结了相关的研究。

4.1.2. Polyphony Generation
4.1.2. 复调音乐的产生

Polyphony is a style of musical composition employing two or more simultaneous but relatively independent melodic lines. These melodic lines intertwine and support each other harmonically, creating a rich musical texture. Each instrument or voice can have its MIDI track that flows and unfolds in a harmonious manner. Typical polyphonic music is polyphonic pieces (e.g., classical music) as well as contemporary musical accom-paniment. Creating polyphonic music is, therefore, more complex than creating a single melody. Generating polyphonic music from text requires the model to correctly extract or understand the music-theoretic knowledge or semantic information contained in the text and to generate harmonized polyphonic music.

复调是一种使用两个或两个以上同时但相对独立的旋律线的音乐创作风格。这些旋律线相互交织，相互支撑，创造出丰富的音乐结构。每一种乐器或声音都可以有自己的 MIDI 音轨，以一种和谐的方式流动和展开。典型的复调音乐是复调片段 (例如，古典音乐) 以及当代音乐伴奏。因此，创作复调音乐比创作单一旋律更为复杂。从文本中生成复调音乐需要模型正确地提取或理解文本中包含的音乐理论知识或语义信息，生成和谐的复调音乐。

1. L:Musical Attribute-based Polyphony Generation

1. l: 基于音乐属性的复调音乐生成

Early studies used strict attribute templates as textual input to accurately generate conforming multi-track symbolic music. Evolving from attribute-conditional controlled generation, this type of research creatively replaced attribute labels with text input. Rutte etal. (2023)[17]proposed FiGARO. This system is based on a Transformer and can generate multi-track symbolic music by combining expert and learning features. They introduce a self-supervised description-to-sequence learning method. This method automatically extracts fine-grained, human-interpretable features from music sequences and trains a sequence-to-sequence model, reconstructing the original music sequence from the de-scription. However, the descriptions are complex attribute templates, such as “expert description’, including three types of musical attributes, namely, instrument, harmony, and meta-information. These”high-level control codes”raise the bar for users while allowing them to precisely control the generation.

早期的研究使用严格的属性模板作为文本输入来精确生成符合规范的多音轨符号化音乐。这类研究从属性条件控制生成发展而来，创造性地用文本输入代替属性标签。Rutte et al.(2023)[17] 提出 FiGARO。该系统基于 Transformer，可以通过结合专家和学习特征来生成多轨符号音乐。他们引入了一种自监督的从描述到序列的学习方法。该方法从音乐序列中自动提取细粒度的、人类可解释的特征，训练一个序列到序列的模型，从描述中重构出原始音乐序列。然而，描述是复杂的属性模板，如 “专家描述”，包括三种类型的音乐属性，即乐器、和声和元信息。这些 “高级控制代码 “在允许用户精确控制生成的同时，也为用户提高了标准。

Table 4. Melody generation tasks.

表 4。旋律生成任务。

Task Type 任务类别	Model Name 型号名称	Year 年份	Framework 框架 MusicModelCategory1Description Musicmodelcategory1 描述 RepresentationArchitectureT.H.L. 代表性建筑。						Large Model 大型模型	Dataset 数据集	Generated Music 生成的音乐	Accessed Link 访问链接
									LLM, Instruction 法学硕士，教学 Following, Next Token 以下为下一代币		A LLM ALLM (全景模型)	
Song Composer 作曲家 [39] [39]	2024	MIDI MIDI 音频接口	Transformer Transformer 变压器	√	designed for 专为 composing 作曲 songs 歌曲	20 k Melodies 20k 旋律 15k Paired 15k Pair	Multiple minutes 好几分钟	//pjlab-songcomposer. //pjlab-songcompose				
										TeleMelody TeleMelody 电话旋律 [35] [35]	2022	MIDI MIDI 音频接口

						ed 15k 对 Lyric-melody 抒情旋律		r//pjlab - 歌曲作曲家。 github.io/ (accessed on Github.io/(在 17 February 2025) 2025 年 2 月 17 日)	Prediction 预测			
SongMass[37] 宋弥	2021	MIDI MIDI 音频接	Transformer Transf	Pre-training, the	/	380 k Lyrics 380 k 歌	Not mentioned 未提	https: 来源： //				

撒[37]		口	ormer变压器	训练前、训练中、		词	及	musicgeneration. 音乐一代。 github.io/SongMASS/(accessed on 17 Github.io/SongMASS/(在17 February 2025)二零二五				
				sentence-level and句子级别和		66k Melodies 66k旋律						
				token-level alignment令牌级别的对齐		8k Paired 8k成对						
				constraints.制约因素。		Lyric-melody抒情旋律						

								年二月)				
https://drive.google.Https://drive.google.												
Lyric-based以歌词为基础	Conditional有条件的	com/file/d/Com/file/d/										
MelodyMelody旋律 Generation	[33][33]	2021	MIDIMIDI 音频接口	LSTM-GANLSTM-GAN	LSTM-GAN,LSTM-GAN, Synch	/	12,197MIDI12,197MIDI12,197	Not mentioned未提及	1ugOwfBsURax1VQ41ugowfbsurax1v			

一代					ronized同步 Lyrics-Note歌词 - 注释		MIDI		q41ugowfbsurax1vq4 jHmI8P3IdE5xdDj0l/图片：jHmI8P3ldE5xdDj0l/ view?usp=sharingUsp = sharing(共享)			
							songs歌曲					
Align	(acces											

ment 对齐	sed on 17 (访问：17											
February 2025) 二零二五年二月)												
SongWriter 词曲作者 [32] [32]	2019	MIDI MIDI 音频接口	RNN 译者：王士杰	√	Seq-to-Seq, 序列到序列， Lyric-Melody 歌词 - 旋律 Alignment	/	18,451 Chinese 18451 Chinese 18451 中国人	Not mentioned未提及	/			
							pop songs					

					对齐		流行歌曲					
ALYSIA[28]阿莉西亚[28]	2017	MusicXML2&MusicXML2&MusicXML2 MIDI MIDI 音频接口	Random随机 Forests森林	Co-creative协同创作 Songwriting Partnert歌曲创作合伙人 Rhythm Model, Melody Rhythm Model，Melody旋律模型	/	/	Not mentioned未提及	http://bit.ly/2eQHado 译自：http://bit.ly/2eQHado (accessed on 17 (访问：17 February 2025) 二零二五年二月)				
									Model模型			

https://www.orpheus-Https://www.orpheus-												
Orpheus[15]俄耳甫斯[15]	2010	MIDI MIDI 音频接口	/	/	Dynamic Programming动态规划	/	Japanese prosody dataset日语韵律数据集	Not mentioned未提及	music.org/index.php Music.org/index.php (accessed on 17 (访问：17 Febru			

									ary 2025) 二零二五年二月)			

Table 4. Cont.

表 4. 续。

Task Type 任务类别	Model Name 型号名称	Year 年份	Music 音乐 Representation 代表	Model 模型 Architecture 架构	Framework 框架 Category1 类别 1		Description 描述	Large Model 大型模型 Relevance 相关性	Dataset 数据集	Generated Music 生成的音乐 Length 长度	Accessed Link 访问链接
BUTTER[42] BUTTER [42]	2020	MIDI,ABCNotation MIDI，abc 符号	VAE 美国航空航天局(VAE)	√	Representation Learning, Bi-directional Music-Text Retrieval 表征学习，	/	16,257 Chinesefolk songs 16257 首中国民歌	Short MusicFragment 短音乐片段	https://github.com/ Https://github.com/ ldzhangyx/BUTTER 来源：		

					双向音乐 - 文本检索				ldzhangyx/BUTTER (accessed on 17 (访问：17 February 2025) 二零二五年二月)		
Musical 音乐剧 attribute-based 基于属性	[41] [41]	2015	MIDI MIDI 音频接口	/	/	Full parse tree, POS Tag 完全解析树，POS 标记	/	/	Not mentioned 未提及	/	

Melody Melody旋律 Generation一代	TransProse TransProse转性 [37] [37]	2014	MIDI MIDI音频接口	Markoy马可 Chains锁链	/	Generate music from生成音乐 Literature, Emotion文学，情感 Density密度	/	Emotional words情感词汇 from the来自 literature文学	Not mentioned未提及	https://www.来源：https://www。 musicfromtext.com/来源：musicfromtext.com/ (accessed on 17 (访问：17 February 2025	

										) 二零二五年二月)	
Description-描述 based Melody基础旋律 Generation一代	ChatMusician聊天音乐家 [45][45]	2024	ABCNotation Abc符号	Transformer Transformer变压器	Music Reasoning, 音乐推理， Repetition Structure 重复结构		An LLM of 一个LLM symbolic象征性的 music under-音乐下 - standing and站立和 generation代	5.17M datas in 517万数据 different formats 不同的格式 (MusicPile)(MusicPile)(音乐堆)	Full Score of ABC美国广播公司总分 Notation Notation符号	https:来源： //shanghaicannon.上海炮。 github.io/Github.io/github.io/ ChatMusician/ChatMusician/ / 聊天音乐家	

										/ (accessed on 17 (访问：17 February 2025) 二零二五年二月)	
[43] [43]	2023	ABCNotation Abc符号	Transformer Transformer变压器	√	Exploring the Efficacy 探索疗效 of Pre-trained 预先训练课程	Using 使用 pre-trained 预先训练 checkpoints 检查站	282,870 text-tune 282,870 text-tune 282,870 pairs 成对	Full Score of ABC 美国广播公司总分 Notation Notation 符号	/		

					Checkpoints. 检查站。						

1 The “Framework Category”column indicates the classification of the model in the technological frameworks in Section 5 (T.: Traditional Learning Framework; H.: Hybrid LLM Augmented Framework; L.: End-to-End LLM System), and the categorization criteria are described in detail in Section 5. 2 MusicXML is a standard XML-based file format for representing sheet music and music information.

1 “框架类别” 一栏表示第 5 部分 (t: 传统学习框架；h: 混合 LLM 增强框架；l: 端到端 LLM 系统) 中技术框架中模型的分类，分类标准在第 5 部分中详细描述。MusicXML 是一个标准的基于 xml 的文件格式，用于表示乐谱和音乐信息。

In order to lower the threshold for users, human natural language is used to describe target-generated music, enabling the re-understanding and re-generalization of natural language to attribute templates to become a new development direction. MuseCoco, proposed by Lu et al. (2023) [18], is a typical representative. Unlike FIGARO, this system extends the set of musical attributes, and its attribute templates cover 12 musical attributes such as instrument, tempo, time, and pitch range. In addition, the system allows for natural language input instead of complex templates. This system also adopts a two-stage framework consisting of text-to-attribute understanding and attribute-to-music generation(Figure 9). MuseCoco leverages ChatGPT's superior performance in text understanding to convert text descriptions into attributes, which allows users to use natural language to generate music. On top of that, A richer set of attribute templates also improves the accuracy of the music generated to meet users'requirements. A richer set of attribute templates also improves the accuracy of the music generated to meet the user's requirements.

为了降低用户使用的门槛，使用人类自然语言来描述目标生成的音乐，使得自然语言对属性模板的重新理解和重新泛化成为一个新的发展方向。Lu 等人 (2023)[18] 提出的 MuseCoco 就是一个典型的代表。与 FIGARO 不同，该系统扩展了音乐属性集，其属性模板涵盖了乐器、节奏、时间和音高范围等 12 个音乐属性。此外，该系统允许自然语言输入，而不是复杂的模板。该系统还采用了由文本到属性理解和属性到音乐生成组成的两阶段框架 (图 9)。MuseCoco 利用 ChatGPT 在文本理解方面的优越性能将文本描述转换为属性，这允许用户使用自然语言来生成音乐。此外，更丰富的属性模板集也提高了生成音乐的准确性，以满足用户的需求。一组更丰富的属性模板也能提高生成的音乐的准确性，以满足用户的需求。

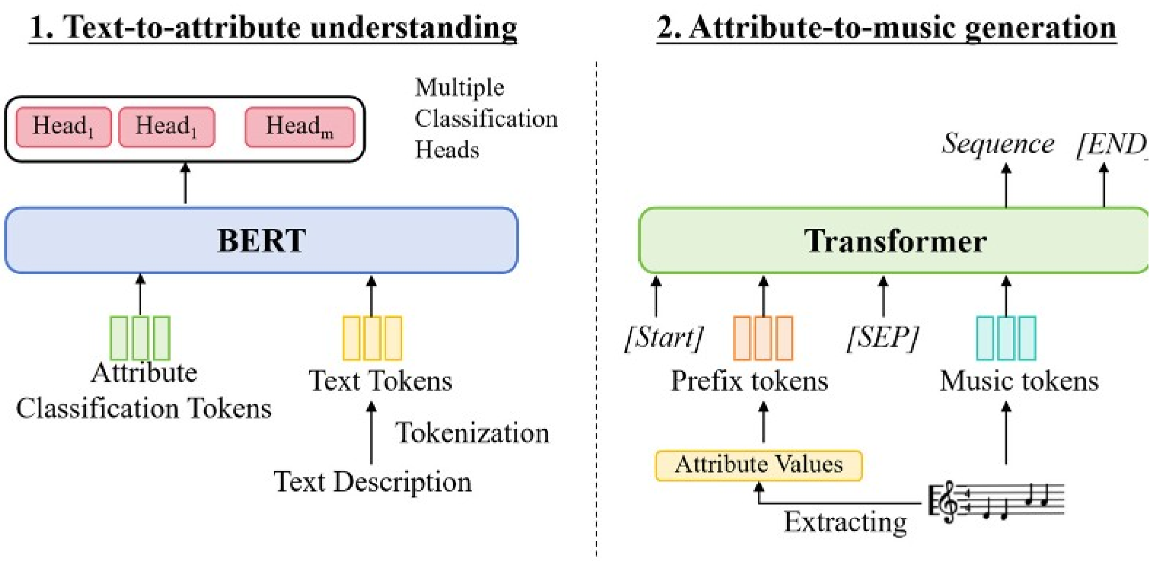


Figure 9. Frameworks of MuseCoco. Based on Transformer. Reproduced from[18].

图 9。 MuseCoco 的框架。基于 Transformer。转载自 [18]。

Description-based Polyphony Generation

基于描述的复调生成

In a recent study, Liang et al. (2024) [46] proposed ByteComposer, which utilizes LLM to simulate mankind's music-composing process. This system adopts a modular design that includes four stages: conceptual analysis, draft generation, self-evaluation and revision, as well as aesthetic selection. Unlike MuseCoco, which only uses ChatGPT to extract attribute information from textual descriptions, ByteComposer embeds LLM as an expert module that not only extracts attribute information but also provides guidance based on a library of music theory knowledge. As a result, ByteComposer allows LLM to play the role of "melody composer". At the same time, a voting module and a memory module are added to ByteComposer, enabling users to subjectively judge the generation results and store the evolution trajectory and interaction data. This system combines the interactive and knowledge-understanding properties of LLMs with existing symbolic music generation models to achieve a melodic composition agent comparable to human creators. In addition, ComposerX proposed by Deng et al. (2024)[47],adopts a multi-agent approach to significantly improve the quality of music composition for LLMs (e.g.,GPT-4). ComposerX can generate coherent polyphonic musical compositions while following users'instructions. This has shown that the multi-agent approach boasts enormous potential in generative tasks. The division of labor of the agents is shown in Table 5.

在最近的一项研究中，Liang 等人 (2024)[46] 提出了 ByteComposer，它利用 LLM 来模拟人类的音乐创作过程。该系统采用模块化设计，包括概念哲学分析、草稿生成、自我评价与修改、审美选择四个阶段。不像 MuseCoco，只使用 ChatGPT 从文本描述中提取属性信息，ByteComposer 嵌入 LLM 作为一个专家模块，不仅提取属性信息，还提供基于音乐理论知识库的指导。因此，ByteComposer 允许 LLM 扮演 “旋律作曲家” 的角色。同时，在 ByteComposer 中增加了投票模块和记忆模块，使用户能够主观地判断生成结果，并存储进化轨迹和交互数据。该系统将 llm 的交互和知识理解属性与现有的符号化音乐生成模型相结合，实现了与人类创作者相媲美的旋律作曲代理。此外，Deng 等人 (2024)[47] 提出的 ComposerX 采用多代理方法来显着提高 llm (例如 gpt-4) 的音乐创作质量。ComposerX 可以在遵循用户指令的同时生成连贯的复调音乐作品。这表明多代理方法在生成任务中拥有巨大的潜力。代理的劳动分工如表 5 所示。

Table 5. Division of Labor.

表 5. 劳动分工。

Agent Name 代理人名称	Task Description 任务描述
Group Leader Agent 组长代理	Responsible for analyzing user input and breaking it down into specific tasks to be 负责分析用户输入并将其分解为具体的任务 assigned to other agents. 分配给其他特工。
Melody Agent 梅洛迪・特工	Generates a monophonic melody under the guidance of the Group Leader. 在组长的指导下生成单声道旋律。
Harmony Agent 和声代理	Adds harmony and counterpoint elements to the composition to enrich

	its structure. 在作品中加入和声和对位元素来丰富作品的结构。
Instrument Agent 仪器代理商	Selects appropriate instruments for each voice part. 为每个声部选择合适的乐器。
Reviewer Agent 评论员代理	Evaluates and provides feedback on the melody, harmony, and instrument choices. 对旋律、和声和乐器的选择进行评估并提供反馈。
Arrangement Agent 编曲代理	Standardizes the final output into ABC notation format. 将最终输出标准化为 ABC 符号格式。

At the same time, ComposerX significantly reduces training costs. The quality of works generated by ComposerX is comparable to polyphonic compositions generated by specialized notated music generation systems [18,43] that require substantial computing resources and data. It is also worth noting that ChatMusician [45] can generate polyphonic music that meets the requirements and maintains good quality. However, it cannot select instruments and only generates polyphonic ABC notation for a single instrument.

同时，ComposerX 显著降低了培训成本。ComposerX 生成的作品的质量可以与专门的符号音乐生成系统 [18,43] 生成的复调作品相媲美，后者需要大量的计算资源和数据。值得注意的是，ChatMusician [45] 可以生成符合要求并保持良好质量的复调音乐。但是，它不能选择乐器，只能为单个乐器生成复调 ABC 符号。

Polyphonic music generation technology has evolved from using structured attribute templates to natural language descriptions, aiming to improve user-friendliness and en-hance the diversity and expressiveness of the generated music. The advantage of early structured attribute templates is their ability to ensure that the generated music adheres to certain musical theory standards. However, these templates have several limitations. First, the forms of text input are constrained by structured templates, requiring the selection of fixed attributes, which makes the generation process less flexible. Second, since the attribute templates essentially define the labels, the generated results often lack personaliza-tion. Additionally, the multitrack melodies generated are relatively independent, lacking coordination. With the development of deep learning and LLMs, modern music generation systems now employ natural language processing to interpret text descriptions and model the collaborative relationships between multiple tracks, making the input process more intuitive and universal while the output sounds more harmonious and pleasant. Relevant studies are summarized in Table 6.

复调音乐生成技术已经从使用结构化属性模板发展到自然语言描述，旨在提高用户友好性，增强生成音乐的多样性和表现力。早期结构化属性模板的优势在于能够保证生成的音乐符合一定的音乐理论标准。然而，这些模板也有一些局限性。首先，文本输入的形式受到结构化模板的约束，需要选择固定的属性，使得生成过程缺乏灵活性。其次，由于属性模板本质上定义了标签，生成的结果往往缺乏个性化。另外，生成的多声道旋律相对独立，缺乏协调性。随着深度学习和线性规划的发展，现代音乐生成系统采用自然语言处理来解释文本描述，并对多首曲目之间的协作关系进行建模，使得输入过程更加直观和通用，输出结果也更加和谐悦耳。相关研究总结在表 6 中。

Table 6. Polyphony Generation Tasks.

表 6。复调生成任务。

Task Type 任务类别	Model Name 型号名称	Framework 框架 MusicModelLarge ModelYearCategoryDescriptionRepresentation Architecture Relevance 音乐模型大型模型类别描述表示架构相关性 T.H.L. T.h.l.								Dataset 数据集	Generated 生成 Music 音乐 Human- 人类	Accessed Link 访问链接
FIGARO[17] 费加罗[17]	2023	REMI+ 译者：王士杰	Transformer Transformer 变压器	interpretable, 可解释的， Expert 专家 Description, 描述，	/	176,581 MIDI 176,581 个 MIDI files (LakhMIDI 文件	Not 不 mentioned 提到	Musical 音乐剧				

						(LakhMIDI Dataset)数据集)						
Attribute-based基于属性的	MIDI MIDI 音频接口	Transformer Transformer变压器	Multi-Track多轨道									
Polyphony复调			947,659 MIDI 947,659 MIDI 947,659 MIDI	https://ai-Https://ai-								
Generation			MuseCoco	Text-to-attri	Textual synt	(Partially (部	muzic.github					

一代			MuseCoco 博物馆	bute 文本到属性 understanding 理解	hesis 文本合成	分	.io/ 原文链接：muzic.github.io/					
2023					and template 和模板 refinement 细化	contains 包含 emotional, genre 情感上的，类型上的 information) 信息)	<=16 bars 译者：王士杰	musecoco/ Museco co / 博物馆 (accessed on 17 (访问：17 February 2025) 二零二五年二				
			[18] 变形金刚	and attribute-to- 以及属性到 music generation 音乐一代								

								月)				
Imitate the 模仿 human creative 人类的创造力												
			MIDI MIDI 音频接口	/								
	ByteComposer ByteComposer 字节作曲家 [46] [46]											
		process, 过程，			A melody 旋律	216,284 216,284	Not 不					
2024		Transformer Transformer 变压器			Multi-step 多步骤 Reasoning, 推理，	composition 组成 LLM agent LLM agent LLM 代理	ABC Notations Abc 符号	mentioned 提到				
Description-描述	Procedural 程序											

based Polyphony基础复调音乐	Control控制											
Generation一代	Significantly值得注意	https:来源：										
ComposerX作曲家 x [47]作曲家	improve the改善	//lllindsey0615.//lllindsey0615.										
	music generation音乐一代	Multi-agent多智能体	324,987324987人	Varied多种多样	github.io/Github.io/github.io							

					/							
	2024	ABC Notation 美国广播公司(ABC)符号	Transformer Transformer 变压器	quality of GPT-4 Gpt-4 的质量 through a 通过	LLM-based 基于 llm 的 framework 框架	ABC Notations Abc 符号	Lengths 长度	ComposerX_作曲家 x demo/(accessed 演示/(访问				
multi-agent 多智能体	on 17 数到 17											
approach 接近	February 2025). 2025 年 2 月)。											

4.2. Audio Domain Methods

4.2 音频域方法

The audio domain text-to-music generation task is a task that automatically generates music segments directly at the audio signal level from input text. The output is usually in the form of time-series sound waveform data rather than a symbolic representation. The research challenge in audio-domain text-to-music generation tasks is establishing a good mapping between text and audio signals. To overcome this challenge, this model must efficiently capture the dependencies between text and music audio and generate high-quality sound outputs. By doing so, the generated music segments will not only follow the textual instructions but also sound smooth and pleasing to the ear.

音频域文本 - 音乐生成任务是一个从输入文本直接在音频信号级别自动生成音乐片段的任务。输出通常是时间序列的声音波形数据，而不是符号表示。音频域文本 - 音乐生成任务的研究难点在于建立文本和音频信号之间的良好映射。为了克服这一挑战，该模型必须有效地捕获文本和音乐音频之间的依赖关系，并生成高质量的声音输出。通过这样做，生成的音乐片段将不仅遵循文本说明，而且声音平滑，悦耳。

Figure 10 illustrates the general workflow of text-to-music generation in the audio domain, highlighting both cross-modal alignment and the audio synthesis process. First, text input (e.g., natural language prompts or labels) is provided to a cross-modal alignment module, which interprets and maps the text into a representation suitable for audio genera-tion. Next, a sequence generation model produces either raw waveforms or spectrograms. For singing voice generation in particular, this output is passed to a vocoder-or another specialized neural synthesis component-to transform the encoded features into a high-quality vocal voice. Finally, the generated audio is played back through a playback device, enabling listeners to hear the resulting music.

图 10 展示了音频领域中文本到音乐生成的一般工作流程，突出了跨模态对齐和音频合成过程。首先，文本输入 (如自然语言提示或标签) 被提供给一个跨模态对齐模块，该模块将文本解释并映射为适合音频生成的表示。接下来，序列生成模型产生原始波形或声谱图。特别是对于歌唱声音的生成，这个输出被传递给声码器 - 或另一个专门的神经合成组件 - 将编码特征转换成高质量的声音。最后，生成的音频通过播放设备回放，使听众能够听到产生的音乐。

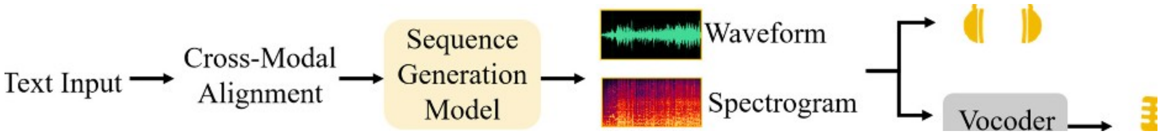


Figure 10. Workflow of Audio Domain Text-to-Music Generation.

图 10 音频领域文本到音乐生成的工作流程。

4.2.1. Instrumental Music Generation
4.2.1 器乐的产生

In recent years, the task of text-to-audio generation has gained significant attention as an important branch of cross-modal tasks. AudioLM [25], as a pioneering work, has made significant breakthroughs in audio modeling. This model maps audio signals to a series of discrete audio representations, transforming the audio generation task into a language modeling problem within this representation space. AudioLM can generate natural and smooth audio from brief prompts, covering human speech, environmental sound effects, and basic piano melodies while maintaining consistency and coherence across long-time sequences. Following the success of AudioLM, researchers have further explored how to use text input to precisely control the audio generation process, leading to models like AudioLDM, Tango, and Tango2 [48-50]. These models combine the advantages of language models and diffusion models, enabling efficient and expressive text-to-audio generation. These advancements have laid a crucial theoretical and technical foundation for music audio generation.

