

第三届  
The Third

# 世界音乐人工智能大会

Summit on Music Intelligence

SOMI 中国·北京  
2026 04.25-26

 中国人工智能学会  
Chinese Association for Artificial Intelligence

 中央音乐学院  
CENTRAL CONSERVATORY  
OF MUSIC

科学与艺术的交融

艺术与科学的盛会

祝世界音乐人工智能大会

圆满成功!

李政道

二〇二一年十月



国际物理学家、诺贝尔物理学奖获得者  
李政道先生 (1926-2024) 给大会的寄语



大会主席  
**于红梅**

---

中央音乐学院院长  
教授、博士研究生导师  
中国民族管弦乐学会胡琴专业委员会会长



共同主席  
**戴琼海**

---

国务院参事  
中国工程院院士  
中国人工智能学会理事长



执行主席  
**李小兵**

---

中央音乐学院音乐人工智能系主任、教授  
中国人工智能学会理事、艺术与AI专委会主任  
中国计算机学会理事、计算艺术分会主任

## 大会名誉主席



俞峰



Jean-Michel Jarre



郭毅可 院士



管晓宏 院士

# 开幕式

# OPENING CEREMONY

## 2026/04/26 Sun.

09:00-11:35 | 中央音乐学院琴房楼演奏厅 Recital Hall of the Practice Building, CCOM

第三届世界音乐人工智能大会开幕式  
SOMI2026 Opening Ceremony

### 大会主旨报告

Keynote Speech



管晓宏

Guan Xiaohong

音乐智能量化与认知的研究进展

Progress on Computational Intelligence and Quantitate Cognition of Music



克里斯·查菲

Chris Chafe

聆听数据：以数据声化为音乐与科学打造定制化计算机音乐网络应用

Listening to Data: Creating Custom Computer Music Webapps for Music and Science through Data Sonification



乔治·海杜

Georg Hajdu

从音乐厅到社会空间：借助技术重新语境化当代音乐

From Concert Hall to Social Space: Recontextualizing Contemporary Music through Technology



李小兵

Li Xiaobing

机文主义：音乐学院的未来在哪里？

Machinism: Where Is the Future of Music Conservatories

# 日程安排

## 2026/04/25 Sat.

10:00-18:00 | 中央音乐学院西门 West Gate, CCOM

报到注册  
Registration

14:00-16:00 | 中央音乐学院教学楼 701 701, Academic Building, CCOM

第三届世界音乐人工智能大会青年论坛  
SOMI2026 Youth Forum



戴琮人  
Dai Congren

音乐全谱理解基准：大模型对完整乐谱理解能力的评测与分析  
Musical Score Understanding Benchmark: Evaluating Large Language Models' Comprehension of Complete Musical Scores



丘治平  
Qiu Zhiping

音频驱动的弦乐演奏动作生成  
ELGAR: Expressive Cello Performance Motion Generation for Audio Rendition



童心怡  
Tong Xinyi

音画共鸣：视频配乐生成的视觉画面、时间节奏与音乐表达对齐  
Video Echoed in Music: Semantic, Temporal, and Rhythmic Alignment for Video-to-Music Generation



吴尚达  
Wu Shangda

CLaMP 3: 跨未对齐模态与未见语言的通用音乐信息检索  
CLaMP 3: Universal Music Information Retrieval Across Unaligned Modalities and Unseen Languages

16:00-17:00 | 中央音乐学院教学楼 717 717, Academic Building, CCOM

中国人工智能学会艺术与人工智能专委会会议  
Meeting of the Art and Artificial Intelligence Technical Committee, Chinese Association for Artificial Intelligence

2026/04/26 Sun.

09:00-11:35 | 中央音乐学院琴房楼演奏厅 Recital Hall of the Practice Building, CCOM

第三届世界音乐人工智能大会开幕式及主旨报告

SOMI2026 Opening Ceremony & Keynote Speech

14:00-15:30 | 中央音乐学院琴房楼演奏厅 Recital Hall of the Practice Building, CCOM

第三届世界音乐人工智能大会学术论坛

SOMI2026 Academic Forum



刘家丰

Liu Jiafeng

统一声学 Token 空间的音乐生成大模型：从深层表征到高质量生成

Large-Scale Music Generation Model with a Unified Acoustic Token Space: From Deep Representations to High-Quality Generation



马军

Ma Jun

音乐脑机接口：概念、研究进展与应用前景

Music Brain-Computer Interfaces: Concepts, Research Progress, and Application Prospects



卢迪

Lu Di

面向 AI 生成音乐工作流的 Web DAW

A Web-based DAW for AI-generated Music Workflow



肯尼斯·菲尔兹

Kenneth Fields

面向网络化电子音乐合奏的 AI 编程助手的实践应用

Practical Applications of AI Coding Assistants for Networked Electronic Music Ensembles



亚伦·威廉姆森

Aaron Williamon

表演科学的未来

The Future of Performance Science



凯特·霍普

Cal Hope

确保音乐领域人工智能政策的包容性与可持续性：一个治理议题

Ensuring Inclusive and Sustainable AI Policy for The Music Sector: A Governance Issue

15:40-17:00 | 中央音乐学院琴房楼演奏厅 Recital Hall of the Practice Building, CCOM

### 第三届世界音乐人工智能大会产业论坛

SOMI2026 Industry Forum



徐帆  
Xu Fan

写歌，正在从创作变成选择吗？  
Is Songwriting Becoming Selection Rather Than Creation?



龚俊民  
Gong Junmin

推动开源音乐生成的边界  
Pushing the Boundaries of Open-Source Music Generation



姜涛  
Jiang Tao

从一杯奶茶到音乐创作和消费的 Agent  
From a Cup of Milk Tea to Agents for Music Creation and Consumption



刘晓光  
Liu Xiaoguang

AI 赋能音乐教育  
AI-Empowered Music Education

17:00-18:00 | 中央音乐学院琴房楼演奏厅 Recital Hall of the Practice Building, CCOM

### 第三届世界音乐人工智能大会主题交流、闭幕式

SOMI2026 Panel Discussion & Closing Ceremony

全天 | 线上 Online

### 第三届世界音乐人工智能大会电子音乐马拉松

SOMI2026 Electronic Music Marathon



管晓宏

Xiaohong Guan

## 音乐智能量化与认知的研究进展

Progress on Computational Intelligence and Quantitate Cognition of Music

## 报告摘要 ABSTRACT

音乐智能量化与认知，特别是 AI 音乐生成受到极大关注。隐藏在优美旋律中的数学物理规律，与众多自然、工程和社会系统中的规律一致，能够定量分析。音乐旋律的三个数学特征，揭示了作曲家追求旋律变化的有约束熵最大，因而存在音乐旋律的普遍规律。报告将讨论音乐智能量化与脑科学认知的研究进展，包括音乐计算智能的分析，建立音乐认知机制的脑科学基础，构建基于脑机测量的音乐智能与认知的量化关联模型，在人工智能作曲等方面的应用。

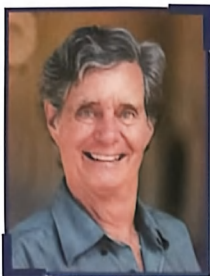
Computational intelligence and quantitate cognition of music especially AI facilitated music generation is a highly attentional research topic. The Power Law hidden in music melody variations is consistent with the common law in many natural, engineering and social systems, and can also be quantified. The three mathematical characteristics are discovered in the melodies of all tonal music. Based on these characteristics it is found that all composers of different styles in the past three hundred years were casually pursuing the constrained entropy maximization in melody composition. The speech will report the research progress on developing the systematic methods for analyzing music computational intelligence, facilitating AI aided composition and establishing the basis on music cognition in brain science.

## 个人简介 BIOGRAPHY

中国科学院院士，IEEE Fellow，分别于1982、1985年获清华大学学士与硕士学位，1993年获美国康涅狄格大学博士学位；1993-1995年任美国PG&E公司高级顾问工程师，1999-2000年任哈佛大学访问科学家，1995年起任西安交通大学教授，2008-2025任电子与信息工程学院院长、电子与信息学部主任；自2001年任清华大学讲席教授组成员，2003-2008年任清华大学自动化系主任；中央音乐学院音乐人工智能与信息科学团队成员。

管晓宏院士主要从事复杂网络化系统的经济性与安全性，电力、能源、制造系统优化，信息物理融合系统，网络空间信息安全等领域的研究，同时开展音乐智能量化和信息处理的研究，曾获2005年、2018年国家自然科学二等奖，2019年何梁何利科技进步奖及多项国际学术奖励。近年来，管晓宏院士与中央音乐学院、西安音乐学院合作，担任中央音乐学院博士生导师，探讨艺术与科学的关系和相互影响，在音乐智能量化领域取得重要研究成果，并创办了“艺术与科学的交汇”系列音乐会。

Guan Xiaohong, the member of Chinese Academy of Science and the Fellow of IEEE, received his B.S. and M.S. degrees from Tsinghua University, Beijing, China, in 1982 and 1985, respectively, and his Ph.D. degree from the University of Connecticut in 1993. He was a senior consulting engineer with Pacific Gas and Electric from 1993 to 1995. He visited the Division of Engineering and Applied Science, Harvard University from 1999 to 2000. From 1985 to 1988 and since 1995 he has been with Xian Jiaotong University, Xian, China as the Cheung Kong Professor of Systems Engineering since 1999, and from 2008 to 2025 as Dean of Faculty of Electronic and Information Engineering. From 2001 he has also been with the Center for Intelligent and Networked Systems, Tsinghua University, Beijing, China, and severed the Head of Department of Automation, Tsinghua University, 2003-2008. Professor Guan is also with Department of Music AI and Information Science, Central Conservatory of Music of China.



克里斯·查菲  
Chris Chafe

## 聆听数据：以数据声化为音乐与科学打造定制化计算机音乐网络应用

Listening to Data: Creating Custom Computer Music Webapps for Music and Science through Data Sonification

### 报告摘要 ABSTRACT

声音乐化 (Sonification) 是将数据转译为声音与音乐的过程。我的视频展示了一些用于科学与艺术目的的实例，这些实例通过“音乐性的聆听”来帮助人们感知数据中的趋势。网页应用 (Webapps) 提供了一种普遍可及的方式来传播实时声音乐化成果，并且可以通过输入给大语言模型 (LLM) 的文本提示，较为方便地完成编程。视频中展示了我近年来在工作坊中使用的一套工作流程，面向的是有兴趣自行创作声音乐化作品的科学家和音乐家。

Sonification is the translation of data to sound and music. My video presents examples for scientific and artistic purposes that use "musical listening" to make a connection to trends in the data. Webapps provide a ubiquitous way of distributing live sonifications and can be easily programmed with text prompts input to LLM's. The video demonstrates a workflow I've been using in workshops for scientists and musicians interested in creating their own sonifications.

### 个人简介 BIOGRAPHY

现任斯坦福大学音乐系主任，杜卡家族音乐教授，并担任斯坦福大学计算机音乐与声学研究中心 (CCRMA) 主任。他是一位作曲家、即兴演奏家和大提琴演奏家，长期致力于音乐、计算技术、现场表演与声学交叉领域的探索，在国际计算机音乐领域具有重要影响力。Chafe 教授的研究涵盖计算机音乐、数字声音合成、实时演奏系统以及网络化音乐表演等方向，尤其在低延迟网络协作演奏与远程音乐表演方面具有开创性贡献。他的工作不断拓展技术条件下音乐创作、演奏协作与听觉感知的边界，同时也涉及声音在科学与医学场景中的应用。作为作曲家与演奏者，Chris Chafe 创作了大量融合器乐实践与技术实验的作品。其大提琴演奏与即兴背景深刻影响了他的艺术语言，使其作品始终保持鲜明的现场性、互动性与探索性。除斯坦福大学外，他还曾在英属哥伦比亚大学、都灵理工大学和柏林工业大学等国际机构担任访问学者或客座教授。凭借多年来在创作、科研、教学与学术平台建设方面的持续投入，Chris Chafe 已成为全球音乐科技与跨学科声音研究领域的重要代表人物之一。

Chris Chafe is the Duca Family Professor of Music at Stanford University and Director of the Center for Computer Research in Music and Acoustics (CCRMA). A composer, improviser, and cellist, he is internationally recognized for his pioneering work at the intersection of music, computation, performance, and acoustic research. At Stanford, his creative and scholarly activities have long explored how computer-based technologies can expand musical expression, collaboration, and listening. Professor Chafe's research spans computer music, digital sound synthesis, real-time performance systems, and networked music performance. He is especially known for developing methods for ultra-low-latency musical interaction over networks, helping to shape new possibilities for distributed ensemble performance and telematic music-making. His work has also engaged questions of auditory perception, human-computer interaction, and sound in scientific and medical contexts. As a composer and performer, Chafe has created a wide range of works that combine instrumental practice with technological experimentation. His background as a cellist and improviser has remained central to his artistic identity, informing a body of work in which live musicianship and computational processes interact closely. In addition to his activities at Stanford, he has held major international visiting appointments, including positions at the University of British Columbia, the Politecnico di Torino, and the Technical University of Berlin. Through decades of creative research, teaching, and institution-building, Chris Chafe has made lasting contributions to the global development of computer music and interdisciplinary sound studies. He continues to be an influential figure in shaping conversations around music technology, artistic innovation, and the future of musical performance.



乔治·海杜  
Georg Hajdu

## 从音乐厅到社会空间：借助技术重新语境化当代音乐

From Concert Hall to Social Space: Recontextualizing Contemporary Music through Technology

### 报告摘要 ABSTRACT

当代音乐正日益从传统音乐厅走向社会化、分布式以及应用化的多元场景。本次报告探讨技术系统如何推动这种“再语境化”（recontextualization）过程，从而重塑演出实践与观众参与方式。其中一个关键案例是“圣保利易北隧道交响曲”，在这一项目中，144位音乐家被分布在长达900米的城市基础设施空间中，并通过基于浏览器的网络化指挥系统实现协同，使每位参与者都能够进行个体化互动，同时促进更具包容性的参与与体验。在此基础上，“疗愈声景”（Healing Soundscapes）项目进一步探索生成音乐在医院环境中的作用，通过自适应系统调节感知、缓解压力，并引导听众以超越传统审美框架的方式参与聆听。这些项目共同表明，技术正在重新定义音乐，使其从单一的艺术呈现形式，转变为嵌入日常生活之中的社会性、空间性与功能性实践。

Contemporary music is increasingly shifting from the concert hall into social, distributed, and applied contexts. This presentation examines how technological systems enable such recontextualization, transforming both performance practice and audience engagement. A key example is the Symphony in the St. Pauli Elbe Tunnel, where 144 musicians were spatially distributed across a 900-meter urban infrastructure and coordinated through a networked, browser-based conducting system, allowing individualized interaction and inclusive participation. Extending this approach, the Healing Soundscapes project explores the role of generative music in hospital environments, using adaptive systems to modulate perception, reduce stress, and engage listeners beyond traditional aesthetic frameworks. Together, these projects illustrate how technology redefines music as a social, spatial, and functional practice embedded in everyday life.

### 个人简介 BIOGRAPHY

德国作曲家、学者，现任德国汉堡音乐与戏剧学院（HfMT）教授，利盖蒂中心主任。他长期致力于音乐、科技与跨学科艺术实践的融合研究，关注多媒体作曲、网络化音乐、生成式音乐系统与数字乐谱等方向。Hajdu 早年在科隆学习分子生物学与作曲，后于美国加州大学伯克利分校完成博士学位，并与伯克利新音乐与音频技术中心（CNMAT）保持长期学术联系，是当代音乐科技领域具有代表性的学者与作曲家之一。

Georg Hajdu is a German composer and scholar, currently serving as Professor of Multimedia Composition at the Hochschule für Musik und Theater Hamburg (HfMT). He also leads the Ligeti Center and serves as the university's Commissioner for Research and Transfer. His work focuses on the intersection of music, technology, and interdisciplinary artistic practice, with particular interests in multimedia composition, networked music, generative systems, and digital score environments. Hajdu studied molecular biology and composition in Cologne and later earned his PhD at the University of California, Berkeley, in close connection with CNMAT.



李小兵

Li Xiaobing

## 机文主义：音乐学院的未来在哪里？

Machinism: Where Is the Future of Music Conservatories ?

### 报告摘要 ABSTRACT

本报告基于中央音乐学院在音乐人工智能领域多年的系统性探索，结合国际音乐科技与人工智能发展的前沿趋势，提出“机文主义”这一面向未来的理论设想。

机文主义并不被理解作为一种关于机器的理论教条，而是被视为一个用于重新思考人类与机器协同创作中主体性的框架。在人文主义长期界定艺术创作主体的历史框架下，人工智能的深度介入，正在动摇“创作从何而来”这一根本问题。报告从哲学层面出发，将“机文主义”界定为一种关于未来艺术主体形态的思想框架——在这一框架中，艺术不再仅仅是“人表达自我”，而逐步演变成为一种由人机协同、乃至多主体共同参与的意义生成结构。在此基础上，报告进一步回到现实制度层面，聚焦音乐学院这一传统艺术教育的核心机构，系统分析人工智能对音乐学院六个二级学科及七个专业方向可能带来的结构性重构。在更大的哲学观视野之下，未来的音乐学院，究竟会消失、重构，还是以一种全新的形态继续存在？

This report is grounded in the Central Conservatory of Music's years of systematic exploration in the field of music artificial intelligence, and integrates cutting-edge international developments in music technology and AI to propose a forward-looking theoretical vision termed Machinism.

Machinism is not understood as a doctrine of machinery, but as a framework for rethinking subjectivity in human-machine co-creation. Within the long-standing humanistic framework that has defined the subject of artistic creation, the deep intervention of artificial intelligence is now unsettling the fundamental question of "where creation comes from." From a philosophical perspective, the report conceptualizes Machinism as a framework for understanding the future form of artistic subjectivity. In this view, art is no longer merely an act of human self-expression, but gradually evolves into a structure of meaning-making shaped by human-machine collaboration, or even by the participation of multiple agents. Building on this theoretical foundation, the report turns to the level of real-world institutions, focusing on the music conservatory as a central pillar of traditional arts education. It systematically examines how artificial intelligence may bring about structural transformations across six sub-disciplines and seven areas of specialization within music academies. From a broader philosophical perspective, will music conservatories in the future disappear, be restructured, or continue to exist in an entirely new form?

### 个人简介 BIOGRAPHY

中央音乐学院教授、博士生导师、音乐人工智能系主任，国家哲学社会科学领军人才、中宣部“四个一批人才”、享受政府特殊津贴专家、国家社科重大项目首席专家、中国人工智能学会艺术与人工智能专委会主任、中国计算机学会计算艺术分会主任，“全国高校黄大年式教师团队”负责人。作曲博士、毕业于中央音乐学院作曲系，师从著名作曲家、中国音乐家协会名誉主席、中央音乐学院名誉院长吴祖强教授，音乐创作涵盖几乎所有音乐类型，部分作品受到群众喜爱具广泛影响力，曾荣获金钟奖、文华大奖、文华作曲奖、全国歌剧、舞剧一等奖、中宣部“五个一工程”奖等国内外奖项。

Professor and Doctoral Supervisor at the Central Conservatory of Music, Director of the Department of Music Artificial Intelligence, National Leading Talent in Philosophy and Social Sciences, recipient of the Central Propaganda Department's "Four Kinds of Talents" award, expert entitled to special government allowances, Principal Investigator of major national social science projects, the Chair of the China Computer Federation (CCF) Computational Art Branch, the Chair of the Chinese Association for Artificial Intelligence (CAAI) Art and Artificial Intelligence Commission. He also leads the "National Huang Danlan-style Faculty Team" in higher education. A Doctor of Composition, Li Xiaobing graduated from the Composition Department of the Central Conservatory of Music, where he studied under the renowned composer Professor Wu Zuqiang, Honorary President of the Chinese Musicians Association and the Central Conservatory of Music. His musical creations span almost all genres, with works enjoying wide popularity and significant influence. He has been honored with numerous domestic and international awards, including the Golden Bell Award, the Wenhua Grand Prize, the Wenhua Composition Award, first prizes in national opera and dance drama competitions, and the "Five One Project" Award from the Central Propaganda Department.



刘家丰

Liu Jiafeng

## 统一声学 Token 空间的音乐生成大模型：从深层表征到高质量生成 Large-Scale Music Generation Model with a Unified Acoustic Token Space: From Deep Representations to High-Quality Generation

### 报告摘要 ABSTRACT

当前音乐生成大模型大多采用分阶段、分表示空间的技术路线：先建模高层音乐结构，再通过扩散模型或神经解码器补全细节。我们的工作探索了一条不同路径，即在同一个统一的声学 Token 空间内，直接完成从粗结构到细节的逐步生成。报告将介绍中央音乐学院自研音乐生成大模型的核心设计，包括深层残差量化声学表征、coarse-to-fine 生成框架，以及面向歌词对齐与高保真音质的 hybrid-attention 训练机制。实验表明，该路线能够在不依赖独立 semantic stage 和 diffusion decoder 的前提下，实现高质量、长时程、可控的音乐生成。报告也将进一步讨论这一技术路线对于未来音乐生成大模型发展的意义，以及其在音乐创作、音乐教育与智能内容生产中的潜在价值。

Most current music generation foundation models adopt a staged, multi-representation-space pipeline: they first model high-level musical structure, and then rely on diffusion models or neural decoders to reconstruct fine acoustic detail. Our work explores a different path: progressively generating music from coarse structure to fine detail within a single unified acoustic token space. This talk will introduce the core design of a music generation foundation model developed at the Central Conservatory of Music, including deep residual-quantized acoustic representations, a coarse-to-fine generation framework, and a hybrid-attention training mechanism for lyric alignment and high-fidelity audio quality. Experiments show that this approach can achieve high-quality, long-form, and controllable music generation without relying on a separate semantic stage or a diffusion decoder. The talk will also discuss the broader significance of this technical route for future music generation foundation models, as well as its potential applications in music creation, music education, and intelligent content production.

### 个人简介 BIOGRAPHY

中央音乐学院音乐人工智能与音乐信息科技系副教授。博士毕业于中央音乐学院，中国首个音乐人工智能博士，师从俞峰教授、孙茂松教授。自幼跟随四川音乐学院钢琴系教授学习，本硕期间曾任校交响乐团首席钢琴。致力于研究多轨道音乐生成、音乐音频信号处理，多模态音乐大模型等前沿方向。提出了世界首个端到端交响乐生成模型，CCOM 声源分离训练与推理框架的研发人，Sound Demixing Challenge 2023 国际音乐声源分离大赛冠军。

Liu Jiafeng is an Associate Professor in the Department of Music AI and Music Information Technology at the Central Conservatory of Music, as well as an AI researcher and pianist. He focuses on multi-track music generation, having proposed the world's first end-to-end symphony generation model. He also conducts in-depth research in music source separation and won first place in the Sound Demixing Challenge 2023. Currently, he devotes his research efforts to large-scale multimodal music generation models.



马军  
Ma Jun

## 音乐脑机接口：概念、研究进展与应用前景

Music Brain-Computer Interfaces: Concepts, Research Progress, and Application Prospects

### 报告摘要 ABSTRACT

音乐脑机接口是音乐、人工智能与脑机接口交叉形成的新兴研究方向，融合脑科学、音乐治疗、情绪计算、信号处理、生物医学工程和人机交互等多学科方法。近年来，相关研究已从脑控音乐交互逐步扩展至情绪调节、睡眠支持、康复训练以及教育与创作等场景，显示出由单一功能研究走向闭环系统构建的发展趋势。本报告将围绕音乐脑机接口的概念界定、研究进展、技术框架与应用方向展开讨论，并结合“央音一号”的相关实践，分析其在本土研究语境中的探索路径。报告旨在从学科交叉与技术演进两个层面，进一步理解音乐脑机接口的研究价值及应用前景。

Music brain-computer interface is an emerging interdisciplinary field formed at the intersection of music, artificial intelligence, and brain-computer interface technologies. It integrates approaches from neuroscience, music therapy, affective computing, signal processing, biomedical engineering, and human-computer interaction. In recent years, related research has expanded from brain-controlled musical interaction to applications in emotion regulation, sleep support, rehabilitation training, as well as education and creative practice, reflecting a broader shift from single-function studies toward the development of closed-loop systems. This presentation will discuss the conceptual definition, research progress, technical framework, and application domains of music brain-computer interfaces, and will further examine the exploratory path of this field in the local research context through the example of "CCOM-1." From the perspectives of interdisciplinary integration and technological evolution, the report aims to further illuminate the research significance and application prospects of music brain-computer interfaces.

### 个人简介 BIOGRAPHY

中央音乐学院音乐人工智能与音乐信息科技系讲师。博士毕业于北京大学神经科学研究所，并于圣路易华盛顿大学麻醉系完成博士后训练。拥有超过 11 年的侵入式与非侵入式脑机接口研发经验。现主要研究方向为音乐脑机接口、基于脑科学的个性化音乐治疗、音乐处理的神经机制。

Ma Jun is a Lecturer in the Department of Music Artificial Intelligence and Music Information Technology at the Central Conservatory of Music. He received his Ph.D. from the Institute of Neuroscience at Peking University and completed postdoctoral training in the Department of Anesthesiology at Washington University in St. Louis. He has more than 11 years of experience in the research and development of both invasive and non-invasive brain-computer interfaces. His current research focuses on music brain-computer interfaces, personalized music therapy based on neuroscience, and the neural mechanisms underlying music processing.