近年来，文本到音频的生成任务作为跨通道任务的一个重要分支受到了广泛的关注。AudioLM [25] 作为一项开创性的工作，在音频建模方面取得了重大突破。该模型将音频信号映射到一系列离散的音频表示，将音频生成任务转化为该表示空间内的语言建模问题。AudioLM 可以从简短的提示生成自然流畅的音频，涵盖人类语言、环境声音效果和基本钢琴旋律，同时保持长时间序列的一致性和连贯性。在 AudioLM 取得成功之后，研究人员进一步探索了如何使用文本输入来精确控制音频生成过程，从而产生了 AudioLDM，Tango 和 tango2 等模型 [48-50]。这些模型结合了语言模型和扩散模型的优点，实现了高效和表达性的文本到音频的生成。这些进展为音乐音频生成奠定了重要的理论和技术基础。

1. Label-based Instrumental Music Generation

1. 基于标签的器乐生成

Early research generated new music by combining audio retrieval with textual labels. A typical example of early commercial implementations is Mubert (https://mubert.com/, accessed on 17 February 2025), which constructs a music database with labels and assigns appropriate labels based on the user’s text input. The appropriate music clips are then selected from the database and combined to create a new piece of music. This approach allows Mubert to respond quickly to user input prompts and to generate musical com-positions with some degree of editing. However, Mubert has some limitations regarding creativity and flexibility because it relies on combining existing music fragments rather than creating entirely new ones;creativity and flexibility because it relies on combining existing music fragments rather

早期的研究通过结合音频检索和文本标签来生成新的音乐。早期商业实现的一个典型例子是 Mubert (https://Mubert.com/，2025 年 2 月 17 日访问) ，它构建了一个带有标签的音乐数据库，并根据用户的文本输入分配适当的标签。然后从数据库中选择合适的音乐片段并组合创建新的音乐片段。这种方法允许 Mubert 快速响应用户输入提示，并通过一定程度的编辑生成音乐作品。然而，穆伯特在创造性和灵活性方面存在一定的局限性，因为它依赖于组合现有的音乐片段，而不是创造全新的音乐片段；

than creating entirely new ones;

而不是创造全新的片段；

2. I Description-based instrumental Music Generation

2. 基于描述的器乐生成

The launch of Riffusion marked the beginning of the use of diffusion models for music generation tasks. Riffusion, developed by Forsgren et al. (2022) [51], is a real-time music generation system based on the stable diffusion model. It features direct noise diffusion on a spectrogram. Riffusion is suitable for live performance or real-time composition as it can rapidly generate short music clips (usually no more than a few seconds) when a specific textual description or lyrics are given. Although Riffusion had significant limitations in terms of music length and complexity, it creatively migrated “text-to-image” technology to the audio domain. Since then, diffusion has become one of the most widely used models for music generation tasks. Huang et al. (2023)[52] proposed a model called Noise2Music. This model uses a two-stage diffusion modeling framework that includes a generator model and a cascade model. This study explored two intermediate representations, i.e., spectrograms and low-fidelity audio (3.2 kHz waveforms). Experimental results show that when low-fidelity audio is used as an intermediate representation, the results are better than when spectrograms are used. Nevertheless, the audio generated by Noise2Music can last for 30 s, and the sampling rate is 24 kHz.

Riffusion 的发布标志着用于音乐生成任务的传播模型的开始。由 Forsgren 等人 (2022) [51] 开发的 Riffusion 是基于稳定扩散模型的实时音乐生成系统。它在声谱图上具有直接的噪声扩散。Riffusion 适用于现场表演或实时作曲，因为当给出特定的文本描述或歌词时，它可以快速生成短音乐片段 (通常不超过几秒钟)。虽然 Riffusion 在音乐长度和复杂度方面有很大的局限性，但它创造性地将 “文本到图像” 技术迁移到了音频领域。从那时起，扩散成为音乐生成任务中使用最广泛的模型之一。Huang 等人。(2023)[52] 提出了一个名为 Noise2Music 的模型。该模型使用包括发生器模型和级联模型的两阶段扩散建模框架。本研究探索了两种中间表示，即语谱图和低保真音频 (3.2 kHz 波形)。实验结果表明，当使用低保真度音频作为中间表征时，结果优于使用语谱图时的结果。尽管如此，在采样率为 24khz 时，Noise2Music 生成的音频可以持续 30s。

Schneider et al. (2023)[53]proposed Mousai. This model also uses a two-stage diffu-sion modeling framework and is capable of generating stereo music at 48 kHz, lasting up to several minutes. The first stage of the model compresses the audio signal by using a diffu-sion amplitude self-encoder, and the second stage generates music using a text-conditional latent space diffusion model (Figure 11). In addition, Mousai achieves a significant break-through in computational efficiency, enabling real-time extrapolation on a consumer-grade graphics processor while maintaining high sound quality and long temporal structural integrity. Recently, Li et al. (2024) [54] proposed JEN-1. Based on a diffusion model, JEN-1can handle multiple types of tasks (music generation, music repair, music continuation, etc.), improving the multitask generalization of music generation models. Also, similar to Noise2Music, JEN-1 processes raw waveform data directly, avoiding the loss of fidelity associated with conversion to spectral formats and generating 48 kHz stereo music. JEN-1incorporates both autoregressive and non-autoregressive structures. The autoregressive mode helps to capture the time-series dependence of music, while the non-autoregressive mode accelerates the process of sequence generation. This hybrid mode overcomes the limitations of a single mode.

Schneider 等人。(2023)[53] 提出了 Mousai。该模型还使用两阶段扩散建模框架，并且能够在 48khz 下产生立体声音乐，持续多达几分钟。该模型的第一阶段使用扩散幅度

自编码器压缩音频信号，第二阶段使用文本条件潜在空间扩散模型生成音乐 (图 11)。此外，Mousai 在计算效率方面取得了重大突破，在消费级图形处理器上实现了实时外推，同时保持了高音质和长时间的结构完整性。最近，Li 等人 (2024)[54] 提出了 JEN-1。Jen-1 基于扩散模型，可以处理多种类型的任务 (音乐生成、音乐修复、音乐延续等) ，提高了音乐生成模型的多任务泛化能力。此外，与 Noise2Music 类似，jen-1 直接处理原始波形数据，避免了与转换到频谱格式和生成 48 千赫立体声音乐相关的保真度损失。Jen-1 包含自回归和非自回归结构。自回归模式有助于捕捉音乐的时间序列相关性，而非自回归模式加速了序列生成过程。这种混合模式克服了单一模式的局限性。

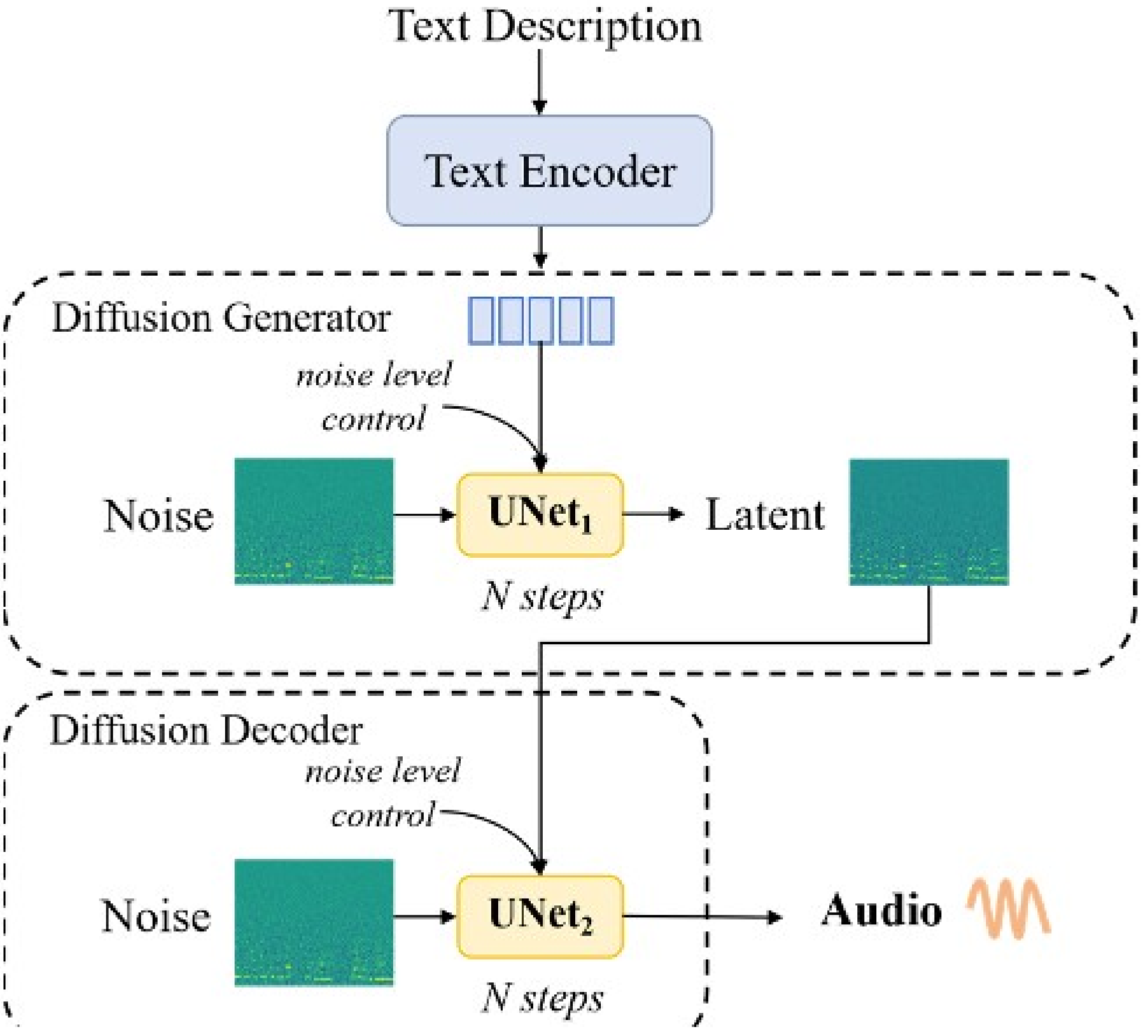


Figure 11. Frameworks of Mousai. Based on Diffusion. Reproduced from [53].

图 11 Mousai 的框架，基于扩散原理，转载自 [53]。

diffusion modeling, Lam et al.(2023)[57] proposed MeLoDy, an approach that combines

扩散建模，Lam 等人。(2023)[57] 提出 MeLoDy，一种结合

Another part of this research explored the application of language models to genera-tive tasks. Almost simultaneously, Agostinelli et al. (2023) [55] proposed MusicLM, which introduces a hierarchical sequence-to-sequence autoregressive modeling approach. This model extends AudioLM to include three levels of language models semantic, coarse acoustic, and fine acoustic-and is able to generate musical audio at 24 kHz (Figure 12). Mu-sicLM addresses the problem of paired audio-text data scarcity by combining MuLan [56]. In addition, MusicLM demonstrates its potential for melodic transformation, being able to stylize a hummed or whistled melody based on a prompt. Considering the advantages of language modeling and diffusion mod-eling, Lam et al. (2023)[57]proposed MeLoDy, an approach that combines the language model with the diffusion model. MeLoDy is an LM-guided diffusion model. It uses the Dual Path Diffusion (DPD) model and an audio VAE-GAN to decode semantic tokens for the fast generation of musical waveforms. The DPD model effectively incorporates semantic information into the underlying representation passages in the denoising step while handling coarse-grained and fine-grained acoustic features. While MeLoDy continues to use the top-level language model in MusicLM for semantic modeling, it significantly reduces the number of forward passes in MusicLM. As well as improving generation efficiency, MeLoDy maintains musicality and text relevance comparable to MusicLM and Noise2Music and exceeds the baseline model in terms of audio quality.

本研究的另一部分探索了语言模型在生成任务中的应用。几乎同时，Agostinelli 等人。(2023)[55] 提出了 MusicLM，它引入了分层序列到序列的自回归建模方法。这个模型扩展了 AudioLM，包括语义、粗声学和细声学三个层次的语言模型，并且能够生成 24 kHz 的音乐音频 (图 12)。Mu-sicLM 通过结合 MuLan 56 解决了成对音频文本数据稀缺的问题。此外，MusicLM 展示了旋律转换的潜力，能够基于提示符风格化哼唱或吹口哨的旋律。考虑到语言建模和扩散建模的优势，Lam 等人 (2023)[57] 提出了 MeLoDy，一种将语言模型和扩散模型相结合的方法。MeLoDy 是一个 lm 引导的扩散模型。它使用双路径扩散 (DPD) 模型和音频 VAE-GAN 来解码语义标记，以快速生成音乐波形。DPD 模型在处理粗粒度和细粒度声学特征的同时，有效地将语义信息融入到基底形式通道中。MeLoDy 在继续使用 MusicLM 中的顶层语言模型进行语义建模的同时，显著减少了 MusicLM 中的前向传递次数。除了提高生成效率，MeLoDy 保持了与 MusicLM 和 Noise2Music 相当的音乐性和文本相关性，并在音频质量方面超过了基准模型。

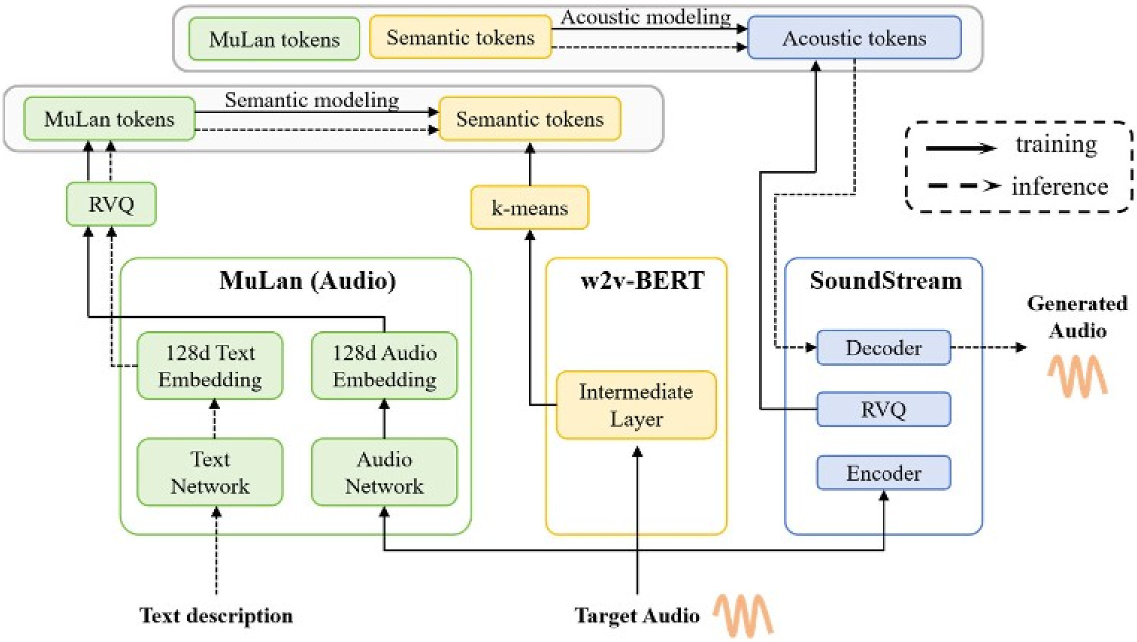


Figure 12. Frameworks of MusicLM. Based on Transformer. Reproduced from[55].

图 12。基于 Transformer 的 MusicLM 框架。转载自 [55]。

Previous models suffer from the limited size of the music dataset, copyright infringe-ment, and plagiarism. To address these problems, Chen et al. (2023)[58]proposed Mu-sicLDM, which aims to address this challenge. Based on the AudioLDM architecture, MusicLDM introduces a beat-synchronized Mixup strategy to enhance the novelty of text-to-music generation. The mixup strategy is a method that restructures existing training samples through linear interpolation, whereby it can augment the training dataset. This approach facilitates MusicLDM to learn via interpolation among training samples rather than simply memorizing a single training instance. Consequently, it helps to reduce overfit-ting resulting from the limited size of the dataset and reduces the risk of plagiarism in the generated content.

以前的模型受到音乐数据集规模有限、版权侵犯和剽窃的困扰。为了解决这些问题，Chen 等人。(2023)[58] 提出了 Mu-sicLDM，旨在解决这一挑战。基于 AudioLDM 架构，MusicLDM 引入了节拍同步 Mixup 策略，以增强文本到音乐生成的新颖性。Mixup 策略通过线性插值重构现有的训练样本，从而扩大训练数据集。这种方法有助于 MusicLDM 通过训练样本间的插值来学习，而不是简单地记忆单个训练实例。因此，它有助于减少由于数据集规模有限而导致的过拟合，降低生成内容中剽窃的风险。

Traditional multi-stage music generation methods usually rely on cascading of mul-tiple models or upsampling steps. This not only increases the complexity of the system but also imposes a high computational overhead. Copet et al. (2023) [59] proposed Music-Gen. MusicGen moves from the traditional multi-stage generation approach to a single autoregressive Transformer decoder, which can simultaneously operate multiple parallel streams of music representations by efficiently interleaving compressed discrete music representations (i.e., music tokens). This approach not only simplifies the music gener-ation process but also significantly reduces the computational costs while maintaining high-quality music output. Notably, unlike MusicLM [55], which relies on supervised data, MusicGen can control melody through unsupervised data.

传统的多阶段音乐生成方法通常依赖于多个模型的级联或上采样步骤。这不仅增加了系统的复杂度，而且计算开销较大。Copet 等人 (2023)[59] 提出了 Music-Gen。MusicGen 从传统的多阶段生成方法转变为单一的自回归 Transformer 解码器，它可以通过有效地交织压缩的离散音乐表示 (即音乐标记) 来同时操作多个并行的音乐表示流。这种方法不仅简化了音乐生成过程，而且在保持高质量音乐输出的同时显著降低了计算成本。值得注意的是，与依赖于监督数据的 MusicLM [55] 不同，MusicGen 可以通过无监督数据控制旋律。

From the above studies, diffusion models and language models have demonstrated their powerful capabilities in music generation. Diffusion models, with their unique noise diffusion and denoising process, have achieved remarkable results in music generation tasks and can generate high-quality and diverse musical works. Language models, based on their mature application in natural language processing, enable effective modeling and generation of music signals by mapping audio signals to discrete representations. Both models can generate music based on textual descriptions, and control such musical attributes as style, melody, rhythm, etc., demonstrating a high degree of controllability and flexibility in music generation. Relevant studies are summarized in Table 7.

从上述研究中，扩散模型和语言模型已经证明了它们在音乐生成方面的强大能力。扩散模型以其独特的噪声扩散和去噪过程，在音乐生成任务中取得了显著的效果，能够生成高质量、多样化的音乐作品。语言模型基于其在自然语言处理中的成熟应用，通过将音频信号映射到离散表示，实现了对音乐信号的有效建模和生成。这两个模型都可以根据文本描述生成音乐，并控制风格、旋律、节奏等音乐属性，表现出高度的可控性和灵活性。相关研究总结在表 7 中。

4.2.2. Singing Voice Synthesis
4.2。2、歌唱声音合成

Singing voice synthesis (SVS) refers to the synthesis of a singing voice according to lyrics and musical scores with the help of speech synthesis techniques. Compared with traditional music generation tasks, SVS is a relatively independent research field because it involves more digital signal processing techniques and audio sampling synthesis techniques. Text-based singing voice generation mainly refers to providing lyrics to gen-erate singing voices. The technical basis of this task is text-to-speech. Like text-to-speech, the mainstream task is divided into three types: splicing synthesis, statistical parameter synthesis, and the current popular neural network synthesis method.

歌唱语音合成是指根据歌词和乐谱，借助语音合成技术对歌唱声音进行合成。与传统的音乐生成任务相比，歌唱语音合成涉及更多的数字信号处理技术和音频采样合成技术，是一个相对独立的研究领域。基于文本的歌声生成主要是指提供歌词来生成歌声。该任务的技术基础是文语转换。像文本到语音的转换一样，主流任务分为三类： 拼接合成、统计参数合成和当前流行的神经网络合成方法。

1. L. Splicing Synthesis

1. 拼接合成

Splicing synthesis first requires creating a sound inventory containing a large number of short vocal units. Then, based on the features of the target vocal, such as pitch, duration, and timbre, the unit with the smallest distance from the target unit is selected for splicing. The duration and pitch of the selected units are then adjusted to match the melody and tempo of the target voice. As early as 1997, Macon et al. (1997) [60] proposed Lyricos, a song synthesizer extended from a text-to-speech synthesizer based on unit concatenation. Since then, a large number of studies have been modeled on this framework, which has developed song synthesis systems of various languages. A successful commercial case is the Vocaloid [61] software released by Yamaha in 2003. This software uses this Splicing synthesis method. Since then, many companies have used Vocaloid as the engine to launch a series of virtual singers, such as Hatsune Miku and LuoTianyi.

拼接合成首先需要创建一个包含大量短声单元的声音库。然后，根据目标声音的音高、音长、音色等特征，选择与目标单元距离最小的单元进行拼接。然后调整所选单元的时长和音高，使其与目标声音的旋律和节奏相匹配。早在 1997 年，梅肯等人 (1997)[60] 提出了 Lyricos，一种从文本到语音合成器扩展而来的基于单位连接的歌曲合成器。从那时

起，大量的研究已经以这个框架为模型，这个框架已经开发了各种语言的歌曲合成系统。一个成功的商业案例是雅马哈在 2003 年发布的 Vocaloid [61] 软件。该软件使用这种 Splicing 合成方法。从那时起，许多公司使用 Vocaloid 作为引擎，推出了一系列的虚拟歌手，如初音美久和骆天一。

Since splicing synthesis synthesizes a song by recording, arranging, and splicing different pronunciations, it has the advantages of a wide range of sounds and a high degree of editorial freedom. However, this method relies heavily on pre-recorded sound libraries, which are expensive to acquire, label, and train; secondly, when splicing different audio segments, the transition between neighboring segments can lead to artifacts at the splices if not handled properly; and finally, it is difficult for the model to generate pitch variations or articulation styles beyond the range of the training data, which limits the effectiveness of the generated singing voice.

由于拼接合成是通过录制、编排和拼接不同的发音来合成一首歌曲，因此它具有发音范围广、编辑自由度高的优点。然而，该方法严重依赖于预先录制的声音库，采集、标注和训练成本较高；其次，在拼接不同音频片段时，如果处理不当，相邻片段之间的过渡会导致拼接处产生伪影；最后，模型难以生成训练数据范围之外的音高变化或发音风格，限制了生成歌唱声音的有效性。

2. Statistical Parameter Synthesis

2. 统计参数合成

Hence, statistical parameter-based synthesis methods have come into being. Saino et al. (2006) [62] extended the application of HMMs in speech synthesis research to song synthesis. In speech synthesis, HMM attaches importance to precisely quantizing time-series variations of speech features into specific statistical parameters. The model treats the textual information as an observable outcome with the acoustic features as its hidden states. The model aims to accurately map from textual to acoustic information through these statistical parameters. When applied to vocal synthesis, HMM needs to record a large number of vocal clips of the same singer and then refine the acoustic feature parameters(e.g,pitch, duration, resonance peaks, etc.) for vocal synthesis through its modeling; finally, the sequence of acoustic features is converted into an audio signal through a vocoder to realize the synthesis of the vocals.

因此，基于统计参数的综合方法应运而生。Saino 等人 (2006)[62] 将 HMMs 在语音合成研究中的应用扩展到歌曲合成。在语音合成中，HMM 重视将语音特征的时间序列变化

精确量化为特定的统计参数。该模型将文本信息作为可观测的输出，声学特征作为其隐含状态。该模型旨在通过这些统计参数准确地将文本信息映射到声学信息。隐马尔可夫模型应用于声音合成时，首先需要记录同一歌手的大量声音片段，然后通过其建模细化用于声音合成的声学特征参数 (如基音、持续时间、共振峰等) ，最后通过声码器将声学特征序列转换为音频信号，实现声音的合成。

The statistical parameter synthesis technique significantly reduces the labor cost in singing voice synthesis compared to the traditional sample splicing method providing more stable and consistent results. This technology has become the basic framework of the current research on singing voice synthesis. However, constrained by statistical laws, statistical models have limitations in capturing complex pitch and rhythmic variations.

与传统的样本拼接方法相比，统计参数合成技术显著降低了歌唱声音合成的人工成本，提供了更稳定、一致的结果。该技术已成为当前歌声合成研究的基本框架。然而，受统计规律的限制，统计模型在捕捉复杂的基音和节奏变化方面存在局限性。

3. Neural Network Synthesis

3. 神经网络综合

With the development of neural networks, some studies have begun to apply neural networks to singing voice synthesis. Nishimura et al. (2016) [63] proposed a DNN-based singing voice synthesis method. Since singing voice synthesis considers more contextual factors than standard TTS synthesis, DNN is used to represent the mapping function from contextual features to acoustic features. Compared to HMM, DNN can better handle complex contextual factors. To address the problem of pitch context sparsity, singing voice synthesis employs note-level pitch normalization and linear interpolation techniques to improve the accuracy of F0 prediction (The fundamental frequency of pitch, commonly used to describe the pitch of a sound). In the subjective listening test, this system signifi-cantly outperforms the HMM-based system. Based on similar neural network frameworks, song synthesis techniques based on various types of neural networks, such as CNN [64], LSTM [65],GAN [66], etc., have been born since then.

随着神经网络的发展，一些研究开始将神经网络应用于歌声合成。Nishimura 等人 (2016)[63] 提出了一种基于 dnn 的歌唱语音合成方法。由于歌唱语音合成考虑了比标准 TTS 合成更多的上下文因素，所以 DNN 被用来表示从上下文特征到声学特征的映射函数。与隐马尔可夫模型相比，深度神经网络能更好地处理复杂的上下文因素。针对基音上下文稀疏的问题，歌声合成采用音符级基音归一化和线性插值技术来提高 F0 (基音的基

频，通常用来描述声音的基音) 预测的准确性。在主观听力测试中，该系统明显优于基于隐马尔可夫模型的系统。基于类似的神经网络框架，基于各种类型的神经网络的歌曲合成技术，如 CNN [64] ，LSTM [65] ，GAN [66] 等，自此诞生。

Table 7. Instrumental Music Generation Tasks.

表 7. 器乐生成任务。

//mubert.com/

来源： mubert.com/

(accessed on 17

(访问： 17

February 2025).

2025 年 2 月)。

(accessed on 17

(访问： 17

Task Type 任务类别	Model 模型	Framework 框架 MusicModelYearCategoryDescriptionLarge ModelRepresentation Architecture Relevance 表示结构相关性 T.H.L. T.h.l. Tag-based control, 基于标签的控制， /Waveform//Music segment/ / 波形 // 音乐段 / combination 组合	Dataset 数据集	Generated 生成 Music 音乐	Accessed Link 访问链接
	Name 名称			Length 长度	https: 来源：
Label-based 基于					
	Variedlengt				
			Mubert	/	

标签 Instrumental工具 Music音乐	hs不同的长度								莫伯特			
JEN-1[54] JEN-1[54]	2024	Diffusion扩散	Omnidirectional全方位 Diffusion Models,扩散模型，	Employ雇佣 FLAN-T5 to Flan-t5至	https:来源：							
						//jenmusic.ai/ //jenmmusic.ai/						
			Waveform波形		Hybrid AR and混合动力AR	provide提供	15k48kHz15k48kHz	Varied多种多样	audio-demosAudio-dede			

					和 NAR Architecture, NAR架构，				mos音频演示			
			(48kHz) (48kHz)	Masked Noise 掩盖噪音	superior text优秀的文本	audios音频	lengths长度					
			embedding嵌入 extraction提取	Robust鲁棒	February 2025). 2025年2月)。							
						Autoencoder自动编码器						
				2024	Diffusion扩散	2.8M 2.8m 2.8米	https:来源：					

			Beat-synchronous 节拍同步			Trained on 训练有素	music-audio 音乐 - 音频	//musicldm. //musicldm 音乐。				
MusicLDM MusicLDM 音乐世界	Mel- 梅尔	mixup, Latent 混淆，潜伏	Broad Data at 宽泛的数据			pairs for CLAP 双人鼓励计划	Varied 多种多样	github.io/ Github.io/github.io/				
[58] [58]	Spectrogram 语谱图	Diffusion, CLAP, 扩散，克拉普，	Scale 比例			(20k hours) (20k hours) (20k hours)	lengths 长度					
AudioLDM AudioL	10 k pairs for 10k											

DM	pairs for											
MusicLDM MusicLDM 音乐世界												
Description-描述												
based 基于												
instrumental 乐器	site/Music-网站 / 音乐											
music Generat	50 k 48 kHz											

ion 音乐一代	50 k 48 kHz 50k 48 kHz											
Mousai[53]穆赛[53]	2023	Waveform波形 (48 kHz @2)(48 kHz @2)(48 kHz)	Diffusion扩散	Latent Diffusion潜在扩散 64× compression 64 × 压缩	/	Text-music pairs文本 - 音乐对 (2.5kh)(2.5千米)	Multiple多个 minutes分钟					
08f6002d8708f6002d87												
(accessed on												

17 (访问：17												
February 2025). 2025 年 2 月)。												
Transformer LM, 变压器 LM,	https: 来源：											
	Trained on 训练有素	390 k 32 kHz 390 k 32 kHz										
MusicGen [59] MusicGen [59]	2023	Discrete tokens(32kHz) 离散	Transformer Transformer	CodebookInterleaving Cod	BroadData atScale 大规模数	audios(20khours) 音频(20	<=5 min <= 5 分钟	audiocraft Audiocraft 音响				

		代币(32kHz)	变压器	ebook交织	据	小时)		设备				
Strategy 策略	(accessed on 17 (访问：17											

(accessed on 17

(访问： 17

February 2025).