卢迪  
Lu Di

## 面向 AI 生成音乐工作流的 Web DAW

A Web-based DAW for AI-generated Music Workflow

### 报告摘要 ABSTRACT

AI 的快速发展推动了 AI 生成音乐的显著突破，现已能够生成 48kHz 采样率的高质量音频。然而，当前 AI 音乐生成技术仍面临一个根本性挑战：端到端的音频生成模型虽然能产生高保真音乐，但其黑箱特性使得音高、时值等音乐元素难以编辑；而符号生成方法（如 MIDI）虽能保留结构化、可编辑的音乐表征，却高度依赖专业后期制作才能实现高保真音频质量。为解决这一不足，本文提出了一款基于网页的、用户可编辑的数字音频工作站（DAW），专门面向符号生成方法所生成的音乐。主要贡献包括：一个支持直接操作音乐元素的直观可视化界面、一个基于浏览器的实时音频渲染器，以及一个用于集成符号音乐生成系统的标准化 API，从而实现了从 AI 生成内容到用户精修的端到端无缝 workflow。

The rapid advancement of artificial intelligence has led to significant break-throughs in AI-generated music, enabling the production of high-quality audio at 48kHz sampling rates. However, a fundamental challenge persists in current AI music generation technologies: while end-to-end audio generation models produce high-fidelity music, their black-box nature makes musical elements (e.g., pitch, duration) difficult to edit. Conversely, symbolic generation methods (e.g., MIDI) retain structured and editable representations of music, yet rely heavily on professional post-production to achieve high-fidelity audio quality. To address this gap, this paper introduces a user-editable web-based Digital Audio Workstation (DAW) specifically designed for music generated by symbolic generation methods. The main contributions include: an intuitive visual interface supporting direct manipulation of musical elements, a browser-based real-time audio renderer, and a standardized API that integrates symbolic music generation systems, enabling a seamless end-to-end workflow from AI-generated content to user refinement.

## 个人简介 BIOGRAPHY

中央音乐学院音乐人工智能与音乐信息科技系助理研究员，东京大学情报理工学系硕士。国内首个商业歌声合成软件及首个自动作曲软件的核心开发者，拥有 15 年音乐 + 计算机交叉领域从科研到商业落地的全流程经验，持有多项软件著作权与专利。

Assistant Researcher at the Department of Music AI and Music Information Technology, Central Conservatory of Music. He holds a Master's degree from the Department of Information Science and Technology, the University of Tokyo. He is the core developer of China's first commercial singing voice synthesis software and the first automatic composition software, with 15 years of full-cycle experience from research to commercial deployment in the interdisciplinary field of music and computer science. He holds multiple software copyrights and patents.



肯尼斯·菲尔兹  
Kenneth Fields

## 面向网络化电子音乐合奏的 AI 编程助手的实践应用

Practical Applications of AI Coding Assistants for Networked Electronic Music Ensembles

### 报告摘要 ABSTRACT

一个涉及中国科学院大学 (UCAS)、中央音乐学院 (CCOM)、北京大学 (PKU) 以及武汉音乐学院 (WHCM) 等多校区之间的网络化电子音乐合奏项目，依托中国高速的 CERNET2 教育科研网络，在今年取得了快速进展，这在很大程度上得益于 AI 聊天式编程助手的高效引入。该项目通过使用基于 p5.js 编写的去中心化动态图形乐谱 (结合 socket.io 库，并部署于 openprocessing.org 平台) 实现协作；与此同时，所有合奏成员均使用名为 SuperCollider 的音频编程应用，并通过开放声音控制 (OSC) 前端进行控制。整体而言，这是一个技术性极强的项目，但其显著变化在于，如今已不再过度依赖复杂代码本身。

从作曲者的角度来看，在 2024 年之前，大多数重要的艺术 / 音乐与科技类项目都会投入大量时间、精力以及机构资源于音频与动态图形编程 (如 SuperCollider、Max/MSP、Java 等)。电子音乐专业的学生在编程能力上始终存在显著差异；对于部分学生而言，编写代码构成了明显的劣势。因此，这一领域在某种程度上演变为“逻辑 / 空间能力优胜者的生存竞争”，而非真正以音乐时间结构与创造性想象力为核心的竞争。电子音乐学生的身份逐渐被“黑客文化”、复杂精密的信号路由 / 补丁结构，或对预制插件的高度依赖所定义。随着诸如 ChatGPT、DeepSeek、Gemini 等 AI 聊天机器人引入课堂，这些课程正在发生转变，使音乐与科技相关项目得以重新将重心回归到音乐 (“有组织的声音”) 以及动态图形的创意本身。这标志着一种重要甚至具有革命性的进展——作品的形式、音色与控制逻辑可以更加直接地由创作者的想象力主导，而不再被编程语法这一狭窄门槛所牵制。

A project involving networked electronic music ensembles between the UCAS, CCOM, PKU and WHCM campuses using China's high-speed CERNET2 network, has made rapid progress this year, due to the expedient introduction of AI chatbot coding assistants. The project allows collaboration using decentralized animated graphic scores coded in p5js using the socket.io library (openprocessing.org), while ensemble members all use the audio coding application called Supercollider, controlled with open sound control (OSC) front-ends. It's a highly technical project, though most notably now, without the disproportionate focus on complex code.

From the composer's point of view, before 2024, much time, effort and institutional space in most major arts/music and technology programs was dedicated to audio and animated graphics coding (supercollider, max/msp, java, etc). Students of electronic music always exhibited a wide differential in their abilities when it came to computer programming skills; writing code proved to be a serious disadvantage for some. The field thus became a survival of the logical/spatially oriented fittest, not necessarily the fittest in terms of creativity and imagination with musical/temporal forms. Electronic music student's identity became wrapped up in the idea of hacking, exquisite complexity in signal routing/patching, or otherwise becoming highly dependent on prefabricated plugins. After AI chatbots were introduced, such as ChatGPT, Deepseek, Gemini, etc., these courses have been transformed, enabling music and tech focused programs to once again focus essentially on musical (organized sound) and animated visual ideas. This is a significant, even revolutionary progression, where form, timbre and control can be led directly by the creative imagination rather than being sidetracked by the narrow gateway of programming syntax.

#### 个人简介 BIOGRAPHY

Kenneth Fields (博士) 是中国科学院大学媒体艺术教授，在音乐与科技领域拥有丰富经验。2003 年至 2023 年间，他曾在中央音乐学院担任外籍教授。2008 年至 2013 年，他担任卡尔加里大学“远程媒体艺术”加拿大研究讲席教授 (Canada Research Chair in Telemedia Arts)。Fields 教授现任《Organised Sound》(剑桥大学出版社出版) 及电子音乐研究亚洲网络 (电子音乐研究亚洲网络, EMSAN) 编委会成员。他还是 2021-2026 年欧洲研究委员会 (ERC) 国际科研项目 “The Digital Score” 的共同负责人 (Co-PI)，该项目旨在探究音乐乐谱的本质及其相关技术。

Kenneth Fields (Ph.D) is a Professor of Media Arts at the University of the Chinese Academy of Sciences, having rich experience in the field of Music and Technology. From 2003 to 2023, he was a Foreign Professor at the Central Conservatory of Music in Beijing. From 2008 to 2013, he was the Canada Research Chair in Telemedia Arts at the University of Calgary. Prof. Fields serves on the editorial boards of the Journal of Organized Sound (Cambridge Univ. Press) and the Electronic Music Studies Asia Network (EMSAN). He is Co-PI in the international 2021-26 European Research Council Grant, entitled: The Digital Score which is an investigation into the nature and technology of music scores.



亚伦·威廉姆森  
Aaron Williamon

## 表演科学的未来 The Future of Performance Science

### 报告摘要 ABSTRACT

表演科学是一个综合性领域，致力于提升我们对人类在表演艺术、体育、教育以及职场等不同领域中表现的理解与优化。近年来，该领域已从对狭义学科成就的关注，转向更广泛地探讨表现如何受到社会塑造，并反过来影响社会。这包括表演与社会价值之间不断演变的关系、表演职业路径的多样化与可持续性、教育与训练体系的转型，以及健康、福祉与心理韧性的的重要性日益凸显。尽管已经取得了显著进展，但在将基础技能与现实世界成果相连接，以及应对界定该领域未来的跨学科挑战方面，仍然存在重要的空白。本次报告将勾勒表演科学未来的发展路线图，展示如何通过表演心理学、教育与健康等领域的洞见，应对快速变化的技术、文化与社会环境。

Performance Science is an integrative field focused on advancing our understanding and enhancement of human performance across domains such as the performing arts, sports, education, and the workplace. Recent years have seen a shift from narrowly defined disciplinary achievements to broader considerations of how performance is shaped by, and in turn shapes, society. This includes the evolving relationship between performance and societal values, the diversification and sustainability of performance careers, the transformation of education and training systems, and the centrality of health, wellbeing and psychological resilience. Despite substantial progress, there remain important gaps in connecting fundamental skills with real-world outcomes and navigating the interdisciplinary challenges that define the future of the field. This talk will outline a roadmap for the future of Performance Science, one that demonstrates how insights from performance psychology, education and health can help respond to rapid technological, cultural and societal changes.

### 个人简介 BIOGRAPHY

Aaron Williamon 是 Royal College of Music 表演科学教授，并担任表演科学中心 (Centre for Performance Science, CPS) 主任。该中心由皇家音乐学院与 Imperial College London 共同合作建立。他于 2000 年加入皇家音乐学院担任研究员，2004 年晋升为高级研究员，并于 2010 年被任命为表演科学教授。他的研究主要关注高水平表演能力，以及将科学研究应用于音乐学习与教学的实践，同时也探讨音乐与艺术对社会的影响。Aaron 是国际表演科学研讨会 (International Symposium on Performance Science) 的创始人之一，同时也是学术期刊 Performance Science (隶属于 Frontiers) 的创刊主编，并担任“健康音乐学院” (Healthy Conservatoires) 国际网络的创始主席。该网络成立于 2015 年，旨在支持学生及专业表演艺术家的健康与福祉。他是 Royal Society of Arts 会士 (FRSA) 以及英国高等教育学会 AdvanceHE 会士 (FHEA)。2008 年，他被授予皇家音乐学院荣誉会员 (HonRCM)。

Aaron Williamon is Professor of Performance Science at the Royal College of Music (RCM) where he directs the Centre for Performance Science (CPS), a partnership of the RCM and Imperial College London. Aaron joined the RCM as Research Fellow in 2000 and was appointed Senior Research Fellow in 2004 and Professor of Performance Science in 2010. His research focuses on skilled performance and applied scientific initiatives that inform music learning and teaching, as well as the impact of music and the arts on society. Aaron is the founder of the International Symposium on Performance Science, founding chief editor of Performance Science (a Frontiers journal), and the founding chair of Healthy Conservatoires, an international network constituted in 2015 to support health and wellbeing among student and professional performing artists. Aaron is a fellow of the Royal Society of Arts (FRSA) and the UK's higher education academy, AdvanceHE (FHEA), and in 2008, he was elected an Honorary Member of the Royal College of Music (HonRCM).



凯特·霍普  
Cat Hope

## 确保音乐领域人工智能政策的包容性与可持续性：一个治理议题 Ensuring Inclusive and Sustainable AI Policy for The Music Sector: A Governance Issue

### 报告摘要 ABSTRACT

我们如何在发挥人工智能变革性潜力的同时，维护创作自主性、文化多样性以及音乐工作者的生计？随着各类人工智能工具在音乐领域不断涌现，本报告指出，当前亟需超越针对单一工具的应对思路，转而思考更加全面的治理框架，并将人的主体性置于技术效率之上。真实性、艺术自由、文化身份与可持续性价值，应当也完全可以被纳入治理与政策机制之中。

How do we harness artificial intelligence's transformative potential while preserving creative autonomy, cultural diversity and musicians livelihoods? As a variety of artificial intelligence tools proliferate across the music sector, this presentation argues that there is an urgent need to move beyond tool-specific approaches and consider comprehensive governance frameworks that prioritise human agency over technological efficiency. The importance of values such as authenticity, artistic freedom, cultural identity and sustainability should, and can, be embedded in governance and policy mechanisms.

### 个人简介 BIOGRAPHY

当代作曲家、电子音乐与数字乐谱研究者，长期从事实验音乐创作、非传统记谱与数字乐谱（Digital Score）研究。其学术与实践强调演奏者能动性、空间化记谱以及新型乐谱界面在当代音乐创作中的作用。曾担任澳大利亚国家级艺术与研究项目负责人，并在国际会议（如 TENOR、ICMC 等）和重要艺术机构中持续推动音乐、技术与文化政策的交叉研究。

Contemporary composition; Electronic music; Digital Score research; Experimental and non-traditional notation; Performer agency and spatialised notation; Music-technology-policy interdisciplinary research.



徐帆  
Xu Fan

## 写歌，正在从创作变成选择吗？ Is Songwriting Becoming Selection Rather Than Creation?

### 报告摘要 ABSTRACT

生成式人工智能正在快速进入音乐创作领域，使“从文本到音乐生成”成为现实。随着音频生成模型能力的提升以及交互方式从复杂工具向自然语言转变，音乐创作的技术门槛显著降低，越来越多非专业用户开始参与创作过程。以 Suno 等 AI 音乐系统的发展为例，可以看到这一领域已经从技术验证阶段走向产品化与规模化应用阶段。

在这一背景下，音乐创作的结构正在发生变化。本报告将从工程视角出发，探讨创作过程如何从“从零生成”转向“在生成空间中搜索与选择”，以及创作者角色如何从“生产者”转变为“判断与引导者”。同时将讨论 AI 创作中“可控性与随机性”的核心张力，并进一步思考在 AI 参与创作的条件下，音乐人、作品与作者关系可能面临的重新定义。

Generative artificial intelligence is rapidly entering the field of music creation, making "text-to-music generation" a reality. With the advancement of audio generation models and the shift in interaction paradigms from complex tools to natural language, the technical barriers to music creation have been significantly lowered, enabling an increasing number of non-professional users to participate in the creative process. The development of AI music systems such as Suno illustrates that this field has already moved beyond the stage of technical validation and is entering a phase of productization and large-scale application.

Against this backdrop, the structure of music creation is undergoing transformation. From an engineering perspective, this report explores how the creative process is shifting from "generation from scratch" to "searching and selecting within a generative space," and how the role of the creator is evolving from a "producer" to a "judge and guide." It also examines the fundamental tension between controllability and randomness in AI-driven creation, and further reflects on how the relationships among musicians, works, and authorship may be redefined under conditions where AI participates in the creative process.

## 个人简介 BIOGRAPHY

前在 Suno 从事生成式人工智能驱动的音乐创作产品与工程工作，负责将核心模型能力转化为面向用户的创作工具。作为早期创始团队一员，他参与了公司网页端与移动端产品从 0 到 1 的设计与开发。在此之前，他在 Meta 担任资深软件工程师，参与大规模数据系统以及 Meta Reality Labs 相关产品的开发。本科毕业于北京大学。

Xu Fan is currently working at Suno, where he focuses on generative AI-driven music creation products and engineering, translating core model capabilities into user-facing creative tools. As an early member of the founding team, he contributed to the design and development of the company's web and mobile products from 0 to 1.



龚俊民

Gong Junmin

## 推动开源音乐生成的边界

Pushing the Boundaries of Open-Source Music Generation

## 报告摘要 ABSTRACT

过去一年，AI 音乐生成领域在很大程度上由闭源商业产品所主导。ACE-Step 项目则从零起步，致力于以完全开源、可本地运行的方式，推动开源音乐生成技术达到可与商业模型相竞争的水平。本次分享将避免过多技术细节的堆砌，而是从更本质的视角出发，探讨 AI 音乐模型背后的核心驱动力——并非单纯的算力或模型架构，而是那些具备音乐素养、能够进行细致聆听与准确表达的参与者。我将介绍 ACE-Step 的研发过程，重点说明为何项目高度依赖具备音乐专业背景的合作伙伴。同时，结合我作为算法工程师与业余作曲者的双重视角，分享对 AI 与音乐人关系的理解与思考。

Over the past year, the field of AI music generation has been largely dominated by closed-source commercial systems. The ACE-Step project was initiated from the ground up with the goal of advancing open-source, locally deployable music generation to a level comparable with commercial models.

## 个人简介 BIOGRAPHY

ACE Studio 合伙人，ACE-Step 开源音乐生成模型系列作者。算法工程师出身，先后就职于多家头部科技公司，长期专注于音频与音乐生成方向。业余编曲、写词，与作词人方文山同门。一直相信做 AI 音乐最重要的不是模型本身，而是和真正懂音乐的人一起工作。

Gong Junmin is a partner at ACE Studio and the creator of the ACE-Step open-source music generation model series. With a background as an algorithm engineer, he has worked at several leading technology companies and has long focused on audio and music generation. In addition to his technical work, he is an amateur composer and lyricist, and is part of the same mentorship lineage as renowned lyricist Fang Wenshan. He firmly believes that the most critical factor in AI music is not the model itself, but collaborating with people who truly understand music.



姜涛

Jiang Tao

## 从一杯奶茶到音乐创作和消费的 Agent

From a Cup of Milk Tea to Agents for Music Creation and Consumption

### 报告摘要 ABSTRACT

Agent 是未来音乐生成和消费的新入口和新交互。千问通过“订奶茶”，证明了 agent 已经可以对衣食住行的各方面都造成影响。本次分享将分享音潮的音乐创作 agent (hitto.ai)，包含音乐创作、mv 创作、音乐 remix 等；畅想 agent 下音乐消费的形态；同时简单介绍一下音潮端到端音乐生成的原理。

Agents will become a new entry point and interaction paradigm for music generation and consumption. Qwen has demonstrated—through use cases like “ordering milk tea”—that agents can already influence many aspects of everyday life. This talk will introduce YinChao’s music creation agent (hitto.ai), which supports music generation, MV creation, and music remixing; explore how music consumption may evolve in an agent-driven paradigm; and provide a brief overview of the principles behind YinChao’s end-to-end music generation system.

### 个人简介 BIOGRAPHY

硕、博毕业于哈尔滨工业大学，有多年的 AI 和音频算法研发经验，及工程团队管理经验。先后在快手、腾讯音乐、昆仑万维组建了国内领先的音乐和音频算法团队，基于相关算法的产品功能已经服务于千万级用户。23 年~24 年期间，带领团队成为国内首家实现类 suno 音乐生成模型，并产品化服务于用户；在腾讯音乐期间，首创了 K 歌多维度评价、臻品音质等核心音乐消费功能，同时塑造了小天、小琴两款现象级虚拟歌手；在快手期间，完成了国内首个端到端 AI 音乐生成 app（小森唱），作品原声、智能配乐、音悦台等核心创作和消费功能。有多段相关方向的创业经历。

Jiang Taoben received his bachelor’s, master’s, and doctoral degrees from Harbin Institute of Technology. He has many years of experience in AI and audio algorithm research and development, as well as in engineering team management. He has successively built leading music and audio algorithm teams in China at Kuaishou, Tencent Music, and Kunlun Wanwei, with products powered by these technologies serving tens of millions of users. Between 2023 and 2024, he led his team to become the first in China to develop a Suno-like music generation model and successfully productize it for users. During his time at Tencent Music, he pioneered core music consumption features such as multi-dimensional evaluation for karaoke and premium audio quality, and helped create two breakout virtual singers, Xiaotian and Xiaoqin. At Kuaishou, he developed China’s first end-to-end AI music generation app (Xiaosen Chang), featuring key creative and consumption functionalities such as original soundtrack generation, intelligent accompaniment, and music video platforms. He also has multiple entrepreneurial experiences in related fields.



刘晓光

Liu Xiaoguang

## AI 赋能音乐教育

AI-Empowered Music Education

### 报告摘要 ABSTRACT

本报告探讨运用 AI 解决音乐教育中学生端和老师端遇到问题的可行性，介绍 DeepMusic 在音乐数据标注与数据集建设上的核心工作，并结合全国多个中小学实践案例，展示 AI 赋能音乐教育的落地成效与应用路径。

This report explores the feasibility of using AI to address challenges faced by both students and teachers in music education. It introduces DeepMusic's core work in music data annotation and dataset construction, and presents practical cases from multiple primary and secondary schools across the country to demonstrate the implementation outcomes and application pathways of AI-empowered music education.

### 个人简介 BIOGRAPHY

清华大学化学系系硕博，清华企业家协会青创会员。编曲师，键盘手，吉他手。有 100+ 首音乐作品创作及制作经验，作品全网播放量数亿次。有多年音基教育经验。

Liu Xiaoguang, CEO of DeepMusic. Ph.D., M.S., and B.S. in Chemistry from Tsinghua University. Member of the Tsinghua Entrepreneur & Executive Club. Arranger, keyboardist, and guitarist. With over 100 original music compositions and production credits, his works have achieved hundreds of millions of streams across major platforms. Has years of experience in music fundamentals education.



戴琮人

Dai Congren

## 音乐全谱理解基准：大模型对完整乐谱理解能力的评测与分析

Musical Score Understanding Benchmark: Evaluating Large Language Models' Comprehension of Complete Musical Scores

### 报告摘要 ABSTRACT

对完整乐谱的理解需要对音高、节奏、和声以及大尺度结构进行综合推理，然而，大语言模型与视觉语言模型在解读完整音乐记谱方面的能力仍缺乏充分研究。我们提出了音乐乐谱理解基准（Musical Score Understanding Benchmark, MSU-Bench），这是首个由人工精心构建的、面向乐谱层级音乐理解的基准，涵盖文本模态（ABC 记谱）与视觉模态（PDF）。MSU-Bench 包含来自巴赫、贝多芬、肖邦、德彪西等作曲家作品的 1,800 组生成式问答对，并按照难度递增划分为四个层级，从起始音信息到织体与曲式结构。对十五个以上最先进模型在零样本与微调设置下的评测表明，不同模态之间存在显著差距，各层级表现不稳定，并且在保持多层次一致性方面面临挑战。微调在保留通用知识的同时，显著提升了跨模态表现，使 MSU-Bench 成为未来多模态推理研究的坚实基础。

Understanding complete musical scores entails integrated reasoning over pitch, rhythm, harmony, and large-scale structure, yet the ability of Large Language Models and Vision-Language Models to interpret full musical notation remains insufficiently examined. We introduce Musical Score Understanding Benchmark (MSU-Bench), the first human-curated benchmark for score-level musical understanding across textual (ABC notation) and visual (PDF) modalities. MSU-Bench contains 1,800 generative Question-Answering pairs from works by Bach, Beethoven, Chopin, Debussy, and others, organised into four levels of increasing difficulty, ranging from onset information to texture and form. Evaluations of more than fifteen state-of-the-art models, in both zero-shot and fine-

tuned settings, reveal pronounced modality gaps, unstable level-wise performance, and challenges in maintaining multilevel correctness. Fine-tuning substantially improves results across modalities while preserving general knowledge, positioning MSU-Bench as a robust foundation for future research in multimodal reasoning.

## 个人简介 BIOGRAPHY

中央音乐学院与清华大学联合培养博士研究生（一年级在读），师从小兵教授与孙茂松教授。此前先后获得纽芬兰纪念大学软件工程学士学位、伦敦国王学院数据科学理学硕士学位，以及伦敦帝国理工学院人工智能与机器学习研究型硕士学位。曾在多家企业从事人工智能与数据相关工作，包括于橡胶机器人有限公司担任大模型算法工程师，在 Google 从事计算机视觉算法研发，并于英国 720 Management Ltd. 与万联证券担任数据分析师，在玉柴股份有限公司担任全栈工程师，具备跨领域的工程与研究经验。

Dai Congren is a first-year joint PhD student at the Central Conservatory of Music and Tsinghua University, supervised by Professor Xiaobing Li and Professor Maosong Sun. He previously received a Bachelor's degree in Software Engineering from Memorial University of Newfoundland, an MSc in Data Science from King's College London, and an MRes in Artificial Intelligence and Machine Learning from Imperial College London.



丘治平  
Qiu Zhiping

## 音频驱动的弦乐演奏动作生成

ELGAR: Expressive Cello Performance Motion Generation for Audio Rendition

### 报告摘要 ABSTRACT

乐器演奏艺术是人类创造力与情感的生动体现。然而，生成乐器演奏动作是一项极具挑战性的任务，它不仅要求捕捉复杂的肢体动作，还需重建“演奏者-乐器”之间复杂的动力学交互。现有演奏动作生成工作主要聚焦于局部身体动作建模。为此，我们提出了 ELGAR（面向音频呈现的表现力大提琴演奏动作生成）。这是一个基于扩散模型的框架，仅需音频输入即可生成全身精细化乐器演奏动作。为强调演奏过程中的交互特性，我们引入了手部交互接触损失（HICL）与琴弓交互接触损失（BICL），有效保障了交互过程的真实性。此外，为了更好地评估生成动作与音乐音频语义背景的对齐程度，我们专为弦乐演奏动作生成设计了全新的评价指标，包括指尖接触距离、弓弦距离及运弓评分。通过详尽的评估与消融实验，验证了所提方法的有效性。同时，我们还推出了从动捕数据集 SPD 整理并归一化而成的动作生成数据集 SPD-GEN。实验证明，ELGAR 在生成具有复杂、快速交互特征的乐器演奏动作方面展现出巨大潜力，将有力推动动画制作、音乐教育及交互艺术创作等领域的发展。

The art of instrument performance stands as a vivid manifestation of human creativity and emotion. Nonetheless, generating instrument performance motions is a highly challenging task, as it requires not only capturing intricate movements but also reconstructing the complex dynamics of the performer-instrument interaction. While existing works primarily focus on modeling partial body motions, we propose Expressive ceLlo performance motion Generation for Audio Rendition (ELGAR), a state-of-the-art diffusion-based framework for whole-body fine-grained instrument performance motion generation solely from audio. To emphasize the interactive nature of the instrument performance, we introduce Hand Interactive Contact Loss (HICL) and Bow Interactive Contact Loss (BICL), which effectively guarantee the authenticity of the interplay. Moreover, to better evaluate whether the generated motions align with the semantic context of the music audio, we

design novel metrics specifically for string instrument performance motion generation, including finger-contact distance, bow-string distance, and bowing score. Extensive evaluations and ablation studies are conducted to validate the efficacy of the proposed methods. In addition, we put forward a motion generation dataset SPD-GEN, collated and normalized from the MoCap dataset SPD. As demonstrated, ELGAR has shown great potential in generating instrument performance motions with complicated and fast interactions, which will promote further development in areas such as animation, music education, interactive art creation, etc.