2025 年 2 月)。

https://diligent-

Https://diligent-

pansy-4cb.notion.

Pansy-4cb. notion 三色堇。

Generation-with-

Generation-with - 一代

Diffusion-ebe6e9

扩散 -ebe6e9

e528984fa1b226d4

E528984fa1b226d4

//github.com/

网址： github.com/

facebookresearch/

脸谱网研究 /

February 2025).

2025 年 2 月)。

Table 7. Cont.

表 7. 续。

Task Type 任务类别	Model 模型	Year 年份	Framework 框架 MusicModelCategoryLarge ModelDescription Musicmodelcategory 大型模型描述 Representation Architecture Relevance 表示架构相关性 T.H.L. T.h.l.								Dataset 数据集	Generated 生成
												Music 音乐
	Name 名称											Length 长度
2023	Dual-path 双重路径	Trained 训练有素	6.4 M 24 kHz 6.4 m 24 kHz 6.4 m 24	10s-30s 10-30 岁								

			kHz									
	MeLoDy旋律	Waveform波形	Diffusion&扩散&		diffusion, language扩散，语言	LLaMA for美洲驼	audios音频					
	[57][57]	(24kHz)(24kHz)	VAE-GAN作者：VAE-GAN		model, Audio Model，Audio 模型，音频	semantic语义学	17	(257k hour)(257khour)(257khour)				
	VAE-GAN作者：VAE-GAN	modeling建模										
2	Tr	B	O	5								

023	ansformer Transformer变压器	ased on基于	ptimize using优化使用	M 24 kHz 5 m 24 kHz 5 m 24 kHz								
		MusicLM MusicLM音乐	Waveform (24波形(24	AudioLM. AudioLM音频模块。	pre-trained预先训练	audios (280k音频(280k	Multiple多个					
		Description-描述	155	kHz)千赫)	multi-stage多阶段	models Mulan Models Mulan模特花木兰	hours)时间)	minutes分钟				
		based	modeling,	and w2v								

		基于	MuLan建模，木兰	-BERT还有w2v-BERT								
instrumental乐器												
music Generation音乐一代	Using for使用											
Description for描述												
Noise2Music噪音2音	2023	Spectrogram语谱图	Diffusion扩散	Cascading瀑布 di	Training训练 Gene	6.8 M 24 kHz and 6.8 m	30s 30秒					
								Embedding嵌入				

乐 152		and Wave-和 Wave- form(better) Form (better)形式(更好)		ffusion, 1D 扩散，1D Efficient U-Net Efficient u-net 高效 u 型网络	ration and 世代和 Text 文字	24 kHz 和 16 kHz audios 16 kHz 音频 (340 k hours) (340 k hours) (340 千小时)						
Extraction 提取												
Riffusion 即兴演奏 [51]	2022	Spectrogram 语谱图	Diffusion 扩散	√	Tag-based control, 基于标签	/	/	~ 10s ~ 10 秒				

[51]					的控制， Music segment 音乐部分 combination 组合							

Accessed Link

访问链接

https://efficient-

Https://efficient-

melody.github.

Melody.github.

io/ (accessed on

Io/(在

February 2025).

2025 年 2 月)。

https://google-

Https://google-

research.github.

Research.github 搜索引擎。

io/seanet/

Io/seanet/io/seanet/

musiclm/

Musiclm / 音乐 /

examples/

示例 /

(accessed on 17

(访问： 17

February 2025).

2025 年 2 月)。

https://google-

Https://google-

research.github.

Research.github 搜索引擎。

io/noise2music

Io/noise2music io/noise2music

(accessed on 17

(访问： 17

February 2025).

2025 年 2 月)。

https://www.

来源： https://www。

riffusion.com/

来源： riffusion.com/

(accessed on 17

(访问： 17

XiaoiceSing [67] is one of the earliest commercially deployed SVS systems driven by deep learning. This system is built on the main architecture of FastSpeech [68] and makes specific adjustments to adapt to singing synthesis tasks. To avoid the out-of-tune issue, XiaoiceSing adds a residual join to the F0 prediction to make the predicted pitch more accurate. In addition, to improve rhythm, XiaoiceSing, in addition to the duration loss of each phoneme, calculates the total duration of all the phonemes that make up a note. Using WORLD[69] as a vocoder, XiaoiceSing is able to ensure that the input F0 contour is consistent with the F0 contour in the generated vocals, ensuring a high level of quality and consistency. During this period, research on using Transformers and WORLD vocoders has been springing up[70-72]. In order to overcome the limitation of the sampling rate, Chen et al. (2020)[73] proposed HiFiSinger. It replaces WORLD with a parallel WaveGAN [74] to generate waveforms at a high-fidelity 48kHz sampling rate, although it utilizes the same FastSpeech-based acoustic model as XiaoiceSing. WaveGAN, unlike WaveRNN, can generate a more realistic audio waveform through a discriminator.

小冰星 [67] 是最早由深度学习驱动的商业部署 SVS 系统之一。该系统建立在 FastSpeech [68] 的主要架构上，并进行具体的调整以适应歌唱合成任务。为了避免走调问题，XiaoiceSing 在 f0 预测中添加了一个残余连接，以使预测的音高更准确。此外，为了改善节奏，小冰心除了计算每个音素的音长损失外，还计算了组成一个音符的所有音素的总音长。使用 WORLD [69] 作为声码器，xiao icesing 能够确保输入的 f0 轮廓与生成的声音中的 f0 轮廓一致，确保高质量和一致性。在此期间，关于使用变形金刚和 WORLD 声码器的研究正在兴起 [70-72]。为了克服采样率的限制，Chen 等人。(2020) [73] 提出了 HiFiSinger。它用一个并行的 WaveGAN [74] 取代 WORLD，以高保真的 48khz 采样率产生波形，尽管它使用与 XiaoiceSing 相同的基于 fast speech 的声学模型。与 WaveRNN 不同，WaveGAN 可以通过鉴别器产生更真实的音频波形。

In addition to the FastSpeech architecture, Tacotron[75]is also widely used for vocal synthesis tasks, with a greater focus on generative detail and expressiveness. Gu et al.(2020)[76]proposed ByteSing, which combines the advantages of a Tacotron-like architec-ture with the neural vocoder of WaveRNN. Neurovocoders are capable of capturing and reproducing more complex acoustic features. This high-fidelity generative capability is much better than the generative ability of conventional vocoders. ByteSing employs an autoregressive decoder to convert the input features (extended by duration information) directly into Mel spectrograms, which are synthesized into waveforms by the vocoder. By using attention-based alignment and the encoder-decoder framework, ByteSing effec-tively manages long-range dependencies and detailed acoustic feature modeling. Auxiliary phoneme duration prediction models are added to enhance ByteSing's ability to handle the complex temporal nuances inherent in singing. This system is capable of a guaranteed sampling rate of 24 kHz.

除了 FastSpeech 架构之外，Tacotron [75] 也被广泛用于声音合成任务，更加注重生成细节和表现力。Gu 等人。(2020)[76] 提出了 ByteSing，它结合了 tacotron 样结构与 WaveRNN 的神经声码器的优点。神经声码器能够捕获和再现更复杂的声学特征。这种高保真的生成能力远远优于传统声码器的生成能力。ByteSing 使用自回归解码器将输入特征 (通过持续时间信息扩展) 直接转换为梅尔语谱图，由声码器将其合成为波形。通过使用基于注意力的对齐和编码器 - 解码器框架，ByteSing 有效地管理了长程相关性和详细的声学特征建模。添加了辅助音素持续时间预测模型，以增强 ByteSing 处理歌唱中固有的复杂时间细微差别的能力。该系统能够保证 24khz 的采样率。

As diffusion models demonstrate enormous potential in generative tasks, Diffsinger proposed by Liu et al. (2021)[77], also based on FastSpeech, employs a denoising dif-fusion probabilistic model to transform generative tasks into parametric Markov chains conditioned on musical scores. This model adds noise to the Mel spectrogram through a diffusion process until it becomes Gaussian and gradually restores the Mel spectrogram during denoising. In order to improve sound quality and inference speed, Diffsinger introduces a shallow diffusion mechanism and utilizes prior knowledge acquired from simple loss to reduce inference steps, allowing the model to close to a real-time generation(Figure 13, Left). The techniques mentioned above all rely on large databases, so studies aiming to reduce data consumption are cropping up, such as LiteSing[71], Sinsy[78], etc.

由于扩散模型在生成任务中表现出巨大的潜力，Liu 等人提出的 Diffsinger。(2021)[77] 也基于 FastSpeech，采用去噪扩散概率模型将生成任务转化为以音乐分数为条件的参数马尔可夫链。该模型通过扩散过程将噪声添加到 Mel 谱图中，直到它变成高斯，并在去噪过程中逐渐恢复 Mel 谱图。为了提高声音质量和推断速度，Diffsinger 引入了浅层扩散机制，并利用从简单损失获得的先验知识来减少推断步骤，使模型接近实时生成 (图 13，左)。上面提到的技术都依赖于大型数据库，因此旨在减少数据消耗的研究如雨后春笋般涌现，如 LiteSing [71]、 Sinsy [78] 等。

As traditional SVS techniques employ a two-stage generation approach, independent training of the acoustic model and vocoder may result in mismatches. However, VIsinger [79] and VISinger 2 [80], proposed by Zhang et al., have significantly reduced these mismatches. They have successfully applied end-to-end speech synthesis techniques to song synthesis and generated song audio directly from lyrics and music scores (Figure 13, Right). This method operates on the main architecture of VITS (variational inference with adversarial learning for end-to-end text-to-speech) [81]. It means that VITS uses a combined end-to-end speech synthesis model that incorporates VAE, normalizing flow, and GAN to improve the encoder following singing characteristics. While modeling the acoustic variations in singing, VITS introduces an F0 predictor to obtain stable singing performance. This system also optimizes rhythm, modifying the traditional duration prediction to the duration ratio of phonemes to notes. Introducing the VIsinger series takes singing synthesis to a new end-to-end model.

由于传统的 SVS 技术采用两阶段生成方法，声学模型和声码器的独立训练可能导致不匹配。然而，Zhang 等人提出的 VIsinger [79] 和 VIsinger 2 [80] 显着减少了这些不匹配。他们已经成功地将端到端的语音合成技术应用于歌曲合成，并直接从歌词和乐谱生成歌曲音频 (图 13，右)。这种方法在 VITS (具有端到端文本到语音对抗学习的变分推理) 的主要架构上运行 [81]。这意味着 VITS 使用一个端到端的语音合成模型，结合 VAE，归一化流和 GAN 来改善编码器的歌唱特性。在对歌唱中的声学变化进行建模时，VITS 引入了 f0 预测器以获得稳定的歌唱性能。该系统还优化了节奏，将传统的时值预测修改为音素与音符的时值比。VIsinger 系列的引入使得歌唱合成成为一种新的端到端模式。

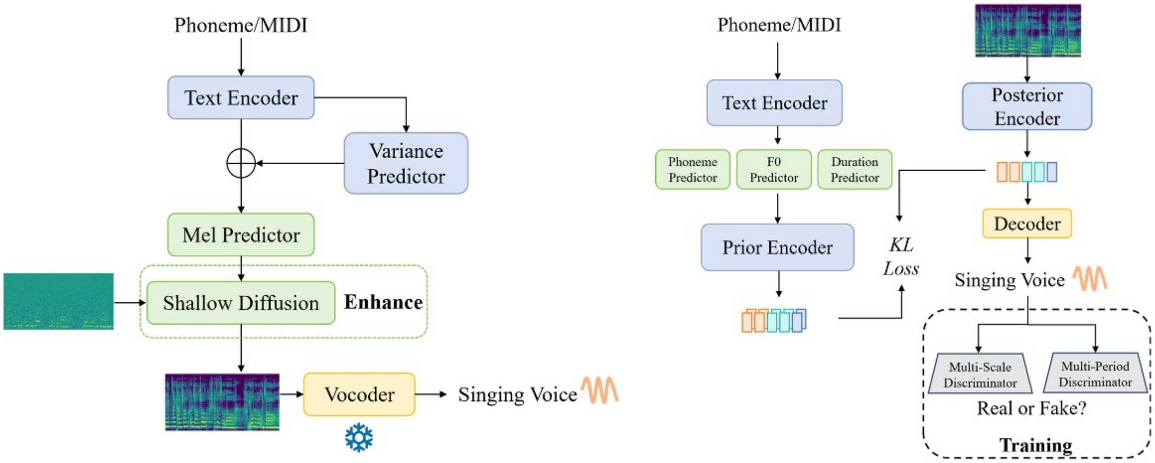


Figure 13. (Left): Frameworks of Diffsinger. Based on Diffusion. (Right): Frameworks of VIsinger Based on VAE and GAN.

图 13。(左) : Diffsinger 框架，基于 Diffusion。(右) : VIsinger 框架，基于 VAE 和 GAN。

The neural network synthesis approach simplifies the system architecture. Firstly, it generates high-quality singing audio from text through an end-to-end modeling ap-proach. Secondly, deep learning models enable the learning of complex acoustic feature representations, generating high-fidelity singing voices. Finally, the improved architectural and training technique can improve computation efficiency, making real-time generation possible and supporting multi-modal information fusion.

神经网络综合方法简化了系统架构。首先，通过端到端的建模方法从文本中生成高质量的歌唱音频。其次，深度学习模型能够学习复杂的声学特征表示，生成高保真的歌唱声音。最后，改进的架构和训练技术可以提高计算效率，使实时生成成为可能，并支持多模态信息融合。

Singing voice synthesis technology has undergone three stages: splicing synthesis, statistical parameter synthesis, and neural network synthesis. Great changes have occurred in the representation of acoustic features, model structure, and other aspects. These techniques have improved the naturalness of synthesized singing and enabled the system to better capture pitch changes and rhythms, generating more vivid and realistic singing audio. Future research will continue to explore new ways to reduce data requirements, increase synthesis speed, and enhance model generalization capabilities. Relevant studies are summarized in Table 8.

歌声合成技术经历了拼接合成、统计参数合成和神经网络合成 3 个阶段。声学特征的表征、模型结构等方面都发生了巨大的变化。这些技术提高了合成歌唱的自然度，使系统能够更好地捕捉音高变化和节奏，生成更加生动逼真的歌唱音频。未来的研究将继续探索降低数据需求、提高合成速度、增强模型泛化能力的新途径。相关研究总结在表 8 中。

4.2.3. Complete Song Generation

4.2。3、完整的歌曲生成

A song is a combination of vocals and accompaniment. Complete song generation synthesizes research on pure music generation and song synthesis with the goal of automat-ing the creation of complete songs. This generation task is a multi-modal one involving multiple types of generation tasks. It requires not only synthesizing the corresponding accompaniment based on the textual content but also generating matching lyrics, vocals, etc.

歌曲是人声和伴奏的结合。完整歌曲生成综合了纯音乐生成和歌曲合成的研究，目标是实现完整歌曲的自动生成。该生成任务是一个涉及多种类型生成任务的多模态任务。它不仅需要根据文本内容合成相应的伴奏，还需要生成匹配的歌词、人声等。

1. L. Staged Generation

1. 分阶段生成

Hong et al. (2024)[82], for the first time, proposed text-to-song, which incorporates both vocal and accompaniment generation. They developed Melodist, a two-stage text-to-song method. Melodist generates singing voice synthesis (SVS) first and then vocal-to-accompaniment (V2A) synthesis based on SVS (Figure 14). Finally, Melodist mixes SVS and V2A together to form a complete song. In the vocal-to-accompaniment synthesis stage, the Melodist adopts the tri-tower contrastive pre-trained framework to learn more efficient text representations and jointly embeds text, vocals, and accompaniment into an aligned space, which enables the model to control accompaniment generation by using natural language cues. This experiment shows that the outputs generated by the Medodist model achieve better performance in terms of subjective and objective metrics assessment, as well as text natural language cues. This experiment shows that the outputs generated by the Medodist consistency. However, as the results generated rely on the quality of the source separation, this method still has limitations-it cannot achieve an end-to-end generation. On top of this, this method also sees the accompaniment as a piece of music, ignoring the complex combinations between instrumental tracks.

Hong et al。(2024)[82] 首次提出了文本到歌曲，包括声乐和伴奏生成。他们开发了旋律，一个两阶段的文本到歌曲的方法。旋律合成器首先生成歌唱声音合成 (SVS) ，然后基于 SVS 生成声音到伴奏 (V2A) 合成 (图 14)。最后，Melodist 将 SVS 和 V2A 混合在一起形成一首完整的歌曲。在人声 - 伴奏合成阶段，旋律师采用三塔对比预训练框架学习更有效的文本表示，将文本、人声和伴奏共同嵌入对齐空间，使模型能够利用自然语言线索控制伴奏生成。实验结果表明，medist 模型的输出结果在主观和客观评价指标以及文本自然语言线索方面都取得了较好的效果。实验结果表明，中介模型生成的输出具有一致性。然而，由于生成的结果依赖于源分离的质量，这种方法仍然有局限性 - 它不能实现端到端的生成。此外，该方法还将伴奏视为一段音乐，忽略了器乐曲之间的复杂组合。

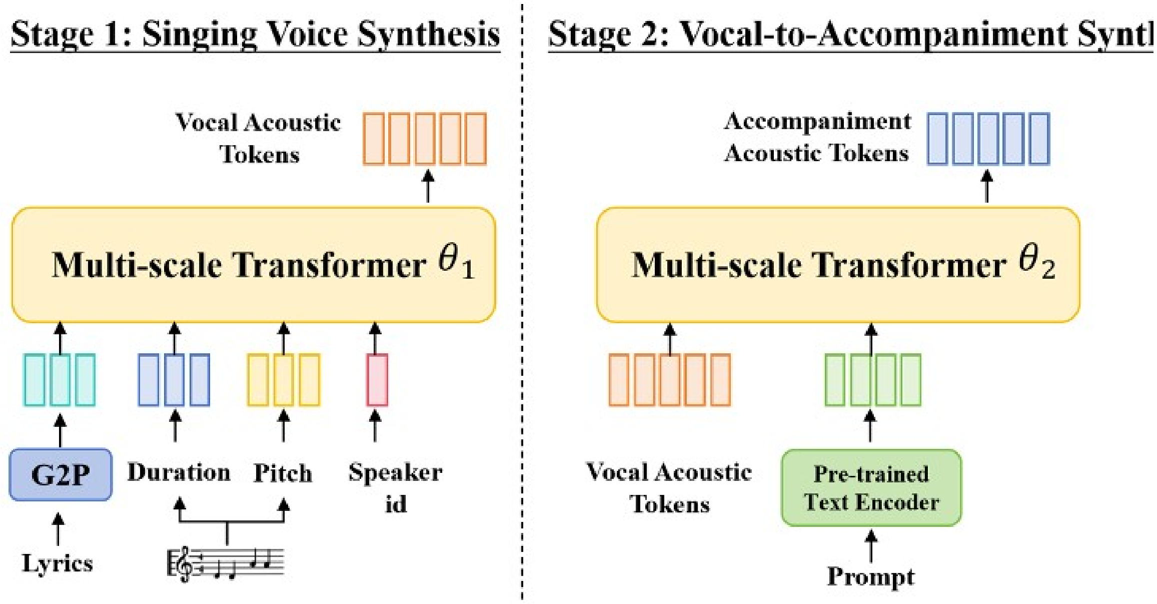

Stage 1: Singing Voice Synthesis
Stage 2: Vocal-to-Accompaniment Syntl
Vocal Acoustic Tokens
Multi-scale Transformer θ_1
G2P
Lyrics
Duration
Pitch
Speaker id
Accompaniment Acoustic Tokens
Multi-scale Transformer θ_2
Vocal Acoustic Tokens
Pre-trained Text Encoder
Prompt

Figure 14. Frameworks of Melodist. Based on Transformer.

图 14。基于 Transformer 的旋律家框架。

Table 8. Singing Voice Synthesis Tasks.

表 8. 歌唱声音合成任务。

Task Type 任务类别	Model Name 型号名称	Year 年份	Music 音乐 Representation 代表	Model 模型 Architecture 架构	Description 描述	Dataset 数据集	Accessed Link 访问链接
Commercial Singing 商业歌唱 Voice Engine 语音引擎	ACE Studio ACE 工作室	2021	/	/	AI synthesis, Auto pitch 人工智能合成，自动变距	/	https://acestudio.ai/ 来源：https://acestudio.ai/ (accessed on 17 (访问：17 February 2025). 2025 年 2 月)。
	Synthesizer V Studio 合成器 v	2018	/	/	WaveNet vocoder, DNN,	/	https://dreamtonics.co

	工作室				AI WaveNet 声码器，DNN，人工智能 synthesis 合成		m/ Https://dreamtonics.com/ synthesizerv/ (accessed on 17 Synthesizerv/(访问于 17 February 2025). 2025 年 2 月)。
	Vocaloid Vocaloid 声带	2004	Waveform&Spectrum 波形与频谱	/	sample concatenation 样品级联	/	https://www.vocaloid.com/ Https://www.vocaloid.com/ (accessed on 17 (访问：17 February 2025).

							2025 年 2 月)。
VISinger 2[80] 视觉 2 [80]	2022	Mel-Spectrogram Mel-Spectrogram 梅尔 - 谱图	VAE + DSP VAE + DSP 公司	conditional VAE, Improved 条件 VAE，改进 Decoder, Parameter 解码器，参数 Optimization, Higher 优化，更高 Sampling Rate(Considering 采样率(考虑 to VISinger) 致 VISinge	100 24 kHz Chinese 10024 kHz Chinese 10024 kHz 中文 singing records 唱片 (5.2h) (5.2 小时)	https://zhangyongmao. Https://zhangyongmao. github.io/VISinger2/ 来源：github.io/VISinger2/ (accessed on 17 (访问：17 February 2025). 2025 年 2 月)。	

				r)			
Singing VoiceSynthesis 歌声合成	VISinger[79] 译者：王士杰	2022	Mel-Spectrogram Mel-Spectrogram 梅尔 - 谱图	VAE + GAN VAE + GAN 公司	end-to-end solution, FO 端到端解决方案，FO predictor, normalizing flow 预测器，归一化流程 based prior encoder and 基于先验编码器和 adversarial decoder 对抗性解码器	100 24 kHz Mandarin 10024 kHz Mandarin 10024 kHz 普通话 singing records 唱片 (4.7h) (4.7 小时)	https: // zhangyongmao. Https:// zhangyongmao. github.io/ VISinger/ (accessed Io/VISinger/(访问 on 17 February 2025). 2025 年 2 月 17 日)。
	DiffSinger[77] 迪弗辛	2021	Mel-Spectrogram Mel-	Diffusion + Diffusion + 扩散	Shallow diffusion	117 24 kHz Mandarin	https: // www.diffsinger.

	格 [77]		Spectrogram 梅尔 - 谱图	+ Neural Vocoder 神经声码器	浅层扩散 mechanism, parameterized 机制，参数化 Markov chain, Denoising 马尔可夫链，去噪 Diffusion Probabilistic 扩散概率 Model ,FastSpeech 模型，Fast Speech	11724 kHz Mandarin 11724 kHz 普通话 singing records 唱片 (5.89h) (5.89 小时)	com/ 来源：https://www.diffsinger.com/ (accessed on 17 (访问：17 February 2025). 2025 年 2 月)。
	HiFiSinger[73] Hifinger [73]	2020	Mel-Spectrogram Mel-Spectrogram 梅	Transformer+ 变形金刚 + Neura	Parallel WaveGAN 平行的 WaveG	6817 48 kHz pieces 681748 kHz	https:// speechresearch. 来源：

			尔 - 谱图	l Vocoder 神经声码器	AN (sub-frequency GAN+ (亚频率 GAN + multi-length GAN), 多长度 GAN) , FastSpeech fastSpeech FastSpeech	singing records 唱片 (11h) (11 小时)	https://speechresearch。 github.io/ hifisinger/ 图片来源：github.io/hifisinger/ (accessed on 17 (访问：17 February 2025). 2025 年 2 月)。
	ByteSing[76] 字节编码 [76]	2020	Mel-Spectrogram Mel-Spectrogram 梅尔 - 谱图	Transformer+ Neural Vocoder 变压器 + 神经声码器	WaveRNN, Auxiliary WaveRNN，辅助 Phoneme DurationPredict	90 24 kHz Mandarinsinging records 9024 kHz 国语唱片	https:// ByteSings.github.io https:// ByteSings.github.io (accessed on

					ion model, Tacotron 音素持续时间预测模型，Tacotron		17 (访问：17 February 2025). 2025 年 2 月)。

Table 8. Cont.