## 个人简介 BIOGRAPHY

中央音乐学院与清华大学联合培养在读博士研究生，师从中央音乐学院俞峰教授与清华大学戴琼海教授，长期致力于探索音乐与具身智能的交叉领域，涵盖从多模态感知、精细化动作生成到具身物理执行的完整路径。

Qiu Zhiping is a joint Ph.D. candidate at the Central Conservatory of Music and Tsinghua University, co-advised by Prof. Feng Yu and Prof. Qionghai Dai. He has long been dedicated to the intersection of music and embodied AI, covering the complete trajectory from multimodal perception and fine-grained motion generation to physical execution.



童心怡  
Tong Xinyi

## 音画共鸣：视频配乐生成的视觉画面、时间节奏与音乐表达对齐 Video Echoed in Music: Semantic, Temporal, and Rhythmic Alignment for Video-to-Music Generation

### 报告摘要 ABSTRACT

视频背景音乐自动生成技术旨在通过创作高度契合的配乐来提升视听沉浸感。然而，现有方法在深度视频语义理解以及视觉转场与音乐节奏的精准同步方面仍面临严峻挑战。为打破这一多模态对齐瓶颈，研究提出了“Video Echoed in Music (VeM)”的创新潜在音乐扩散模型，并构建了高质量音画对齐数据集及配套评估指标。该模型开创性地将层级化视频解析作为“音乐指挥”深度融入生成框架，通过引入分镜引导的交叉注意力机制与帧级转场-节拍对齐适配器，在宏观上保障了跨模态语义与时序的连贯融合，在微观层面实现了视觉场景切换与音乐节拍的动态硬同步。本研究为影视、广告及短视频工业的智能化音乐创作奠定了坚实的技术基石。相关工作已被人工智能顶级国际会议 AAAI 2026 录用为 Oral 报告。

Automatic video-to-music generation aims to enhance audiovisual immersion by crafting perfectly matched soundtracks. However, existing methods encounter significant challenges in deep video semantic understanding and the precise synchronization of musical beats with visual transitions. To overcome these multimodal alignment bottlenecks, we propose "Video Echoed in Music (VeM)," an innovative latent music diffusion model, and introduce a high-quality audiovisual alignment dataset alongside novel evaluation metrics. The framework pioneeringly integrates hierarchical video parsing as a "music conductor" into the generation process. By employing a storyboard-guided cross-attention mechanism and a frame-level transition-beat aligner and adapter, the model ensures coherent cross-modal semantic and temporal fusion at a macro level, while achieving dynamic synchronization between scene transitions and musical beats at a micro level. This research establishes a robust technical foundation for intelligent music creation across the film, advertising, and short-video industries. This work has been accepted as an Oral presentation at AAAI 2026.

中央音乐学院与北京大学联合培养在读博士，师从北京大学朱松纯教授与中央音乐学院俞峰教授，主要研究方向为多模态生成音乐生成，并致力于探索人工智能对音乐概念的建模，以及跨模态对齐艺术表达对齐问题。以第一作者身份在 AAAI (Oral Presentation)、CVPR 及 IEEE TCSS 等国际学术会议与期刊发表多篇论文，并作为参与撰写出版教材《音乐的人工智能 U-V 理论》，参与讲授北京大学通选课程《人工智能与音乐》。曾荣获首届国际通用人工智能大会优秀成果奖、教育部中美青年创客大赛主赛道一等奖等，持续在人工智能与艺术交叉前沿探索深层融合的可能。

Tong Xinyi is a joint Ph.D. candidate at the Central Conservatory of Music and Peking University, co-advised by Prof. Song-Chun Zhu and Prof. Feng Yu. Her primary research focuses on multimodal music generation, with a strong dedication to the AI-driven modeling of musical concepts and the cross-modal alignment of artistic expressions. She has published multiple papers in premier international conferences and journals, including AAAI (Oral Presentation), CVPR, and IEEE TCSS. Beyond her research, she co-authored the textbook *Artificial Intelligence in Music: U-V Theory* and co-lectures the general elective course "Artificial Intelligence and Music" at Peking University.



吴尚达

Wu Shangda

### CLaMP 3: 跨未对齐模态与未见语言的通用音乐信息检索

CLaMP 3: Universal Music Information Retrieval Across Unaligned Modalities and Unseen Languages

#### 报告摘要 ABSTRACT

CLaMP 3 是一个旨在解决音乐信息检索 (MIR) 中跨模态与跨语言泛化挑战的统一框架。该框架采用对比学习技术，将乐谱、演奏信号及音频录制等所有主要音乐模态与多语言文本对齐到统一的共享表征空间中，从而实现了以文本为桥梁的非对齐模态间的跨模态检索。CLaMP 3 配备了一个能够适应未知语言的多语言文本编码器，展现出强大的跨语言泛化能力。利用检索增强生成 (RAG) 技术，我们构建了大规模数据集 M4-RAG，其中包含 231 万个音乐-文本对。该数据集富含详细的元数据，涵盖了广泛的全球音乐传统。为了推动后续研究，我们发布了 WikiMT-X 基准测试集，由 1,000 组涵盖乐谱、音频及丰富文本描述的三元组组成。实验表明，CLaMP 3 在多项 MIR 任务中均达到了当前最先进 (SOTA) 的水平，显著超越了以往的强基准模型，并在多模态及多语言音乐语境下表现出卓越的泛化性能。

CLaMP 3 is a unified framework developed to address challenges of cross-modal and cross-lingual generalization in music information retrieval. Using contrastive learning, it aligns all major music modalities—including sheet music, performance signals, and audio recordings—with multilingual text in a shared representation space, enabling retrieval across unaligned modalities with text as a bridge. It features a multilingual text encoder adaptable to unseen languages, exhibiting strong cross-lingual generalization. Leveraging retrieval-augmented generation, we curated M4-RAG, a web-scale dataset consisting of 2.31 million music-text pairs. This dataset is enriched with detailed metadata that represents a wide array of global musical traditions. To advance future research, we release WikiMT-X, a benchmark comprising 1,000 triplets of sheet music, audio, and richly varied text descriptions. Experiments show that CLaMP 3 achieves state-of-the-art performance on multiple MIR tasks, significantly surpassing previous strong baselines and demonstrating excellent generalization in multimodal and multilingual music contexts.

## 个人简介 BIOGRAPHY

吴尚达博士现任职于国内领先的互联网企业，致力于语音大模型领域的算法研究。他于 2025 年 6 月毕业于中央音乐学院，获音乐人工智能与信息科技博士学位，师从清华大学孙茂松教授与中央音乐学院俞峰教授。此前，他分别于 2021 年和 2019 年获得中山大学软件工程硕士学位及星海音乐学院钢琴表演学士学位。

他的研究长期深耕人工智能与音乐的交叉领域，尤其在音乐生成与音乐信息检索 (MIR) 方向产出了多项成果。作为第一或共同第一作者，他在 ACL、NAACL、IJCAI、ICASSP 及 ISMIR 等人工智能与音乐领域的国际顶尖会议及期刊上发表了多篇学术论文。其代表性工作包括 CLaMP 系列多模态检索模型、NotaGen 及 ChatMusician 等。凭借出色的科研能力，他曾荣获 2023 年国际音乐信息检索大会 (ISMIR 2023) 最佳学生论文奖，并先后获评 2024 年国家研究生一等学业奖学金及 2025 年北京市优秀毕业生。在正式投身工业界研究之前，吴尚达博士曾先后在微软亚洲研究院 (MSRA)、微软 Azure Cloud 及字节跳动 Seed-Music 实验室担任研究实习生。他致力于深化跨学科研究，旨在通过与领域专家的紧密合作，持续推动音乐人工智能这一前沿科技领域的发展与突破。

Dr. Shangda Wu is currently a research scientist at a leading domestic internet company, specializing in algorithm research for speech large language models. He obtained his Ph.D. in Music Artificial Intelligence and Information Technology from the Central Conservatory of Music in June 2025, where he was co-advised by Prof. Maosong Sun (Tsinghua University) and Prof. Feng Yu (Central Conservatory of Music). Dr. Wu holds a unique interdisciplinary background, having earned a Master of Science in Software Engineering from Sun Yat-sen University in 2021 and a Bachelor of Music in Piano Performance from the Xinghai Conservatory of Music in 2019. His research is deeply rooted in the intersection of AI and music, with significant contributions in music generation and music information retrieval (MIR). As a first or co-first author, he has published multiple papers in top-tier international conferences and journals in the fields of AI and acoustics, including ACL, NAACL, IJCAI, ICASSP, and ISMIR. His representative works include the CLaMP series of multimodal retrieval models, NotaGen, and ChatMusician. Recognized for his academic excellence, he was honored with the Best Student Paper Award at ISMIR 2023 and has been a recipient of the First-Class National Graduate Academic Scholarship (2024) and the title of Outstanding Graduate of Beijing (2025). Prior to his current role in the industry, Dr. Wu gained extensive experience as a research intern at Microsoft Research Asia (MSRA), Microsoft Azure Cloud, and ByteDance (Seed-Music). He is dedicated to advancing interdisciplinary research and aspires to push the frontiers of music AI through collaborative innovation with global experts.

# AI

## 第三届 世界音乐人工智能大会

---

第三届世界音乐人工智能大会 (The Third Summit on Music Intelligence) 将于 2026 年 4 月 25 日至 26 日在北京中央音乐学院举办。大会汇聚全球音乐人工智能领域的学术领军人物及音乐大模型领域的代表性企业,凝聚智慧、拓展视野,共同探索未来音乐的发展图景。会议将聚焦前沿技术进展与产业热点,搭建高水平交流平台,推动音乐人工智能在“产、学、研、用”各环节的深度融合,服务北京发展、助力国家战略,与世界携手一同开创音乐与智能融合的未来。

# 组委会

---

## 大会主席

于红梅

## 共同主席

戴琼海

## 执行主席

李小兵

## 名誉主席

俞 峰 Jean-Michel Jarre 郭毅可 院士 管晓宏 院士

## 程序委员会（按姓氏笔画排序）

于 阳 方恒健 王志鸥 孙茂松 邱志杰

吴玺宏 杨 丽 栾 家 钱 琦

## 外事统筹

陶 倩

## 工作委员会（按姓氏笔画排序）

于海波 马 军 王晓庆 王雪莹 王文潇 卢 迪 刘家丰 孙宇明

李茜茜 张 渊 张昕然 谷美莲 周晴雯 周麟一 周昊天 赵艺璇

柴 扉 高 妍 栾 家

## 志愿者（按姓氏笔画排序）

元佳宁 卜禹翔 王楚旖 王文楚 王 茜 王 紫 王鑫琛 冯子懿

丘悦欣 刘俊汝 刘 毅 刘恩洋 孙静茹 芦乐妍 肖 翔 杨佳一

杨婷絮 张 博 段 晨 龚颖芸 金 戈 林向彬 林雨声 洪若希

海 纳 赵雪丹妮 许玥童晖 陈 菲 李小宁 李 頔 李思麒 何紫怡

黄千倪 黄 都 黄文杰 彭 晨 蓝善美 魏圣普 隋林木

## 承办单位

中国人工智能学会艺术与人工智能专委会

中央音乐学院音乐人工智能与音乐信息科技系

## 主办单位



中国人工智能学会  
Chinese Association for Artificial Intelligence



中央音乐学院  
CENTRAL CONSERVATORY  
OF MUSIC

第三届  
The Third

# 世界音乐人工智能大会

Summit on Music Intelligence



第三届  
The Third

# 世界音乐人工智能大会

Summit on Music Intelligence



主持人：管晓宏

Host: Guan Xiaohong

中国科学院院士

Academician of the Chinese Academy of Sciences

SOMI  
2026

中国·北京  
2026  
04.25-26

CAAI 中国人工智能学会  
Chinese Association for Artificial Intelligence

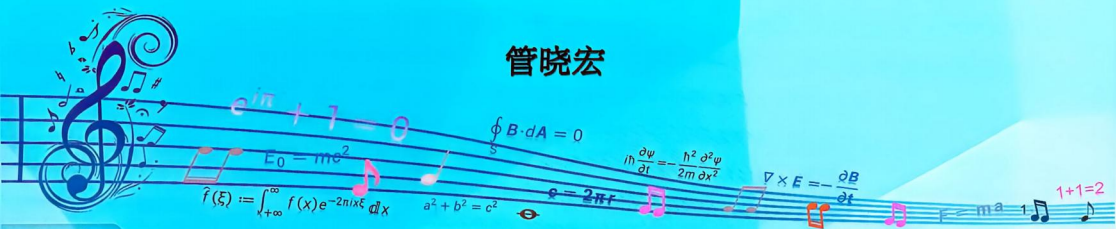


西安交通大学  
XI'AN JIAOTONG UNIVERSITY

# 音乐智能量化与认知的研究进展与艺术实践

西安交通大学

管晓宏



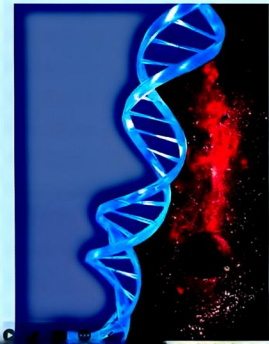


## 音乐旋律计算智能量化模型研究的起源

---

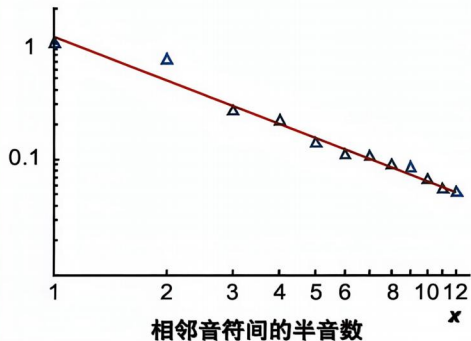
---

自然与工程系统中的幂律关系



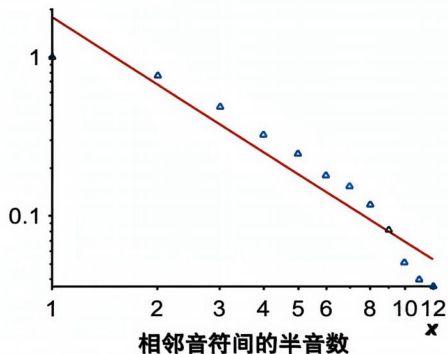
## 音乐旋律变化的幂律 (2)

$T(x)$  CCDF



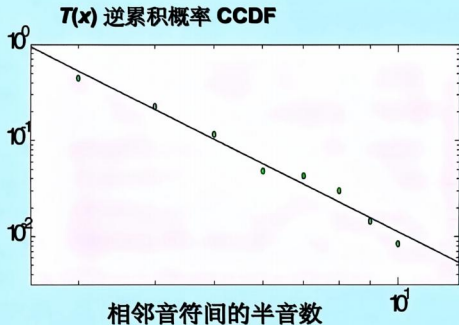
莫扎特: 《C大调长笛竖琴协奏曲》

$T(x)$  CCDF

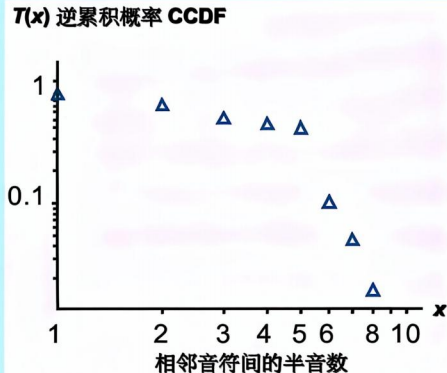


巴赫: 《d小调双小提琴协奏曲》

# 音乐旋律变化的幂律 (3)



莫扎特：D大调第二长笛协奏曲

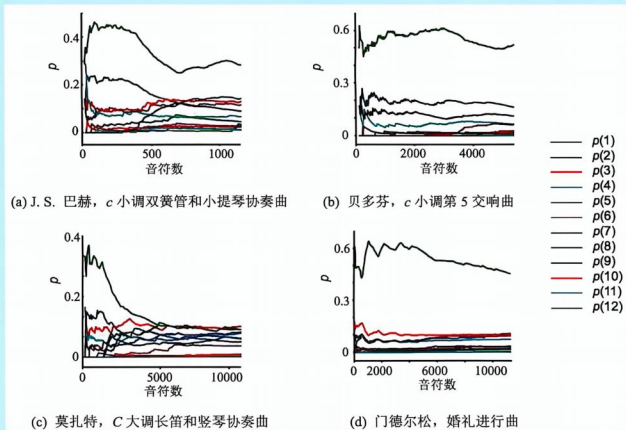


斯托克豪森：摩羯座  
Stockhausen: Capricorn

# 旋律的三个数学特征 (1)

旋律结构的重复使用以及高度统一的组织，使得旋律音程分布为稳态分布：

$$\lim_{N \rightarrow \infty} p(i, N) = p(i)$$



随着旋律的发展，以半音为单位的旋律音程的概率分布的变化

# 主要参考文献

- [1] J. G. Donald, V. P. Claude, *A History of Western Music* (W W Norton & Co Ltd, New York, 2014).
- [2] A. Schoenberg, *Fundamentals of Musical Composition* (Faber & Faber, London, 1970).
- [3] P. Hindemith, *The Craft of Musical Composition, I. Theoretical Part* (Schott & Co., Ltd., London, 1945).
- [4] 赵晓生, 《传统作曲技法》, 上海音乐出版社, 2003
- [5] S. Kostka, *Materials and Techniques of Twentieth-Century Music* (Pearson Prentice Hall, Upper Saddle River, 2006).
- [6] E. Toch, *Melodics* (People's Music Publishing House, 1984).
- [7] H. Schenker, *Free Composition* (Longman, London, 1979).

## 问题求解 (2)

泛函极值条件:

$$\frac{\partial L_i}{\partial p(i)} \delta p(i) = 0$$

求解上式, 得到调性音乐的旋律音程逆累积概率分布函数服从幂律:

$$T(i) = \frac{C}{i^D}$$

其中  $c > 0, D > 0, i = \{k, k + 1, \dots, n\}$ .

Nan N, Guan XH, Wang YX, Du YT, Common quantitative characteristics of music melodies — pursuing the constrained entropy maximization casually in composition, *SCIENCE CHINA Information Sciences*, 2022, 65 (7): 174201:1-3.

Nan N, Guan XH, Goal orientation in music composition and other social behaviors leading to the common quantitative law, *IEEE Transactions on Computational Social Systems*, 10 (1), 2023, 388-402.

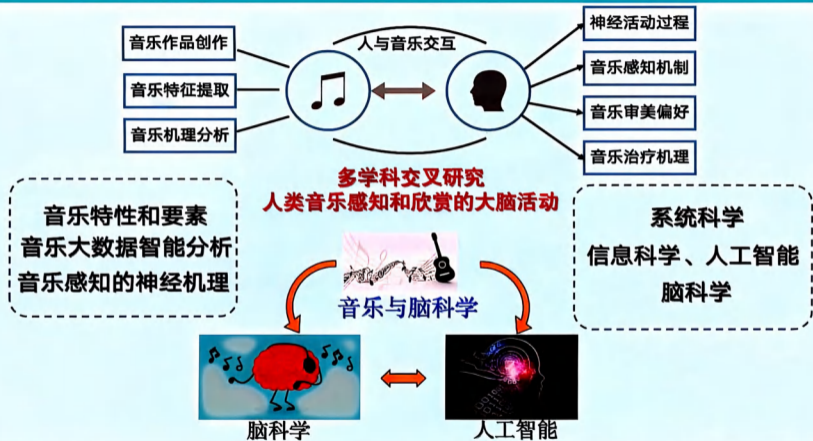
## 问题二：音乐感知的愉悦感

---

没有受过专业音乐训练的非音乐人  
为什么喜欢优美的旋律？

旋律变化符合幂律的音乐  
更能够使人产生愉悦感？

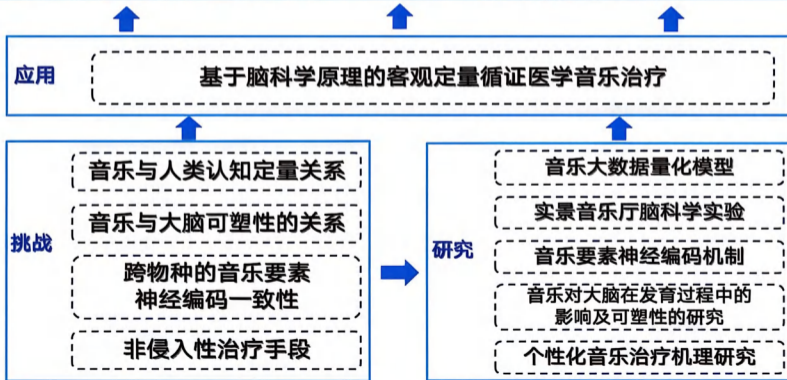
# 音乐脑科学认知研究的重大意义



多学科交叉研究音乐感知和欣赏的大脑活动，能够以全新范式，揭示音乐感知和审美的神经机理，发现音乐治疗的机理，促进音乐艺术创作和脑科学的共同发展，具有重大理论和应用价值。

# 研究目标

建立机理与数据融合的音乐与脑认知量化关联模型，平行开展人和非人灵长类动物的多层次跨物种音乐脑认知研究，从传统宏观层次拓展到微观层次构建脑科学测量方法，提出基于脑科学原理的客观定量循证医学音乐治疗方法，奠定音乐智能与脑认知量化基础理论与实用方法



# 研究内容

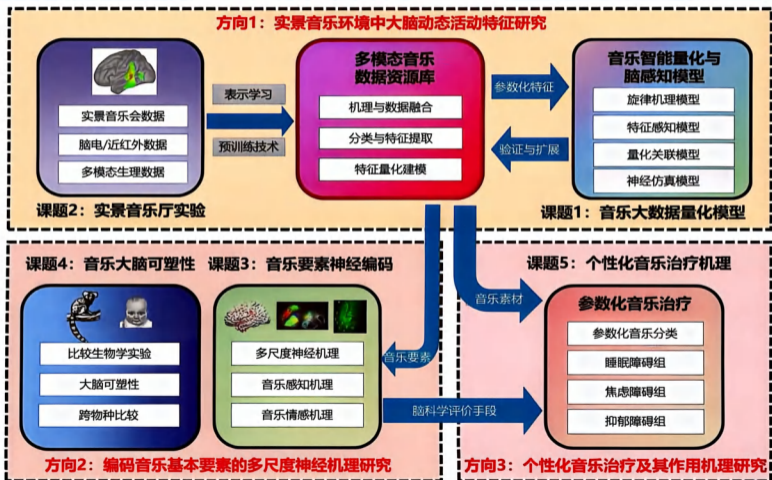
● 课题一：  
音乐大数据量化模型

● 课题二：  
实景音乐厅脑科学实验

● 课题三：  
音乐要素神经编码机制

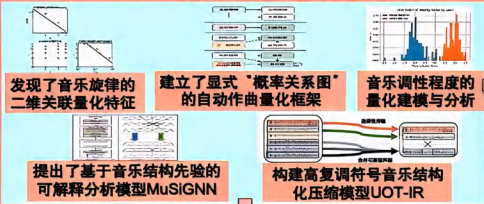
● 课题四：  
音乐对大脑在发育过程中的影响及可塑性

● 课题五：  
个性化音乐治疗机理

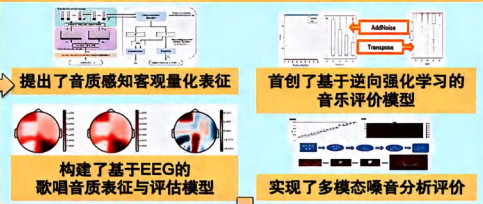


# 课题一进展：音乐量化与感知关联性建模

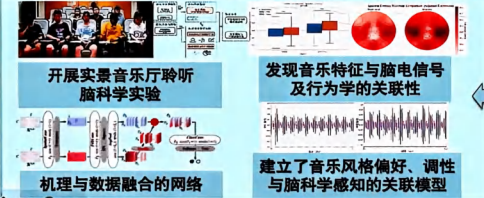
## 一、构建了机理与数据融合的音乐量化模型



## 二、提出了融合脑认知的音频量化表征方法



## 三、建立了音乐旋律与脑科学感知的关联性模型



## 四、研发了多模态音乐理解与生成大模型



# 研究进展1.1: 旋律音高节奏二维量化模型

## □ 西方调性音乐旋律中音高与节奏的二维量化模型

- 音高音程和音符时值的边际统计均服从幂律分布

$$P(X \geq x) = F(x) = \left(\frac{1}{a}x\right)^{-\alpha}, x > a$$

$$P(Y \geq y) = F(y) = \left(\frac{1}{b}y\right)^{-\beta}, y > b$$

- 音高 - 节奏存在稳健的下尾强耦合、上尾近独立特征

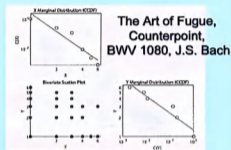
$$\lambda_U = \lim_{q \rightarrow 1} P(Y > F_Y^{-1}(q) | X > F_X^{-1}(q))$$

$$\lambda_L = \lim_{q \rightarrow 0} P(Y \leq F_Y^{-1}(q) | X \leq F_X^{-1}(q))$$

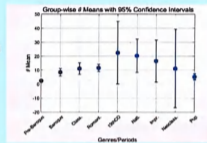
- 音高与节奏的相关性完全由Copula函数确定，其依赖参数 $\theta$ 可作为量化作曲理念演化指标，其值呈非单调演化

基于以上量化特征，得到音乐旋律音高与节奏的双变量联合分布：

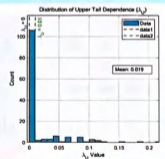
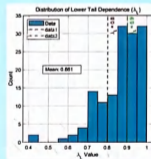
$$F(x, y) = P(X \leq x, Y \leq y) = \begin{cases} \left\{ \frac{1}{[1 - a(x+1)^{-\alpha}]^\beta} + \frac{1}{[1 - b(y+1)^{-\beta}]^\alpha} \right\}^{-1/\theta}, & x, y = 1, \dots, M-1, \\ 1 - a(x+1)^{-\alpha}, & x = 1, \dots, M-1; y = M, \\ 1 - b(y+1)^{-\beta}, & x = M; y = 1, \dots, M-1, \\ 1, & x = y = M. \end{cases}$$



音程和音符时值的幂律分布



不同音乐时期/风格的 Clayton 依赖系数

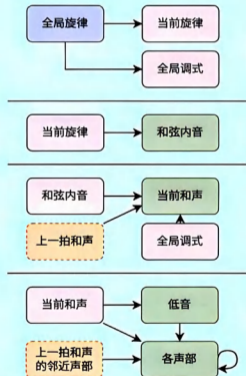


上下尾部系数分布

该模型的量化特征可为音乐的理解与生成（课题一）提供量化理论

# 研究进展1.2: 和声走向生成式量化概率模型

## 建立显式“概率关系图”的自动生成和声量化框架



### ➤ 变量关系建模

“旋律-和弦内音-和声-邻近声部走向”的多级变量关系

### ➤ 量化作曲规则

- 对音乐中的旋律轮廓、和弦、邻近声部连接等样本进行统计
- 基于概率量化写作规则

### ➤ 解决少样本问题

针对多条件变量下样本数不足的问题，进行适当的条件独立假设拆解为多个概率



若  $A, B$  关于  $Z, y$  条件独立, 即  $P(A, B|Z, y) = P(A|Z, y)P(B|Z, y)$ , 则  $P(y|A, B, Z) \propto P(y|A, Z)P(y|B, Z)/P(y|Z)$ .

✓ 该框架可为多声部音乐的生成（课题一的研究方向四）提供和声生成的基础模型



# 研究进展1.4：音乐调性程度的量化建模与分析

- 背景**
- 音乐调性的发展：传统调性-调性式微-无调性；调性与音乐感知息息相关。
  - 音乐的调性程度缺乏量化标准与建模分析，制约音乐调性的感知与脑科学研究。

## □ 基于音乐机理提出音乐调性的三个量化特征

### ➤ 量化特征一：无调性程度 (A)

理论依据：十二音体系核心思想。

$$A = \frac{Profile \cdot 1}{||Profile|| \cdot ||1||}$$

其中Profile表示调性轮廓，是十二个音级的权重向量。

### ➤ 量化特征二：调性中心稳定性 (S)

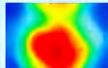
理论依据：调性音乐调性中心长期稳定；转调少，多为近关系调。无调性音乐“调性中心”频繁跳跃，呈“闪烁不定”状态。

$$S = \frac{1}{N-1} \sum_{i=1}^{N-1} corr(Profile_i, Profile_{i+1})$$

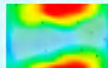
### ➤ 量化特征三：调性清晰度(C)

理论依据：人类对调性的感知来自音乐片段与“最匹配调性模板”的相似程度。调性音乐清晰、持续；无调性音乐：模糊、分散。

$$C = \frac{1}{N} \sum_{i=1}^N \max_{k \in Keys} \{corr(Profile_i, Profile_k)\}$$



调性音乐

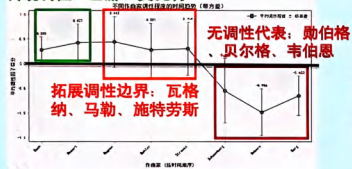


无调性音乐

## □ 音乐调性程度模型的验证与分析

- ✓ 构建了包含不同调性音乐程度的8个作曲家、613首作品的数据集
- ✓ 因子分析：提出三个量化特征由潜在调性程度因子T决定，分类准确率达0.9847，显著超越各项基线模型

传统调性：巴赫、莫扎特



不同作曲家的作品调性程度规律分析

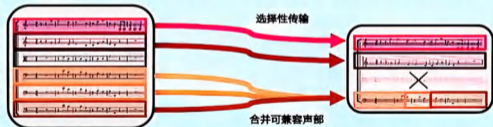
- 成果**
- 为分析不同作曲家、不同历史时期的音乐作品调性程度提供统一的量化模型
  - 可以作为音乐脑科学实验（课题一、二和三）的刺激材料提供量化标准和理论依据

# 研究进展1.5: 高复调符号音乐结构化压缩模型

## 研究背景与挑战

- ✓ **预算受限:** 高复调符号音乐常包含大量轨道, 而下游任务通常要求固定数量的轨道或槽位。
- ✓ **结构缺失:** 传统启发式删减方法容易直接丢弃重要声部, 难以保持音乐结构角色的一致性。
- ✓ **约束不足:** 通用降维或聚类方法缺乏对配器兼容性、音域限制与可演奏性的显式建模。

## 建立了符号音乐的最优传输压缩模型UOT-IR



- ✓ **提出了 UOT-IR 框架:** 重构为固定预算下的结构化路由, 构建可解释的解耦式框架。
- ✓ **构建了结构感知代价:** 融合配器先验、局部描述符、语义角色信息与物理可演奏性约束。

## 模型评估、验证与分析

- ✓ 在统一固定槽位预算下评估了两类任务: 模板标准化与自适应保留。
- ✓ 提出的模型UOT-IR能在有限预算下更好地兼顾内容保留、结构一致性与音乐合理性。

Method	Content Fidelity			Distributional and Structural Statistics			
	Note-F1 $\uparrow$	Note-P $\uparrow$	Note-R $\uparrow$	PTED $\downarrow$	PC-JSD $\downarrow$	Dir-JSD $\downarrow$	IR-JSD $\downarrow$
Direct	0.7709	0.9999	0.6563	0.0184	0.1049	0.1282	0.1328
Random	0.7157	0.9999	0.6545	0.0099	0.0768	0.0790	0.1094
Greedy	0.9113	0.9998	0.6458	0.0207	0.0376	0.0502	0.0411
Skyline	0.8778	0.9975	0.7982	0.0138	0.0369	0.0618	0.0481
PCA	0.2992	0.9998	0.5813	0.0071	0.0781	0.0918	0.1244
KMeans	0.7338	0.9999	0.6046	0.0854	0.0645	0.0711	0.1139
BM3	0.8989	0.9998	0.8270	0.0172	0.0356	0.0554	0.0444
NMF	0.7401	0.9999	0.6152	0.0060	0.0672	0.0782	0.1102
Vanilla-UOT	0.8593	0.8896	0.6119	0.0351	0.0109	0.0234	0.0130
UOT-IR-Core	0.9053	0.9104	0.9082	0.0228	0.0041	0.0142	0.0056
UOT-IR	0.9138	0.9120	0.9208	0.0250	0.0071	0.0178	0.0078

Method	Fidelity	Standardization and Structural Compatibility			
	Note-F1 $\uparrow$	PR-Diff $\downarrow$	PC-Diff $\downarrow$	FTED $\downarrow$	SC $\downarrow$   BC $\downarrow$
Direct	0.6712	0.1421	0.0699	0.0186	20.3849   0.6681
Random	0.6732	0.0953	0.0854	0.0100	18.1475   0.4247
Greedy	0.7385	0.0514	0.0745	0.0209	19.1898   0.4645
Skyline	0.7301	0.0568	0.0864	0.0140	18.2123   0.4451
PCA	0.6078	0.1218	0.1210	0.0073	16.6987   0.4116
KMeans	0.6327	0.1009	0.1041	0.0056	16.5824   0.3939
BM3	0.7315	0.0501	0.0799	0.0173	18.8827   0.4570
NMF	0.6334	0.1099	0.1077	0.0062	16.8215   0.4077
Vanilla-UOT	0.6335	0.0461	0.0842	0.0294	30.2076   0.7965
UOT-IR-Core	0.9334	0.0196	0.0400	0.0225	18.1452   0.4466
UOT-IR	0.9370	0.0116	0.0285	0.0238	14.7165   0.3406

**自适应保留任务:** Note-F1=0.9120, 该模型在有限预算下更有效地保留关键音乐内容

**模板标准化任务:** 结构兼容性指标表现最优, 该模型在结构一致性和冲突控制方面更具优势。

## 成果

- 为符号音乐压缩构建了一种基于路由的研究范式, 实现了兼具结构化和连贯性的音乐表示
- 该模型可以作为音乐生成、编曲等多个下游任务 (如课题一的研究方向四) 提供预处理模型

# 研究进展2.1: 音乐质量表征与智能评估模型

面临问题: 现有的音乐评价体系十分依赖专家主观打分, 缺乏客观机制

## 音乐旋律质量评估模型

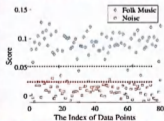
- 首創了基于逆向强化学习的音乐旋律评价模型: 音乐特征 $\vec{x}$ 到得分 $s$ 的映射 $F: \vec{x} \in R^n \rightarrow s \in R^p$

打分机制  $\xrightarrow{\text{逆向强化学习}}$  奖励函数搜索

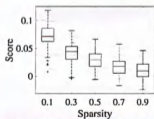
$$\max_{\bar{\theta}} F(\bar{\theta}) = \frac{1}{N_{(+)}} \sum_{i=1}^{N_{(+)}} \bar{\theta}^T \vec{x}_{(+),i} - \frac{1}{N_{(-)}} \sum_{i=1}^{N_{(-)}} \bar{\theta}^T \vec{x}_{(-),i}$$

s. t.  $\|\bar{\theta}\| \leq 1$

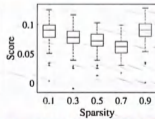
- $\vec{x}_{(+),i}$  是 '高质量音乐' 特征, 即93首中国民乐特征
- $\vec{x}_{(-),i}$  是 '低质量音乐' 特征, 即100首噪声序列特征



准确识别音乐与噪声



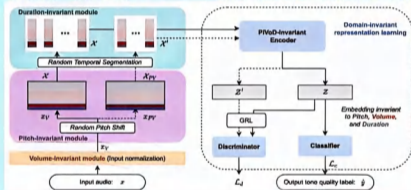
识别音乐中的噪声水平



模拟人的听觉美感

## 演奏与乐器音质表征与评估模型

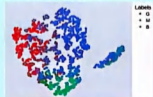
- 基于音质具有音高、音量、音长不变性的感知理论, 构建了PiVoD不变性音质量化表征模型



演奏音质分类  
准确率达到92.3%



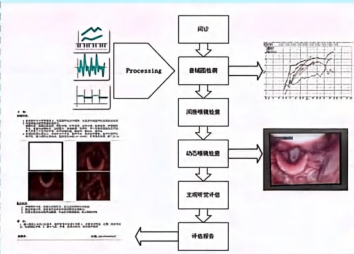
乐器音质分类  
准确率达到81.3%



成果 为AI作曲、演奏教学提供审美标尺, 提升音乐要素评估的客观性与人类偏好契合度

# 研究进展2.2: 多模态嗓音分析与智能评估

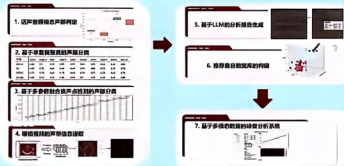
中央音乐学院嗓音研究中心  
嗓音检测专家临床经验  
为研究背景和理论支撑



通过人工智能多模态方法精准识别  
嗓音特质，辅助教师制定个性化训练  
方案，提升声乐人才培养效率

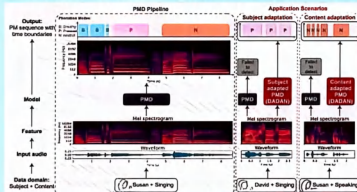
成果

## 多模态演唱声部分析与智能评估



声部分类准确率达到92%  
(男高音、男低音、女高音、  
女低音)

## 歌唱和语音发声模式音质分析与个体自适应检测系统



歌唱和语音中发声模式检测  
(漏气式发声、挤卡式发声、  
正常发声) 跨被试平均F-score  
分别提高了44.7%和6.8%

# 研究进展2.3: 融合脑认知的乐音音质量化表征与评估模型



传统专家打分评价:  
存在主观性、时间成本高、  
标准不一致等问题

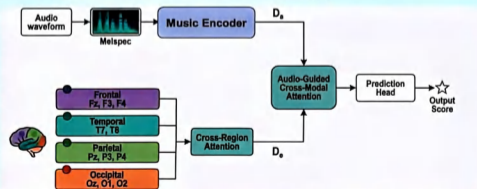


建立神经生理信号辅助  
音质评价新范式:  
客观、节省时间、标准一致

## 融合声学 & 脑电感知的自监督音质评估模型

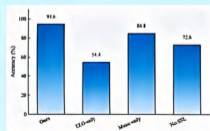
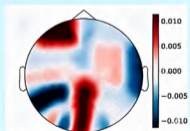
### 创新成果

构建了首个EEG歌唱质量评估多模态数据集  
提出了基于音频特权学习的声学-EEG跨模态特征融合框架  
设计了脑区感知的时空联合EEG编码器



### 主要结论

验证了EEG神经客观评估的可行性: 音频特权学习框架证明了仅依赖 EEG 信号可实现对歌唱音质的有效且精准的评估。  
明确了歌唱评价神经客观评估的核心维度: 从神经生理学视角证实, 音准、节奏与和谐度是歌唱评价最具决定性的维度。  
揭示了不和谐音质感知的左右脑不对称性: 大脑在处理谐波比不同的刺激时, 左右脑alpha波呈现不对称。



成果 实现主观听感与大脑响应的关联建模, 使音质评估从听觉经验迈向脑科学认知层面的科学量化

# 研究进展3.1：音乐风格偏好与脑科学感知的关联性模型

## 基于实景音乐凝听量化与感知相关性分析

### 实景音乐厅聆听音乐脑科学实验

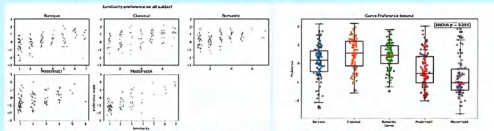
传统音乐

现代音乐

分期	风格流派	时长	作曲家姓氏	作品名称
1	Baroque	05:15	Pachelbel	Canon in D
1	Baroque	04:10	Bach	Contrapunctus VI
2	Baroque	04:43	Bach	Contrapunctus XI
2	Baroque	04:54	Bach	Contrapunctus VIII
1	Classical	05:41	Boccherini	Quartet in C minor
1	Classical	04:25	Beethoven	String Quartet No.5 in A major Op.18
2	Classical	04:41	Haydn	String Quartet in D major
2	Classical	05:44	Mozart	Serenade No.13 in G major Eine Kleine Nachtmusik
1	Romantic	04:37	Schubert	String Quartet No.14 in D minor D810
1	Romantic	05:57	Schumann	String Quartet No.2 in F major Op.41b
2	Romantic	04:13	Mendelssohn	String Quartet No.1 in E-flat major Op.12
2	Romantic	05:12	Tchaikovsky	String Quartet No.1 in D major Op.11
1	Tonal	05:14	Bartok	String Quartet No.4
1	Tonal	04:52	Kodaly	String Quartet No.2, Op.10
2	Tonal	05:06	Bridge	Novelletten, H.44
2	Tonal	05:19	Bartok	String Quartet No.5
1	Atonal	04:08	Webern	5 Movements, Op.5
1	Atonal	04:08	Milhaud	String Quartet No.8, Op.121
2	Atonal	03:53	Webern	String Quartet 1
2	Atonal	04:27	Honegger	String Quartet No.3, H.114

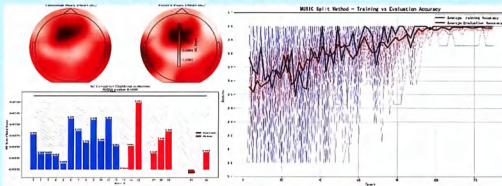


### 音乐偏好的行为学关联分析



音乐风格偏好与音乐熟悉度正相关

### 传统/现代音乐偏好的脑电信息熵差异显著



成果 揭示音乐风格偏好的脑科学机制，为个性化音乐治疗提供的神经科学依据

# 研究进展3.2: 音乐调性与脑科学感知的关联模型

## 对比聆听实验设计

1. 针对脑电非稳态问题  
无调性和调性片段配对  
连续播放
2. 针对情绪耦合问题,  
每个无调性与两类情绪的片段  
配成两对

无调性  
音乐片段

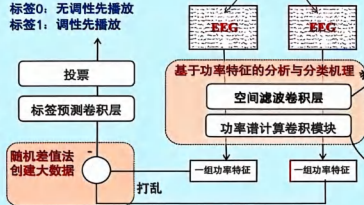
忧伤的调性  
音乐片段

欢快的调性  
音乐片段



## 机理与数据融合的可解释网络

标签0: 无调性先播放  
标签1: 调性先播放



## 基于可解释网络的音乐调性与脑信号关联分析

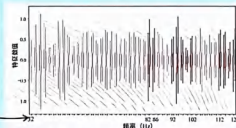
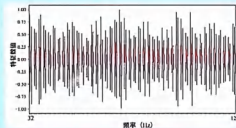
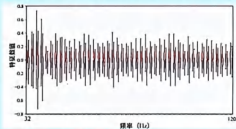
顶叶



右颞叶



左颞叶—颞叶  
连接区



红色: 无调性音乐  
蓝色: 调性音乐

1. 顶叶和颞叶, 无调性音乐  
激发的脑电 $\gamma$ 波 ( $>30\text{Hz}$ ) 能量  
大于调性音乐激发的脑电  
(被试间共性)

原因: 听众在聆听时会预测  
音乐结构。无调性音乐的难  
预测性, 要求多脑区高度协  
作以应对更复杂的信息处理  
需求, 表现为 $\gamma$ 波能量增强

2. 左颞叶, 部分频带内无调  
性音乐的脑电能量小于调性  
音乐 (个性)  
体现无调性音乐处理中  
左右颞叶的非对称性

**成果** 阐明调性感知的脑响应规律, 深化对音乐情感与认知加工的理解

# 研究进展4.1：基于3D层次化架构的交响乐自动作曲与编排系统

## 研究背景

**缺乏可控性：**现有模型在捕捉交响乐复杂的结构方面缺乏可控性。

**编排复杂性：**交响乐涉及大量乐器与轨道，传统扁平化模型计算开销大。

**出现不协和音：**现有模型在生成符号音乐时，常出现听感不适的“不协和音”。

## 核心创新

**层次化模型：**将音乐表示为“小节、轨道、音符”三个维度，通过逐层解码实现高效生成。

**和声骨架控制：**引入多声部和声走向作为生成条件，确保音乐生成围绕给定的乐理约束展开。

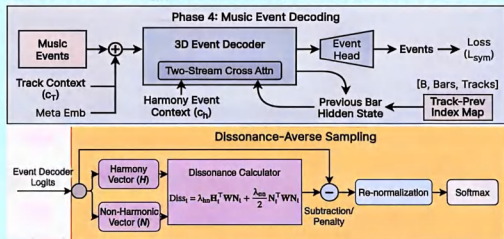
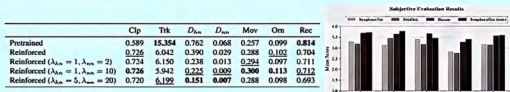
**交叉注意力：**针对旋律与和声的关联进行优化，确保多轨道生成时，旋律与和弦的高度契合。

**基于感知的采样：**引入不协和音规避采样机制，实现音乐听感的本质优化。

## 研究成果

**客观评价：**通过多项乐理指标找到不协和音规避平衡点。

**主观听感：**与业内主流模型对比，本系统在整体上表现更优。



**成果** 基于研究进展1.2，突破交响乐多声部结构建模瓶颈，实现符合专业范式的可解释自动化作曲、编曲

## 研究进展4.2：多模态音乐理解与生成大模型

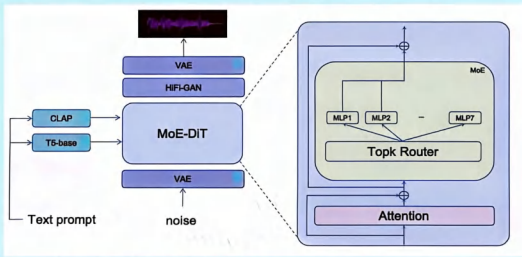
### 提出了大规模的音频多模态数据集和细粒度处理方法

- **研究背景**：现有音频-文本数据集规模有限，文本细粒度不足，制约音频大模型发展。
- **核心创新**：提出Colla Data数据集（5.4万小时，624万高质量音频-文本对），规模显著超越已知开源数据集；设计三阶段细粒度文本处理流程，结合大模型标注与幻觉率抑制。
- **成果价值**：基于Colla Data的生成模型性能显著提升，为音频多模态研究提供高质量数据基础。

### 提出了基于MoE-DiT的音频音乐生成模型

- 提出Colla Music MoE模型，采用Flow-matching的MoE-DiT结构，总参数量0.7 B。
- 引入混合专家系统（MoE），通过多专家模型分工协作，有效缓解多任务冲突。
- 首次验证多个小专家模型的良好耦合，可超越1B~2B级单一稠密大模型的音乐生成性能。

Model	MusicCaps			AudioCaps		
	FD <sub>↓</sub>	KL <sub>↓</sub>	CLAP <sub>↑</sub>	FD <sub>↓</sub>	KL <sub>↓</sub>	CLAP <sub>↑</sub>
Stable Audio	119	1.24	0.32	83.1	2.23	0.28
Colla Music (Ours)	98.15	1.20	0.35	70.52	2.20	0.35



# 近期研究计划

聚焦音乐智能量化与脑科学感知的关联模型、智能分析与实验验证：  
**音乐风格、音质、情感的脑科学感知量化**

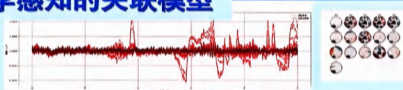
## 音乐智能量化与脑科学感知的关联模型



音乐调性与音乐偏好的关联量化分析

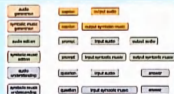


音乐声学与信号处理的音乐感知量化

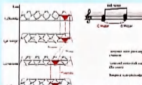


整合音乐结构、行为反应与神经复杂度的分析路径

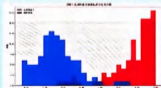
## 机理与数据融合的音乐量化模型



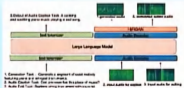
建立高质量音乐结构量化数据库



基于音乐结构的音乐量化建模与分析



音乐调性程度量化模型建模与分析



构建多模态多任务音乐大模型

# 课题二进展：多种场景下的实景音乐厅脑科学实验



单人、多人聆听音乐实验 (2023.10, 清华大学隔声间)

现场采集多人聆听音乐的神经生理数据 (2023.10.28, 中央音乐学院歌剧厅)



实验一，钢琴二重奏



实验二，小型乐队合奏

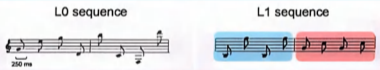
清华大学艺术博物馆 (2023.12.18)

# 课题三进展：发现音乐要素神经编码机制

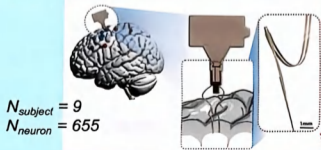
## • 音乐结构加工的神经机制

- 发现：人类前额叶存在音乐层级结构敏感的神经元
- 提示：**前额叶神经元**可能支持了**复杂音乐结构**的神经编码

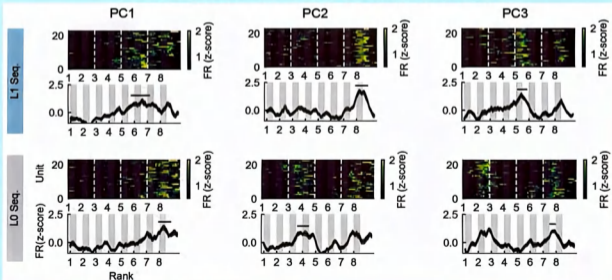
### 音乐抽象结构研究范式



### 人脑神经元记录

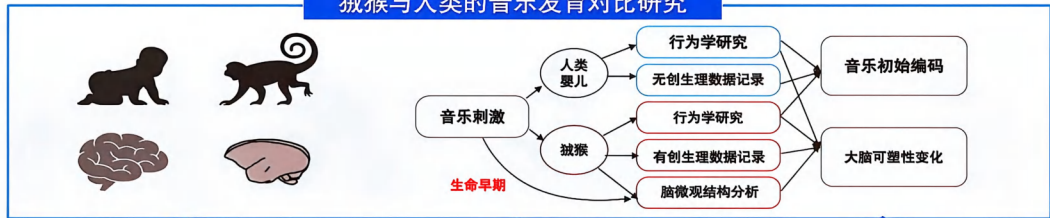


### 人类额叶皮层神经元对音乐结构信息的编码响应



# 课题四进展：音乐对大脑发育的影响及可塑性实验

## 猿猴与人类的音乐发育对比研究



### 大脑最初如何编码音乐



音乐编码

### 初始音乐偏好是否具有普遍性



音乐偏好

### 猿猴在不同音乐刺激环境中的行为功能和大脑结构发育过程



音乐发育

# 研究进展五：发现个性化音乐治疗机理

形成常见精神障碍的临床治疗决策需要兼顾有效性和安全性

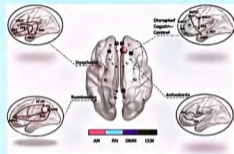
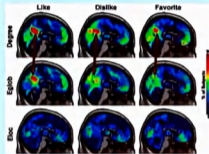