表 8。

Task Type 任务类别	Model Name 型号名称	Year 年份	Music Representation MusicRepresentation 音乐展示	Model Architecture ModelArchitecture 模型架构	Description 描述	Dataset 数据集	Accessed Link 访问链接
XiaoiceSing [67] 小冰心 [67]	2020	Acousticparameters 声学参数	Transformer+ WORLD Transformer + WORLD 变压器 + 世界	integrated network, ResidualF0, syllable durationmodeling, FastSpeech 集成网络，残	229748 kHz Mandarinsinging records(74h) 229748 kHz 国语歌唱记录 (74h)	https:// xiaoicesing.github.io/ (accessed on 17February 2025). Https:// xiaoicesing.github.io/(2	

				差 f0，音节时长建模，快速语音		025 年 2 月 17 日访问)。	
Singing Voice 歌声 Synthesis 合成	[63] [63]	2016	Acoustic 原声 parameters 参数	DNN 作者：DNN	musical-note-level pitch 音符级别的音高 normalization, 归一化， linear-interpolation 线性插值	70 48 kHz Japanese 7048 kHz Japanese 7048 kHz 日语 singing records 唱片	/
	[62] [62]	2006	Acousticparameters 声学参数	HMM 嗯	Context-dependent HMMs, 上下文相关的 hmm, duration models,	6044.1kHz Japanesesinging records 6044.1 kHz 日本歌唱记录	https:// www.sp.nitech.ac.jp/ 来源：https:// www.sp.nitech.ac.jp/ ~k-

					and 持续时间模型 time-lag models 时滞模型		saino/ music/ (accessed on ~ k-saino/music/(在 17February 2025). 2025 年 2 月 17 日)。
Lyricos[60] 抒情诗[60]	1997	Waveform 波形	Sinusoidal model 正弦模型	ABS/OLA sinusoidal model, ABS/OLA 正弦模型， vibrato, phonetic modeling 颤音，语音建模	10 min singing records 10 分钟歌唱唱片	/	

https://www.sp.nitech.ac.jp/
~k-saino/music/ (accessed on

2. H End-to-end Generation

H 端到端的一代

Jukebox [83], proposed by the OpenAI team, is among the first to explore complete song generation. Rather than generating complete songs based entirely on text, Jukebox generates complete songs by modeling the raw audio domain while providing a way to use lyrics to control the generated content. It uses a multi-layered VQ-VAE architecture capable of compressing audio into discrete spaces while retaining as much musical information as possible. Jukebox uses an encoder-decoder model to implement conditional control of lyrics and uses the NUS AutoLyricsAlign tool to align lyrics and music. In addition to lyrics, Jukebox also allows users to control artists and genres.

由 OpenAI 团队提出的 Jukebox [83] 是首批探索完整歌曲生成的项目之一。Jukebox 不是完全基于文本生成完整的歌曲，而是通过对原始音频域进行建模来生成完整的歌曲，同时提供了一种使用歌词来控制生成内容的方法。它使用多层 VQ-VAE 架构，能够将音频压缩到离散空间，同时尽可能多地保留音乐信息。Jukebox 使用编码器 - 解码器模型来实现歌词的条件控制，并使用 NUS 自动歌词对齐工具来对歌词和音乐进行对齐。除了歌词，Jukebox 还允许用户控制艺术家和流派。

As a representative of commercial projects for end-to-end text-to-song tasks, Suno(https://suno.com/,accessedon17February2025)iscurrentlyoneofthemostinfluentialsoftware.Itiscapableofgeneratingcompletesongswithlyricsvianaturallanguagedescriptions,oritcanusenaturallanguagedescriptionstocontrolthegenerationofaccompanimentontheconditionthatlyricsareprovided.Itusesheuristicsforaudiotokenization andthetransformerarchitecture,butitisanunofficialopen-sourceprojectnow(Relevantcontentreferencedfromthepodcast:https://www.latent.space/p/suno,accessedon17February2025).Theteam’sotheropen-sourceprojectisatext-to-audiogenerationmodelcalledBark(https://github.com/suno-ai/bark,accessed on 17 February 2025), which is ca-pable of generating near-human-level speech and can be used to generate music by adding tokens. This project’s excellence in text-to-audio generation also laid the groundwork for the creation of Suno.

作为端到端文本到歌曲任务的商业项目代表，Suno (https://Suno.com/，accessedon17 february2025) 是目前最具影响力的软件之一。它可以通过自然语言描述生成完整的歌曲，也可以通过自然语言描述来控制歌词的生成。它使用音频标记和变压器架构，但现在是一个非官方的开源项目 (Relevantcontentreferencedfromthepodcast: https://www.latent.space/p/suno，accessedon17 february2025)。Theteam 的另一个开源项目是 text-to-audiogenerationmodelcalledbark (https://github.com/suno-ai/bark，于 2025 年 2 月 17 日访问) ，它能够生成接近人类水平的语音，并可以通过添加令牌来生成音乐。这个项目在文本到音频生成方面的卓越表现也为 Suno 的创建奠定了基础。

Recently, the ByteDance team proposed Seed-Music [84], a multi-modal music genera-tion end-to-end large model (Figure 15). This is a comprehensive framework designed to generate high-quality music through fine-grained style control. It integrates autoregressive language modeling and diffusion methods to support two key workflows: controlled music generation and post-editing. The controlled generation workflow harmonically unified vocals and accompaniment (accompaniment in MIDI format) to be created through multimodal inputs (e.g., lyrics, stylistic descriptions, audio references, scores, and voice cues), providing a high degree of customization and adaptability. For another thing, post-production editing features enable users to interactively modify elements of existing music tracks, including vocal lyrics, melody, and timbre.

最近，ByteDance 团队提出了 Seed-Music [84] ，一个多模态音乐生成端到端的大型模型 (图 15)。这是一个全面的框架，旨在通过细粒度的风格控制生成高质量的音乐。它集成了自回归语言建模和扩散方法，支持两个关键工作流程： 受控音乐生成和后期编辑。受控生成工作流程通过多模态输入 (例如歌词，风格描述，音频参考，分数和语音提示) 创建和谐统一的人声和伴奏 (MIDI 格式的伴奏) ，提供高度的定制和适应性。另一方面，后期制作编辑功能使用户能够交互式地修改现有音乐轨道的元素，包括歌词、旋律和音色。

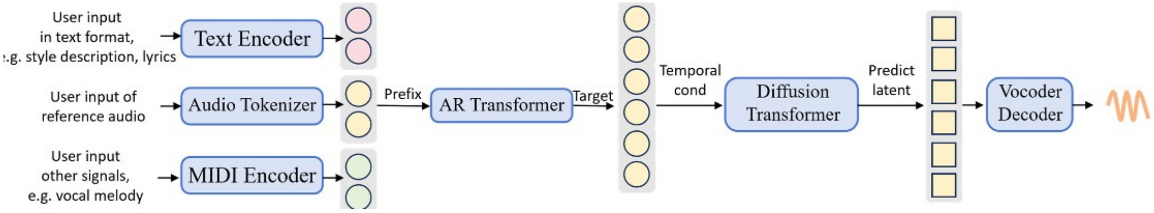


Figure 15. Frameworks of Seed-Music. Based on Transformer and Diffusion reproduced from [84].

图 15 种子音乐的框架，基于 Transformer 和 Diffusion 转载自 [84]。

Currently, the research on text-to-music generation has expanded from pure audio generation to more complex and comprehensive tasks, and complete song generation is a great challenge with a generation process that integrates various types of tasks. In the future, with the development of multimodal large models as well as generative models, it is possible to provide richer contexts and details for song generation and to further improve the quality as well as the diversity of generations. Relevant studies are summarized in Table 9.

目前，文本到音乐的生成研究已经从单纯的音频生成扩展到更复杂、更全面的任务，而完整的歌曲生成是一个融合了多种类型任务的生成过程，具有很大的挑战性。未来，随着多模态大模型和生成模型的发展，有可能为歌曲生成提供更丰富的语境和细节，进一步提高生成质量和生成多样性。表 9 总结了相关研究。

Table 9. Complete Song Generation Methods.

表 9。完整的歌曲生成方法。

Task Type 任务类别	Model 模型 Name 名称	Year 年份	Music 音乐 Representation 代表	Model 模型 Architecture 架构	Framework 框架 Category 类别		Description 描述	Large Model 大型模型 Relevance 相关性	Dataset Name 数据集名称	Generated 生成 Music 音乐 Length 长度	Accessed Link 访问链接
Staged 分阶段 Genera	Melodist 旋律家 [82]	2024	Waveform 波形	Transformer Transformer 变压	√	Tri-Tower 三塔 Contra	Using LLM to 使用 LLM 来	5 k Chinese 5k 中文 songs	Not 不 mentioned	https://text2 Https://text2	

tion一代	[82]			器		stive对比 Pre-training, 培训前， Cross-Modality交叉培训 Information信息 Matching, Lyrics and匹配，歌词和 Prompt-based	generate natural自然生成 language prompts语言提示	with歌曲 attributes属性 (180h)(180小时)	提到	songMelodist.github.songMelodist.githubsongMelodist.github. io/Sample/图片来源：io/Sample/ (accessed on 17(访问：17 Februa	

						基于提示符				ry 2025). 2025年2月)。	
Seed-Music Seed-Music 种子 - 音乐 [84] [84]	2024	Waveform & 波形 & amp MIDI MIDI 音频接口	Transformer& 变形金刚 & Diffusion 扩散		Multi-modal Inputs, 多模态输入， Auto-regressive 自回归模型 Language Modeling, 语言建模， Vo	Large multi-modal 大型多模态 language models 语言模型 for understanding 用于理解 and generati	Not mentioned 未提及	Varied 多种多样 Length 长度	https://team. Https://team. doubao.com/en/ 豆宝网 /en/ special/ seed-music 特别 / 种子音乐		

					coder Latents, 潜在的声码器， Zero-shot Singing Zero-shot Singing 零拍子歌唱 Voice Conversion 声音转换	on 以及一代			(accessed on 17 (访问：17 February 2025). 2025 年 2 月)。		
end-to-end 端到端 Ge	Suno AI Suno AI 太阳人工	2023	Waveform 波形	Transformer Transformer	√	Heuristic method, 启发	/	Not mentioned 未提	<= 4min 4 分钟	https://alpha.suno.ai/ 来	

neration一代	智能			变压器		法， Audio Tokenization, 音频标记， Zero threshold零阈值 for use供使用		及		源：https://alpha.suno.ai/ (accessed on 17February 2025). (于2025年2月17日检索)。	
	Jukebox点唱机 [83] [83]	2020	Waveform波形	VQ-VAE + VQ-VAE + VQ-VAE + Transf	√	Multiscale VQ-VAE, 多尺度VQ-VAE, Autoregres	Trained on Broad训练有素 Data at Scale Data	1.2M English 1.2m 英语 songs with 歌曲	Multiple多个 minutes分钟	https://jukebox. Https://jukebox. openai.com/	

				ormer Transformer变压器		sive自回归 Transformer, 变压器， Conditional有条件的 Generation, 一代， Hierarchical等级制度 Modeling建模	at Scale规模数据	attributes属性 (600k hours)(600khours)(600khours)		Openai.com/openai.com/ (accessed on 17 (访问：17 February 2025). 2025年2月)。	

4.3. Comments on Existing Techniques

4.3. 对现有技术的评论

4.3.1. Comparative Analysis of Symbolic and Audio Domains

4.3.1. 符号域和音频域的比较分析

1. L+Editability

***1. l + e* 可转换性**

The results generated by symbolic domain methods are editable. Based on the proper-ties of the symbolic representation, the generated results are discrete sequences of symbols. When the user is not satisfied with the generated result, or when further editing is needed to produce music for MIDI files, professional music production software (e.g., Cubase, FL studio, etc.) can be utilized to conveniently modify the pitch, intensity, and duration of each note; for text format or other symbol format files, there is also specialized software for parsing. In fact, in the traditional music production process, the symbol format file itself is an intermediate product, not the final result. After the symbolic format file is generated, it often needs tone rendering or real instrument recording to obtain the final work. However, it plays a decisive role in determining the melody of the music and the way the multiple voices work together, and it is an indispensable format for the music production process. The main purpose of symbolic domain music generation is to assist the musician in creating the music, not to generate a complete work.

符号域方法生成的结果是可编辑的。根据符号表示的性质，生成的结果是离散的符号序列。当用户对生成的结果不满意时，或者当需要进一步编辑 MIDI 文件来制作音乐时，可以利用专业的音乐制作软件 (如 Cubase、 FL studio 等) 来方便地修改每个音符的音高、强度和持续时间；对于文本格式或其他符号格式文件，也有专门的解析软件。事实上，在传统的音乐制作过程中，符号格式文件本身是一个中间产品，而不是最终结果。符号格式文件生成后，往往需要进行音调渲染或真实乐器录音才能获得最终作品。然而，它对音乐的旋律和多种声音的协同作用起着决定性的作用，是音乐创作过程中不可或缺的一种形式。符号域音乐生成的主要目的是辅助音乐家进行音乐创作，而不是生成一部完整的作品。

At present, there are also some symbolic domain music generation projects that incorporate the post-adjustment process into the project as well. For example, NetEase Cloud’s “NetEase Tianyin” allows users to adjust the pitch and timbre of the melody after generating it; AIVA’s “AIVA” also integrates melody-editing functions.

目前，也有一些符号域音乐生成项目将事后调整过程纳入到项目中。例如，网易云的 “网易天音” 允许用户在生成后调整旋律的音高和音色；AIVA 的 “AIVA” 也集成了旋律编辑功能。

Audio domain methods generate complete audio as a result, and it is not possible to edit the melody, rhythm, etc., of a specific part. The audio can only be processed by post-production software, such as noise reduction and de-reverberation. The post-processing of audio essentially “destroys” the original waveform or spectrum rather than “adjusting” it, and once the waveform or spectrum is distorted by the processing, it cannot be restored to its initial state. Adjustment of audio involves mixing and mastering in the music production process, and the purpose of these adjustments is to make the music work more harmonious and sound better. Although there is software (such as Suno) that can “edit” a segment, it is essentially a regeneration rather than an edit of the original audio.

音频域方法生成完整的音频，无法对特定部分的旋律、节奏等进行编辑。音频只能通过后期制作软件进行处理，如降噪、去混响等。音频的后处理实质上是对原始波形或频谱的 “破坏” 而不是 “调整”，一旦处理后的波形或频谱失真，就无法恢复到原始状态。音响调整是音乐制作过程中的混音和掌握，这些调整的目的是使音乐作品更加和谐，音质更好。虽然有软件 (如 Suno) 可以 “编辑” 片段，但它本质上是对原始音频的再生而不是编辑。

2. Expressiveness

表现力

The symbolic domain approach regulates expressiveness by explicitly controlling note parameters (e.g., tempo, duration, chords). However, discrete sequences of symbols are limited in the information they can store. Their lack of acoustic details means that the system has to rely heavily on the Renderer’s ability to generate nuanced acoustic characteristics for musical performance. In the traditional music production process, the dynamic information of real instruments or vocals cannot be perfectly reproduced, even with the use of high-quality sound sources and through professional production. The generated results are even less effective. In addition, current symbol domain methods make it difficult to model complex rhythmic transformations. Most of the notes generated by current methods are basic notes such as integer time-valued notes and dotted notes. Although ChatMusician has been able to generate simple ornamental notes, the effect is still far from real music if it is not processed or played.

符号域方法通过明确地控制音符参数 (例如，节奏、时值、和弦) 来调节表现力。然而，离散的符号序列能够存储的信息是有限的。它们缺乏声学细节意味着系统必须严重依赖渲染器的能力来产生音乐表演的微妙的声学特征。在传统的音乐制作过程中，即使使用高质量的声源，通过专业的制作，也无法完美地再现真实乐器或人声的动态信息。所生成的效果更是差强人意。此外，现有的符号域方法难以对复杂的韵律变换进行建模。现有方法生成的音符大多为基本音符，如整数时值音符和点音符。尽管 ChatMusician 已经能够生成简单的装饰音符，但是如果不进行处理或播放，其效果仍然与真正的音乐相去甚远。

Since the audio domain methods directly model waveform and spectrogram, which are trained using real music audio, it is able to preserve the details (e.g., guitar harmonies, vocal breath) and dynamics of real instruments. In terms of sound quality, models such as JEN-1 and Mousai have been able to generate 48 kHz two-channel audio comparable to the sound quality of real recordings. High-fidelity timbre and dynamics are the features and strengths of the audio domain approach.

由于音频域方法直接对波形和声谱图进行建模，并使用真实的音乐音频进行训练，因此能够保留真实乐器的细节 (如吉他和声、人声呼吸) 和动态特性。在声音质量方面，jen-1 和 Mousai 等型号已经能够产生 48khz 的双通道音频，与真实录音的声音质量相当。高保真的音色和动态性是音频领域方法的特点和优势。

3. Integrity

完整性

The symbolic domain excels in structural integrity, mainly due to its discrete data representation and explicit control of musical rules. The symbolic model is able to accurately capture timing relationships between notes, such as beat alignment, chord transitions, and motive development. Through the use of templates or rule constraints, music generated by symbolic domains is able to maintain clear segmentation(e.g.,intro, chorus) and repetition patterns (e.g., ABA structures) and is also often harmonically plausible. For example, ChatMusician can generate classical music with complex weave structures.

符号领域在结构完整性方面表现突出，主要是由于其离散的数据表示和对音乐规则的明确控制。符号模型能够准确捕捉音符之间的时序关系，如节拍对齐、和弦转换、动机发展等。通过使用模板或规则约束，由符号域生成的音乐能够保持清晰的分割 (例如，介绍，

合唱) 和重复模式 (例如，ABA 结构) ，并且通常也是合理的和声。例如，ChatMusician 可以生成具有复杂编织结构的古典音乐。

The audio domain is relatively weak in terms of structural integrity, limited primarily by the difficulty of modeling high-dimensional continuous signals. Although audio models are capable of generating natural and smooth audio, they are prone to rhythmic breaks and thematic deviations in long sequence generation. The structural control of the audio domain usually relies on textual cues (e.g., “repeating chorus”), but the lack of explicit segmentation and repetition mechanisms leads to the generation of compositions that do not perform as well as the symbolic domain in terms of long-term dependencies. Meanwhile, the audio domain is weak in multi-track synergy, with post-generation mixing (e.g,Jukebox) and one-shot generation (e.g.,Noise2Music), which can easily lead to vocal conflicts.

音频领域在结构完整性方面相对较弱，主要受限于高维连续信号建模的难度。虽然音频模型能够生成自然和平滑的音频，但是它们在长序列生成中容易出现节奏中断和主题偏差。音频域的结构控制通常依赖于文本线索 (如 “重复合唱”) ，但缺乏明确的分段和重复机制导致产生的作品在长期依赖性方面表现不如符号域。与此同时，音频领域在多声道协同方面较弱，存在后生成混音 (如自动点唱机) 和一次性生成 (如噪音 2music) ，容易导致声乐冲突。

While the structural coherence of generated results in both the symbolic and audio domains decreases with duration, the long-term modeling capabilities of symbolic-domain generation methods are much higher than those of audio-domain generation methods due to the high compression ratios of information stored in symbolic data, e.g., as early as 2019, deep-learning-based MuseNet (Payne, Christine. “MuseNet”. OpenAI, ope-nai.com/blog/musenet, accessed on 25 April 2019) will be able to generate scores up to4 min long, and the same length is already almost the current upper limit of audio domain music generation.

虽然符号域和音频域生成结果的结构一致性随时间延长而降低，但由于符号数据中存储信息的高压缩比，符号域生成方法的长期建模能力远高于音频域生成方法，例如早在 2019 年，基于深度学习的 MuseNet (Payne，Christine。“ MuseNet”。OpenAI，ope-nai.com/blog/musenet，2019 年 4 月 25 日访问) 将能够生成长达 4 分钟的乐谱，同样的长度已经几乎是目前音频域音乐生成的上限。

In terms of the integrity of the work, the result of the symbolic domain is not a complete musical work, and further production is required, while the result of the audio domain is already a complete piece of music or song, which reduces the difficulty for the public to create music.

在作品的完整性方面，符号域的结果不是完整的音乐作品，需要进一步制作，而音频域的结果已经是完整的音乐或歌曲，降低了公众创作音乐的难度。

4. Data Efficiency

4. 数据效率

The symbolic domain demonstrates significant advantages in data efficiency. Due to its highly structured representations (e.g.,MIDI, ABC Notation), symbolic data achieve far higher information compression ratios than raw audio. For instance, Yuan et al. quantified that ABC Notation exhibits 2.6× and 43× higher tokens per song (Tok./Song) and tokens per second (Tok./Sec.) compared to MIDI and WAV formats, respectively, enabling efficient musical encoding with minimal data volume and accelerated processing [45].

符号域在数据效率方面显示出显著的优势。由于其高度结构化的表示 (如 MIDI，ABC 符号) ，符号化数据比原始音频获得了更高的信息压缩比。例如，Yuan 等人量化了 ABC 记数法每首歌曲的标记数分别高出 2.6 倍和 43 倍 (Tok。/Song) 和每秒令牌 (Tok。/ 秒)与 MIDI 和 WAV 格式相比，分别实现了最小数据量和加速处理的高效音乐编码 [45]。

However, symbolic domains face challenges in acquiring high-quality datasets. Most existing datasets rely on automated MIDI transcription, which introduces pitch drift and rhythmic errors, while professional manual annotation remains prohibitively expensive, limiting model generalizability. In contrast, audio domains leverage vast streaming plat-forms for dataset construction, yet still require professional studio recordings and prepro-cessing (e.g., source separation, alignment) for specialized tasks like singing voice synthesis.

然而，符号域在获取高质量数据集方面面临挑战。大多数现有的数据集依赖于自动 MIDI 转录，这引入了音高漂移和节奏错误，而专业的手动注释仍然非常昂贵，限制了模型的普遍性。相比之下，音频领域利用巨大的流媒体平台来构建数据集，但仍然需要专业的录音室录音和预处理 (例如，源分离，对齐) 来完成像歌唱声音合成这样的专业任务。

Despite disparities in data efficiency, audio domains are rapidly advancing in real-time generation. Early models like Jukebox required 9 h to generate 1-min audio, whereas recent approaches (e.g.,MeLoDy) achieve faster-than-real-time synthesis on a single V100 GPU through model lightweighting and algorithmic optimizations. Continued advancements in computational power and architectural innovations will further bridge the efficiency gap between audio and symbolic generation.

尽管在数据效率方面存在差异，但音频领域在实时生成方面正在迅速发展。像 Jukebox 这样的早期模型需要 9 小时才能生成 1 分钟的音频，而最近的方法 (例如 MeLoDy) 通过模型轻量化和算法优化在单个 V100 GPU 上实现比实时合成更快的合成。计算能力和架构创新的持续进步将进一步弥合音频和符号生成之间的效率差距。

4.3.2. Critical Evaluation of Technical Approaches
4.3.2 技术方法的批判性评价

The field of text-to-music generation has made great progress over the years, with advancements in rule-based systems, statistical models, generative approaches, and LLMs. However, each method has its strengths and limitations. Challenges such as data depen-dency, model controllability, and generalization remain significant. This section reviews these techniques and highlights the key issues that require attention.

随着基于规则的系统、统计模型、生成方法和 llm 的发展，文本到音乐的生成领域近年来取得了巨大的进步。然而，每种方法都有其优势和局限性。诸如数据依赖性、模型可控性和泛化性等挑战依然显著。本节回顾这些技术并强调需要注意的关键问题。

1. Rule-Based and Template Methods

基于规则和模板的方法

Rule-based and template-driven methods are among the earliest approaches in text-to-music generation. These methods follow predefined musical rules, such as chord progres-sions or rhythmic patterns, to generate melodies. Their simplicity and reliability make them highly interpretable and consistent, useful in structured applications like educational tools or composition guides. However, their deterministic nature severely limits their ability to adapt to the diversity and complexity of real-world music. For example, such methods struggle to create dynamic and expressive outputs when handling lyrics with varying emotional tones. While they are useful for tasks requiring fixed structures, their lack of creativity makes them unsuitable for tasks that demand innovation and diversity.

基于规则和模板驱动的方法是最早的文本到音乐的生成方法之一。这些方法遵循预定义的音乐规则，如和弦进度或节奏模式，以生成旋律。它们的简单性和可靠性使它们具有高度的可解释性和一致性，在教育工具或作曲指南等结构化应用中非常有用。然而，它们的确定性本质严重限制了它们适应现实世界音乐的多样性和复杂性的能力。例如，当处理带有不同情感基调的歌词时，这些方法很难创造出动态的和富有表现力的输出。虽然它们对需要固定结构的任务很有用，但是缺乏创造力使得它们不适合需要创新和多样性的任务。

2. Statistical Models

统计模型

Statistical approaches, such as Markov chains and n-gram models, introduced ran-domness that improved over rule-based systems. By analyzing patterns in training data, these models can generate melodies that exhibit some variability and complexity. However, they cannot capture long-term dependencies, which are crucial for creating coherent music. For instance, while a Markov chain might produce locally plausible note sequences, it often fails to generate melodies with meaningful global structures. Additionally, statistical models are highly dependent on the quality of the training data, often overfitting to specific patterns and failing to generalize to new musical styles or datasets. This limits their ability to produce diverse and innovative outputs across broader applications.

统计方法，如马尔可夫链和 n-gram 模型，引入了随机性，这比基于规则的系统有所改进。通过分析训练数据中的模式，这些模型可以生成表现出一些可变性和复杂性的旋律。然而，它们不能捕捉长期的依赖关系，而这对于创造连贯的音乐是至关重要的。例如，虽然马尔可夫链可能产生局部合理的音符序列，但它往往无法产生具有有意义的全局结构的旋律。此外，统计模型高度依赖于训练数据的质量，往往过度拟合特定的模式，无法推广到新的音乐风格或数据集。这限制了他们在更广泛的应用中产生多样化和创新性产出的能力。

3. Generative Models

生成模型

Generative models represent a significant leap in text-to-music generation, offering powerful tools for creating realistic, diverse, and dynamic outputs. GANs are one of the most prominent methods, leveraging an adversarial framework where a generator produces melodies and a discriminator evaluates their quality. Models like LSTM-GANs combine sequential modeling with GANs to improve the coherence of generated music. However, GANs often face issues such as mode collapse, where the generator produces limited variations and instability during training, making it challenging for them to optimize effectively. Furthermore, GANs generally lack fine control over specific musical attributes, which limits their use in tasks requiring precise alignment with textual inputs.

生成模型代表了文本到音乐生成的一个重大飞跃，为创建现实的、多样的和动态的输出提供了强大的工具。GANs 是最突出的方法之一，利用对抗框架，生成器生成旋律，鉴别器评估旋律质量。像 LSTM-GANs 这样的模型将顺序建模与 GANs 结合起来，以提高生成

音乐的连贯性。然而，gan 经常面临模式崩溃等问题，其中生成器在训练过程中产生有限的变化和不稳定性，使得它们难以进行有效的优化。此外，gan 通常缺乏对特定音乐属性的精细控制，这限制了它们在需要与文本输入精确对齐的任务中的使用。

In addition to GANs, Variational Autoencoders (VAEs) have gained attention for their ability to learn structured latent representations of music. VAEs map input data (e.g, melodies) to a latent space, allowing for smooth interpolation between musical features and generating new, coherent outputs. Their probabilistic framework ensures stable training and enables control over the diversity of generated melodies. However, VAEs tend to produce less sharp or vivid outputs than GANs, which can affect the perceptual quality of the music.

除了 gan 之外，变分自动编码器 (vae) 因其学习音乐的结构化潜在表示的能力而受到关注。Vae 将输入数据 (例如旋律) 映射到潜在空间，允许在音乐特征之间进行平滑插值并产生新的连贯输出。他们的概率框架确保稳定的训练，并能够控制生成的旋律的多样性。然而，与 gan 相比，vae 往往产生不那么尖锐或生动的输出，这会影响音乐的感知质量。

More recently, diffusion models have emerged as a powerful alternative in genera-tive tasks, including music generation. These models progressively transform noise into structured outputs through a reverse diffusion process guided by a learned probability distribution. Diffusion models excel at generating high-quality outputs with precise control over attributes, making them suitable for tasks that require both diversity and coherence.

最近，扩散模型已经成为生产任务 (包括音乐生成) 中的一个强大的替代方案。这些模型通过学习概率分布指导的反向扩散过程逐步将噪声转换为结构化输出。扩散模型擅长生成精确控制属性的高质量输出，使其适合需要多样性和一致性的任务。

For example, text-to-music diffusion models can effectively map lyrics to melodies by cap-turing nuanced relationships in a step-by-step generation process. While these models are computationally intensive and require careful tuning, they have demonstrated significant potential in addressing the limitations of earlier generative approaches.

例如，文本到音乐的传播模型可以有效地将歌词映射到旋律，通过逐步生成过程中的 cap-turing 微妙关系。虽然这些模型是计算密集型的，需要仔细调整，但它们在解决早期生成方法的局限性方面已经显示出巨大的潜力。

4. Transformer-Based Models

4. 基于变压器的模型

The introduction of Transformer-based models has significantly advanced music generation by addressing many of the limitations of earlier methods. Transformers excel in modeling long-range dependencies and capturing intricate relationships between musical elements, such as aligning lyrics with melodies. Models like SongMASS leverage these capabilities by using separate encoders and decoders for lyrics and melody, combined with pre-training techniques to improve generation quality. These models effectively handle the complexity of musical structures by focusing on parallel processing and self-attention mechanisms, enabling more coherent and contextually aligned outputs.

Transformer-based 模型的引入解决了早期方法的许多局限性，极大地提高了音乐生成的质量。Transformer 擅长模拟长距离的依赖关系，捕捉音乐元素之间复杂的关系，比如歌词和旋律的对齐。像 SongMASS 这样的模型通过对歌词和旋律使用独立的编码器和解码器来利用这些能力，结合预训练技术来提高生成质量。这些模型通过关注并行处理和自我注意机制，有效地处理了音乐结构的复杂性，使得输出结果更加连贯和符合上下文。

In addition to task-specific Transformer models, language models based on the Trans-former architecture have been adapted for music generation tasks. These models learn representations of sequences, whether text or symbolic music and can generate music by treating it as a language. For instance, in some approaches, musical notes and rhythms are tokenized into sequences akin to words in natural language, allowing the models to predict the next note or phrase based on the preceding context. This adaptation of language models provides a flexible and scalable framework for melody generation, where the system can leverage transfer learning from vast datasets of natural language or symbolic music.

除了特定任务的 Transformer 模型之外，基于 Transformer 架构的语言模型已经适用于音乐生成任务。这些模型学习序列的表示，无论是文本还是符号音乐，并且可以通过将其视为语言来生成音乐。例如，在某些方法中，音符和节奏被标记为类似于自然语言中的单词的序列，从而允许模型根据前面的上下文预测下一个音符或短语。这种语言模型的适应性为旋律生成提供了一个灵活的、可扩展的框架，在这个框架中，系统可以利用来自自然语言或符号音乐的大量数据集的迁移学习。

Despite these strengths, Transformer-based architectures and language models share several limitations. They are computationally intensive, requiring large amounts of memory and processing power. Their performance is also highly sensitive to the quality and diversity of training data, which can restrict their ability to handle underrepresented musical styles or genres. Additionally, self-attention mechanisms, while powerful, may struggle with extremely long sequences, a common challenge in music generation tasks. To address these issues, some researchers have introduced hierarchical Transformer architectures that split sequences into smaller, more manageable chunks, improving computational efficiency without sacrificing output quality.

尽管有这些优势，基于 transformer 的架构和语言模型共享一些限制。它们是计算密集型的，需要大量的内存和处理能力。它们的性能也对训练数据的质量和多样性高度敏感，这会限制它们处理未被充分代表的音乐风格或流派的能力。此外，自我注意机制，虽然强大，可能会与极长的序列斗争，在音乐生成任务中的常见挑战。为了解决这些问题，一些研究人员引入了分层 Transformer 架构，将序列分割成更小、更易管理的块，在不牺牲输出质量的情况下提高了计算效率。

Integrating language models into music generation represents a bridge between nat-ural language processing and musical creativity. While these models effectively handle sequential data and enable creative outputs, they often require fine-tuning and additional preprocessing to adapt to the unique characteristics of music data, such as time signa-tures and harmonic progressions. Despite these challenges, language models have opened new possibilities for using textual descriptions to guide music generation, providing a foundation for more advanced systems.

将语言模型整合到音乐生成中，是自然语言处理和音乐创作之间的桥梁。虽然这些模型可以有效地处理顺序数据并实现创造性输出，但它们往往需要进行微调和额外的预处理，以适应音乐数据的独特特征，例如时间特征和和声级数。尽管存在这些挑战，语言模型为使用文本描述来指导音乐生成打开了新的可能性，为更先进的系统提供了基础。

5. Large Language Models

大型语言模型

Large language models, such as SongComposer, represent the latest advancements in text-to-music generation. LLMs are highly flexible and capable of learning complex patterns from vast datasets, enabling them to produce high-quality music that aligns closely with textual inputs. Unlike traditional methods that rely on fixed rules or templates, LLMs learn to infer relationships between lyrics and melodies in a data-driven manner. This allows them to handle tasks like melody generation, lyrics-melody alignment, and even song continuation with impressive results. However, LLMs also face challenges. Their reliance on massive computational resources makes them less accessible for smaller-scale applications. Additionally, they lack fine-grained control over specific musical attributes, which can lead to outputs that deviate from user expectations. The “black-box” nature of LLMs also makes it difficult to interpret their decisions, which can be problematic in applications requiring transparency or adherence to strict musical guidelines.

大型语言模型，如 SongComposer，代表了文本到音乐生成的最新进展。Llm 非常灵活，能够从庞大的数据集中学习复杂的模式，使它们能够生成高质量的音乐，与文本输入紧密结合。与依赖于固定规则或模板的传统方法不同，llm 学习以数据驱动的方式推断歌词和旋律之间的关系。这使得他们能够处理像旋律生成，歌词 - 旋律对齐，甚至歌曲延续这样的任务，并获得令人印象深刻的结果。然而，llm 也面临着挑战。它们对于大规模计算资源的依赖使得它们对于小规模应用程序来说更加难以访问。此外，它们缺乏对特定音乐属性的细粒度控制，这可能导致输出偏离用户期望。Llm 的 “黑匣子 “性质也使得解释他们的决定变得困难，这在需要透明度或严格遵守音乐准则的应用程序中可能是有问题的。

A key challenge across all methods is balancing creativity and control. Rule-based and statistical methods excel in providing structure and interpretability but fail to produce diverse and expressive music. On the other hand, generative models, such as GANs, VAEs, and diffusion models, along with LLMs, offer unparalleled creativity and flexibility but often lack fine control over the outputs. Another significant issue is the dependency on high-quality datasets. Many models require extensive, diverse, and well-annotated training data to perform well, yet such datasets are often scarce, especially for underrepresented musical styles or languages. This limitation hinders the generalization of these models to broader and more diverse applications.

所有方法的一个关键挑战是平衡创造力和控制。基于规则和统计的方法擅长于提供结构和可解释性，但无法产生多样化和富有表现力的音乐。另一方面，生成模型，如 gan、vae 和扩散模型，以及 llm，提供了无与伦比的创造性和灵活性，但通常缺乏对输出的精细控制。另一个重要的问题是对高质量数据集的依赖。许多模型需要广泛的、多样的和注释良好的训练数据才能表现良好，然而这样的数据集往往是稀缺的，特别是对于代表性不足的音乐风格或语言。这种限制阻碍了这些模型推广到更广泛和更多样化的应用。

To address these challenges, future research should focus on hybrid approaches that combine the strengths of different methods. For example, integrating rule-based templates with Transformer-based architectures could provide better control over specific musical features while retaining the flexibility of deep learning models. Similarly, LLMs could be enhanced with interpretable mechanisms or user-guided controls to improve alignment with specific requirements. Exploring self-supervised learning and transfer learning techniques could also help mitigate the dependency on large labeled datasets, making models more versatile and adaptable across diverse scenarios.

为了应对这些挑战，未来的研究应该集中在结合不同方法优势的混合方法上。例如，将基于规则的模板与基于 transformer 的架构相结合，可以更好地控制特定的音乐特征，同时保留深度学习模型的灵活性。类似地，llm 可以通过可解释的机制或用户引导的控件来增强，以改善与特定要求的一致性。探索自我监督学习和迁移学习技术也可以帮助减轻对大型标记数据集的依赖，使模型在不同的场景中更加通用和适应。

5. Frameworks

5. 框架

Text-to-music generation frameworks are evolving rapidly, driven by advancements in LLMs. This section categorizes existing approaches into three paradigms based on their integration of LLMs: Traditional Rule-Driven Frameworks, Hybrid LLM-Augmented Frameworks, and End-to-End LLM-Centric Frameworks. Each paradigm addresses distinct challenges in semantic-text-to-music alignment, controllability, and scalability, offering unique trade-offs between interpretability and generative flexibility. This framework taxon-omy serves as a technical summary of the task-oriented methods discussed in Section 4. This dual classification (task + technical framework) aims to help readers simultaneously grasp the application scenarios and technical evolution of the methods. By organizing models in this way, we provide a comprehensive understanding of how different they are.

在 llm 的推动下，文本到音乐的生成框架正在迅速发展。本节根据 llm 的集成将现有的方法分为三种范例： 传统的规则驱动框架、混合的 llm 增强框架和端到端的以 llm 为中心的框架。每个范例都解决了语义 - 文本 - 音乐对齐、可控性和可伸缩性方面的不同挑战，在可解释性和生成灵活性之间提供了独特的平衡。这个框架分类可以作为第 4 节中讨论的面向任务方法的技术总结。这种双重分类 (任务 + 技术框架) 旨在帮助读者同时掌握方法的应用场景和技术演进。通过以这种方式组织模型，我们可以全面了解它们之间的差异。

5.1. Traditional Learning-Based Frameworks

5.1 传统的基于学习的框架

Traditional learning-based frameworks in text-to-music generation typically rely on machine learning or deep learning models designed for sequence generation. These models treat music and text as sequences, using neural networks to capture relationships between the two modalities. By training on paired datasets of text and music, they aim to generate musical outputs that align with textual inputs. These methods usually employ encoder-decoder architectures or recurrent structures (e.g.,LSTM, RNN) to model dependencies within and across the modalities. The general pipeline for traditional methods in text-to-music generation can be divided into three main stages:

传统的基于学习的文本 - 音乐生成框架通常依赖于为序列生成设计的机器学习或深度学习模型。这些模型将音乐和文本视为序列，使用神经网络捕捉两种模态之间的关系。通过对文本和音乐的配对数据集进行训练，他们的目标是生成与文本输入一致的音乐输出。这些方法通常使用编码器 - 解码器架构或循环结构 (例如 LSTM，RNN) 来模拟模态内部和跨模态的依赖性。文本到音乐生成的传统方法的一般流水线可以分为三个主要阶段：

Text Encoding: The input text (e.g.,lyrics) is converted into numerical representations using embedding layers, capturing semantic and rhythmic information.

文本编码： 使用嵌入层将输入文本 (例如歌词) 转换为数值表示，捕获语义和节奏信息。

Sequence Generation: Deep learning models (e.g.,LSTM, RNN) generate musical sequences (e.g., melody, chords, or rhythm) based on the encoded text.

序列生成： 深度学习模型 (如 LSTM、 RNN) 基于编码文本生成音乐序列 (如旋律、和弦或节奏)。

Output Synthesis: The generated musical sequences are converted into symbolic music formats (e.g.,MIDI) or synthesized into audio.

输出合成： 生成的音乐序列被转换成符号音乐格式 (如 MIDI) 或合成成音频。

Case 1: TeleMelody

Case 1: TeleMelody 案例 *1: TeleMelody*

Ju et al. (2022) [38] proposed TeleMelody, a two-stage lyric-to-melody generation system that leverages a carefully designed template to bridge the gap between lyrics and melodies. This system decomposes the task into a lyric-to-template module and a template-to-melody module, significantly reducing the complexity of learning the corre-lation between lyrics and melodies. This template includes key musical elements such as tonality, chord progression, rhythm pattern, and cadence, which are extracted from melodies in a self-supervised manner. This approach allows for better controllability and data efficiency, as the system can generate melodies without requiring large amounts of paired lyrics-melody data. The template-to-melody module uses a Transformer-based model with alignment regularization, guided by musical knowledge, to ensure that the generated melodies align well with the template. This design enables users to control the generated melodies by adjusting the musical elements in the template, offering a more flexible and user-friendly approach to melody generation.

Ju 等人 (2022)[38] 提出了 TeleMelody，一个两阶段的歌词到旋律的生成系统，利用一个精心设计的模板来弥合歌词和旋律之间的差距。该系统将任务分解为歌词 - 模板模块和模板 - 旋律模块，大大降低了学习歌词与旋律相关性的复杂度。这个模板包括关键的音乐元素，如调性，和弦进程，节奏模式，和抑扬顿挫，这是从旋律中提取自我监督的方式。这种方法允许更好的可控性和数据效率，因为系统可以生成旋律，而不需要大量的成对歌词 - 旋律数据。模板到旋律模块使用基于 transformer 的对齐规则化模型，以音乐知识为指导，确保生成的旋律与模板对齐良好。这种设计允许用户通过调整模板中的音乐元素来控制生成的旋律，为旋律生成提供了一种更加灵活和用户友好的方法。

Traditional text-to-music generation methods, including deep learning and neural network-based models such as sequence-to-sequence models or Generative Adversarial Networks (GANs), suffer from several key limitations. Firstly, these methods typically rely on large amounts of high-quality paired lyrics-melody data for training. The scarcity of such data significantly restricts the model’s generalization ability, particularly in low-resource languages or niche musical genres. Secondly, the generated results often lack diversity and creativity: models tend to learn common patterns from the training data, leading to repetitive or overly conservative melodic structures that struggle to break free from established frameworks. Additionally, traditional methods exhibit weak user control-lability, lacking fine-grained adjustments for musical elements such as chord progressions, rhythmic variations, or emotional expression, making it difficult to meet the needs of professional music creation. Although some studies like TeleMelody have mitigated these issues by introducing templates or rules, they remain constrained by the trade-off between data dependency and flexibility.

传统的文本到音乐的生成方法，包括深度学习和基于神经网络的模型，如序列到序列模型或生成对抗网络 (GANs) ，受到几个关键的限制。首先，这些方法通常依赖于大量高质量的成对歌词 - 旋律数据进行训练。这类数据的稀缺性极大地限制了模型的泛化能力，尤其是在低资源语言或小众音乐流派中。其次，产生的结果往往缺乏多样性和创造性： 模型往往从训练数据中学习共同的模式，导致旋律结构重复或过于保守，难以摆脱既定框架。此外，传统方法的用户可控性较弱，缺乏对和弦进行、节奏变化、情感表达等音乐元素的

细粒度调整，难以满足专业音乐创作的需求。虽然像 TeleMelody 这样的研究通过引入模板或规则缓解了这些问题，但是他们仍然受到数据依赖性和灵活性之间的权衡的限制。

5.2. Hybrid LLM-Augmented Frameworks

5.2 混合 LLM-Augmented 框架

Hybrid approaches in text-to-music generation integrate LLMs as a core module alongside traditional sequence generation models. In this framework, the LLM plays a versatile role by processing text in various ways, such as extracting musical attributes, generating lyrics, or reconstructing descriptions. The LLM enriches the input text, which is fed into a subsequent music generation model (e.g.,LSTM, Transformer) to produce musical outputs. By acting as a powerful intermediary, the LLM helps bridge the gap between complex textual input and the generated music, ensuring better alignment and context preservation. The general pipeline for hybrid approaches can be summarized in the following stages:

文本到音乐生成的混合方法将 llm 作为核心模块与传统的序列生成模型相结合。在这个框架中，LLM 通过各种方式处理文本，如提取音乐属性、生成歌词或重建描述，扮演着多功能的角色。LLM 丰富了输入文本，并将其输入到后续的音乐生成模型 (如 LSTM、Transformer) 中以生成音乐输出。通过作为一个强大的中介，LLM 有助于弥合复杂的文本输入和生成的音乐之间的差距，确保更好的对齐和上下文保存。混合方法的一般流程可以总结为以下几个阶段：

●Text Encoding: With traditional methods;

●文本编码： 使用传统方法；

LLM Module: Extracts key semantic features and contextual information from the input text and generates new content, such as lyrics or expanded descriptions, based on the input;

LLM 模块： 从输入文本中提取关键语义特征和上下文信息，并根据输入生成新内容，如歌词或扩展描述；

Sequence Generation: With traditional methods;

序列生成： 使用传统方法；

Output Synthesis: With traditional methods. Sometimes, LLM is used to give feedback.

输出综合： 用传统的方法，有时用 LLM 给出反馈。

Case 2: MuseCoCo

Case 2: MuseCoCo 案例 2: MuseCoCo

MuseCoCo [18] is an innovative hybrid model for generating music from text descriptions. It combines the power of pre-trained language models (LLMs) with traditional sequence generation models to enhance text-to-music generation. In the MuseCoCo system, templates are pre-prepared for the LLM. For example, a template could be, “The music is imbued with [EMOTION]”. When the input prompt is “write a happy four-beat pop song’, the LLM extracts the relevant attributes, such as emotion and time signature, and refines the template by filling in values. This results in a description like”The music is imbued with [happiness] and the [4/4] time signature is used in the music. The genre of the music is [pop]“. These templates guide the generation process, ensuring the music aligns with the user’s description.

MuseCoCo [18] 是一个从文本描述生成音乐的创新混合模型。它将预先训练的语言模型 (llm) 的功能与传统的序列生成模型相结合，以增强文本到音乐的生成。在 MuseCoCo 系统中，模板是为 LLM 预先准备的。例如，一个模板可以是 “The music is imbueed with [EMOTION]”。当输入提示为 “写一首快乐的四拍流行歌曲” 时，LLM 提取相关属性，如情感和时间签名，并通过填入值来精炼模板。这会导致类似于 “音乐充满了 [快乐] ，并且在音乐中使用了 [4/4] 时间签名。音乐的类型是 [流行]”。这些模板指导生成过程，确保音乐符合用户的描述。

While integrating the LLM provides greater flexibility and control over the music generation process, it also introduces an additional layer of complexity. The model’s per-formance depends heavily on the quality of the LLM’s text processing and its ability to accurately extract or generate relevant musical attributes. Moreover, the system’s effective-ness is contingent on the quality of the attribute templates used to guide the generation process. Poorly defined or overly rigid templates can limit the system’s creativity and adaptability, preventing the generation of truly innovative or diverse music.

虽然整合 LLM 为音乐生成过程提供了更大的灵活性和控制，但它也引入了额外的复杂层。模型的性能在很大程度上取决于 LLM 文本处理的质量和它准确提取或生成相关音乐属性的能力。此外，系统的有效性取决于用于指导生成过程的属性模板的质量。定义不好或过于死板的模板会限制系统的创造力和适应性，阻碍真正创新或多样化音乐的生成。

5.3. End-to-End LLM-Centric Frameworks

5.3 端到端以 *llm* 为中心的框架

End-to-End LLM-Centric Frameworks treat music as a second language, applying sequence-based models, typically used in natural language processing (NLP), to generate music directly from text. In this approach, the LLM processes textual input (e.g.,lyrics, prompts, or descriptions) and generates corresponding musical elements (such as melody, rhythm, and harmony), considering these elements analogous to linguistic structures like words and sentences. This eliminates the need for separating music theory modules or templates, offering a unified framework for text-to-music generation. The general pipeline for end-to-end LLM-based systems in text-to-music generation consists of the following stages:

端到端以 llm 为中心的框架将音乐视为第二语言，应用基于序列的模型 (通常用于自然语言处理 (NLP)) 直接从文本生成音乐。在这种方法中，LLM 处理文本输入 (如歌词、提示或描述) 并生成相应的音乐元素 (如旋律、节奏和和声) ，考虑这些元素类似于语言结构，如单词和句子。这消除了分离音乐理论模块或模板的需要，为文本到音乐的生成提供了一个统一的框架。在文本到音乐的生成中，基于 llm 的端到端系统的一般流水线包括以下几个阶段：

●Text Encoding: With traditional methods;

●文本编码： 使用传统方法；

LLM Processing: The encoded text is processed by the language model, which treats music as a sequence similar to text. This model predicts the next musical element (e.g., note, rhythm, or harmony) based on the current context, generating a complete musical sequence in an iterative manner. In this stage, the LLM is able to use its extensive pre-trained knowledge of language and patterns to generate musically coherent sequences that align with the input description;

LLM Processing: 编码后的文本由语言模型处理，语言模型将音乐视为与文本类似的序列。该模型根据当前上下文预测下一个音乐元素 (例如，音符、节奏或和声) ，以迭代的方式生成一个完整的音乐序列。在这个阶段，LLM 能够使用其广泛的预先训练的语言和模式知识来生成与输入描述一致的音乐连贯序列；

Output Synthesis: Extract symbol information from textual music sequences and synthesize them.

输出合成： 从文本音乐序列中提取符号信息并合成它们。

Case 3: SongComposer

Case 3: SongComposer 案例 3: SongComposer

SongComposer [39] is a specialized LLM for generating lyrics and melodies directly from textual input. It is trained using a high-quality lyrics-melody pairing dataset, which fine-tunes the LLM to understand the relationship between lyrics and melody more ef-fectively. This fine-tuning step, along with the introduction of innovative encoding rules, enables the model to process melody sequences, ensuring that the generated music is both contextually appropriate and musically coherent. SongComposer operates by accepting a text description, generating both the lyrics and matching melody, including information like note pitch, duration, and rest duration. This dual output facilitates the generation of complete musical pieces and allows the extracted information to be used for further music creation.

SongComposer [39] 是一个专门的 LLM，用于直接从文本输入生成歌词和旋律。它使用一个高质量的歌词 - 旋律配对数据集进行训练，该数据集对 LLM 进行微调，以更有效地理解歌词和旋律之间的关系。这个微调步骤，以及创新编码规则的引入，使模型能够处理旋律序列，确保生成的音乐既适合上下文，又在音乐上连贯。SongComposer 通过接受文本描述进行操作，生成歌词和匹配的旋律，包括音高、持续时间和休息时间等信息。这种双重输出有助于生成完整的音乐片段，并允许提取的信息用于进一步的音乐创作。

LLM-based systems for text-to-music generation offer creativity and flexibility, as they can generate diverse and contextually relevant music across a wide range of genres, styles, and emotional tones. By integrating text processing and music generation into a single framework, these systems ensure a seamless alignment between the input text and the generated music. However, there are limitations, particularly in the lack of fine-grained control over musical features such as tempo, dynamics, and instrumentation. While LLMs can create highly creative and musically coherent compositions, achieving precise control over these elements is difficult. Additionally, these models are computationally intensive, requiring significant resources for both training and inference. Their performance is also computationally intensive, requiring significant resources for both training and inference.

基于 llm 的文本 - 音乐生成系统提供了创造性和灵活性，因为它们可以生成多种多样且与上下文相关的音乐，涵盖了广泛的流派、风格和情感基调。通过整合文本处理和音乐生成到一个单一的框架，这些系统确保了输入文本和生成的音乐之间的无缝对齐。然而，这些系统也存在局限性，特别是缺乏对音乐特征的细粒度控制，例如节奏、动态和乐器。虽然 LLMs 可以创造高度创造性和音乐连贯的作品，实现对这些元素的精确控制是困难的。

此外，这些模型是计算密集型的，需要大量的训练和推理资源。它们的性能也是计算密集型的，需要大量的训练和推理资源。

While LLMs can create highly creative and musically coherent compositions, achieving highly dependent on the quality and diversity of the training data, making them less effective for underrepresented genres or musical styles.

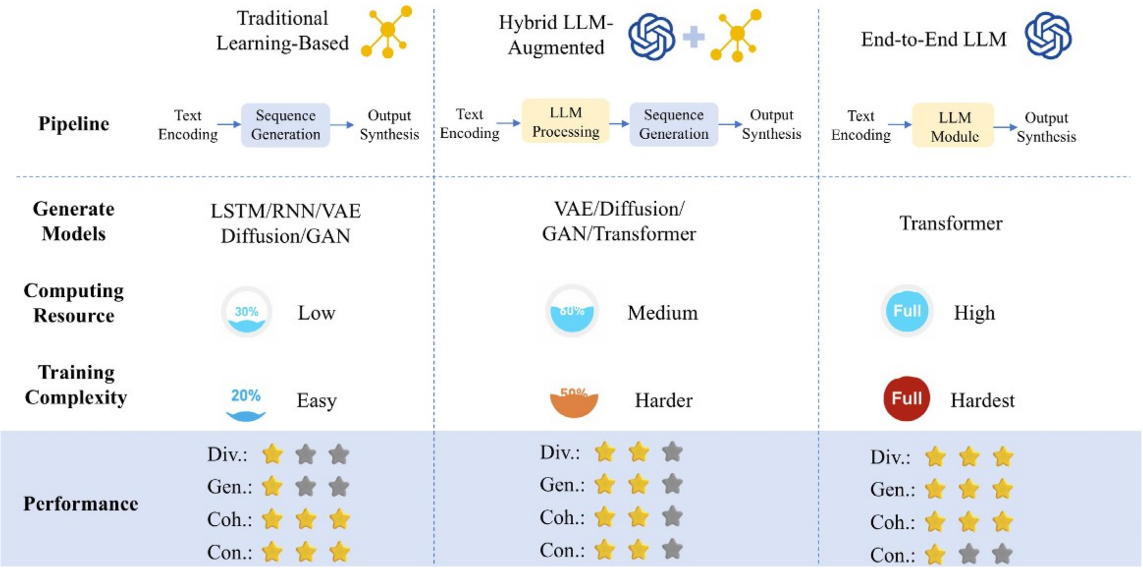
虽然 LLMs 可以创造高度创造性和音乐连贯的作品，实现高度依赖的质量和多样性的训练数据，使他们不太有效的代表性流派或音乐风格。

5.4. Comparative Analysis and Limitations

5.4 比较分析和局限性

Figure 16 compares three typical cases of general frameworks for text-to-music generationTraditional Methods, Hybrid Approaches, and End-to-End LLM Systems—across key aspects, emphasizing their strengths and limitations.

图 16 比较了文本到音乐生成的传统方法、混合方法和端到端 LLM 系统的通用框架的三个典型案例，强调了它们的优势和局限性。



	Traditional Learning-Based	Hybrid LLM-Augmented	End-to-End LLM
Pipeline	Text Encoding → Sequence Generation → Output Synthesis	Text Encoding → LLM Processing → Sequence Generation → Output Synthesis	Text Encoding → LLM Module → Output Synthesis
Generate Models	LSTM/RNN/VAE Diffusion/GAN	VAE/Diffusion/ GAN/Transformer	Transformer
Computing Resource	30% Low	80% Medium	Full High
Training Complexity	20% Easy	50% Harder	Full Hardest
Performance	Div.: ★ ☆ ☆ Gen.: ★ ☆ ☆ Coh.: ★ ★ ★ Con.: ★ ★ ★	Div.: ★ ★ ☆ Gen.: ★ ★ ☆ Coh.: ★ ★ ☆ Con.: ★ ★ ☆	Div.: ★ ★ ★ Gen.: ★ ★ ★ Coh.: ★ ★ ★ Con.: ★ ☆ ☆

Figure 16. Comparison between frameworks. Div.:Diversity; Gen.: Generalizability; Coh: Coherence; Coherence; Con: Controllability.