药物治疗



非药物治疗

神经生物学发现给个性化音乐治疗的开发提供了神经学基础



启发1：人脑对不同音乐有独特的响应模式启发2：焦虑障碍发生发展中的脑环路机制

脑活动监测技术和人工智能为个性化音乐治疗开发提供工具



+脑科学监测技术



+人工智能

需要解决的几个关键问题

- 音乐治疗的脑响应特征（主观的、客观的）如何？常见精神障碍患者的特异性脑响应特征又如何？（**基础**）
- 基于对音乐治疗的脑响应特征的深入理解，如何整合脑科学监测技术和人工智能技术构建个性化音乐治疗方案？（**方法**）
- 个性化音乐治疗是否能在临床实例的解决中真正体现出疗效和安全性优势？（**验证**）
- 能否通过构建疗效预测系统指导个性化音乐治疗的临床实践（**实践**）

# 艺术实践：艺术与科学的交汇系列音乐会

艺术与科学深度融合、科普讲座与专业演奏交替穿插、面向受众定制创作

## 常规版

面向高校师生和公众 内容创新



## 英文版

代表“丝绸之路大学联盟”科普品牌



## 绿色发展版

面向双碳目标 服务国家重大需求



## 中学版

面向中学生做科普艺普 现场音准的数学物理实验



## 小学版

面向小学生做科普艺普 分数运算题打开音高之门



# 音乐智能与认知科学研究与艺术实践未来可期

- 促进信息科学、脑科学、生物医学与人文艺术学科的交叉融合，开创全新研究方向
- 艺术形象思维与科学逻辑思维相互启迪，推动“五育并举”，培养拔尖创新人才
- 创新形式的艺术实践和科学普及，提升公众艺术与科学素养，服务社会

第三届  
The Third

# 世界音乐人工智能大会

Summit on Music Intelligence



从音乐厅到社会空间：借助技术重新语境化当代音乐  
From Concert Hall to Social Space: Recontextualizing  
Contemporary Music through Technology

乔治·海杜

Georg Hajdu

德国利盖蒂中心主任，ICMC25主席

Director of Ligeti Center (Germany); Chair of ICMC25

somi  
2026

中国·北京  
04.25-26





# From Concert Hall to Social Space

## Recontextualizing Contemporary Music through Technology

Summit on Music Intelligence, Beijing, April 26, 2026



Georg Hajdu | *ligeti center* | Hamburg

# Introduction

## Computer Music - Pandora's Box

- **Max Mathews:** “There are no theoretical limits to the performance of the computer as a source of musical sounds.”
- In a 2011 eulogy, the author (falsely) claimed that Mathews was the author of a story about the small and self-enclosed elite music world of virtuosos dedicated to classical repertoire and the rest of the world would be involved in musical activities mediated by digital interfaces.
- While there is no evidence for the existence of this text, the claim captures the essence of his musical vision and the role that software and hardware play in democratizing music and moving it beyond the cultural temples of the nineteenth century.
- My presentation will be about elucidating this path and the contributions my collaborators and I were able to make.



# A Few Words About Myself I

## Background

- **Born in Göttingen into a Hungarian family**  
father a physicist; grandmother a concert pianist (student of Leo Weiner at the Franz Liszt Academy, later docent there)
- **Raised in Cologne**  
a major center for New Music, shaped by the activities of WDR, one of Europe's leading broadcasting institutions
- **Academic formation**  
studies in molecular biology and composition in Cologne
- **PhD at the University of California, Berkeley**
- **Professor of Multimedia Composition**  
Hamburg University of Music and Drama (since 2002)
- **Founding Director of the ligeti center**  
interdisciplinary platform for music, technology, and research (since 2023)



# A Few Words About Myself II

## Inspiration

- **Cologne & Hamburg**
  - Susy Krebs – first composition teacher; American composer and painter who moved to Cologne to study with Stockhausen and Kagel
  - György Ligeti – visiting student in the 1980s; formative encounter
  - Clarence (Klarenz) Barlow – met in the early 1980s; major artistic inspiration
  - Karlheinz Stockhausen – met in 1975; key figure shaping the Cologne music scene
- **San Francisco Bay Area (CNMAT, Mills College, CCRMA)**
  - David Wessel – mentor at CNMAT (early 1990s); focus on “music and mind”
  - Alvin Curran – introduction to site-specific composition; collaboration as assistant on a Max project
  - The HUB – first exposure to networked music performance and collective electronic practice
  - John Chowning – role model for establishing a world-class research environment



# Four stages of computer music composition

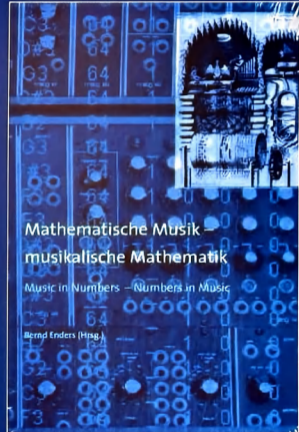
- **Three stages in 2004 essay “The computer as a source of inspiration for composers • a historical overview” (appeared in the somewhat obscure book *Mathematische Musik - musikalische Mathematik*):**
  - **Speculative (1950s–1975s)**
  - **Exploratory (1975s–2000s)**
  - **Interactive (2000s–2020s)\***
- **Fourth stage: Latent\*\* (since 2016)**

\*

Negroponte 1998: Beyond digital  
Cascone 2000: post-digital

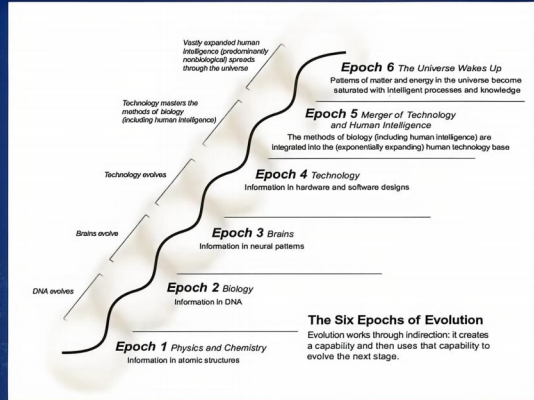
\*\*

model-based  
post-symbolic  
co-creative / symbiotic



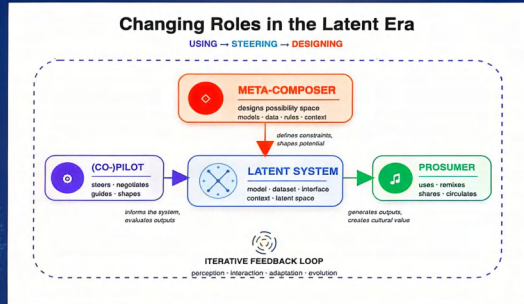
# The Larger Picture

- **McLuhan (Gutenberg Galaxy, 1962):** The role of media in human development; abilities are lost, others are gained
- **Kurzweil (2005):** Singularity is near. Prediction of the role of AI. Process of technological innovation seems to ever speed up
  - Humanity is currently transitioning from Epoch 4 to 5
- **Redefining the role of music making in the era of machine intelligence:**
  - Genres
  - Locations / social spaces (where does music production and performance take place?)



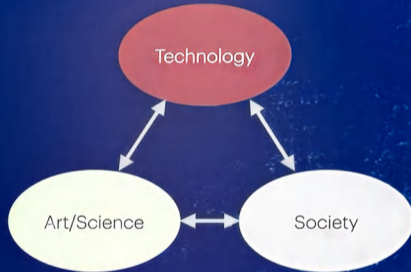
# Changing Roles and Cultural Identities

- **Prosumer (since Web 2.0):** digital music-production and consumption of online digital content.
- **(Co-)pilot:** steering generative systems such as Sumo (who is pilot or co-pilot)
- **Meta composer:** designing models, datasets, constraints and platforms; shift from object-based to relational authorship



# Cultural Shift: Policy, Funding, and Expectations

- **Changing funding logics:** from artistic autonomy toward measurable outcomes
- **Politicization of art production:** art increasingly expected to respond to social, technological, or geopolitical agendas
- **Impact as key criterion:** art must demonstrate relevance, accessibility, participation, transformation, or innovation
- **Alignment with high-tech agendas:** funding schemes prioritizes AI, energy, health, semiconductors, and infrastructure/mobility
- **Art inside innovation agendas:** productive when art is not marginal, but understood as a method of research, translation, and cultural imagination (notably visible in some Chinese policy frameworks that place art within technological innovation)
- **Return on investment:** cultural value increasingly framed through visibility, societal effect, knowledge transfer, and economic or strategic benefit
- **Changing role of work in society:** policy frameworks will shape how expanded cultural participation and artistic practice evolve given the effects of increasing automatization of labor achieved by AI



# Networked Music Performance (NMP)

## Artistic trajectory

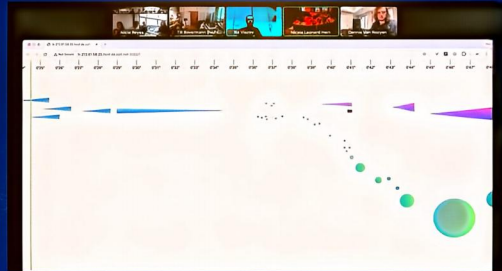
- **Origin in multimedia opera Der Sprung - Beschreibung einer Oper: Interest in distributed dramaturgy and mediated presences**
- **Development of Quintet.net & European Bridges Ensemble (1999-2015)**
  - **real-time interaction across large distances**
  - **early platform for distributed co-creation**
  - **performers act both locally and as remote agents**
- **Focus on embodiment / agency / situatedness**
  - **Embodiment and Disembodiment in networked music performance (2017)**
- **Pedagogical implementation**
  - **Master program Multimedia Composition (since 2004)**
    - **“no studio please”: mobility, flexibility, and networked thinking**
  - **resonance with Randall Packer’s concept of the total artwork**



# Networked Music Performance (NMP)

Why NMP reflects changing roles in music practice

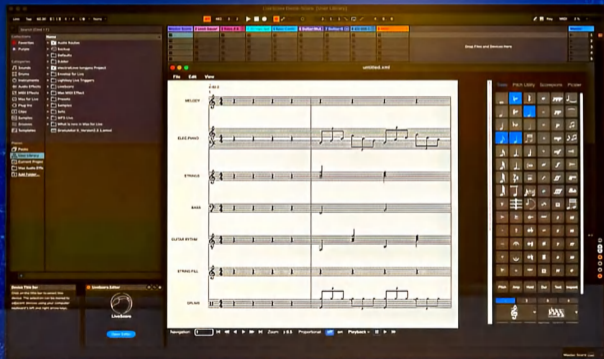
- **From performer → networked agent**
  - musicians operate within systems rather than local ensembles
- **From composer → system designer**
  - designing interaction frameworks instead of fixed works
- **From audience → distributed participant**
  - listening becomes spatially and socially extended
- **From work → process**
  - music as continuous, situated interaction rather than fixed object



# Notation for Everybody and Everything

MaxScore: tools for real-time notation and audience participation

- **Implementations**
  - **Quintet.net (NMP environment)**
  - **Max patching (MGraphics and Jitter)**
  - **Ableton Live**
  - **Drawsocket (Rama Gottfried)**
- **Non-standard and microtonal notation**
- **Proportional notation**
- **Animated scores**
- **Interactive scores**
- **... to facilitate novel types of interaction with symbolic content**



# Site-Specific Multimedia Performance as a Contemporary Artform

Dissertation by Zhou Dong

- **Environment as active agent**
  - **Site contributes materially (acoustics, architecture, social context), not just symbolically**
- **Decentering of artistic hierarchies:**
  - **No fixed dominance of music, narrative, or image; structural authority becomes fluid**
- **Non-repeatability as aesthetic condition:**
  - **Works are inherently tied to specific situations and cannot be fully reproduced elsewhere**
- **Influence of game logic and interactivity:**
  - **Meaning emerges through user choices, navigation, and engagement rather than fixed dramaturgy**
- **Expansion of compositional competence:**
  - **Technological accessibility allows practitioners from different fields to operate across disciplines**
- **Concert as one option among many:**
  - **Listening shifts from a fixed institutional setting to mobile, situational, and socially embedded contexts**

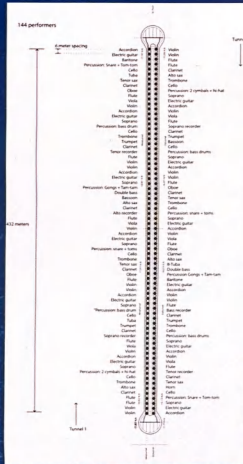
## Table of Contents

Preface	4
<b>Chapter 1: Introduction</b>	<b>6</b>
Significance of Topic	6
Thesis	6
Methodology	7
Delimitation	9
Background	9
Chapter Organization	12
<b>Chapter 2: The Genealogy of Site-specific Multimedia Performance</b>	<b>13</b>
Understanding Multimedia Performance: Body, Technology, and Integration	14
Early Trajectory from the 1870s to 1960s	20
Non-everything: Happening and Fluxus	24
Three Tributaries: Music, Theater, and Computer Technology	29
Computer Technology and Game	34
Composers Include Multimedia	36
Conclusion	40
<b>Chapter 3: Sound and Environment</b>	<b>42</b>
Early Concerts in Unconventional Spaces	43
Involving the Environment Through the Composition of Electronic Music	45
Conclusion	60
<b>Chapter 4: Mobile Technologies and Logistics</b>	<b>62</b>
Devices In Spaces	64
Portable Loudspeakers	66
Portable Sound-gathering Devices	72
Multifunctional Devices	73
Communication Devices	79
Logistics	80
Conclusion	83
<b>Chapter 5: Case Studies</b>	<b>85</b>
Alexander Schubert: <i>Control</i> (2018), Oslo	86
Georg Hajdu: <i>Symphony in St. Pauli Elbe Tunnel</i> (2018), Hamburg	100
Manos Tsangaris: <i>City Pieces (Window Piece)</i> (2016), Hongkong	123
Kirsten Reeser: <i>Berlin Rosenthaler Platz</i> (2018), Berlin	134
Luna Valentin: <i>Quatuor pour la grotte Saint Marcal</i> (2020), Biden	141
Dong Zhou: <i>Sound Ritual</i> (2023), from Hannover to Leiria	150
Conclusion	156
<b>Chapter 6: Influences on Other Art Forms</b>	<b>159</b>
Contemporary Classical Music	160
Contemporary Dance	162
Telematic Performance	165
Film	167
Conclusion	169
<b>Concluding Remarks: Expanding Artistic Approaches</b>	<b>171</b>



# Site-specific Performance: Symphony in the St. Pauli Elbe Tunnel

- **Large-scale distributed performance system**
  - 144 musicians arranged across 864 meters, coordinated via a browser-based networked score system (Drawsocket)
- **Integration of space as compositional parameter**
  - architecture and extreme acoustics (long reverberation, delay) fundamentally shape musical perception and structure
- **Network replaces traditional conducting**
  - central server + bidirectional communication system enable coordination, feedback, and collective timing
- **Animated / networked notation**
  - individual scrolling scores on iPads synchronize performers in real time, redefining the role of the score
- **Hybrid communication ecology**
  - interplay of digital network, acoustic space, and local human interaction creates multiple layers of coordination
- **Perception as situated and subjective**
  - each listener experiences a different “present” due to spatial delay and movement, producing a thickened, distributed sense of time



# Healing Soundscapes – A New Gesamtkunstwerk

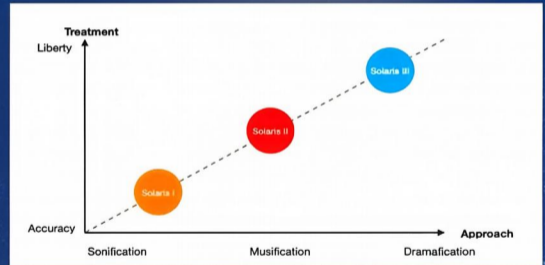
- **Interdisciplinary project: composition + neuroscience + design**
- **Immersive sound systems create zones of acoustic empathy that transform sterile hospital spaces into calming sonic environments**
- **A model for technologically embedded well-being**
- **AI generated musical structures (DJster, based on Clarence Barlow's probabilistic event generator Autobusk)**
- **Real-time notation for choir and organ via MaxScore DJster Scorepion (plugin)**



# Multimedia Performance: A Space Journey - Perspectives on the Unknown

## Solaris by Georg Hajdu

- **Collaboration between Hamburg University of Music and Drama and the Astrophysics Department of the Hamburg University**
- **Working groups consisting of 8 composers, theater directors and scientists each**
- **Sonification of astronomical data**
  - **Solaris: cosmic web**
- **Bohlen-Pierce Ensemble (11 players) and choir**
- **Empowerment of the choir by providing acoustic cues in addition to the animated score**



# Conclusion & thank you for listening

- **From work to system**
  - contemporary music shifts from composing objects to designing interactive, technological, and social systems
- **From fixed roles to fluid agency**
  - composer, performer, and audience merge into distributed roles
- **From space to networked/social environments**
  - concert hall expands into hybrid spaces: site-specific, networked, and data-driven contexts
- **From medium to infrastructure**
  - technology is no longer a tool, but the condition shaping perception, interaction, and creation
- **From autonomy to embeddedness**
  - music operates within policy, research, and innovation frameworks emphasizing impact, participation, and relevance
- **From division of labor to participation**
  - AI and technological systems reshape creative practice, expanding access while redefining artistic work



第三届  
The Third

# 世界音乐人工智能大会

Summit on Music Intelligence



## 机文主义：音乐学院的未来在哪里？

Machinism: Where Is the Future of Music Conservatories?

### 李小兵

Li Xiaobing

中央音乐学院音乐人工智能系主任

Director of the Dept. of Music AI, CCOM

中国人工智能学会艺术与AI专委会主任

Chair of the CAAI Art and AI Commission.

中国计算机学会计算艺术分会主任



中国人工智能学会  
Chinese Association for Artificial Intelligence



中央音乐学院  
CENTRAL CONSERVATORY  
OF MUSIC

Chair of CCF Computational Art Branch

SOMI  
2026

中国·北京  
2026  
04.25-26

# 人工智能将影响世界 AI Will Impact the World



1951年

“一旦机器思维方法开始，似乎很可能很快就会超越我们微弱的力量……因此，在某个阶段，我们应该期待机器能够被掌控。”

—— 艾伦·麦席森·图灵



人脑

140-160亿脑细胞

1000亿神经元

VS

数据

2020年GPT训练数据集  
已达48TB

算法

GPT-3参数量1750亿  
谷歌BERT参数量4810亿

算力

摩尔定律  
每18个月算力翻倍

存储

PTB级存储

无限迭代

7×24小时，全年无休

并行感知

多模态输入输出

形而上的意识形态：艺术、宗教、哲学……

人类的意义也将受重新被定位

人工智能或将实现共产主义

# 人工智能将影响世界 AI Will Impact the World

- 所有人的工作将会受到影响，所有学科都会受到影响（以音乐为例）
- Everyone's work will be affected, and all disciplines will be impacted (taking music as an example)



音乐表演（学科代码：1302Z1）

Music Performance



音乐学（学科代码：1302Z2）

Musicology



作曲与作曲技术理论（学科代码：1302Z3）

Music Composing Theory



音乐教育学（学科代码：1302Z4）

Music Education



音乐与科技（学科代码：1302Z5）

Music Technology



音乐管理与传播（学科代码：1302Z6）

Music Management and Communication



音乐表演（专业代码：130201）

Music Performance



音乐学（专业代码：130202）

Musicology



作曲与作曲技术理论（专业代码：130203）

Music Composing Theory



音乐教育（专业代码：130212T）

Music Education



音乐科技（专业代码：130215T）

Music Technology



流行音乐（专业代码：130209T）

Pop Music



音乐治疗（专业代码：130210T）

Music Therapy

研究生音乐（1352）一级学科的  
六个二级学科

本科音乐与舞蹈学类的七个专业

艺术，是人表达自身经验与情感的方式。

Art is a way for human beings to express their own experiences and emotions.

创作者，是唯一主体；

The creator is the sole subject;

意义，来源于人的内心世界；

Meaning arises from the inner world of the human mind;

作品，是表达的结果。

The work is the result of that expression.

# 新的现实

A new reality

人工智能已经可以写歌、作曲、生成图像，甚至参与表演。

AI can already write songs, compose music, generate images, and even participate in performances.

今天的艺术，还是“人表达自己”吗？

Is art today still “human self-expression” ?

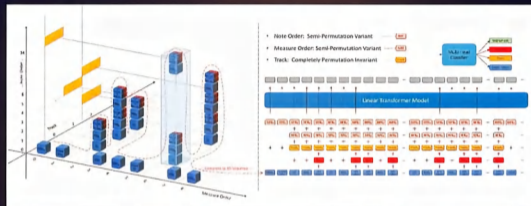
还是已经进入了一种新的生成结构？

Or has it already entered a new kind of generative structure?

- 中央音乐学院研发的歌声合成技术
- Singing Voice Synthesis Technology by CCOM



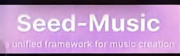
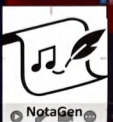
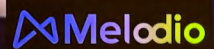
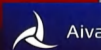
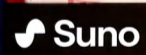
- 中央音乐学院研发的音乐大模型与交响乐生成
- Music LLM and Symphony Generation by CCOM



哲学断裂：  
艺术的主体

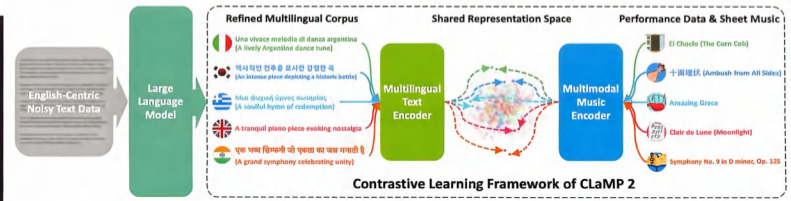
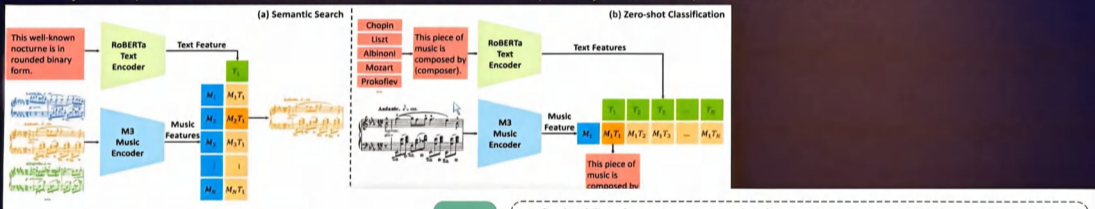
可能正在变得不再单一。

A philosophical rupture:  
The subject of art  
may no longer be singular.



# 人工智能音乐创作与分析 AI Music Composing and Analysis

- ❑ 音乐理解模型CLaMP获ISMIR 2023唯一“最佳学生论文奖”，CLaMP2被人工智能顶会NAACL接收CLaMP3被人工智能顶会ACL接收。
- ❑ The music understanding model CLaMP won the sole "Best Student Paper Award" at ISMIR 2023. CLaMP2 was accepted by the top AI conference NAACL, and CLaMP3 was accepted by another top AI conference, ACL.



J = 60  
III. Largo

Violino

g

*Andante con moto dolce, misterioso*

Flauto I & II  
Oboi I & II  
Clarinetto I & II  
Fagotto I & II  
Soprano Fagotto  
Corno in Fc, Hc, Bc  
Frottoia in A & C  
Tromba  
Tromba in Bc  
Tromba in Fc  
Violino I  
Violino II  
Viola  
Vcllo  
Violoncello  
Basso in Bc

# 人工智能音乐创作与分析 AI Music Composing and Analysis

- 音乐大模型与复调音乐生成，可能会影响未来的作曲四大件的教学。
- Music LLMs and polyphonic music generation may impact the teaching of the four core components of composition.



中央音乐学院  
CENTRAL CONSERVATORY  
OF MUSIC

复调改题系统

课程

您现在使用的是世界首个由AI驱动的复调音乐生成与教学辅助系统。该系统由中央音乐学院音乐人工智能系自主研发，嵌入了世界领先的复调音乐生成大模型。在这里，您仅需按下按钮，就能获取具有鲜明巴洛克风格且富于音乐性的复调音乐片段。同时，学生可以通过在线题库完成基础对位法的有效训练，实时提交给系统进行批改。这一辅助教学功能是从面向未来的角度，对复调教学方式的做出的积极探索。

立即体验

Composite  
Subject

复合主体

人

Human

AI

AI

系统

System

人，不再是唯一主体

The human is no longer the sole subject.

AI，不再只是工具

AI is no longer merely a tool.

系统，成为意义生成的结构性条件

The system becomes the structural condition for the generation of meaning.

创作，不再是“表达”，

Creation is no longer “expression,”

而是“对生成结构的调度”

but “the orchestration of generative structures.”

围绕这个问题，我提出了一个理论设想——  
机文主义。

Building on this question, I propose a theoretical hypothesis—  
Machinism.

我们面对的，不只是“AI能不能创作”，

What we are facing is not merely the question of whether AI can create,

而是创作这件事情本身，已经发生了结构性的变化。

but that the very structure of creation itself has already undergone a fundamental transformation.

# 机文主义 Machinism

## (1) 现实层面的机文主义

所谓机文主义，是指：  
在人的价值设定与伦理约束之下，  
将智能系统视为意义生产的协同结构，  
使其成为情感表达与审美组织能力的外延。

✘ 技术取代人 Technology replaces humans

✘ 机器凌驾人类 Machines dominate humanity

## (1) Machinism at the level of

reality

What is called Machinism refers to this:  
under human-defined values and ethical constraints,  
intelligent systems are regarded as a collaborative structure  
for the production of meaning,  
extending the capacities of emotional expression and  
aesthetic organization.

✔ 人机协同生成意义 Humans and machines  
collaboratively generate meaning

✔ 技术成为表达能力的放大器 Technology  
becomes an amplifier of expressive capacity

# 机文主义 Machinism

## (1) 现实层面的机文主义

如果说人文主义强调

🎯 “人是意义的唯一来源”

那么机文主义则强调

🎯 “意义可以通过人机协同生成”

在更大的理论框架中，我们甚至可以

把它理解为

🎯 一种“扩展人文主义”

## (1) Machinism at the level of reality

If humanism emphasizes

🎯 “humans are the sole source of meaning,”

then Machinism emphasizes

🎯 “meaning can be generated through human – machine collaboration.”

Within a broader theoretical framework, it

can even be understood as

🎯 a form of “extended humanism.”

# 机文主义 Machinism

## (2) 未来层面的机文主义（开放问题）

(2) Machinism at the level of the future (open questions)

如果机器未来具备认知能力，

If machines were to develop cognitive capacities in the future,

是否会出现一种以机器为主体的机文主义？

might a form of Machinism emerge in which the machine itself becomes the subject?

这不是结论，而是一个需要持续讨论的问题。

This is not a conclusion, but a question that calls for ongoing inquiry and discussion.

# 人工智能对音乐表演的影响

The impact of artificial intelligence on music performance

## □ 歌声合成 Singing Voice Synthesis

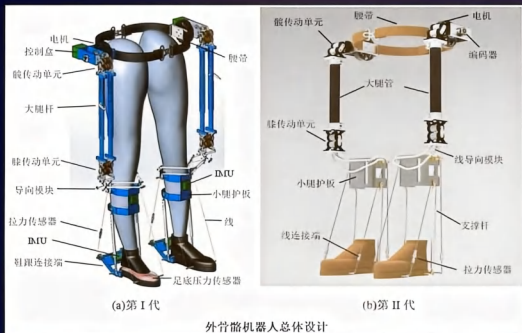
- 基于深度神经网络实现歌声合成，合成音质好、自然度高，同时支持云端合成和本地合成
- 最优质的歌声数据集，海量知名歌手的歌声采样



# 人工智能对音乐表演的影响

The impact of artificial intelligence on music performance

- Bixo——世界首创的双向线驱动下肢康复外骨骼机器人包揽4项国内顶尖创新创业赛事大奖。
- Bixo—the world's first bidirectional cable-driven lower-limb rehabilitation exoskeleton robot—has won four top domestic innovation and entrepreneurship competition awards.

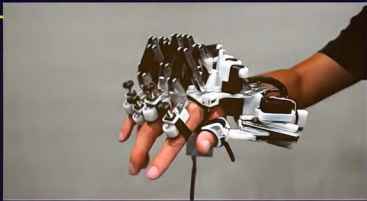


- “中美青年创客大赛”总决赛一等奖第一名（2023）
- “挑战杯科技作品赛”全国一等奖（2022）
- “挑战杯创业赛”全国金奖（2021）及银奖（2023）
- “互联网+”全国银奖（2023）

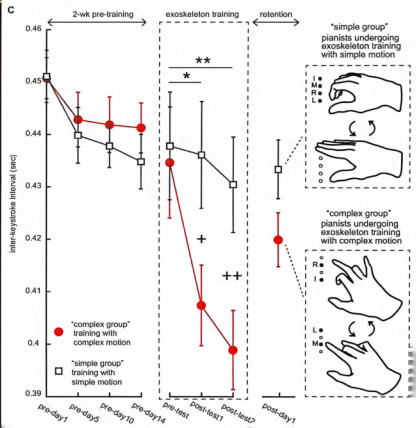
2022级音乐人工智能专业博士生王天乐入选第十八届全国大学生年度人物

# 人工智能辅助音乐表演训练

## AI-assisted Music Performance Training

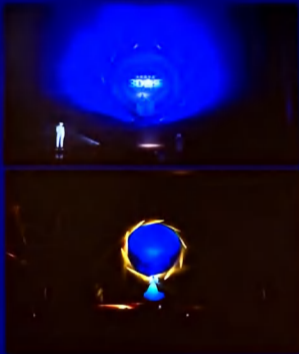


国际学术顶刊《科学·机器人》：使用一只手进行的外骨骼训练也提高了未经训练的对侧手的运动技能，体现了双手间的转移效应。



# 人工智能对音乐表演的影响

The impact of artificial intelligence on music performance



•3D音乐范式研究，建立3D音乐创作、制作和表演过程中，从科研到实践、从硬件到软件的范式与标准。

•Research on the 3D music paradigm: establishing paradigms and standards for 3D music creation, production, and performance—spanning from scientific research to practical application, and from hardware to software.

•音乐科技亮点主打原创3D音乐，搭配虚拟主持人，机器人，人工智能伴奏，5G+VR/AR共享平台，大型多媒体等诸多现代元素，融合中国民歌、戏曲等多种元素，结合计算机音乐、人工智能技术、新媒体艺术等创新性表达形式。

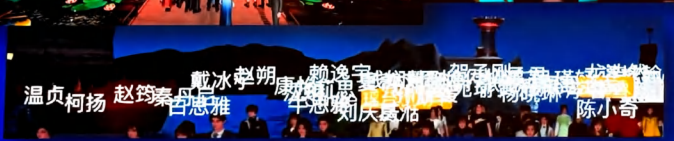
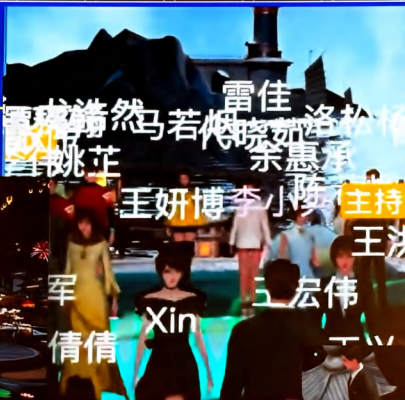
•The highlights of music technology focus on original 3D music, integrated with virtual hosts, robots, AI accompaniment, 5G + VR/AR shared platforms, and large-scale multimedia. These modern elements are combined with Chinese folk songs, opera, and other traditional forms, and further enriched through innovative modes of expression such as computer music, artificial intelligence technologies, and new media art.

“未来音乐会” 3D音乐  
“Future Concert” 3D Music

# 人工智能音乐表演 AI Music Performing

□ 元宇宙音乐会

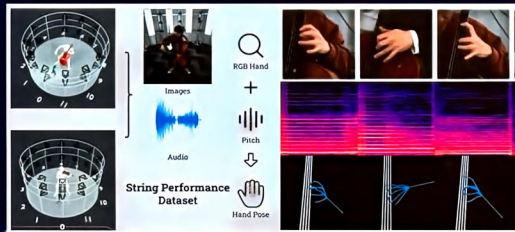
□ Metaverse Concert



# 人工智能对音乐教育的影响

The impact of artificial intelligence on music education

中央音乐学院研发弦乐动作捕捉技术及数据集，被计算机图形学“奥斯卡”之称的SIGGRAPH2024、ACM Trans. on Graphics接收。在情感计算、演奏家风格研究等领域应用广泛。



# 人工智能对音乐教育的影响

The impact of artificial intelligence on music education

- 针对音频演绎的富表现力大提琴演奏动作生成框架 ELGAR再次被SIGGRAPH接收，基于扩散模型，仅依赖音频即可生成全身精细化乐器演奏动作的最新方法。



在动画制作、表演教学等应用前景广泛。

It has broad application prospects in areas such as animation production and performance pedagogy.

# 人工智能对音乐教育的影响

The impact of artificial intelligence on music education

## □ 视唱练耳教学与训练系统 Solfeggio Teaching and Training System



智能识别歌唱者的音高、节奏、歌唱速度、演唱技法  
Intelligently identifies a singer's pitch, rhythm, tempo, and vocal techniques

实现对视唱者速度、音高及节奏方面的追踪和评价  
enabling real-time tracking and evaluation of a sight-singer's speed, pitch accuracy, and rhythmic performance.

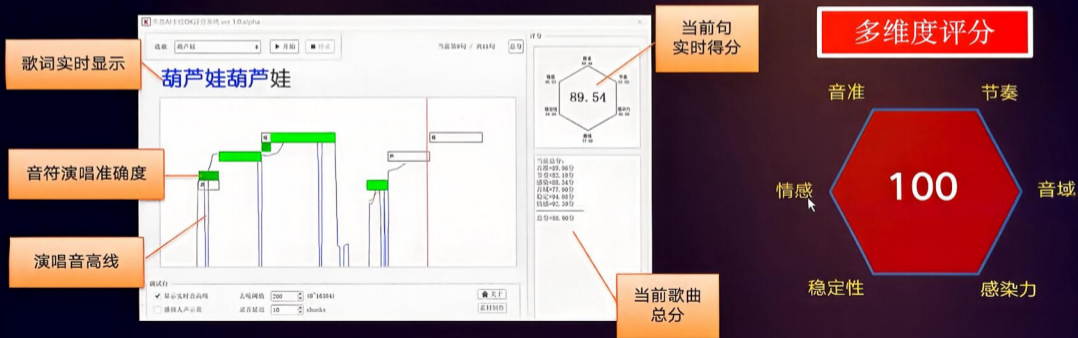
## □ 智能出题系统已经在管弦系招生考试中使用4年。

□ An intelligent question-generation system has already been used in orchestral department entrance examinations for four years.

# 人工智能对音乐教育的影响

The impact of artificial intelligence on music education

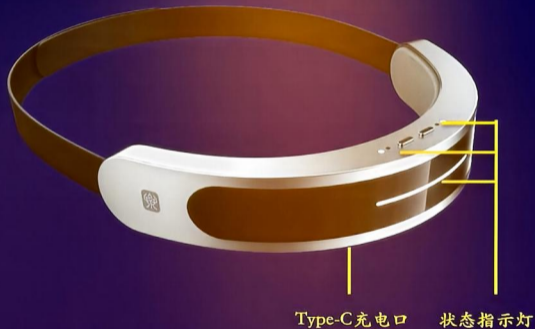
- 卡拉OK评分 (CCTV音乐频道《渴望现场》合作课题)
- AI Karaoke Scoring (In collaboration with CCTV)



# 开放的问题：脑科学、人工智能对音乐治疗、音乐创作、音乐教育、音乐表演、音乐学等的影响

An open question: The impact of brain science and AI on music therapy, music creation, music education, music performing and musicology.

蓝牙键 电源键



## 音乐脑机接口（央音一号）

Music Brain - Computer Interface (CCOM-1)

# 开放的问题：脑科学、人工智能对音乐治疗、音乐创作、音乐教育、音乐表演、音乐学等的影响

An open question: The impact of brain science and AI on music therapy, music creation, music education, music performing and musicology.



对话界面

理解—评估—处方—记录（开放环）



实验现场

因人选曲与处方播放

语音交互：语音识别与自然语言生成

情绪识别：文本/语音情感分析与状态评估

音乐生成/选择：大模型驱动的曲目生成与推荐

策略反馈：根据用户主观反馈滚动优化对话与音乐

# 未来发展方向

Future directions

央音二号：缓解疼痛、睡眠、情绪



央音三号：青少年发育、注意力



终极目标：提高身心健康，延长健康寿命

# 音乐人工智能哲学与美学

## AI Music Philosophy and Aesthetics

### □ 音乐人工智能美学问题 AI Music Aesthetics



音乐人工智能与“艺术感知”



音乐人工智能与审美



音乐人工智能的评价体系



“新文艺复兴”？

### □ 音乐人工智能的伦理问题 AI Music Ethics



音乐人工智能的“生命及文化经验”



音乐家身份的问题



音乐人工智能作品的版权归属、音乐数据合规性

未来创作权可能不仅仅是版权那么简单，不是一种“AI盗版人类”，也不是“人类VS AI”，可能法律界人士、伦理方面的专家得用更高的哲学维度重新定义、规划人类版权的未来。

# 音乐人工智能哲学与美学

## AI Music Philosophy and Aesthetics

- 最近经常看到一些专家的言论，有说AI音乐没有感情，有说AI音乐都是同质化的音乐，有说AI音乐都是基于数据库东拼西凑出来的，完全没有个性...甚至下结论说AI音乐是负面的？
- 我这里不做对与错的评价，时间会给予正确的回答，但我对这些专家的说法有些担忧。如果少数专家对某些事情在没有深入了解前，就急于下结论，有可能会让错误的信息影响大众。
- 人类的作品和AI作品一样，是不是也存在大量没有感情、没有个性、同质化的音乐作品呢。音乐大模型是学习了大量的数据进行预测，和人类学习后的创作过程很像，而且大模型的泛化能力也会产生非常创新的音乐。
- 人类无法区别AI，目前的技术手段下，机器也不能识别AI。未来AI加速发展，更加无法鉴别。鉴别AI可能是一个伪命题，从更大的哲学观来说，可能完全没有必要。

# 音乐人工智能哲学与美学

## AI Music Philosophy and Aesthetics

- 另外，人类在自然、在宇宙面前是比较渺小的，我们也说不清楚大模型里面是怎么盲算出来的结果，这跟我们人类是怎么创作出来的结果是一样的，人类自己也说不清楚，因为就面前全世界人类对大脑的认知，可能知道的还超不过5%，更谈不上解释清楚。
- 讲一些题外话，至于AI音乐有没有感情，会不会感动人？我觉得没有争论的必要。动画片《哪吒》也不是真人演的，也一样票房超过真人演的电影，一样让很多人感动。我认为，人类未来要用更大的哲学观思考AI音乐的一切。

# 新型音乐工作方式也将改变音乐学院的未来

Emerging modes of musical practice are poised to transform the future of music conservatories

“新质生产力”正深刻重塑文化与艺术的生产方式。音乐作为高度依赖情感表达、创意、技术与协作的综合性艺术，正面临可能被AI驱动的结构性的转变。

## 核心公式

新质生产力 · 音乐  $AI \times OPC = OPMC$

针对“音乐  $AI \times OPC$ ”这一新型组合，我们提出

One Person Music Company (OPMC)

探讨人工智能如何赋能超级个体，推动音乐产业组织形态的根本变革。

关键词：新型音乐生产组织

# 音乐学院怎么面向未来

How Music Conservatories Can Prepare for the Future

---

第一、传统意义上的音乐专业将会继续发展；

First, traditional music disciplines will continue to evolve.

第二、科技赋能传统意义的音乐专业创新发展；

Second, technological empowerment will drive innovation within traditional music disciplines.

第三、新科技产生新的音乐专业独立发展。

Third, emerging technologies will give rise to new and independent music disciplines.

# AI不再是工具，它开始陪你理解“活着”这件事

AI is no longer just a tool. It is beginning to accompany us in understanding what it means to be alive.

- ❑ AI正从工具变为伙伴，未来它将会成为世界的一部分，有些人可能不喜欢它的变化，但技术不会因为你的不喜欢而停止。
  - ❑ AI的风格越来越像人，但人类还没准备好如何与他共创这个时代。
  - ❑ 最革命性的事情可能还没开始。
  - ❑ 虽然目前的人工智能还存在幻觉，但是我们要心存敬畏、重新思考人类对自己的定义，拥抱人工智能。
1. 学习问题：基础教学（小学、初中、高中）是否还需要12年，有可能未来体育和美育更重要。
  2. 就业问题：旧金山街上的黄色出租车大部分是无人驾驶，只有很少的老人。AI交作业：美国教授对AI交作业问题的反应。
- 都是我们人类需要面对的问题。

2025年9月，一则新闻在全球头条和社交媒体刷屏：“世界三大顶级高级翻译学院之一、被称为“翻译界哈佛”的明德大学蒙特雷国际研究学院（MIIS），宣布2027年6月停招其唯一的研究生”。

# 音乐学院怎么面向未来

## How Music Conservatories Can Prepare for the Future

---

1. 守住本源，坚守音乐的艺术性和人文关怀。

Stay grounded in the essence of music, upholding its artistic integrity and humanistic values.

2. 重点培养我们的审美意识、情感表达能力和社会责任感。

Place emphasis on cultivating aesthetic awareness, emotional expressivity, and a strong sense of social responsibility.

3. 面对未来，不要害怕，要多关注、思考人工智能技术的最新发展动态和对未来音乐发展的影响带来的机遇和挑战，是自己成为适应未来发展的优秀音乐家，要和人工智能交朋友，驾驭她。

In facing the future, do not be afraid. Actively engage with and reflect on the latest developments in artificial intelligence and their impact on music, both the opportunities and the challenges. Strive to become a musician who can adapt to the future: befriend AI and learn to harness it.

人工智能是帮助人类、赋能人类的，不是代替人类的。

AI is meant to assist and empower humans, not replace them.

**未来的音乐学院将站在AI巨人肩上前行!**

The music conservatories of the future will advance on the shoulders of AI giants!

**人类未来的艺术将站在AI巨人肩上前行!**

The future of human art will advance on the shoulders of AI giants!



中央音乐学院  
CENTRAL CONSERVATORY  
OF MUSIC

# Khala: 统一声学 Token 空间的音乐生成大模型从深层表征到高质量生成

Khala: Large-Scale Music Generation Model with a Unified Acoustic Token Space: From Deep Representations to High-Quality Generation

# 为什么高质量音乐生成很难?

Why is high-quality music generation so difficult?



## Musicality

音乐性



Melody

旋律



Rhythm

节奏



Long-range Form

段落与结构组织



Lyric / Prompt Consistency

歌词与提示词的一致性



## Fidelity

音质



Natural Timbre

自然的音色



Fine Detail

丰富的细节



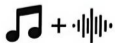
Clarity

清晰度



Realistic Vocal / Instrument Quality

真实的人声与乐器质量

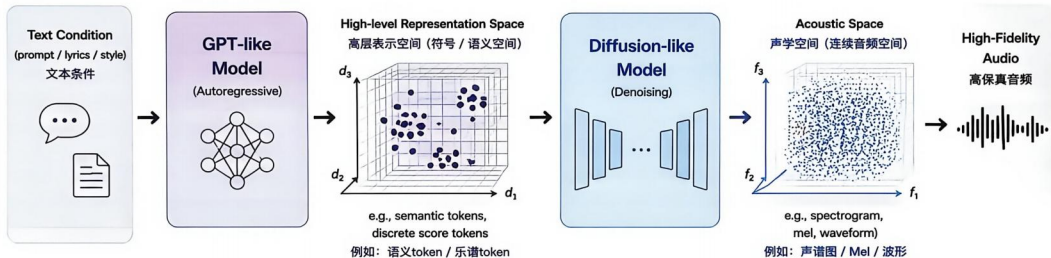


**Both must hold at the same time.**

两者必须同时成立，缺一不可。

# 主流方案：离散空间 vs 连续空间

## The Mainstream Paradigm: Discrete vs Continues



### The Cost 代价



**Different spaces**  
(semantic vs. acoustic)  
不同表示空间  
(语义空间 vs. 声学空间)



**Different model modules**  
(GPT vs. Diffusion)  
不同模型模块  
(GPT vs. Diffusion)



**Longer training & inference pipeline**  
训练和推理链路更长



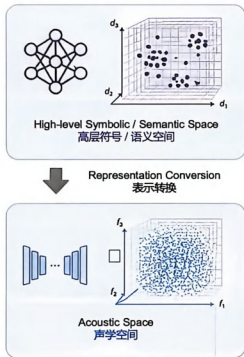
**Information loss & harder alignment**  
信息损失，结构与声学  
对齐更难

# 我们想探索的核心问题是什么？

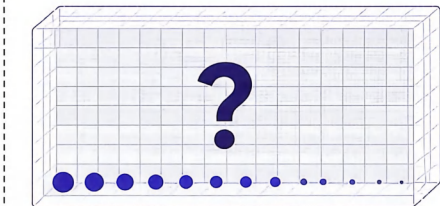
What is the core problem we are looking to investigate?

Mainstream: Two Different Spaces

主流方法：两个完全不同的空间



Can we generate music within one acoustic token hierarchy?  
能不能始终在同一个声学 Token 空间里完成音乐生成？



Coarse Structure  
粗粒度结构

Fine Detail  
精细声学细节



One Representation  
统一的表示



Hierarchical Tokens  
层级化 Token



Progressive Generation  
递进 / 渐进式生成

Why Explore This?  
为什么要探索这条路？



More Unified System  
系统更统一



Less Information Gap  
系信息损失更小

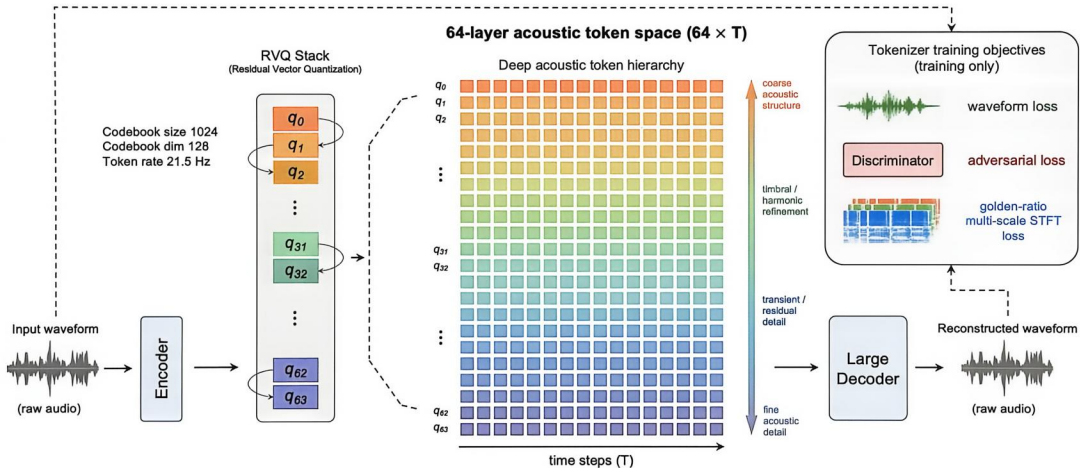


Easier Scaling Path  
更清晰的扩展路径



Better Alignment Potential  
更好的对齐潜力

# 64 层声学 Token 架构

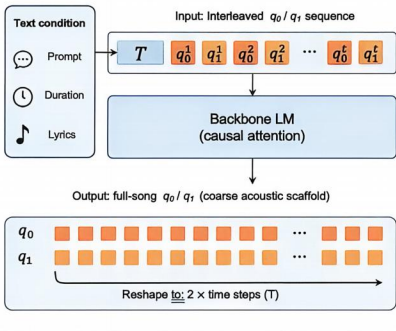


# 两阶段声学 Token 生成

## Two-Stage Acoustic Token Generation

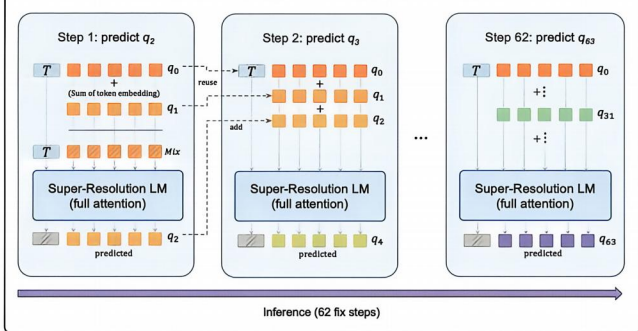
### Backbone Model (Stage 1)

Autoregressive generation of coarse acoustic scaffold



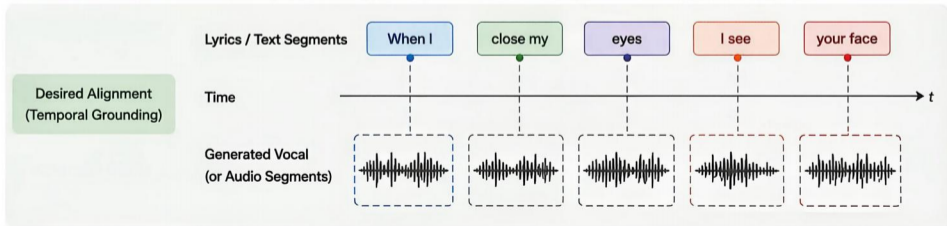
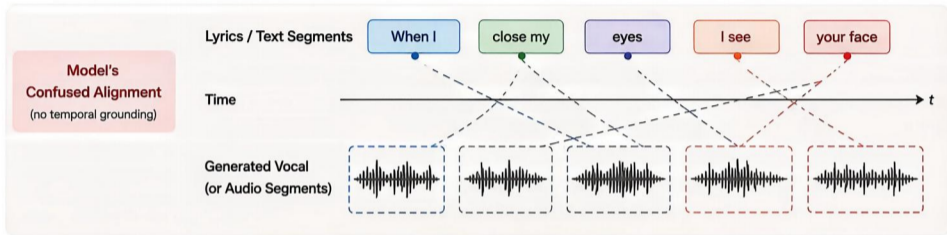
### Super-Resolution Model (Stage 2)

Layer-wise refinement to higher acoustic layers (parallel across time)



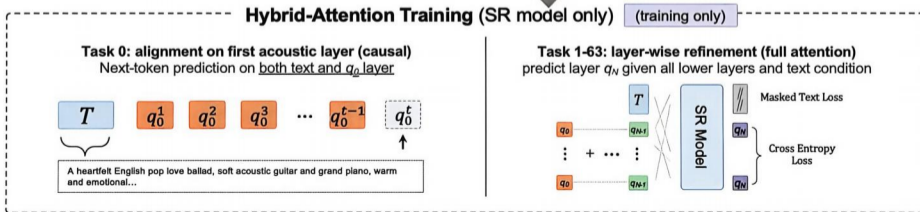
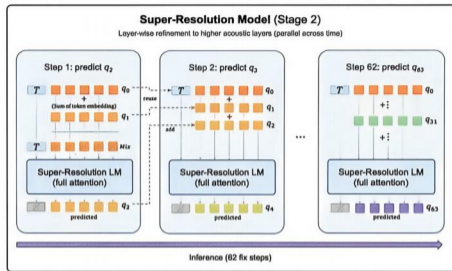
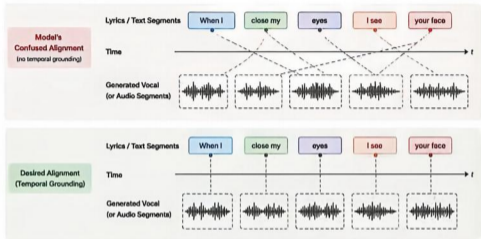
# 核心发现：时序绑定

## The Key Insight: Temporal Grounding



# 核心发现：时序绑定

## The Key Insight: Temporal Grounding



### Controls

Settings · Sampling · Generate

⋮ Advanced

#### PROMPT MODE

Natural

Tags

Pick tags; prompt is assembled automatically.