图 16。框架之间的比较。：多样性；一般： 概括性；Coh: 一致性；一致性；缺点： 可控性。

Traditional Methods:

传统方法：

Traditional methods often rely on models like LSTM, VAE, and diffusion. These models follow a linear pipeline from text encoding to sequence generation and music synthesis. They focus on symbolic music generation, typically using MIDI data for training.

传统方法通常依赖于 LSTM、 VAE 和扩散模型。这些模型遵循从文本编码到序列生成和音乐合成的线性流程。他们专注于符号音乐生成，通常使用 MIDI 数据进行训练。

●Strengths: These methods offer high control and reliable output quality due to their use of high-quality paired datasets. They excel in structured tasks, like generating music based on fixed templates or rhythms, offering good controllability over tonalities, chords, rhythm, and cadence. Additionally, these methods typically have lower computational demands compared to newer, more complex models; ●Limitations: Their creativity and generalization abilities are limited. They are often constrained by predefined templates, restricting them to a narrower range of tasks. Furthermore, they require labeled data for training and can struggle to adapt to more diverse, dynamic input.

优势： 由于使用了高质量的成对数据集，这些方法提供了高控制和可靠的输出质量。它们擅长于结构化任务，比如基于固定模板或节奏生成音乐，对音调、和弦、节奏和韵律提供良好的可控性。此外，与更新、更复杂的模型相比，这些方法通常具有更低的计算需求；●局限性： 它们的创造力和概括能力是有限的。它们经常受到预定义模板的限制，将它们限制在一个较窄的任务范围内。此外，它们需要标记数据进行训练，并且难以适应更多样化、动态的输入。

Hybrid LLM-Augmented Frameworks:

混合 *LLM-Augmented* 框架：

Hybrid approaches like MuseCoco combine traditional models with LLMs to enhance creativity and adapt to more varied input types. These frameworks utilize BERT and ChatGPT for text synthesis and attribute-conditioned generation, which allows for more flexible control over music attributes such as instruments, emotion, and rhythm.

像 MuseCoco 这样的混合方法结合了传统模型和 llm 来提高创造力和适应更多不同的输入类型。这些框架利用 BERT 和 ChatGPT 进行文本合成和属性条件生成，允许对乐器、情感和节奏等音乐属性进行更灵活的控制。

Strengths: Hybrid systems strike a balance between creativity and control. They can integrate multiple sources of data, offering higher diversity in the generated music.

优势： 混合系统在创造力和控制之间取得了平衡。它们可以整合多种数据来源，为生成的音乐提供更高的多样性。

These systems are more flexible and can accommodate more complex input, including unstructured text descriptions;

这些系统更加灵活，可以容纳更加复杂的输入，包括非结构化的文本描述；

Limitations: Hybrid approaches still rely on traditional techniques to maintain some degree of control, which can limit the degree of innovation. The model size and com-putational demands are higher than traditional methods, though still more efficient than end-to-end LLM systems.

限制： 混合方法仍然依赖于传统技术来保持一定程度的控制，这可能会限制创新的程度。模型规模和计算需求高于传统方法，但仍然比端到端的 LLM 系统更有效。

End-to-End LLM Systems:

端到端 LLM 系统：

End-to-end LLM systems like SongComposer utilize large-scale pre-trained models and integrate symbolic music representations to directly synthesize music from text in-structions. These systems are designed to handle complex, multi-modal input and offer end-to-end learning, where the model is trained on a vast amount of paired data (e.g.,lyrics and melody).

像 SongComposer 这样的端到端 LLM 系统利用大规模的预训练模型和集成符号化的音乐表示从文本指令直接合成音乐。这些系统被设计用来处理复杂的、多模态的输入，并提供端到端的学习，其中模型是基于大量的配对数据 (如歌词和旋律) 进行训练的。

●Strengths: These systems push the boundaries of creativity and generalization. They can generate highly diverse outputs, incorporating various musical elements (such as melody, harmony, and rhythm) and adapt to cross-task and cross-genre settings. With the ability to use multi-modal data, they enhance the emotional and thematic depth of the generated music;

优势： 这些系统拓展了创造性和概括性的边界。他们可以产生高度多样化的输出，结合各种音乐元素 (如旋律，和声，节奏) ，并适应跨任务和跨流派的设置。通过使用多模态数据的能力，他们增强了生成音乐的情感和主题深度；

Limitations: The control over specific attributes, such as rhythm or instrumentation, can be weaker in these systems. Additionally, computational resource efficiency is often a concern, as they require large-scale datasets and extensive processing power for both training and inference.

限制： 在这些系统中，对节奏或乐器等特定属性的控制可能较弱。此外，计算资源的效率往往是一个问题，因为它们需要大规模的数据集和广泛的处理能力的训练和推理。

6. Challenges and Future Directions

6. 挑战和未来方向

6.1. Challenges

6.1 挑战

6.1.1.Technical Level

6.1 技术水平

Although breakthroughs have been made, text-to-music generation tasks still face the following technical challenges.

尽管已经取得了突破性进展，文本到音乐的生成任务仍然面临以下技术挑战。

1. L. Dataset Scarcity and Representation Limitations

数据集稀缺性和表示限制

High-quality datasets are the foundation for training effective text-to-music generation models. However, current datasets often lack diversity in musical styles and emotional expressions, resulting in generated outputs that are overly homogeneous. Furthermore, the accuracy of dataset labeling directly impacts model training, as incorrect labels may mislead models into learning faulty musical patterns. Large-scale datasets, essential for training complex models, also pose significant challenges in terms of data collection, processing, and representation. Symbolic datasets (e.g., MIDI) may fail to capture the expressive nuances of music, while audio-based datasets are computationally demanding and challenging to align with textual semantics. Addressing these limitations requires innovative approaches to dataset design, multi-modal alignment, and data augmentation;

高质量的数据集是训练有效的文本 - 音乐生成模型的基础。然而，目前的数据集往往缺乏音乐风格和情感表达的多样性，导致生成的输出过于同质化。此外，数据集标注的准确性直接影响模型训练，不正确的标注可能会误导模型学习错误的音乐模式。大规模数据集对于复杂模型的训练至关重要，同时也给数据的收集、处理和表示带来了巨大的挑战。符号数据集 (如 MIDI) 可能无法捕捉音乐表达的细微差别，而基于音频的数据集在计算上要求很高，并且很难与文本语义保持一致。解决这些限制需要数据集设计，多模态对齐和数据增强的创新方法；

2. Model Training and Generalization

2. 模型训练和泛化

The generalization ability of current models remains a key limitation, especially when dealing with unseen data. Many existing systems struggle to produce coherent and contextually appropriate music outside their training data distribution. Moreover, training large-scale models demands extensive computational resources, which limits accessibility for researchers and developers. Additionally, model interpretability is a significant concern; understanding how models make decisions during music generation is crucial for improving their performance and providing more guided outputs. Enhancing interpretability can also aid in debugging and refining models to better align with the intended tasks;

当前模型的泛化能力仍然是一个关键的限制，特别是在处理未知数据时。许多现有的系统难以在其训练数据分布之外生成连贯的和上下文适当的音乐。此外，训练大规模模型需要大量的计算资源，这限制了研究人员和开发人员的可用性。此外，模型的可解释性是一个值得关注的问题；理解模型在音乐生成过程中如何做出决策对于提高其性能和提供更多的指导性输出至关重要。增强可解释性还可以帮助调试和完善模型，以更好地与预期的任务保持一致；

3. Evaluation Metrics for Creativity

3. 创造力的评估指标

The limitation in model generalization is a key factor restricting the creativity of gen-erated outputs. Creativity inherently involves successful extrapolation beyond the dataset distribution, whereas current machine learning methods mainly address interpolation rather than extrapolation [85,86]. Models need to strike a balance between imitating exist-ing musical styles and generating novel music.

Additionally, the lack of effective methods for quantifying and evaluating musical creativity limits objective assessments of innovation in generated music;

模型泛化的局限性是制约生成输出创造性的关键因素。创造力本质上涉及超越数据集分布的成功外推，而目前的机器学习方法主要解决内插而不是外推 [85,86]。模型需要在模仿现有的音乐风格和生成新颖的音乐之间取得平衡。此外，缺乏量化和评估音乐创造力的有效方法限制了对生成音乐创新性的客观评估；

4. Song Structure and Long-Term Coherence

4. 歌曲结构和长期连贯性

Music often relies on complex short-term and long-term structures, such as the verse-bridge-chorus format in popular music or the thematic development in classical composi-tions. Capturing and generating such structures poses a significant challenge, particularly for long-sequence modeling tasks. Current models struggle to simultaneously manage local coherence (e.g., smooth transitions between notes or measures) and global struc-ture (e.g., thematic development across an entire piece). Achieving this balance requires advanced techniques that can effectively handle hierarchical dependencies in musical compositions [21];

音乐往往依赖于复杂的短期和长期结构，如流行音乐中的韵文 - 桥梁 - 合唱形式或古典作品中的主题发展。捕捉和生成这样的结构提出了一个重大的挑战，特别是对于长序列建模任务。当前的模型难以同时管理局部一致性 (例如，音符或度量之间的平滑过渡) 和全局结构 (例如，整个片段的主题发展)。实现这种平衡需要先进的技术，可以有效地处理音乐作品中的层次依赖性 [21] ;

5. D. Emotion Representation and Modeling

5. 情感表征和建模

Although emotion is a vital component of music, representing and modeling emotion poses a complex challenge. The limitations of existing models lie in how to effectively analyze emotional representations in text and model emotional features. Furthermore, the relationship between emotion and musical elements is a complex issue that involves both psychology and musicology, requiring models to understand and leverage these associations to generate music with specific emotional qualities. Only a few studies have addressed the emotional aspect of music [87-92];

虽然情绪是音乐的重要组成部分，但是情绪的表征和建模是一个复杂的挑战。现有模型的局限性在于如何有效地分析文本中的情感表征并建模情感特征。此外，情感和音乐元素之间的关系是一个复杂的问题，涉及心理学和音乐学，需要模型来理解和利用这些关联，以生成具有特定情感特质的音乐。只有少数研究涉及音乐的情感方面 [87-92]；

6. Interactivity between Human and Computer

6. 人机交互

While end-to-end modeling has enabled systems to generate complete musical com-positions seamlessly, there is a growing demand for interactive generation systems. Users often prefer to engage with AI as a “musical partner’”, adjusting outputs dynamically dur-ing the generation process. Existing interactive systems [93,94]have demonstrated promise, but they are far from widespread adoption. Key challenges include designing interfaces that allow for intuitive user interaction, enabling real-time feedback without compromising the coherence of the generated music, and addressing the balance between user input and model autonomy. Further exploration of human-AI interaction in the context of music generation is essential to creating systems that are not only functional but also user-friendly and adaptable to diverse creative workflows;

尽管端到端的建模使得系统能够无缝地生成完整的音乐作品，但是对于交互式生成系统的需求却在不断增长。用户往往更喜欢与人工智能作为 “音乐伙伴”，在生成过程中动态调整输出。现有的交互式系统 [93,94] 已经显示出了希望，但是它们还远远没有被广泛采用。关键的挑战包括设计界面，允许直观的用户交互，实现实时反馈而不损害生成音乐的一致性，以及解决用户输入和模型自主性之间的平衡。进一步探索音乐生成环境中的人类 - 人工智能互动对于创建不仅具有功能性而且对用户友好和适应多种创造性工作流程的系统至关重要；

7. Lack of Commercial Applications

7. 缺乏商业应用

Despite the promising results of state-of-the-art (SOTA) models, many of these ap-proaches struggle to be implemented at scale. Most existing models, although capable of generating high-quality music, require substantial computational resources and training datasets, which makes them difficult to deploy in commercial applications. Additionally, the lack of user-friendly, accessible software limits the practical use of these technolo-gies outside of research labs. This gap between cutting-edge research and real-world applications needs to be addressed for the technology to reach its full potential;

尽管最先进的模型 (SOTA) 取得了令人鼓舞的成果，但是许多方法难以大规模实施。大多数现有的模型，虽然能够生成高质量的音乐，但需要大量的计算资源和训练数据集，这使得它们很难在商业应用中部署。此外，缺乏用户友好，可访问的软件限制了这些技术在研究实验室之外的实际应用。需要解决尖端研究和现实世界应用之间的这种差距，以使技术充分发挥其潜力；

8. AI Security and Adversarial Attacks

8. 人工智能安全和对抗性攻击

As AI-driven music generation becomes more advanced, security concerns, particu-larly adversarial attacks, have emerged as a critical issue. Adversarial examples refer to small, intentionally crafted perturbations in input data that can mislead AI models into making incorrect predictions or generating faulty content. These attacks are a significant concern in creative fields like music generation, where adversarial inputs could lead to manipulated or unauthorized content creation. Recent studies have explored AI security in the context of adversarial examples, proposing strategies for defending AI models against such attacks. These research efforts are essential to ensure that AI tools, particularly those used in music generation, are robust, reliable, and resistant to malicious manipulation[95].

随着人工智能驱动的音乐生成技术越来越先进，安全问题，尤其是对抗性攻击，已成为一个关键问题。对抗性例子指的是在输入数据中有意制造的小扰动，这些扰动可能会误导人工智能模型做出错误的预测或生成错误的内容。这些攻击在音乐生成等创意领域备受关注，对抗性输入可能导致被操纵或未经授权的内容创作。最近的研究探索了对抗性例子背景下的人工智能安全，提出了防御此类攻击的人工智能模型策略。这些研究工作对于确保人工智能工具，特别是那些用于音乐生成的工具，是健壮的，可靠的，并且能够抵抗恶意操纵 [95]。

6.1.2. Social Level

6.1。2. 社会层面

Due to the unique nature of artistic works, the text-to-music generation faces several social challenges:

由于艺术作品的独特性，从文本到音乐的一代面临着一些社会挑战：

1. L+Copyright Issues

1. l + c 版权问题

The music generation task now faces three main challenges, including the legality of training datasets, originality of generated content, and copyright ownership. The music industry is most concerned that AI learning from songs to generate new con-tent could infringe on the copyrights of original artists [95,96]. Major music companies, such as Universal Music Group, have begun taking steps and demanding that stream-ing platforms prevent AI tools from scraping lyrics and melodies from copyrighted songs. The Recording Industry Association of America has submitted a list of AI de-velopers to the U.S. government and filed a lawsuit against AI music companies, aiming to prevent the unauthorized use of copyrighted recordings to “train” generative AI mod-els (https://www.riaa.com/record-companies-bring-landmark-cases-for-responsible-ai-againstsuno-and-udio-in-boston-and-new-york-federal-courts-respectively/,accessed on17 February 2025). Additionally, the “deep fake”of generated content also deserves atten-tion[97,98], as it poses a serious threat to the originality and personal style of artists;

音乐生成任务目前面临三个主要挑战，包括训练数据集的合法性、生成内容的原创性和版权所有权。音乐行业最担心的是，从歌曲中学习生成新内容的人工智能可能侵犯原创艺术家的版权 [95,96]。环球音乐集团 (Universal Music Group) 等主要音乐公司已开始采取措施，要求流媒体平台阻止人工智能工具从受版权保护的歌曲中删除歌词和旋律。美国唱片业协会向美国政府提交了一份人工智能开发者名单，并对人工智能音乐公司提起诉讼，旨在防止未经授权使用版权录音来 “训练” 生成性人工智能模型 (https://www.riaa.com/record-companies-bring-landmark-cases-for-responsible-AI-againstsuno-and-udio-in-boston-and-new york-federal-courts-respectively/，access on 2025)。此外，生成内容的 “深度假” 也值得关注 [97,98] ，因为它对艺术家的原创性和个人风格构成严重威胁；

2. Privacy Concerns

2. 隐私问题

Despite years of research in artificial intelligence, privacy concerns remain unre-solved [99]. Privacy is especially pronounced in singing voice generation. Bai et al. (2024) [84] noted that they have recognized “the singing voice evokes one of the strongest expressions of individual identity”. Therefore, it becomes a burning issue to ensure data collection and usage do not infringe on personal privacy in training;

尽管对人工智能进行了多年的研究，但隐私问题仍然没有得到解决 [99]。隐私在歌声生成中尤其明显。Bai 等人 (2024)[84] 指出，他们已经认识到 “歌唱的声音唤起个人身份最强烈的表达之一”。因此，确保数据收集和使用不侵犯培训中的个人隐私成为一个亟待解决的问题；

3.). Impact on Human Musicians

3. 对人类音乐家的影响

The rise of AI music may have some social implications for human musicians. On the one hand, AI music lowers the barriers for ordinary people to create music, allowing for more people to be involved; but on the other hand, it could threaten the livelihoods of professional musicians and composers. With the proliferation of low-cost or even free AI music tools, the traditional music industry may be impactedthe demand for artificially created music may be reduced, and the work of professional musicians may be undervalued. More worryingly, the increasing automation of the music creation process could lead to unemployment for some musicians, and the professional skills they have honed over the years could be devalued as a result [100].

人工智能音乐的兴起可能会对人类音乐家产生一些社会影响。一方面，人工智能音乐降低了普通人创作音乐的门槛，允许更多的人参与其中；但另一方面，它也可能威胁到专业音乐家和作曲家的生计。随着低成本甚至免费的人工智能音乐工具的激增，传统音乐产业可能受到冲击，人工创作音乐的需求可能减少，专业音乐人的工作可能被低估。更令人担忧的是，音乐创作过程的日益自动化可能会导致一些音乐家失业，他们多年磨练出来的专业技能可能会因此贬值 [100]。

4. Bias in Music Datasets

4. 音乐数据集的偏见

Another critical ethical issue is the bias present in many music datasets used for training AI systems. Most datasets tend to be Western-centric, often reflecting a narrow range of musical genres, styles, and cultural contexts. This bias can lead to AI models that favor certain musical traditions while marginalizing others, perpetuating a lack of diversity in AI music. As AI music becomes more widespread, there is an increasing concern that such biases may influence the kinds of music produced and potentially result in cultural misrepresentation or appropriation. This underscores the importance of creating more inclusive and diverse datasets that reflect global musical traditions and avoid reinforcing existing stereotypes [101];

另一个关键的伦理问题是许多用于训练人工智能系统的音乐数据集中存在的偏见。大多数数据集倾向于以西方为中心，往往反映了狭窄范围内的音乐流派、风格和文化背景。这种偏见会导致人工智能模型偏爱某些音乐传统，而边缘化其他音乐传统，使人工智能音乐缺乏多样性。随着人工智能音乐越来越普及，人们越来越担心这种偏见可能会影响音乐的种类，并可能导致文化上的不正当手法引诱或挪用。这强调了创建更具包容性和多样性的数据集的重要性，这些数据集反映了全球音乐传统，并避免强化现有的刻板印象 [101]；

5. Concerns Regarding Cultural Representation

5. 关于文化代表性的担忧

As AI music becomes more prominent, issues surrounding cultural representation and authenticity become increasingly important. AI models trained on datasets dominated by certain cultures may fail to capture the nuances of music from other cultures, potentially leading to cultural appropriation. For example, if an AI model trained predominantly on Western classical music generates music based on non-Western texts or traditions, it may misinterpret or oversimplify the original cultural context, leading to outputs that lack authenticity and respect for the source material. This raises critical ethical questions about how AI systems are trained and the extent to which they can authentically represent diverse musical traditions.

随着人工智能音乐变得越来越突出，围绕文化表征和真实性的问题变得越来越重要。在某些文化主导的数据集上训练的人工智能模型可能无法捕捉到来自其他文化的音乐的细微差别，这可能导致文化挪用。例如，如果主要以西方古典音乐为基础的人工智能模型生成基于非西方文本或传统的音乐，则可能会曲解或过度简化原始文化背景，导致输出缺乏真实性和对原始材料的尊重。这提出了关于 AI 系统如何训练以及它们在多大程度上可以真实地代表不同音乐传统的关键伦理问题。

6.2. Future Directions

6.2 未来发展方向

Text-to-music generation represents a revolutionary advancement in music creation, applying natural language processing and machine learning techniques to the composition process and opening new pathways for music creation. From early rule-based methods to today’s deep learning models, music generation technology has made great strides, enabling researchers to produce music with considerable artistic value and emotional depth. Given the challenges outlined above, several future directions for text-to-music generation are proposed.

文本到音乐的生成代表了音乐创作的革命性进步，将自然语言处理和机器学习技术应用于作曲过程，为音乐创作开辟了新的途径。从早期的基于规则的方法到今天的深度学习模型，音乐生成技术取得了长足的进步，使得研究人员能够创作出具有相当艺术价值和情感深度的音乐。鉴于上述挑战，提出了文本到音乐生成的几个未来方向。

1. Enhancing Data Quality and Diversity
1. 提高数据质量和多样性

Future developments will prioritize enhancing data quality and diversity. Building comprehensive music datasets that cover a wider range of styles, genres, and cultures will enable models to learn broader musical characteristics and improve generalization. High-quality datasets enriched with detailed annotations—such as emotional content, structural markers, and performance techniques—will be essential for refining models'learning capabilities. The incorporation of synthetic data generation using LLMs could also serve as a supplementary approach to address data scarcity by creating realistic textual and musical annotations;

未来的发展将优先考虑提高数据质量和多样性。构建涵盖更广泛的风格、流派和文化的综合音乐数据集将使模型能够学习更广泛的音乐特征并提高泛化能力。具有详细注释的高质量数据集——如情感内容、结构标记和性能技术——对于完善模型的学习能力至关重要。使用 llm 合成数据生成的结合也可以作为通过创建现实的文本和音乐注释来解决数据稀缺性的补充方法；

2. Addressing Bias in Music Datasets
2. 解决音乐数据集中的偏见

A key future direction will be the enhancement of music datasets to reduce bias, particularly the Western-centric bias in many current datasets. Future models should be trained on datasets that reflect a broader range of global music traditions, ensuring that AI-generated music is more inclusive and representative of diverse cultures. Developing datasets that include diverse musical styles, languages, and genres will help mitigate cultural appropriation concerns and ensure more accurate and respectful representations of different musical traditions:

未来的一个重要方向是增强音乐数据集以减少偏差，尤其是当前许多数据集中的西方中心偏差。未来的模型应该在反映更广泛的全球音乐传统的数据集上进行培训，确保人工智能生成的音乐更具包容性，更能代表不同的文化。开发包括不同音乐风格、语言和流派的数据集将有助于减轻文化挪用的担忧，并确保对不同音乐传统的更准确和尊重的表述：

3.+Optimizing Training Efficiency
3. + 优化训练效率

Another key direction is the optimization of model training methods, which can reduce training time and costs through the application of distributed computing platforms and improve algorithm efficiency to facilitate a more efficient learning process;

另一个关键方向是模型训练方法的优化，通过分布式计算平台的应用减少训练时间和成本，提高算法效率，促进更高效的学习过程；

4. Improving the Quality and Personalization
4. 提高质量和个性化

Future models will increasingly adopt innovative mechanisms to improve the quality and personalization of generated music. Techniques such as attention mechanisms and style transfer will make compositions more adaptable to specific user requirements, producing works that are highly artistic and personalized;

未来的模型将越来越多地采用创新机制，以提高生成音乐的质量和个性化。注意力机制和风格转移等技术将使作品更适应特定的用户需求，生成高度艺术化和个性化的作品；

5. I Deepening Understanding of Musical Structures
5. 加深对音乐结构的理解

Model designs will focus on better understanding musical structures, such as seg-ment divisions, motif development, and long-term coherence. Future systems will aim to generate compositions that exhibit complex structures and rich variations, enhancing the ability to capture both local and global patterns in music. The emphasis will be on improving the integration of musical theory and computational methods to better align with human creativity:

模型设计将致力于更好地理解音乐结构，例如音段划分、主题发展和长期连贯性。未来的系统将致力于生成展示复杂结构和丰富变化的作品，增强捕捉音乐中局部和全局模式的能力。重点将放在改善音乐理论和计算方法的整合，以更好地与人类的创造力保持一致：

6. Bridging Music and Emotion
6. 连接音乐与情感

Emotion modeling will also become a central area of focus, enabling models to generate music that evokes emotional resonance. Future systems will aim to achieve a deeper understanding of the relationships between linguistic expressions and musical emotions by integrating multi-modal data, such as text and audio. These improvements will empower models to create emotionally expressive compositions that resonate with listeners on a profound level;

情感建模也将成为关注的中心领域，使模特能够产生唤起情感共鸣的音乐。未来的系统将致力于通过整合文本和音频等多模态数据，更深入地理解语言表达和音乐情感之间的关系。这些改进将使模型能够创作出能够与听众产生深刻共鸣的情感表达作品；

7. Advancing Multi-Modal Music Generation
7. 推进多模态音乐生成

The rise of large-scale models will push advancements toward model integration and cross-modal capabilities. Future systems with multi-modal inputs—such as text, images, and video 一 will pave the way for generating music inspired by diverse input types, significantly expanding the application scenarios of music generation. For example, a model could generate soundtracks for a video or a painting, seamlessly bridging artistic domains and enriching creative workflows;

大规模模型的兴起将推动模型集成和跨模态能力的进步。未来具有多模态输入的系统一一例如文本、图像和视频一一一将为生成受多种输入类型启发的音乐铺平道路，极大地拓展音乐生成的应用场景。例如，一个模型可以生成视频或绘画的配乐，无缝连接艺术领域，丰富创造性的工作流程；

8. Developing Commercial Applications
8. 开发商业应用程序

Developing commercial applications can bridge the gap between research and real-world applications. Successful examples like Suno and Seed Music demonstrate the poten-tial for high-quality text-to-music generation while maintaining unique product features. These efforts should focus on making the technology more accessible while ensuring that it remains efficient and scalable for diverse applications, not only in music production but also in music therapy, metaverse, etc. Addressing these challenges will unlock new possibilities for the industry and will help transform the landscape of music;

开发商业应用程序可以缩小研究和实际应用之间的差距。像 Suno 和 Seed Music 这样的成功案例展示了在保持产品独特功能的同时生成高质量文本到音乐的潜力。这些努力应侧重于使该技术更易于获取，同时确保其不仅在音乐制作方面，而且在音乐治疗、元宇宙等方面的各种应用中保持高效和可扩展性。应对这些挑战将为行业开启新的可能性，并将有助于改变音乐的格局；

9. Establishing Clear Copyright Ownership
建立清晰的版权所有权

From a social perspective, it is imperative to set forth more clear-cut rules of copyright ownership and record the process of creation with the help of blockchain technology in the future. This initiative will ensure the legality of creations and promote the development of open copyright music databases, encouraging data sharing while protecting artists’rights and interests;

从社会角度来看，未来有必要制定更为明确的著作权归属规则，并借助区块链技术记录创作过程。这一举措将确保创作的合法性，促进开放版权音乐数据库的发展，在保护艺术家权益的同时鼓励数据共享；

10. Strengthening Privacy Protection
十、加强隐私权保护

In terms of privacy protection, stronger encryption and anonymization of user data will become a defining trend. This means that users’ privacy will not be invaded and that specific, transparent service terms will be rolled out to build users’ trust;

在隐私保护方面，加强用户数据的加密和匿名化将成为一个决定性的趋势。这意味着用户的隐私不会受到侵犯，将推出具体、透明的服务条款来建立用户的信任；

11.. Fostering Collaboration Between Technology and Artists
促进技术和艺术家之间的合作

In the music industry, there will be a greater focus on collaboration between technology platforms and artists, exploring innovative applications, and developing fairer revenue distribution mechanisms to balance the conflicting interests between automated generation and traditional manual creation.

在音乐产业中，将更加注重技术平台和艺术家之间的合作，探索创新应用，发展更加公平的收益分配机制，以平衡自动生成和传统手工创作之间的利益冲突。

7. Conclusions

7. 结论

This paper provides a comprehensive overview of recent advancements in text-to-music generation, focusing on the classification and methodologies of symbolic and audio domains. It systematically introduces key improvements in text-to-music generation across various tasks (melody generation, polyphony generation, instrumental music generation, singing voice synthesis, and complete song generation). By introducing a taxonomy of text types (lyrics, musical attributes, and natural descriptions) and musical representations(MIDI, spectrograms, ABC notation), we establish a structured framework for evaluating cross-modal alignment challenges.

本文全面综述了文本到音乐生成的最新进展，重点介绍了符号和音频领域的分类和方法。它系统地介绍了文本到音乐生成在各种任务 (旋律生成，复调生成，器乐生成，歌唱声音合成，和完整的歌曲生成) 的关键改进。通过引入文本类型 (歌词、音乐属性和自然描述) 和音乐表征 (MIDI、谱图、 ABC 符号) 的分类，我们建立了一个评估跨模态对齐挑战的结构化框架。

The primary contribution of this work lies in its critical review and classification of existing frameworks for text-to-music generation. By categorizing approaches into tradi-tional methods, hybrid techniques, and end-to-end LLM-centric frameworks, this paper provides a detailed comparison of their strengths, limitations, and applicability to different tasks. Unlike prior surveys focused on single-modality generation, we highlight how LLMs enhance controllability and generalization by integrating semantic understanding with musical structure modeling, addressing limitations of rule-based systems in creativity and data-driven models in interpretability.

本文的主要贡献在于对现有的文本 - 音乐生成框架进行了批判性的回顾和分类。通过将方法分为传统方法、混合技术和以端到端 llm 为中心的框架，本文详细比较了它们的优势、局限性和对不同任务的适用性。不同于以往关于单一模态生成的研究，本文重点介绍了 llm 如何通过将语义理解与音乐结构建模相结合，解决基于规则的系统在创造力方面的局限性以及数据驱动模型在可解释性方面的局限性，从而增强可控性和泛化能力。

Key technical challenges are identified, including dataset scarcity for underrepresented genres,long-term coherence in multi-track compositions, and the need for emotion-aware generation. Social challenges, such as copyright ambiguity and AI-generated content origi-nality, are also discussed. Future advancements are expected to improve the quality and diversity of generated music while simplifying the generation process to make it more intuitive and accessible. Key areas for exploration include developing more sophisticated algorithms to better interpret textual semantics, reducing dependence on large labeled datasets through innovative data processing techniques, and enhancing model general-ization to produce more creative and personalized outputs. Additionally, integrating multi-modal large models will enable systems to incorporate diverse information sources, such as images, videos, and environmental sounds, fostering the creation of richer and more multidimensional musical experiences.

确定了关键的技术挑战，包括代表性不足的流派的数据集稀缺，多轨作品的长期一致性，以及情感感知生成的需要。还讨论了版权模糊和人工智能生成内容原创性等社会挑战。未来的发展有望提高生成音乐的质量和多样性，同时简化生成过程，使其更加直观和易于理解。需要探索的关键领域包括开发更复杂的算法以更好地解释文本语义，通过创新的数据处理技术减少对大型标记数据集的依赖，以及增强模型泛化以产生更具创造性和个性化的输出。此外，集成多模态大型模型将使系统能够纳入不同的信息来源，如图像，视频和环境声音，促进创建更丰富和更多维的音乐体验。

This work provides a critical roadmap for advancing text-to-music generation by systematizing methodologies, clarifying cross-modal alignment challenges, and highlight-ing the transformative role of LLMs in enhancing controllability and interpretability. By bridging gaps between semantic understanding and structural modeling and prioritizing ethical and technical challenges, this paper lays the groundwork for future innovations that balance creativity, technical rigor, and societal impact in AI-generated music, empowering researchers to develop more accessible, diverse, and socially responsible systems.

这项工作提供了一个关键的路线图，通过系统化的方法，澄清跨模态对齐的挑战，并突出 llm 在提高可控性和可解释性的变革性作用，推进文本到音乐的生成。通过弥合语义理解和结构建模之间的差距，并优先考虑伦理和技术挑战，本文为未来的创新奠定了基础，在人工智能生成的音乐中平衡创造力、技术严谨性和社会影响，使研究人员能够开发更易访问、多样化和社会责任感更强的系统。

Author Contributions: Conceptualization,Y.Z. and M.Y.; methodology, Y.Z.,J.D. and M.Y.; formal analysis,Y.L. and X.Z.; investigation, F.S. and Y.Z.; resources, Y.Z., Z.W. and H.N.; data curation, M.Y. and F.S.; writingoriginal draft preparation, Y.Z. and Y.L.; writingreview and editing, Y.Z. and M.Y.;visualization,Y.Z. and M.Y.; supervision, Z.W. and H.N.; project administration, Z.W., H.N. and J.D. All authors have read and agreed to the published version of the manuscript.

作者贡献： 概念化，y.z. 和 m.y. ; 方法论，y.z. ，j.d. 和 m.y. ; 形式分析，y.l. 和 x.z. ; 调查，f.s. 和 y.z. ; 资源，y.z. ，z.w. 和 h.n. ; 数据管理，m.y. 和 f.s. ;。

Funding: This research received no external funding.

资金： 这项研究没有收到外部资金。

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable.

数据可用性声明： 本研究未创建或分析新数据，数据共享不适用。

Acknowledgments: The authors thank the Cyberspace Data and Intelligence Lab at the School of Computer and Communications Engineering, University of Science and Technology Beijing, for providing the research facilities.

鸣谢： 作者感谢北京科技大学计算机与通信工程学院网络空间数据与情报实验室提供的研究设备。

Conflicts of Interest: Author Mingzhi Yang was employed by Guangxi Tourism Development One-Click Tour Digital Cultural Tourism Industry Co., Ltd. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

利益冲突： 作者 Mingzhi Yang 受雇于广西旅游发展一键式旅游数字文化旅游产业有限公司。其余作者声明，研究是在没有任何商业或金融关系，可以被解释为潜在的利益冲突的情况下进行的。

References

参考文献

1. Ji,S.; Luo, J.; Yang,X. A Comprehensive Survey on Deep Music Generation: Multi-Level Representations, Algorithms, Evaluations, and Future Directions. arXiv 2020, arXiv:2011.06801.

罗，j。杨，x。深度音乐生成的综合调查： 多层次的表征、算法、评估和未来的方向。arXiv 2020，arXiv: 2011.06801.

2. Ma,Y.;Oland, A.; Ragni, A.; Sette, B.M.D.; Saitis, C.; Donahue, C.; Lin, C.; Plachouras, C.; Benetos, E.; Shatri, E.; et al. Foundation Models for Music: A Survey. arXiv 2024, arXiv:2408.14340.

奥兰岛；Ragni，a; Sette，b.m.d; Saitis，c; Donahue，c; Lin，c; Plachouras，c; Benetos，e; Shatri，e; et al。音乐基础模型： 一项调查。arXiv 2024，arXiv: 2408.14340 arXiv: 2408.14340.

3. Briot, J.-P.;Pachet, F. Music Generation by Deep Learning-Challenges and Directions. Neural Comput. Appl. 2020, 32, 981-993.[CrossRef]

通过深度学习产生音乐 —— 挑战与方向。神经计算，应用，2020,32,981-993。[交叉参考]

4. Ji, S.; Yang,X.; Luo, J. A Survey on Deep Learning for Symbolic Music Generation: Representations, Algorithms, Evaluations, and Challenges. ACM Comput. Surv. 2023, 56,1-39.[CrossRef f] 56Hernandez-Olivan, C.; Beltran, J.R. Music Composition with Deep Learning: A Review; Springer: Cham, Switzerland, 2021.

Ji，s。 ; Yang，x。 ; Luo，j。符号音乐生成的深度学习研究： 表征、算法、评估与挑战。ACM 计算。生存。2023 年 5 月 1 日至 39 日。[crossre f f] 56 hernandez-olivan，c. ; Beltran，j.r。音乐创作与深度学习： 评论；Springer: Cham，瑞士，2021。

Civit, M.; Civit-Masot,J.;Cuadrado, F; Escalona, M.J. A Systematic Review of Artificial Intelligence-Based Music Generation:Scope, Applications, and Future Trends. Expert Syst. Appl. 2022,209, 118190. [CrossRef]

西维特，m. ; 西维特 - 马索特，j. ; 瓜达拉多，f. ; 埃斯卡洛纳，M.j。基于人工智能的音乐生成： 范围，应用和未来趋势的系统综述。专家系统。应用程序。2022,209,118190. [crossre f]

7. Herremans, D.; Chuan, C.-H.; Chew, E. A Functional Taxonomy of Music Generation Systems. ACM Comput. Surv. 2017, 50,69. ICrossRef

音乐生成系统的功能分类学。 ACM 计算机杂志，2017,50,69

8. Zhu,Y.;Baca, J.; Rekabdar,B.; Rawassizadeh, R. A Survey of AI Music Generation Tools and Models. arXiv 2023, arXiv:2308.12982.[CrossRef]

人工智能音乐生成工具与模型调查，arXiv 2023，arXiv: 2308.12982

9. Wen, Y.-W.; Ting,C.-K. Recent Advances of Computational Intelligence Techniques for Composing Music. IEEE Trans. Emerg. Top. Comput. Intell. 2023,7,578-597.[CrossRef]

音乐创作的计算智能技术的最新进展。电气与电子工程师协会译。涌现。顶部。计算机。英特尔。2023,7,578-597。[交叉参考]

10. Briot,J.-P.;Hadjeres,G.; Pachet,F.-D. Deep Learning Techniques for Music Generation-A Survey. arXiv 2019, arXiv:1709.01620.

音乐生成的深度学习技术 - 调查，arXiv 2019，arXiv: 1709.01620。

11. Xenakis, I. Formalized Music: Thought and Mathematics in Composition; Pendragon Press: Hillsdale, NY, USA, 1992.

《形式化的音乐： 作曲中的思想与数学》 ; 彭德拉贡出版社： 希尔斯代尔，纽约，美国，1992。

12. Schot, J.W.; Hiller, L.; Isaacson, L.M. Experimental Music Composition with an Electronic Computer. In Proceedings of the Mathematics of Computation;

American Mathematical Society: Providence, RI, USA, 1962; Volume 16, p. 507.

肖特，j.w。；希勒，l。；艾萨克森，L.m。实验音乐创作与电子计算机。美国数学学会：普罗维登斯，罗德岛，美国，1962 年，第 16 卷，第 507 页。

Biles, J.A. GenJam: A Genetic Algorithm for Generating Jazz Solos. In Proceedings of the International Conference on Mathematics and Computing; Rochester Institute of Technology: Rochester, NY, USA, 1994.

拜尔斯，j.a。 GenJam: 生成爵士独奏的遗传算法。罗彻斯特理工学院： 罗彻斯特，纽约，美国，1994。

14. Cope, D. Experiments in Music Intelligence (EMI). In Proceedings of the International Conference on Mathematics and Computing; University of California: Berkeley, CA, USA, 1987.

科普，d。音乐智能实验 (EMI)。数学与计算国际会议论文集；加州大学伯克利分校，加州，美国，1987。

15. Fukayama, S.;Nakatsuma, K.; Sako, S.;Nishimoto, T.; Sagayama, S. Automatic Song Composition From The Lyrics Exploiting Prosody OfJapanese Language; Zenodo: Genève, Switzerland, 2010.

深山，s。；nakatuma，k。；Sako，s。；Nishimoto，t。；Sagayama，s。《利用日语韵律的歌词自动作曲》；Zenodo: 日内瓦，瑞士，2010。

16. Scirea, M.; Barros, G.A.B.; Shaker, N.; Togelius, J. SMUG: Scientific Music Generator; Brigham Young University: Provo, UT, USA, 2015.

自鸣得意： 科学音乐生成器；杨百翰大学： 普罗沃，犹他州，美国，2015。

17. Von Riutte, D.; Biggio, L.; Kilcher, Y.; Hofmann, T. FIGARO: Controllable Music Generation Using Learned and Expert Features. arXiv 2022, arXiv:2201.10936v4.

冯 ritte，d。；Biggio，l。；Kilcher，y。；Hofmann，t。 FIGARO: 可控音乐生成使用学习和专家特征。arXiv 2022，arXiv: 2201.10936 v4.

18. Lu, P.; Xu,X.; Kang, C.; Yu, B.; Xing, C.; Tan, X.; Bian, J. MuseCoco: Generating Symbolic Music from Text. arXiv 2023, arXiv:2306.00110.[CrossRef]

吕平、徐晓、康长江、余斌、邢长江、谭晓、卞建国。《从文本生成象征音乐》。arXiv 2023，arXiv: 2306.00110.[crossre f]

19. Herremans, D.; Weisser, S.; Sorensen, K.; Conklin, D. Generating Structured Music for Bagana Using Quality Metrics Based on Markov Models. Expert Syst. Appl. 2015,42,7424-7435. [CrossRef]

使用基于马尔可夫模型的质量指标为 Bagana 生成结构化音乐。专家系统。应用程序。2015,42,7424-7435.[crossre f]

20. Wu,J.;Hu,C.;Wang,Y.;Hu,X.; Zhu,J. A Hierarchical Recurrent Neural Network for Symbolic Melody Generation. IEEE Trans. Cybern. 2020,50,2749-2757. [CrossRef][PubMed] 21.. Guo, Z.; Dimos, M.; Dorien, H. Hierarchical Recurrent Neural Networks for Conditional Melody Generation with Long-Term Structure.arXiv 2021, arXiv:2102.09794.

吴，j。；胡，c。；王，y。；胡，x。；朱，j。符号旋律生成的分层递归神经网络。美国电气电子工程师学会。赛博恩。2020,50,2749-2757.[crossre f][PubMed]21..Dorian，h. 用于长期结构的条件旋律生成的分层递归神经网络.arXiv 2021，arXiv: 2102.09794。

22. Choi,K.; Fazekas, G.; Sandler, M. Text-Based LSTM Networks for Automatic Music Composition. arXiv 2016, arXiv:1604.05358.

自动音乐创作的基于文本的 LSTM 网络，arXiv 2016，arXiv: 1604.05358。

23.2-3Dong, H.-W.; Hsiao, W.-Y.;Yang, L.-C.;Yang,Y.-H. MuseGAN: Multi-Track Sequential Generative Adversarial Networks for Symbolic Music Generation and Accompaniment. Proc. AAAI Conf. Artif. Intell. 2018,32. [CrossRef]

23.2-3Dong，h.-w。；Hsiao，w.-y。；Yang，l.-c。MuseGAN: 用于象征音乐生成和伴奏的多轨序列生成对抗网络。原文链接：。AAAI 会议。Artif 艺术。英特尔。2018 年 32 岁。[crossre f]

24. Huang, C.-Z.A.; Vaswani, A.; Uszkoreit, J.; Shazeer, N.M.; Simon, I.; Hawthorne, C.; Dai, A.M.; Hoffman, M.; Dinculescu, M.; Eck, D. Music Transformer: Generating Music with Long-Term Structure. arXiv 2018, arXiv:1809.04281.

音乐转换器： 用长期结构生成音乐。arXiv 2018，arXiv: 1809.04281.

25. Borsos, Z.; Marinier, R.; Vincent, D.; Kharitonov, E.; Pietquin, O.; Sharifi, M.; Roblek, D.; Teboul, O.; Grangier, D.; Tagliasacchi,M.; et al. AudioLM: A Language Modeling Approach to Audio Generation. arXiv 2023, arXiv:2209.03143. [CrossRef]

Kharitonov，e. ; Pietquin，o. ; Sharifi，m. ; Roblek，d. ; Teboul，o. ; Grangier，d. ; Tagliasacchi，m. ; et al.AudioLM: 音频生成的语言建模方法。arXiv 2023，arXiv: 2209.03143.[crossre f]

26. Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; Ommer, B. High-Resolution Image Synthesis with Latent Diffusion Models. arXiv 2022, arXiv:2112.10752.

采用潜在扩散模型的高分辨率图像合成，arXiv 2022，arXiv: 2112.10752。

27.. Evans, Z.; Carr, C.; Taylor, J.; Hawley, S.H.; Pons, J. Fast Timing-Conditioned Latent Audio Diffusion. In Proceedings of the Forty-First International Conference on Machine Learning, Vienna, Austria, 21-27 July 2021.

27..埃文斯，z。；卡尔，c。；泰勒，j。；霍利，s.h。；Pons，j。在 2021 年 7 月 21 日至 27 日在奥地利维也纳举行的第四十一届国际机器学习会议论文集中。

28. C OpenAI; Achiam, J.; Adler, S.;Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F.L.; Almeida, D.; Altenschmidt,J.;Altman, S.; et al. GPT-4 Technical Report. arXiv 2024, arXiv:2303.08774.

阿德勒，s。；阿加瓦尔，s。；艾哈迈德，l。；阿卡亚，i。；阿勒曼，佛罗里达州；阿尔梅达，d。；Altenschmidt，j。Gpt-4 技术报告。arXiv 2024，arXiv: 2303.08774.

29. Huang,Y.-S.; Yang,Y.-H. Pop Music Transformer: Beat-Based Modeling and Generation of Expressive Pop Piano Compositions. In Proceedings of the MM ’20,28th ACM International Conference on Multimedia, Seattle, WA, USA, 12-16 October 2020; Association for Computing Machinery: New York, NY, USA, 2020;pp. 1180-1188. [CrossRef]

黄，y-s。流行音乐变形金刚： 基于节拍的建模和表现流行钢琴作品的生成。MM’20，第 28 届 ACM 国际多媒体会议论文集，美国华盛顿州西雅图，2020 年 10 月 12-16 日；计算机械协会： 纽约，纽约，美国，2020 年；第 1180-1188 页。[crossre f]

30. Monteith, K.;Martinez, T.R.; Ventura, D. Automatic Generation of Melodic Accompaniments for Lyrics. In Proceedings of the ICCC, Dublin, Ireland, 30 May-1 June 2012; pp. 87-94.

歌词旋律伴奏的自动生成。In Proceedings of the icc，Dublin，Ireland，30 may-June 12012; pp. 87-94 爱尔兰都柏林，2012 年 5 月 30 日至 6 月 1 日；。

31. Ackerman, M.; Loker, D. Algorithmic Songwriting with ALYSIA. In Proceedings of the Computational Intelligence in Music, Sound, Art and Design; Correia, J., Ciesielski, V., Liapis, A., Eds.; Springer International Publishing: Cham, Switzerland,2017;pp.1-16.

Ackerman，m。; Loker，d。与 ALYSIA 的算法歌曲创作。音乐，声音，艺术和设计的计算智能学报；Correia，j. ，Ciesielski，v. ，Liapis，a. ，编辑；Springer International Publishing: Cham，Switzerland，2017; pp. 1-16。

32. Bao, H.; Huang, S.; Wei,F; Cui, L.; Wu,Y; Tan,C.; Piao, S.; Zhou, M. Neural Melody Composition from Lyrics. arXiv 2019, arXiv:1809.04318.[CrossRef]

鲍、黄、魏、崔、李、吴、谭、朴、周。歌词中的神经旋律作品。arXiv 2019，arXiv: 1809.04318.[crossre f]

33. Yu,Y.; Srivastava, A.; Canales, S. Conditional LSTM-GAN for Melody Generation from Lyrics. ACM Trans. Multimedia Comput. Commun. Appl. 2021,17,1-20. [CrossRef]

从歌词生成旋律的条件 LSTM-GAN。 ACM 翻译。多媒体计算机。公共应用。2021,17,1-20。[交叉参考]

34. Srivastava, A.; Duan, W.; Shah, R.R.; Wu,J.; Tang,S.; Li, W.; Yu,Y. Melody Generation from Lyrics Using Three Branch Conditional LSTM-GAN. In Proceedings of the MultiMedia Modeling;Por Jonsson, B., Gurrin,C., Tran, M.-T.,Dang-Nguyen, D.-T., Hu, A.M.-C., Huynh Thi Thanh, B., Huet,B., Eds., Springer International Publishing: Cham, Switzerland,2022; pp.569-581.

Srivastava，a. ; Duan，w. ; Shah，r.r. ; Wu，j. ; Tang，s. ; Li，w. ; Yu，y. 使用三分支条件 LSTM-GAN 从歌词生成旋律。多媒体模型学报；Por
Jonsson，b。，Gurrin，c。，Tran，m.-t。，Dang-Nguyen，d.-t。，Hu，a.m.-c。，Huynh Thi Thanh，b。，Huet，b。，编辑，Springer International Publishing: Cham，Switzerland，2022; pp. 569-581。

35. Yu,Y.; Zhang,Z.; Duan, W.; Srivastava, A.; Shah, R.; Ren, Y. Conditional Hybrid GAN for Melody Generation from Lyrics. Neural Comput. Appl. 2023,35,3191-3202. [CrossRef]

Yu，y. ; Zhang，z. ; Duan，w. ; Srivastava，a. ; Shah，r. ; Ren，y. 从歌词生成旋律的条件混合 GAN。神经计算。应用程序。2023,35,3191-3202.[crossre f]

36. Zhang,Z.;Yu,Y.; Takasu, A. Controllable Lyrics-to-Melody Generation. Neural Comput. Appl. 2023, 35, 19805-19819. [CrossRef] 37.. Sheng, Z.; Song, K.; Tan,X.; Ren, Y.; Ye, W.; Zhang, S.; Qin, T. SongMASS: Automatic Song Writing with Pre-Training and Alignment Constraint. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtual, 19-21 May 2021; Volume 35, pp.13798-13805.

张，z。；余，y。；高须，a。可控歌词旋律生成。神经计算。应用程序。2023,35,19805-19819 年。[交叉参考] 37。.盛，z。；宋，k。；谭，x。；任，y。；叶，w。；张，s。；秦，t。宋弥斯： 预训练和校准约束下的自动歌曲创作。2021 年 5 月 19-21 日，AAAI 人工智能会议论文集，虚拟版，第 35 卷，第 13798-13805 页。

38. Ju,Z.; Lu,P.; Tan,X.; Wang, R.; Zhang, C.; Wu, S.; Zhang, K.; Li,X.-Y.; Qin, T.; Liu, T.-Y. TeleMelody: Lyric-to-Melody Generation with a Template-Based Two-Stage Method. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Abu Dhabi, United Arab Emirates, 7-11 December 2022; pp. 5426-5437. [CrossRef]

谭，x; 王，r; 张，c; 吴，s; 张，k; 李，x-y; 秦，t; 刘，t-y。TeleMelody: Lyric-to-Melody Generation with a Template-Based Two-Stage Method.2022 年自然语言处理经验方法

会议论文集，阿布扎比，阿拉伯联合酋长国，2022 年 12 月 7-11 日；第 5426-5437 页。[crossre f]

39. Ding, S.; Liu,Z.; Dong,X.; Zhang,P.; Qian, R.;He, C.; Lin, D.; Wang,J. SongComposer: A Large Language Model for Lyric and Melody Composition in Song Generation. arXiv 2024, arXiv:2402.17645. [CrossRef]

作曲家： 宋代抒情与旋律创作的大语言模式。arXiv 2024，arXiv: 2402.17645.[crossre f]

40. Davis, H.; Mohammad, S. Generating Music from Literature. In Proceedings of the 3rd Workshop on Computational Linguistics for Literature (CLFL), Gothenburg, Sweden, 27 April 2014; Association for Computational Linguistics: Gothenburg, Sweden,2014;pp.1-10.

穆罕默德・戴维斯，《从文学中产生音乐》。2014 年 4 月 27 日，瑞典戈森堡，计算机语言学协会： 瑞典戈森堡，第三届文学计算语言学研讨会论文集，第 1-10 页。

41... Rangarajan, R. Generating Music from Natural Language Text. In Proceedings of the 2015 Tenth International Conference on Digital Information Management (ICDIM), Jeju, Republic of Korea, 21-23 October 2015;pp. 85-88. [CrossRef] 42.. Zhang, Y; Wang, Z.; Wang, D.;Xia, G. BUTTER: A Representation Learning Framework for Bi-Directional Music-Sentence Retrieval and Generation. In Proceedings of the 1st Workshop on NLP for Music and Audio (NLP4MusA), Online, 16 October 2020; Oramas, S., Espinosa-Anke, L., Epure, E.,Jones, R., Sordo, M., Quadrana, M., Watanabe, K., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020;pp. 54-58.

41... Rangarajan，r。从自然语言文本生成音乐。2015 年第十届数字信息管理国际会议论文集 (ICDIM) ，大韩民国济州，2015 年 10 月 21-23 日；第 85-88 页。[crossre f]42..张，y; 王，z。；王，d。；夏，g。巴特： 双向音乐句子提取和生成的表征学习框架。在音乐和音频的 NLP (NLP4MusA) 第一次研讨会的会议记录，在线，2020 年 10 月 16 日；Oramas，s。，Espinosa-Anke，l。，e。，Jones，r。，Sordo，m。，Quadrana，m。，Watanabe，k。，编辑；计算机语言学协会： 斯特劳兹堡，宾夕法尼亚州，美国，2020; 第 54-58 页。

43. V Wu, S.;Sun, M. Exploring the Efficacy of Pre-Trained Checkpoints in Text-to-Music Generation Task. arXiv 2023, arXiv:2211.11216. Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Roziere, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al.

探索文本到音乐生成任务中预训练检查点的功效。arXiv 2023，arXiv: 2211.11216.马丁内，x。; Lachaux，m.-a。; Lacroix，t。; Roziere，b。; Goyal，n。; Hambro，e。; Azhar，f。

LLaMA: Open and Efficient Foundation Language Models. arXiv 2023, arXiv:2302.13971.

LLaMA: 开放和高效的基础语言模型.arXiv 2023，arXiv: 2302.13971。

45. Yuan, R.; Lin, H.; Wang,Y; Tian, Z.; Wu, S.; Shen, T.; Zhang,G.; Wu,Y.; Liu, C.; Zhou, Z.; et al. ChatMusician: Understanding and Generating Music Intrinsically with LLM. arXiv 2024, arXiv:2402.16153. [CrossRef]

袁，r。; 林，h。; 王，y。; 田，z。; 吴，s。; 沈，t。; 张，g。; 吴，y。; 刘，c。; 周，z。; 等。ChatMusician: 用 LLM 从本质上理解和生成音乐。arXiv 2024，arXiv: 2402.16153.[crossre f]

46. Liang,X.; Du, X.; Lin,J.; Zou, P.; Wan,Y.; Zhu, B. ByteComposer: A Human-like Melody Composition Method Based on Language Model Agent. arXiv 2024, arXiv:2402.17785.