#### MODE

Vocal

Instrumental

Lyrics section is visible.

#### DURATION

Common presets

3 min

2m

3m

4m

5m

8m

10m

Generate

Reset

Output defaults to 2 samples

### Prompt

Tag builder

#### Tag builder

Pick tags; we join them directly.

Copy

Clear

#### ASSEMBLED PROMPT

Total tags: 6

Mandopop, Female Vocals, Acoustic Guitar, Piano, Atmospheric, Modern Production

#### CATEGORIES

Genre

1

Style backbone

Sound

3

Vocals & Instruments

Mood

1

Emotion & vibe

Prod

1

Rhythm & production

Search in Rhythm/Production/Other...

No matches.

Keep 3-8 tags for clean prompts.

Back to top

Clear Prod

#### SELECTED

Mandopop X

Female Vocals X

Acoustic Guitar X

Piano X

Atmospheric X

Modern Production X

### Tag builder

Link tags; we join them directly.



Copy



Clear

### ASSEMBLED PROMPT

Total tags: 8

Mandopop, Female Vocals, Acoustic Guitar, Piano, Atmospheric, Modern Production

### CATEGORIES

Genre 1

Style backbone

Sound 3

Vocals & instruments

Mood 1

Emotion & vibe

Prod 1

Rhythm & production

Search in Rhythm/Production/Other...

No matches.

Keep 3-8 tags for clean prompts.

Back to top

Clear Prod

### SELECTED

Mandopop X

Female Vocals X

Acoustic Guitar X

Piano X

Atmospheric X

Modern Production X

### Lyrics

Auto-grow → then scroll

#### Verse 1]

I used to walk alone in the dark  
Searching for a light to guide my heart  
Every path I took just led to nowhere  
Until I looked up and saw you standing there

#### [Chorus]

You are the sun breaking through the rain  
You are the melody that heals my pain  
When the world is spinning out of control  
You are the safe place for my soul  
Yeah, you are the one I've been waiting for

#### [Verse 2]

I trace the lines upon your hand  
You make me feel like a better man (如果是女声可  
You make me finally understand)  
We don't need a castle or a crown  
Just a quiet place to settle down

#### [Chorus]

You are the sun breaking through the rain  
You are the melody that heals my pain  
When the world is spinning out of control  
You are the safe place for my soul  
Yeah, you are the one I've been waiting for

# 人类天梯对战盲评 Large-Scale Bline Pairwise Evaluation (1v1)



Human Arena Leaderboard

Rank	Model	BT Elo	Camp
1	Muraka	1689.3	Commercial
2	Suno	1644.2	Commercial
3	<b>Khala(ours)</b>	1510.9	Open-source
4	Minimax	1509.2	Commercial
5	Ace-step	1470.9	Open-source
6	Heartmula	1421.8	Open-source
7	Levo	1173.1	Open-source
...	...	...	...



**#1** among  
**open-source models**  
in blind pairwise evaluation (1v1)



Khala achieves state-of-the-art performance among open-source models, and is highly competitive with commercial systems.

# 人类天梯对战盲评 Large-Scale Bline Pairwise Evaluation (1v1)



Human Arena Leaderboard

Rank	Model	BT Elo	Camp
1	Muraka	1689.3	Commercial
2	Suno	1644.2	Commercial
3	<b>Khala(ours)</b>	1510.9	Open-source
4	Minimax	1509.2	Commercial
5	Ace-step	1470.9	Open-source
6	Heartmula	1421.8	Open-source
7	Levo	1173.1	Open-source
...	...	...	...



**#1** among  
**open-source models**  
in blind pairwise evaluation (1v1)



Khala achieves state-of-the-art performance among open-source models, and is highly competitive with commercial systems.

# 音乐脑机接口：概念、研究进展与应用前景

## Music Brain-Computer Interfaces: Concepts, Research Progress, and Application Prospects

马 军

中央音乐学院 音乐人工智能与音乐信息科技系

# 为什么今天要讨论音乐脑机接口？

音乐人工智能的“下一步”，不是更强的内容生成，而是更好地理解人、响应人、调节人。

## 趋势一 | 音乐AI快速发展

- 生成、推荐、教育和交互场景持续扩展
- 新挑战：从“生成好音乐”走向“可控、实时、个性化”

## 趋势二 | 脑机接口走向状态感知

- 非侵入式 EEG / fNIRS 等降低使用门槛
- 情感BCI关注情绪、疲劳、认知负荷等状态识别

## 趋势三 | 音乐进入反馈闭环

- 音乐神经反馈强调把音乐放进调节回路
- 从单向播放走向“状态—音乐—反馈—再调节”

### 过去的音乐AI主要回答

“这是什么音乐？” “如何生成一段音乐？” “用户可能喜欢什么？”

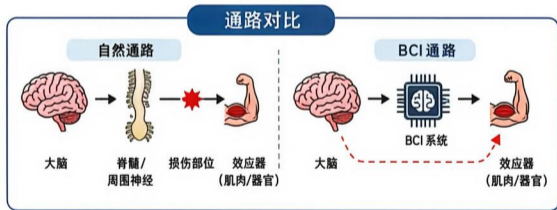
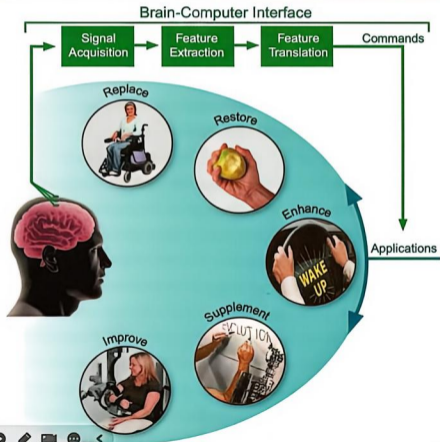
### 音乐脑机接口进一步回答

“用户此刻处于什么状态？” “什么音乐此刻最合适？” “如何动态调整？”

音乐脑机接口是音乐人工智能由“内容智能”迈向“人本交互与闭环调节”的关键入口

# 什么是脑机接口 (BCI) ?

**定义:** 脑机接口 (Brain - Computer Interface, BCI) 是一个通过检测中枢神经系统活动, 并将其转化为人工输出来代替、恢复、增强、补充或改善中枢神经系统正常输出, 由此改变中枢神经系统与内外环境之间持续交互作用的系统



Wolpaw & Wolpaw, *Brain-Computer Interfaces*, 2012

# 什么是音乐脑机接口?

**定义:** 音乐脑机接口是一种通过大脑信号与音乐进行交互的脑机接口 (BCI) 系统, 它能用于音乐创作、演奏、治疗和情感调节。该技术通过读取用户的大脑信号, 将其转换为音乐指令, 或根据大脑信号反馈来控制外部设备和音乐内容。



音乐脑机接口是脑机接口在音乐交互、音乐反馈与音乐干预场景中的一个特定分支。

Miranda & Castet, *Guide to BCMI*, 2014; 钟慧等, 2022

# 音乐脑机接口是一个多学科交叉方向

音乐脑机接口不是单一学科的自然延伸，而是围绕“音乐如何进入人机闭环”形成的交叉领域。

## 音乐与计算音乐学

定义音乐结构、风格、声学参数与可计算表征。

## 音乐治疗 / 康复科学

定义干预目标、适应证与疗效评价。

## 信号处理 / 模式识别

从噪声脑电中提取可靠特征。

## 音乐脑机接口

Music × AI × BCI

## 脑科学与BCI

解释音乐感知、注意、情绪与节律的脑机制，并提供信号通路。

## 生物医学工程 / 可穿戴技术

保证采集硬件、舒适性与稳定性。

## 人机交互 / 音频工程

把反馈设计为可理解、可接受、可持续的体验。

## 人工智能与情绪计算

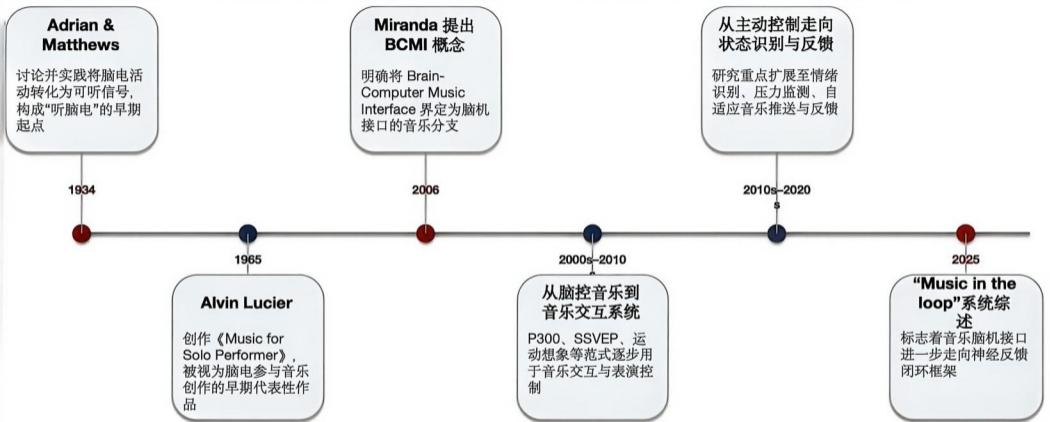
完成状态识别、多模态融合、个体建模与音乐决策。

## 数字疗法 / 智能健康

连接真实场景、风险分级与长期应用。

核心问题：什么样的音乐在什么样的状态下对什么样的人有效？

# 音乐脑机接口的发展历史



总体上，音乐脑机接口的发展经历了：脑电声音化 → 脑控音乐交互 → 状态识别与音乐闭环反馈的演进过程。

# 音乐脑机接口发展：1. 脑电声音化与脑控音乐

## 一、核心流程



关键点：脑电波本身并不是“音乐”，而是先被提取为可计算特征，再转换成可听的声音或音乐结构。

## 二、常见映射关系

脑电特征	可映射音乐参数	常见表达效果
频率	音高	高频→高音；低频→低音
振幅	音量	振幅大→声音更强
节律变化	节奏/密度	波动越快→节奏越密
复杂度	音色/和声张力	复杂度高→更丰富或更紧张

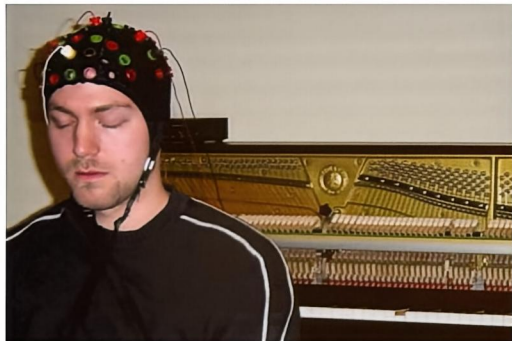
## 三、这种方式的特点

**直接声音化** 把脑电特征直接变成声音，是最早期、最直观的脑波音乐路径

**规则映射** 通过预设规则把脑电映射为音高、节奏、音色等音乐要素

**不等于“读心作曲”** 它更多是“参数控制”或“声音化”，并不能直接读取完整音乐思想

## 音乐脑机接口发展：2. 从脑电声音化到脑控演奏



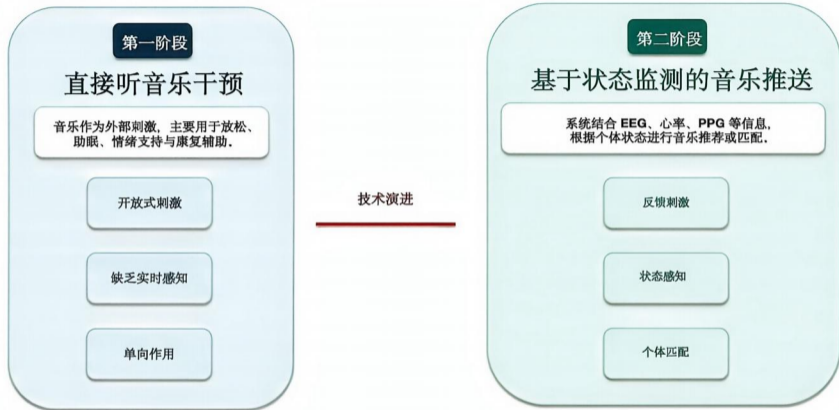
脑信号驱动的音乐交互与表演

### BCMI-Piano 的核心逻辑



这一路径标志着 **BCMI** 从“声音化”进入“可理解的音乐交互与演奏”。

## 音乐脑机接口发展：3.基于状态识别的音乐反馈



发展逻辑：开放式音乐刺激 → 状态监测后的音乐推送 → (下一步) 实时闭环个性化干预

## 音乐脑机接口发展：4. 实时个性化音乐闭环反馈干预



音乐不再是一次性播放内容，而是可实时调整的干预变量

# 音乐脑机接口推动音乐AI走向以人为本的闭环智能

核心判断：BCMI 让音乐AI从“会生成内容”走向“会理解状态、会反馈决策、会动态调节”。

## 1 | 从内容智能到人本智能

过去侧重

生成 / 检索 /  
推荐音乐

BCMI推动

识别情绪、注意、  
压力、唤醒等状态

## 2 | 从单向输出到闭环反馈

过去侧重

一次性播放  
或推荐

BCMI推动

根据脑与生理反应  
持续更新音乐策略

## 3 | 从作品生成到干预变量

过去侧重

音乐作为作品  
或内容

BCMI推动

速度、节奏、音色、  
和声成为可调节变量

更高水平的音乐AI，不只是“生成什么音乐”，而是“在合适的时机，用合适的音乐，作用于合适的人”。

### 情绪调节与压力支持

基于情绪/压力识别进行音乐反馈，目标是改善唤醒与主观体验。

### 睡眠支持与节律调节

从助眠音乐走向脑状态感知和闭环听觉刺激，是最具现实需求的方向之一。

### 康复训练与特殊人群支持

残障辅助、职业治疗、音乐治疗与神经康复场景，为特殊需要群体提供新的参与方式。

### 教育、创作与智能交互

脑控表演、沉浸式交互、音乐学习与创作辅助，让音乐接口进入更广泛的人机交互场景。

# “央音一号”：音乐脑机接口的本土探索案例

政策与场景背景：脑机接口正在从实验室研究走向技术突破、应用场景建设和产业生态培育。



央音一号：音乐脑机接口实践案例

系统路径：感知—评估—反馈—验证

1 多模态感知  
EEG / PPG / 心率 / 呼吸等  
捕捉个体脑与生理状态

2 AI状态评估  
情绪、压力、睡眠、唤醒等  
形成个体状态画像

3 音乐反馈设计  
音乐选择、生成、参数调节  
把音乐作为可设计变量

4 闭环效果验证  
主观量表 + 生理指标 + 行为表现  
持续更新干预策略

从“音乐内容提供者”到“音乐干预变量设计者”

本土探索的价值

- 把音乐学院优势转化为“音乐变量设计”能力。
- 从脑控演奏拓展到情绪支持、睡眠调节等真实场景。
- 连接音乐AI、脑科学、音乐治疗与智能健康。

需要强调的边界

- 当前更适合作为研究平台和案例说明。
- 临床和健康场景仍需规范验证。

# 谢谢！

音乐从内容走向接口，从播放走向调节，从表达走向闭环交互。

From content to interface, from playback to regulation, from expression to closed-loop interaction.

Thank you



第三届  
The Third

# 世界音乐人工智能大会

Summit on Music Intelligence



## 面向 AI 生成音乐工作流的 Web DAW

A Web-based DAW for AI-generated Music Workflow

卢迪

Ma Jun

中央音乐学院 助理研究员

Assistant Researcher, Central Conservatory of Music

somi  
2026

中国·北京  
2026.04.25-26



中国人工智能学会  
Chinese Association for Artificial Intelligence



中央音乐学院  
CENTRAL CONSERVATORY  
OF MUSIC

# 一. 研究背景

- AI音乐生成领域快速发展，但可编辑性不足
- 音乐生成算法

- 音频生成

提示词



音频生成模型



音乐音频

- 😊 高音质
- 😞 难以编辑

- 符号生成

提示词



符号生成模型



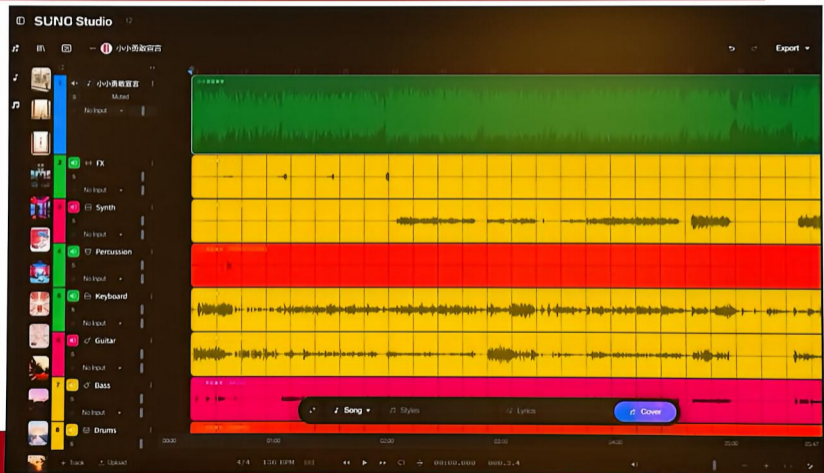
曲谱 (MIDI等)

- 😊 可编辑
- 😞 试听音质差

演奏、虚拟乐器  
混音、母带...

## 二. 相关研究 (1/3)

- 现有方案不足
  - 音频生成



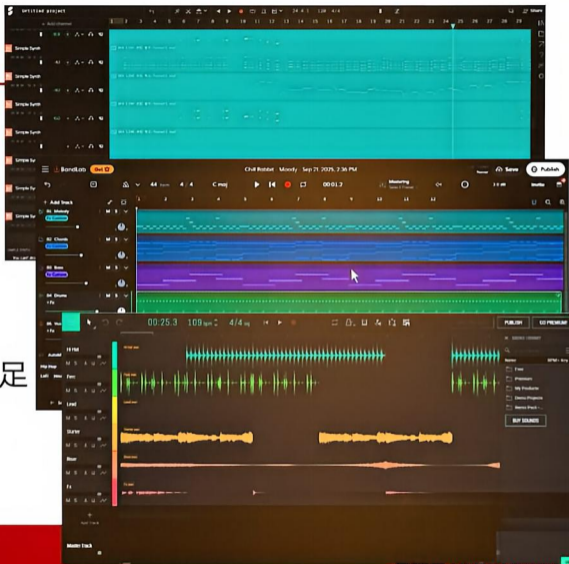
## 二. 相关研究 (2/3)

- 现有方案不足
  - 符号生成



## 二. 相关研究 (3/3)

- 现有方案不足
  - 网页端DAW
    - 商业DAW不提供API
    - 开源DAW效果受限
    - 结构、和声的编辑支持不足





## 二. 本研究的贡献

---



- 提出适用于AI音乐生成工作流的Web DAW（网页端数字音频工作站）
  - 可视化界面：支持音符、和声、曲式结构等音乐元素的直观操作
  - 实时试听：浏览器内实时渲染、48kHz/16bit，低延迟
  - 标准化API：与符号音乐生成模型（或系统）无缝对接