一种基于语言模型 Agent 的类人旋律创作方法。arXiv 2024，arXiv: 2402.17785.

47.7. Deng,Q.; Yang,Q.; Yuan, R.; Huang,Y.; Wang,Y.; Liu,X.; Tian, Z.; Pan, J.; Zhang,G.; Lin, H.; et al. ComposerX: Multi-Agent Symbolic Music Composition with LLMs. arXiv 2024, arXiv:2404.18081

47.7.邓，q。; 杨，q。; 袁，r。; 黄，y。; 王，y。; 刘，x。作曲家 x: 具有 llm 的多代理象征性音乐创作。arXiv 2024，arXiv: 2404.18081

48. Liu, H.; Chen,Z.;Yuan,Y.;Mei, X.; Liu,X.; Mandic, D.; Wang, W.; Plumbley, M.D. AudioLDM: Text-to-Audio Generation with Latent Diffusion Models. arXiv 2023, arXiv:2301.12503.

陈，z。; 袁，y。; 梅，x。; 刘，x。; Mandic，d。; Wang，w。; Plumbley，m.d。arXiv 2023，arXiv: 2301.12503.

49. Ghosal, D.;Majumder, N.; Mehrish, A.; Poria, S. Text-to-Audio Generation Using Instruction-Tuned LLM and Latent Diffusion Model. arXiv 2023, arXiv:2304.13731.

使用指令调优的 LLM 和潜在扩散模型的文本到音频的生成.arXiv 2023，arXiv: 2304.13731。

50. Majumder, N.; Hung, C.-Y; Ghosal, D.; Hsu, W.-N.; Mihalcea, R.; Poria, S. Tango 2: Aligning Diffusion-Based Text-to-Audio Generations through Direct Preference Optimization. In Proceedings of the 32nd ACM International

Conference on Multimedia, Melbourne, Australia, 28 October-1 November 2024; pp. 564-572.

探戈2: 通过直接偏好优化对齐基于扩散的文本到音频的生成。在2024年10月28日至11月1日在澳大利亚墨尔本举行的第32届ACM国际多媒体会议论文集中；第564-572页。

51. L. Forsgren,S.;Martiros, H. Riffusion-Stable Diffusion for Real-Time Music Generation. 2022. Available online: https://riffusion.com(accessed on 17 February 2025).

L. Forsgren，s. ; Martiros，h. riffusion - 实时音乐生成的稳定扩散。2022年。可在线获得： https://riffusion.com (2025年2月17日访问)。

52. Huang, Q.; Park, D.S.; Wang, T.; Denk, T.I.; Ly, A.; Chen, N.; Zhang, Z.; Zhang, Z.; Yu, J.; Frank, C.; et al. Noise2Music:Text-Conditioned Music Generation with Diffusion Models. arXiv 2023, arXiv:2302.03917.

王，t。 ; Denk，T.i。 ; Ly，a。 ; Chen，n。 ; Zhang，z。 ; Zhang，z。 ; Yu，j。Noise2Music: 使用扩散模型的文本条件音乐生成。arXiv 2023，arXiv: 2302.03917.

53. Schneider,F; Kamal, O.; Jin, Z.; Scholkopf, B. Mousai: Text-to-Music Generation with Long-Context Latent Diffusion. arXiv 2023, arXiv:2301.11757.[CrossRef]

施耐德，f; 卡迈勒，o. ; 金，z. ; 肖尔科普夫，b. 穆赛： 长上下文潜在扩散的文本到音乐的生成。 arXiv 2023，arXiv: 2301.11757。[交叉参考]

54. Li, P.P.;Chen,B.;Yao,Y.; Wang,Y.; Wang, A.; Wang, A. JEN-1: Text-Guided Universal Music Generation with Omnidirectional Diffusion Models. In Proceedings of the 2024 IEEE Conference on Artificial Intelligence (CAI), Singapore, 25-27 June 2024; pp.762-769.[CrossRef]

李，p.p。 ; 陈，b。 ; 姚，y。 ; 王，y。 ; 王，a。 ; 王，a。 JEN-1: 文本引导的全向扩散模型的通用音乐生成。2024年IEEE人工智能会议论文集 (CAI) ，新加坡，2024年6月25-27日；第762-769页。[crossre f]

55. Agostinelli, A.; Denk, T.I.; Borsos, Z.; Engel, J.; Verzetti, M.; Caillon, A.;Huang,Q;Jansen, A.; Roberts, A.; Tagliasacchi, M.;etal. MusicLM: Generating Music From Text. arXiv 2023, arXiv:2301.11325.

Borsos，z。 ; Engel，j。 ; Verzetti，m。 ; Caillon，a。 ; Huang，q。 ; Jansen，a。 ; Roberts，a。 ; Tagliasacchi，m。MusicLM: 从文本生成音乐。arXiv 2023，arXiv: 2301.11325.

56. Huang,Q.;Jansen, A.; Lee, J.; Ganti, R.; Li, J.Y.; Ellis, D.P.W. MuLan: A Joint Embedding of Music Audio and Natural Language. arXiv 2022, arXiv:2208.12415.

木兰： 音乐音频与自然语言的联合嵌入。arXiv 2022，arXiv: 2208.12415.

57. Lam,M.W.Y.; Tian, Q.; Li, T.; Yin, Z.; Feng, S.; Tu, M.; Ji,Y.; Xia, R.; Ma, M.; Song,X.; et al. Efficient Neural Music Generation. Adv. Neural Inf. Process. Sust. 2023,36, 17450-17463.

林，M.w.y. ; 田，q. ; 李，t. ; 尹，z. ; 冯，s. ; 屠，m. ; 纪，y. ; 夏，r. ; 马，m. ; 宋，x. ; 等人。高效的神经音乐生成。Neural in f 神经音乐。过程。苏斯特。2023,36,17450-17463.

58. Chen, K.; Wu,Y.; Liu , H.; Nezhurina, M.; Berg-Kirkpatrick, T.; Dubnov, S. MusicLDM: Enhancing Novelty in Text-to-Music Generation Using Beat-Synchronous Mixup Strategies. In Proceedings of the ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Seoul, Republic of Korea, 14-19 April 2024; pp. 1206-1210.

Chen，k。; Wu，y。; Liu，h。; Nezhurina，m。; Berg-Kirkpatrick，t。; Dubnov，s。在 ICASSP 2024-2024 IEEE 声学，语音和信号处理国际会议 (ICASSP) 的会议记录，首尔，大韩民国，2024 年 4 月 14-19 日；第 1206-1210 页。

59.. Copet,J.;Kreuk, F; Gat, I.; Remez, T.; Kant, D.; Synnaeve, G.; Adi, Y.; Defossez, A. Simple and Controllable Music Generation. Adv. Neural Inf. Process. Sust.2023, 36,47704-47720.

59..科佩，j。; 克鲁克，f。; 盖特，i。; 雷米兹，t。; 康德，d。; 提奈夫，g。; 阿迪，y。; 德福塞兹，a。《简单可控的音乐生成》。Neural in f 神经信息。过程。Sust 2023,36,47704-477202023 年 3 月 36 日。

60. N Macon, M.W.;Jensen-Link, L.; George,E.; Oliverio, J.C.;Clements, M. Concatenation-Based MIDI-to-Singing Voice Synthesis. J. Audio Eng. Soc. 1997, 4591. Available online: https://secure.aes.org/forum/pubs/conventions/?elib=7188 (accessed on 17February 2025).

梅肯，M.w。; 詹森 - 林克，l。; 乔治，e。; 奥利弗里奥，j.c。; 克莱门茨，m。基于串联的 MIDI-to-Singing 声音合成。音频工程。社会科学。1997 年 4591。在线可用：https://secure.aes.org/forum/pubs/conventions/? elib = 7188 (2025 年 2 月 17 日访问)。

61.. Kenmochi, H.; Ohshita, H. VOCALOID-Commercial Singing Synthesizer Based on Sample Concatenation. In Proceedings of the INTERSPEECH 2007, 8th Annual

Conference of the International Speech Communication Association, Antwerp, Belgium, 27-31August 2007; Volume 2007, pp. 4009-4010.

61..Vocaloid - 基于采样连接的商业歌唱合成器。国际言语交流协会第八届年会，比利时安特卫普，2007 年 8 月 27-31 日；2007 卷，第 4009-4010 页。

62. Saino, K.; Zen, H.; Nankaku,Y.; Lee, A.; Tokuda, K. An HMM-Based Singing Voice Synthesis System. In Proceedings of the Interspeech 2006, Pittsburgh, PA, USA, 17-21 September 2006; ISCA: Singapore, 2006.

佐野，k。；禅，h。；南库，y。一个基于 hmm 的歌声合成系统。In Proceedings of the interspeak 2006，Pittsburgh，PA，USA，17-21 September 2006; ISCA: Singapore，2006 美国宾夕法尼亚州匹兹堡，2006 年 9 月 17-21 日；。

63. Nishimura, M.; Hashimoto, K.; Oura, K.;Nankaku,Y.; Tokuda, K. Singing Voice Synthesis Based on Deep Neural Networks. In Proceedings of the Interspeech 2016, San Francisco, CA, USA, 8-12 September 2016; ISCA: Singapore, 2016;pp. 2478-2482.

基于深度神经网络的歌声合成。In Proceedings of the interspeak 2016，San Francisco，CA，USA，8-12 September 2016; ISCA: Singapore，2016; pp. 2478-2482.

64. Nakamura, K.; Hashimoto, K.; Oura, K.; Nankaku,Y.; Tokuda, K. Singing Voice Synthesis Based on Convolutional Neural Networks. arXiv 2019, arXiv:1904.06868.

基于卷积神经网络的歌声合成。

65. Kim, J.;Choi, H.;Park, J.; Kim, S.; Kim,J.;Hahn,M. Korean Singing Voice Synthesis System Based on an LSTM Recurrent Neural Network. In Proceedings of the Interspeech, Hyderabad, India, 2-6September 2018; pp. 1551-1555.

基于 LSTM 递归神经网络的韩国歌声合成系统。In Proceedings of the Interspeech，Hyderabad，India，2018,2-6 september; pp. 1551-1555.

66. Hono,Y.; Hashimoto, K.; Oura, K.;Nankaku, Y.; Tokuda, K. Singing Voice Synthesis Based on Generative Adversarial Networks. In Proceedings of the ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12-17 May 2019;pp. 6955-6959. [CrossRef]

基于生成对抗网络的歌声合成。在 ICASSP 2019-2019 IEEE 声学，语音和信号处理国际会议 (ICASSP) 的会议记录，英国布莱顿，2019 年 5 月 12 日至 17 日；第 6955-6959 页。[crossre f]

67. Lu,P.; Wu,J.; Luan, J.; Tan,X.;Zhou, L. XiaoiceSing: A High-Quality and Integrated Singing Voice Synthesis System. arXiv 2020, arXiv:2006.06261.

陆平；吴建军；栾建军；谭晓晓；周丽。肖冰星： 一个高质量的综合性歌唱声音合成系统。

68. Ren, Y; Ruan, Y.; Tan, X.; Qin, T.; Zhao , S.; Zhao, Z.; Liu, T.-Y. FastSpeech: Fast, Robust and Controllable Text to Speech. arXiv2019,arXiv:1905.09263.

任，y。；阮，y。；谭，x。；秦，t。；赵，s。Fast Speech: 快速、健壮、可控的文本语音转换。arXiv2019，arXiv: 1905.09263.

69. Morise,M.; Yokomori, F; Ozawa, K. World: A Vocoder-Based High-Quality Speech Synthesis System for Real-Time Applications. IEICE Trans. Inf. Syst. 2016,99, 1877-1884. [CrossRef]

世界： 基于声码器的实时高质量语音合成系统。IEICE Trans.译者： 王士杰。系统。2016 年 9 月 9 日，1877-1884 年。[crossre f]

70. Blaauw, M.; Bonada, J. Sequence-to-Sequence Singing Synthesis Using the Feed-Forward Transformer. In Proceedings of the ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4-8May 2020;pp.7229-7233.[CrossRef] 71.. Zhuang, X.; Jiang, T.; Chou, S.-Y.; Wu, B.; Hu, P.; Lui, S. Litesing: Towards Fast, Lightweight and Expressive Singing Voice Synthesis. In Proceedings of the ICASSP 2021–2021 IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP), Toronto, ON, Canada, 6-11 June 2021; p. 7082.

Blauw，m。；Bonada，j。使用前馈变换器的序列到序列的歌唱合成。ICASSP 2020 一 2020 IEEE 声学、语音和信号处理国际会议 (ICASSP) 论文集，西班牙巴塞罗那，2020 年 5 月 4-8 日；第 7229-7233 页。[crossre f]71..Zhuang，x。；Jiang，t。；Chou，s.-y。；Wu，b。；Hu，p。；Lui，s。在 ICASSP 2021-2021 IEEE 声学，语音和信号处理国际会议 (ICASSP) 的会议记录，多伦多，ON，加拿大，2021 年 6 月 6 日至 11 日；第 7082 页。

72. Lee, G.-H.; Kim, T.-W.; Bae, H.; Lee, M.-J.; Kim,Y.-I.; Cho, H.-Y. N-Singer: A Non-Autoregressive Korean Singing Voice Synthesis System for Pronunciation Enhancement. arXiv 2022, arXiv:2106.15205v2.

李、李、金、裴、李、李、金、赵、李。N-Singer: 一个用于发音增强的非自回归韩语歌唱声音合成系统。arXiv 2022，arXiv: 2106.15205 v2.

73. Chen, J.; Tan,X.; Luan, J.; Qin, T.; Liu, T.-Y. HiFiSinger: Towards High-Fidelity Neural Singing Voice Synthesis. arXiv 2020, arXiv:2009.01776.

陈，陈，x。；栾，j。；秦，t。；刘，T.-Y。 hiisinger: 走向高保真神经歌唱声音合成。arxiv2020，arXiv: 2009.01776。

74. Yamamoto, R.; Song, E.; Kim,J.-M. Parallel WaveGAN: A Fast Waveform Generation Model Based on Generative Adversarial Networks with Multi-Resolution Spectrogram. In Proceedings of the ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4-8 May 2020.

山本，r。; 宋，e。并行 WaveGAN: 基于多分辨率谱图生成对抗网络的快速波形生成模型。在 ICASSP 2020-2020 IEEE 声学，语音和信号处理国际会议 (ICASSP) 的会议记录，巴塞罗那，西班牙，2020 年 5 月 4 日至 8 日。

75. Wang,Y.; Skerry-Ryan, R.J.; Stanton, D.; Wu,Y.; Weiss, R.J.; Jaitly, N.; Yang,Z.;Xiao,Y.; Chen, Z.; Bengio, S.; et al. Tacotron:Towards End-to-End Speech Synthesis. arXiv 2017, arXiv:1703.10135.

Skerry-Ryan，r.j。; Stanton，d。; Wu，y。; Weiss，r.j。; Jaitly，n。; Yang，z。Tacotron: 走向端到端的语音合成。arXiv 2017，arXiv: 1703.10135.

76. Gu,Y.;Yin,X.; Rao,Y.; Wan,Y.; Tang, B.;Zhang,Y.; Chen, J.; Wang,Y.; Ma, Z. ByteSing: A Chinese Singing Voice Synthesis System Using Duration Allocated Encoder-Decoder Acoustic Models and WaveRNN Vocoders. arXiv 2021, arXiv:2004.11012.

顾勇；殷晓；饶勇；万勇；唐斌；张勇；陈建；王勇；马志。字节编码： 一个使用时长分配的编码器 - 解码器声学模型和 WaveRNN 声码器的中文歌唱声音合成系统。arXiv 2021，arXiv: 2004.11012.

77. Liu,J.; Li,C.; Ren,Y.; Chen, F.; Zhao, Z. DiffSinger: Singing Voice Synthesis via Shallow Diffusion Mechanism. arXiv 2022, arXiv:2105.02446.[CrossRef]

刘建；李春；任永；陈芳；赵春。迪夫辛格： 基于浅扩散机制的歌唱声音合成。 arXiv 2022，arXiv: 2105.02446。[交叉参考]

78. Hono, Y.;Hashimoto, K.; Oura, K.; Nankaku, Y; Tokuda, K. Sinsy: A Deep Neural Network-Based Singing Voice Synthesis System. Proc. IEEE/ACM Trans. Audio Speech Lang. Process. 2021,29,2803-2815.[CrossRef]

基于深度神经网络的歌唱声音合成系统。过程。IEEE/ACM Trans IEEE/ACM Trans.音频语音朗。过程。2021,29,2803-2815 年。[crossre f]

79. Zhang,Y;Cong,J.;Xue, H.; Xie, L.; Zhu, P.; Bi,M. VISinger: Variational Inference with Adversarial Learning for End-to-End Singing Voice Synthesis. arXiv 2022, arXiv:2110.08813.

张，y; 丛，j. ; 薛，h. ; 谢，l. ; 朱，p. ; 毕，M.VISinger: 端到端歌唱声音合成的对抗学习变分推理。arXiv 2022，arXiv: 2110.08813.

80. Zhang,Y.; Xue, H.; Li, H.; Xie, L.; Guo, T.; Zhang, R.; Gong, C. VISinger 2: High-Fidelity End-to-End Singing Voice Synthesis Enhanced by Digital Signal Processing Synthesizer. arXiv 2022, arXiv:2211.02903.

张云；薛海；李海；谢丽；郭涛；张红；龚春。《视觉 2: 数字信号处理合成器增强的高保真端到端歌声合成》。arXiv 2022，arXiv: 2211.02903.

81. Kim, J.; Kong, J.; Son, J. Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech. In Proceedings of the 38th International Conference on Machine Learning, Virtual, 18-24 July 2021; pp. 5530-5540.

用于端到端文本到语音的对抗学习的条件变分自动编码器。In Proceedings of the 38th International Conference on Machine Learning，Virtual，18-24 July 2021; pp. 5530-5540.

82.. I Hong,Z.;Huang,R.; Cheng,X.;Wang,Y.; Li,R.; You, F;Zhao,Z.;Zhang,Z. Text-to-Song: Towards Controllable Music Generation Incorporating Vocals and Accompaniment. arXiv 2024, arXiv:2404.09313.

82. i Hong，z. ; Huang，r. ; Cheng，x. ; Wang，y. ; Li，r. ; You，f. ; Zhao，z. ; Zhang，z. 文本到歌曲： 走向结合声乐和伴奏的可控音乐生成，arXiv 2024，arXiv: 2404.09313。

83. Dhariwal, P.; Jun, H.; Payne, C.; Kim, J.W.; Radford, A.; Sutskever, I. Jukebox: A Generative Model for Music. arXiv 2020, arXiv:2005.00341.

点唱机： 音乐的生成模型。2020 年，arXiv: 2005.00341。

84. Bai, Y.; Chen, H.; Chen, J.; Chen,Z.; Deng,Y.; Dong, X.; Hantrakul, L.; Hao, W.; Huang, Q.;Huang, Z.;et al. Seed-Music: A Unified Framework for High Quality and Controlled Music Generation. arXiv 2024, arXiv:2409.09214.

白，y。；陈，h。；陈，j。；陈，z。；邓，y。；董，x。；汉特拉库尔，l。；郝，w。；黄，q。；黄，z。；等。Seed-Music: a Unified Framework for High Quality and Controlled Music Generation 种子音乐： 高质量和受控音乐生成的统一框架。arXiv 2024，arXiv: 2409.09214.

85.. Chen,G.; Liu,Y.; Zhong, S.; Zhang,X. Musicality-Novelty Generative Adversarial Nets for Algorithmic Composition. In Pro-ceedings of the 26th ACM international conference on Multimedia, Seoul, Republic of Korea, 22-26 October 2018; pp. 1607-1615.[CrossRef]

85..算法音乐创作的新颖生成对抗网络。在 2018 年 10 月 22 日至 26 日在韩国首尔举行的第 26 届 ACM 多媒体国际会议上；第 1607-1615 页。[crossre f]

86. Hakimi, S.H.; Bhonker, N.; El-Yaniv, R. BebopNet: Deep Neural Models for Personalized Jazz Improvisations. In Proceedings of the ISMIR, Online, 11-15 October 2020; pp. 828-836.

BebopNet: 个性化爵士即兴创作的深度神经模型。In Proceedings of the ISMIR，Online，11-15 October 2020; pp. 828-836.

87. Monteith, K.; Martinez, T.R.; Ventura, D. Automatic Generation of Music for Inducing Emotive Response. In Proceedings of the ICCC, Lisbon, Portugal,7-9January 2010; pp. 140-149.

蒙泰斯，k。；马丁内斯，t.r。；文图拉，d。用于诱导情绪反应的音乐的自动生成。2010 年 1 月 7-9 日，葡萄牙里斯本，《国际刑事法院会议记录》第 140-149 页。

88. Hung, H.-T.;Ching,J.;Doh, S.; Kim,N.;Nam,J.;Yang,Y.-H. EMOPIA: A Multi-Modal Pop Piano Dataset For Emotion Recognition and Emotion-Based Music Generation. arXiv 2021, arXiv:2108.01374.

洪，h.-t。；Ching，j。；Doh，s。；Kim，n。；Nam，j。；yang，y。 EMOPIA: 用于情绪识别和基于情绪的音乐生成的多模态流行钢琴数据集。 arXiv 2021，arXiv: 2108.01374。

89. Zheng,K.; Meng,R.;Zheng,C.;Li,X.; Sang,J;Cai,J.; Wang,J; Wang,X. EmotionBox: A Music-Element-Driven Emotional Music Generation System Based on Music Psychology. Front. Psychol. 2022, 13, 841926. [CrossRef] [PubMed]

基于音乐心理学的音乐元素驱动的情感音乐生成系统。前沿。疯子。2022,13,841926. [crossre f][PubMed]

90. Neves, P.; Fornari, J;Florindo, J. Generating Music with Sentiment Using Transformer-GANs. arXiv 2022, arXiv:2212.11134.

使用 Transformer-GANs 生成带有情感的音乐.arXiv 2022，arXiv: 2212.11134。

91. Dash, A.; Agres, K.R. AI-Based Affective Music Generation Systems: A Review of Methods, and Challenges. arXiv 2023, arXiv:2301.06890.[CrossRef]

基于人工智能的情感音乐生成系统： 方法与挑战的回顾，arXiv 2023，arXiv: 2301.06890

92. Ji, S.;Yang,X. EmoMusicTV: Emotion-Conditioned Symbolic Music Generation with Hierarchical Transformer VAE. IEEE Trans. Multimed. 2024,26, 1076-1088.[CrossRef]

情绪音乐电视： 情绪条件符号音乐生成与等级变压器 VAE。 IEEE 翻译。 Multimed。2024,26,1076-1088。[crossre f]

93. Jiang,N.;Jin,S.;Duan, Z.;Zhang, C. RL-Duet: Online Music Accompaniment Generation Using Deep Reinforcement Learning. In Proceedings of the AAAI

Conference on Artificial Intelligence, Hilton, NY, USA, 7-12 February 2020; Volume 34, pp. 710-718.

江，n. ; 金，s. ; 段，z. ; 张，c. rl-duo: 使用深度强化学习的在线音乐伴奏生成。2020 年 2 月 7-12 日在美国纽约希尔顿召开的 AAAI 人工智能会议论文集，第 34 卷，第 710-718 页。

94. Louie,R.; Coenen, A.;Huang, C.Z.; Terry, M.;Cai,C.J. Novice-AI Music Co-Creation via AI-Steering Tools for Deep Generative Models. In Proceedings of the CHI 20,2020 CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA,25-30 April2020; Association for Computing Machinery: New York, NY, USA, 2020;pp. 1-13. [CrossRef]

科南，a。；黄，c.z。；特里，m。；蔡，c.j。 novice - 人工智能音乐协同创作通过人工智能 - 深层生成模型的指导工具。在 CHI 20,2020 关于计算系统中的人为因素的会议论文集，火奴鲁鲁，HI，美国，2020 年 4 月 25-30 日；计算机械协会： 纽约，纽约，美国，2020 年；第 1-13 页。[crossre f]

95. Kwon, H. AudioGuard: Speech Recognition System Robust against Optimized Audio Adversarial Examples. Multimed. Tools Appl. 2024, 83, 57943-57962. [CrossRef]

AudioGuard: 语音识别系统对优化音频对抗示例的鲁棒性。 Multimed.Tools Appl.2024,83,57943-57962。[crossre f]

96. Surbhi, A.; Roy, D. Tunes of Tomorrow: Copyright and AI-Generated Music in the Digital Age. AIP Conf. Proc. 2024, 3220, 050003.[CrossRef] 97.. Hsiao, W.-Y.; Liu, J.-Y.; Yeh,Y.-C.; Yang, Y.-H. Compound Word Transformer: Learning to Compose Full-Song Music over Dynamic Directed Hypergraphs. Proc. AAAI Conf. Artif. Intell. 2021,35, 178-186.[CrossRef]

Surbhi，a. ; Roy，d. Tunes of Tomorrow: 数字时代的版权与人工智能生成音乐。AIP 会议。过程。2024,3220,050003.[crossre f]97..萧伟；刘继扬；叶永庆；杨永庆。复合词转换器： 在动态有向超图上学习创作全歌曲音乐。译者注：。AAAI 会议。Artif 艺术。英特尔。2021 年，35 年，178-186 年。[crossre f]

98. Josan, H.H.S. AI and Deepfake Voice Cloning: Innovation, Copyright and Artists’Rights; Centre for International Governance Innovation: Waterloo, ON, Canada, 2024.

人工智能与深假声音克隆： 创新、版权与艺术家权利；国际治理创新中心： 滑铁卢，加拿大，2024。

99. Zhang,Z.;Ning, H.; Shi,F.; Farha,F;Xu,Y.;Xu,J.; Zhang,F; Choo, K.-K.R. Artificial Intelligence in Cyber Security: Research Advances, Challenges, and Opportunities. Artif. Intell. Reo. 2022, 55,1029-1053. [CrossRef]

网络安全中的人工智能： 研究进展、挑战与机遇。人工智能。英特尔。里奥。2022,55,1029-1053 年。[crossre f]

100. Fox, M.;Vaidyanathan,G.; Breese, J.L. The Impact of Artificial Intelligence on Musicians. Issues Inf. Syst. 2024, 25, 267-276.

人工智能对音乐家的影响〉 ，《信息系统》2024,25,267-276。

101. Liu,X.; Dong,Z.; Zhang,P. Tackling Data Bias in Music-Avqa: Crafting a Balanced Dataset for Unbiased Question-Answering. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3-8 January 2024; pp.4478-4487.

Tackling Data Bias in Music-Avqa: Crafting a Balanced Dataset for Unbiased Question-Answering.IEEE/cv f 计算机视觉应用冬季会议论文集，Waikoloa，HI，USA，2024 年 1 月 3-8 日；第 4478-4487 页。

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

免责声明 / 出版商声明： 所有出版物中的声明、观点和数据仅为个人作者和投稿人的声明、观点和数据，而非 MDPI 和 / 或编辑的声明、观点和数据。MDPI 和 / 或编辑拒绝对内容中提到的任何思想、方法、指示或产品造成的人员或财产损失承担责任。