# 三. 系统架构


## • 分层-模块化设计

可视化界面   
用户交互 

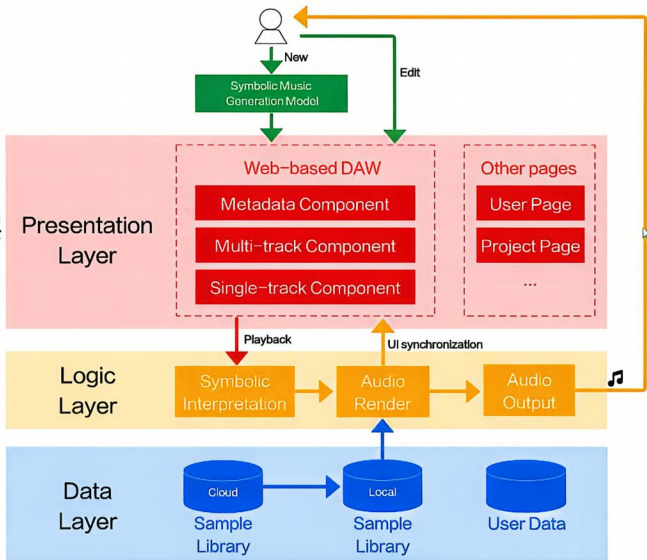
表现层

音色渲染   
音频输出 

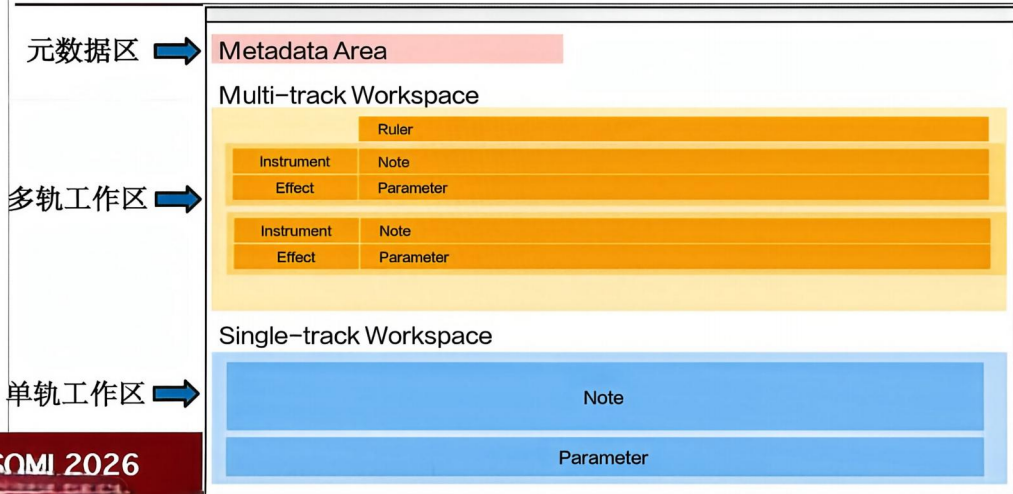
逻辑层

乐器音色采样 

数据层



### 三. 系统架构 — 表现层

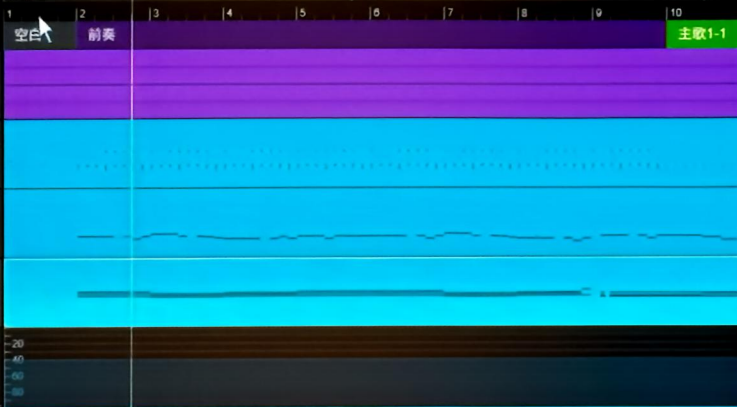


Second Bpm Beat  
00:03:708 114.00 4/4

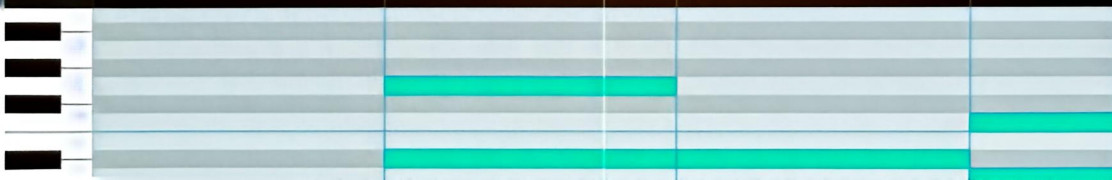
TRK BUS



00:00 | 00:10 | 00:20

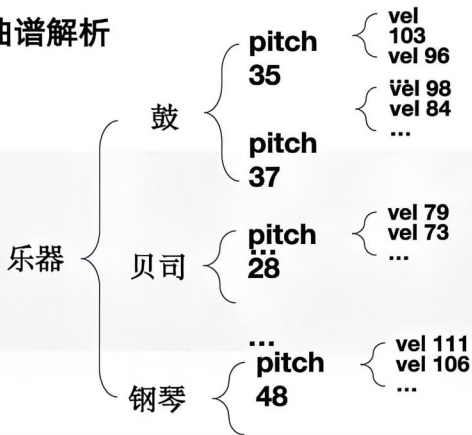


调	bE大调			
和声	bE	bE	bA	bB

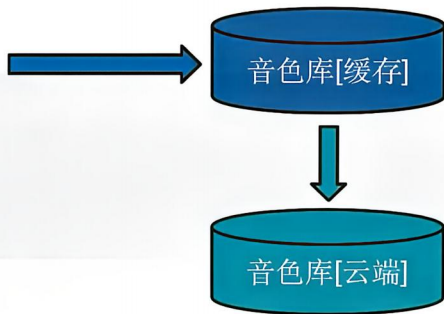


### 三. 系统架构 — 逻辑层 (1/2)

#### (1) 曲谱解析

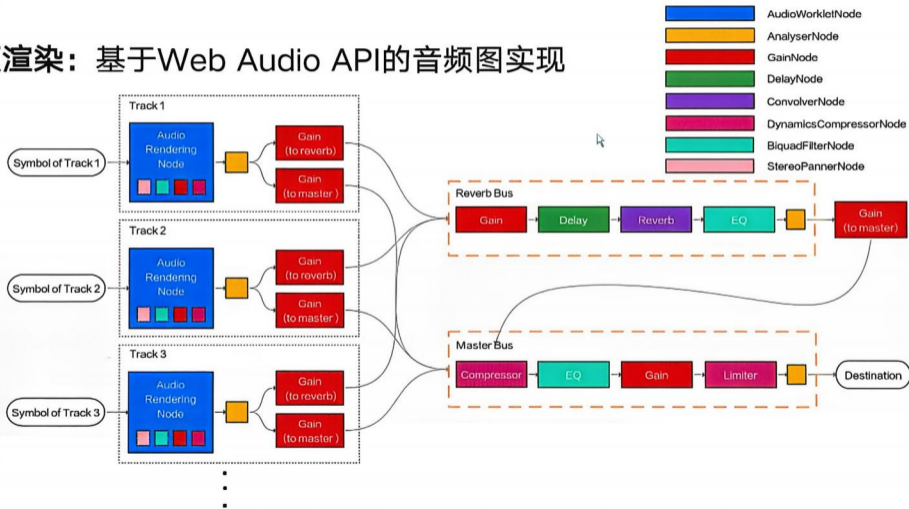


#### (2) 加载音色



# 三. 系统架构 — 逻辑层

## (3) 音频渲染：基于Web Audio API的音频图实现



## 三. 系统架构 — 数据层

---

- 数据层:
  - 采用GM2.0标准作为基础
  - 128种标准乐器音色+8种中国民族乐器+8套打击乐组
  - 每种乐器:
    - 音高按半音阶间隔采样
    - 动态层级 (4-8层) 覆盖力度范围 (ppp/pp/p/mp/mf/f/ff/fff)
    - 样本时长包含短音、中长音和可循环音

## 四. 系统整合

---

- 适用于AI符号音乐生成模型的API接口：

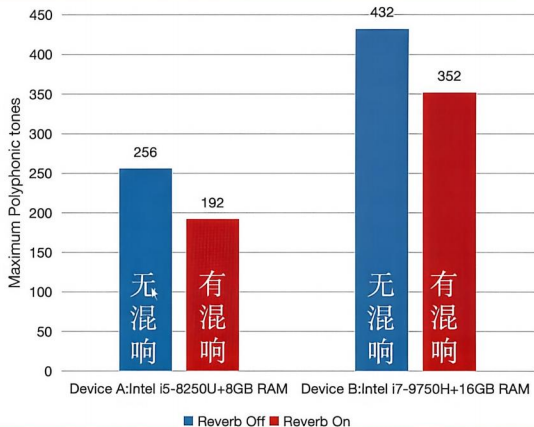
1	载入曲谱	/api/ai/import	POST
2	重新生成指定小节	/api/ai/regenerate_bar	POST
3	重新生成指定乐段	/api/ai/regenerate_section	POST
4	重新生成指定音轨	/api/ai/regenerate_track	POST

## 五. 实验与评估

- 性能测试：可同时播放的复音数

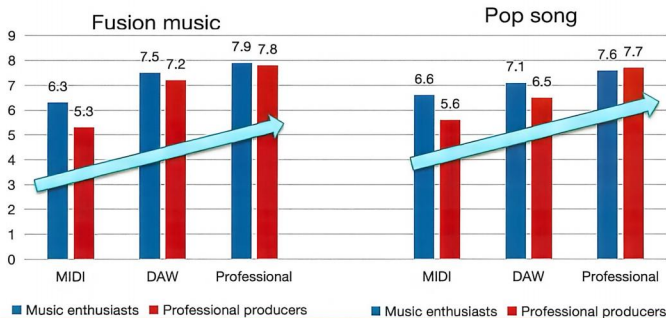
### 实验结果

- 至少支持192复音（带混响）



## 五. 实验与评估

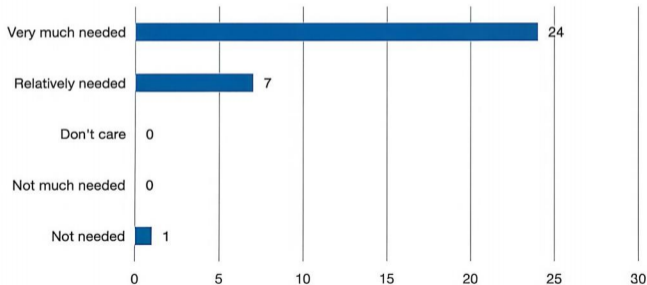
- 主观音质评估 by 16名普通爱好者 + 16名专业音乐人



## 五. 实验与评估

- 主观评估

- 用户需求：97%用户愿意使用该DAW



## 六. 总结与展望

---

- 总结：
  - 面向AI音乐生成工作流的可编辑Web DAW
  - 通过API整合AI符号音乐生成模型
  - 实现“AI生成-用户编辑-实时回放”创作闭环
- 未来工作：
  - 性能优化与跨平台适配
  - AI符号音乐生成系统整合优化
  - WebDAW插件标准兼容

第三届  
The Third

# 世界音乐人工智能大会

Summit on Music Intelligence



面向网络化电子音乐合奏的 AI 编程助手的实践应用

Practical Applications of AI Coding Assistants for Networked  
Electronic Music Ensembles

肯尼斯·菲尔兹

Kenneth Fields

中国科学院大学 教授

Professor, University of Chinese Academy of Sciences

SOMI  
2026

中国·北京  
2026 04.25-26



中国人工智能学会  
Chinese Association for Artificial Intelligence



中央音乐学院  
CENTRAL CONSERVATORY  
OF MUSIC



# Networked China University Laptop Ensembles

UCAS, CCOM, PKU and WHCM

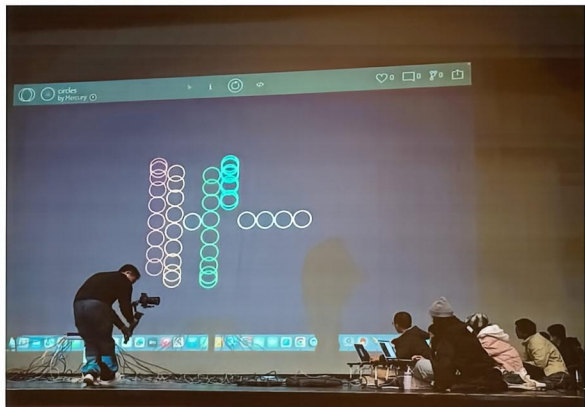
A project involving networked electronic music ensembles:

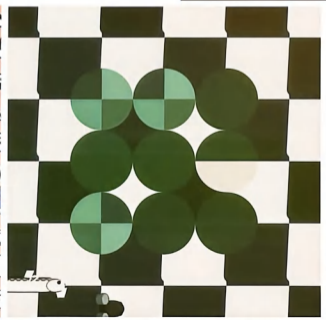
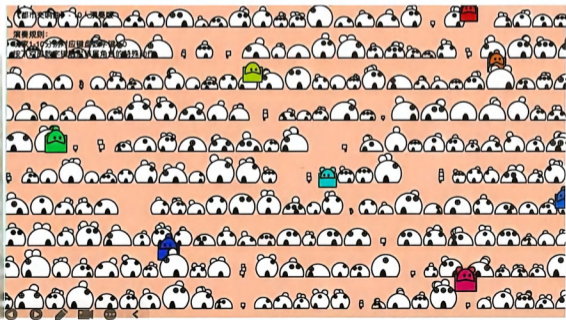
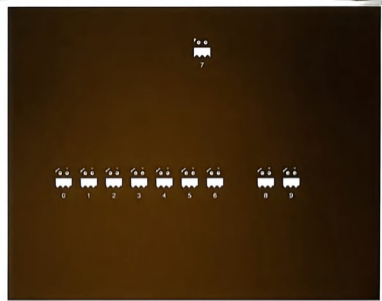
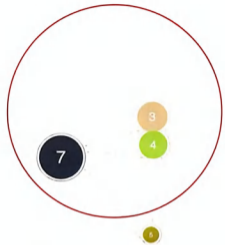
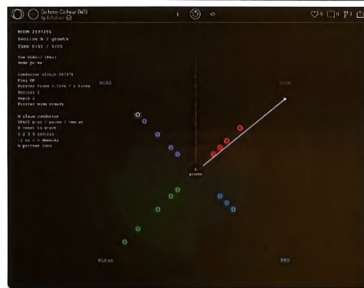
UCAS, CCOM, PKU and WHCM campuses

Using China's high-speed CERNET2 network,

Introduction of AI chatbot coding assistants.

- Animated graphic scores coded in Java p5js
- Openprocessing.org
- Socket.io library
- Supercollider Audio Programming Environment
- TouchOSC: Open Sound Control





[+ Compose](#)

active jacktrip

exited jack

active **jacktrip**
[Edit](#)
[Restart](#)
[Update](#)
[Stop](#)
[Delete](#)

## Containers

### jacktrip

jacktrip/jacktrip: sha-2d409d4

running
4464
61000-61100

&gt;\_ Bash

## Terminal



## compose.yaml

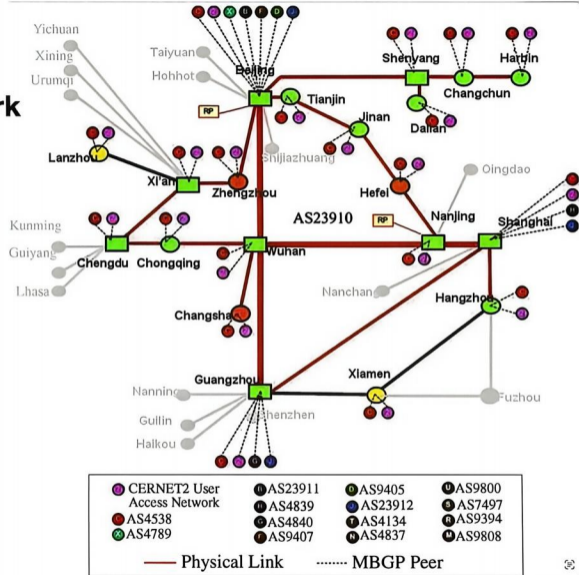
```

1 services:
2   jacktrip:
3     image: jacktrip/jacktrip:sha-2d409d4
4     container_name: jacktrip2
5     environment:
6       - JACKTRIP_OPTS="-V -b 24 -n 2 -S -t -z --
7 hubpatch 2 --bufstrategy 3 -q 24"
8       - JACK_OPTS="-d dummy -C 0 -P 0 --rate 48000
9 --period 128"
10    ports:
11      - 4464:4464/tcp
12      - 61000-61100:61000-61100/udp
13    shm_size: 128M
14    privileged: true
15    restart: unless-stopped
networks: {}
    
```

# CERNET2

## China Education Research Network

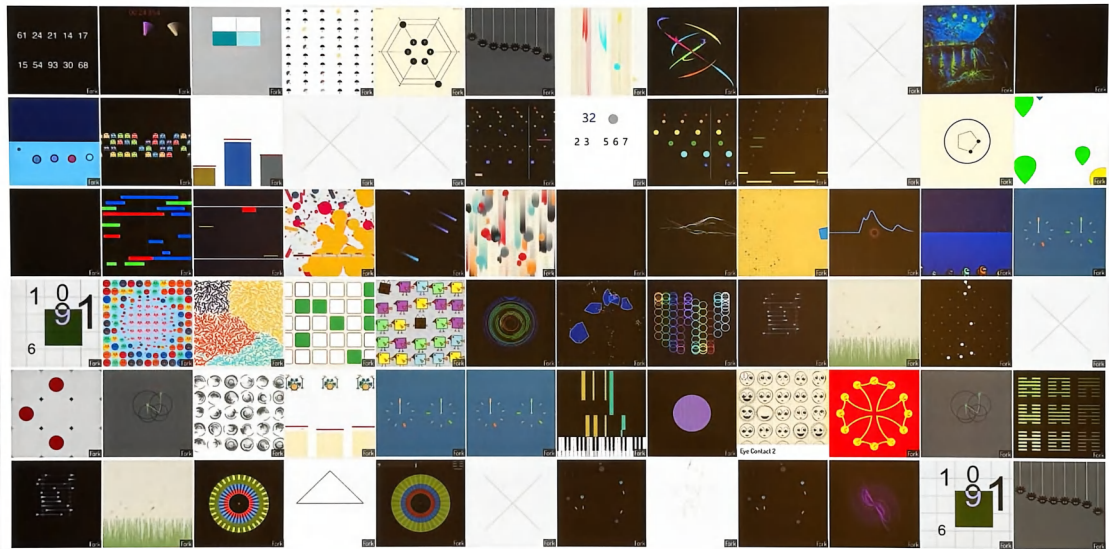
- Audio Signal Path
- IPV6
- Expanding each year
- 4-5 nodes currently
- 48k/24bit Multichannel
- Part of Global Edu Network



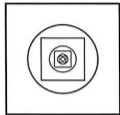
# **The Urklangwerk, Primal Sound-Work**

*The archetypal sound-work containing all possible sound-work forms*

- In our application of the Goethe Formula to organised sound, we observe the multiple instances of sonic potentiae as evidenced in the variation encoded into the latent space of the animated score.
- The variations reflect the multiple facets of the one holistic form underlying all the sound-works.



# Christopher Alexander 15 Archetypes of Architectural Form: The Nature of Order, Vols. 1-IV



Levels of Scale



Strong Centers



Thick Boundaries



Alternating Repetition



Positive Space



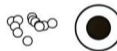
Good Shape



Local Symmetries



Deep Interlock



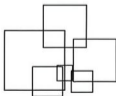
Contrast



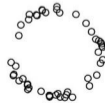
Gradients



Roughness



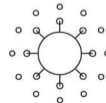
Echoes



The Void



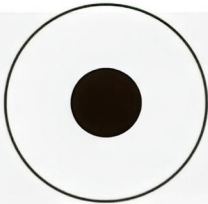
Simplicity



Not-Separateness

# ***Sonus/Soni* - Sonic Center**

Spatial form is order in extension; temporal form is order in succession



**Strong Centers**



Property	Character	Pinyin
1. Levels of Scale	层	Céng
2. Strong Centers	中 or 声中	Zhōng
3. Thick Boundaries	口	Kǒu
4. Alternating Repetition	叠	Dié
5. Positive Space	凹	Āo
6. Good Shape	凸	Tū
7. Local Symmetries	双	Shuāng
8. Deep Interlock	互	Hù
9. Contrast	明	Míng
10. Gradients	川	Chuān
11. Roughness	丑	Chǒu
12. Echoes	从	Cóng
13. The Void	间	Jiān
14. Simplicity	一	Yī
15. Not-Separateness	合	Hé

# Post AI ChatBot

After AI chatbots were introduced, such as ChatGPT, Deepseek, Gemini, etc., EA music courses have been transformed, enabling music and tech focused programs to once again focus essentially on musical (organized sound) and animated visual music representation. This is a significant, even revolutionary progression, where form, timbre and control can be led directly by the creative imagination rather than being sidetracked by the narrow gateway of programming syntax.

# Phenomenologies of Electroacoustic Music and the Sonic Arts

Coordinated Issue of OS and EMS Conference Theme



Call for Submissions – Volume 33, Number 1

Thematic Issue Title:

Phenomenologies of Electroacoustic Music and the Sonic Arts

Date of Publication: April 2028

Publisher: Cambridge University Press

Issue co-ordinators: Blake Stevens ([blakecstevens@pku.edu.cn](mailto:blakecstevens@pku.edu.cn))

and Annie Yen-Ling Liu ([yliuannie@pku.edu.cn](mailto:yliuannie@pku.edu.cn))

**Deadline for submission: 15 May 2027**

**Electroacoustic Music Studies Network**

**2027 EMS Conference, 24-27 June**

**Beijing, Peking University**

**Annie Yen-Ling Liu**

**Blake Stevens**

第三届  
The Third

# 世界音乐人工智能大会

Summit on Music Intelligence



## 表演科学的未来

The Future of Performance Science

## 亚伦·威廉姆森

Aaron Williamon

英国皇家音乐学院表演科学中心主任

Director, Centre for Performance Science, Royal College of Music (UK)

SOMI  
2026

中国·北京  
2026 04.25-26



中国人工智能学会  
Chinese Association for Artificial Intelligence



中央音乐学院  
CENTRAL CONSERVATORY  
OF MUSIC



# The Future of Performance Science

Aaron Williamon

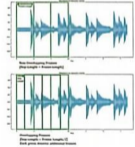
Centre for Performance Science

A partnership of the

Royal College of Music | Imperial College London



Performance science is the systematic study of *how, what, when, where* and *why* people perform, as well as who performs and the consequences that performances hold.



2000

RCM  
2000

🔍 Zoning In: Motivating the Musical  
Leverhulme Trust, 2000–02



BMus  
2000

🔍 Mastery through Imitation  
Swiss National Science Foundation, 2001–02

🔍 Mapping Musical Movement  
British Academy, 2004–05

Musical Excellence 📖  
Oxford University Press, 2004

20 05

PhD  
2006


International Symposium on  
Performance Science 🧠  
2007

🔍 The Listening Gallery  
Arts & Humanities Research Council, 2008–10

2009


2010




The Practice of Practising   
Leuven, 2011

2010  
MSc  
2011

 Rhythm for Life  
Esmée Fairbairn Foundation, 2010–12

 Creative Practice as Mutual Recovery  
Arts & Humanities Research Council, 2013–18

Frontiers | Performance Science   
2014

 Musical Impact  
Arts & Humanities Research Council, 2013–18

 Performance Simulator  
Peter Sowerby Foundation, 2014–15

Healthy Conservatoires   
2015

Imperial  
2015

 Music and Motherhood  
Arts Council England, 2015–17

Musicians in the Making   
Oxford University Press, 2017

 TELMI  
Horizon 2020, 2016–19

 HEartS  
Arts & Humanities Research Council, 2018–21

 ProMiMiC  
Dutch Research Council, 2018–21

 ZikMus  
British Academy, 2019–21

2019

2020

Performing Music Research 

Oxford University Press, 2021

Collaborative Insights 

Oxford University Press, 2022

Musical Care International Network 

2022

Music and Parental Wellbeing Alliance 

2022

Music and Parental Mental Wellbeing 

Oxford University Press, 2024

20 20

 HEartS Professional  
Arts & Humanities Research Council, 2020–22

 Songs from Home  
Arts & Humanities Research Council, 2021–22

 Arts in the City  
British Academy, 2022–23

 RCM Performance Lab  
Arts & Humanities Research Council, 2022–23

 RCM Performance VR  
Arts & Humanities Research Council, 2023–24

 HEartS Follow On  
Arts & Humanities Research Council, 2024–25

20 25



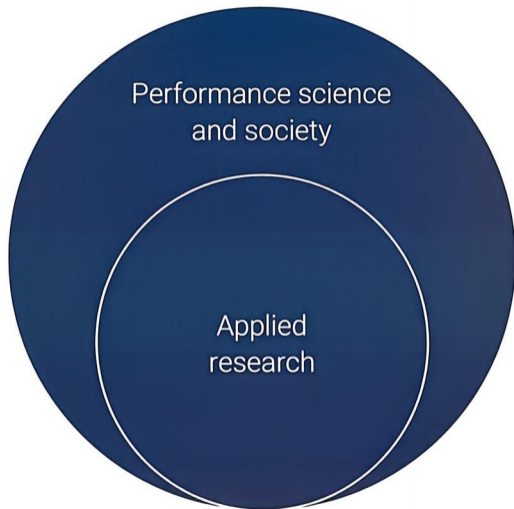
Aaron Williamon

2025



Aaron Williamon

Why?





Aaron Williamon





Aaron Williamon

# MUSICAL IMPACT

[www.musicalimpact.org](http://www.musicalimpact.org)



Aaron Williamon



# HEarts

Health, Economic, and  
Social impact of the ARTs

[www.PerformanceScience.ac.uk/hearts](http://www.PerformanceScience.ac.uk/hearts)

What?

## RESEARCH THEMES

Experiential learning

Performers' health & wellbeing

Music in health & care

Others in development...

Aaron Williamon



# How?

Radical interdisciplinarity

Co-creation

Analytics

International alliances

Policy, knowledge exchange  
and commercialisation

第三届  
The Third

# 世界音乐人工智能大会

Summit on Music Intelligence



确保音乐领域人工智能政策的包容性与可持续性：  
一个治理议题

Ensuring Inclusive and Sustainable AI Policy for The Music Sector: A  
Governance Issue

凯特·霍普

Cat Hope

澳大利亚莫纳什大学 教授

Professor, Monash University (Australia)

somi  
2026

中国·北京  
04.25-26



中国人工智能学会  
Chinese Association for Artificial Intelligence



中央音乐学院  
CENTRAL CONSERVATORY  
OF MUSIC

# ENSURING INCLUSIVE AND SUSTAINABLE AI POLICY FOR THE MUSIC SECTOR: A GOVERNANCE ISSUE

PROFESSOR CAT HOPE PhD  
Monash University, Australia.

# Building AI to Simulate the World

Get Started >

RUNWAY CHARACTERS

MEDIA AND ENTERTAINMENT

ROBOTICS AND AUTONOMY

GENERAL WORLD MODELS

MPREIA  
LOWHNEY

## Chat GPT: AI Chat Online

ChatGPT is an AI-powered language model developed by OpenAI. It has been trained on a massive amount of text data from the internet and can generate human-like text responses to a given prompt. It can answer questions, converse on a variety of topics, and generate creative writing pieces.

Open Chat

Image Generator

**Transversal**  
from EC Foresights**Territorial dimensions**  
From local to international**Digital space**  
AI, Metaverse, hybrid, etc.**Rights**  
Human rights, transparency, liability,  
etc.**Resource-  
consciousness**  
Water, biodiversity, energy, etc.**Culture & media  
domains**  
EU framework**Art**  
Visual, performing, etc.**Culture and heritage**  
Tangible, intangible**Creative industries**  
Incl. audiovisual**Media & news**  
Journalism, editors, etc.**EU cultural policy**  
Context**Supranationality**  
more than the sum of each entity**Subsidiarity & multilevel  
governance**  
Cooperation frameworks**Transversality**  
Culture is a part of all**Public dimensions**  
including offer & demand side

1. Conditional permission to train
2. A creative sovereignty levy
3. Provenance and transparency requirements:.
4. A purpose-based creative sovereignty trust
5. Independent audit and verification

Javanshir, Sackely & Irvine, 2026, p4

## WEB SCIENCE INSTITUTE

### Creative Sovereignty as National Security: A Governance Framework for Generative AI

WSI Position Paper 2026-02  
March 2026

Matt Javanshir, Alistair Sackley, Thomas Irvine



第三届  
The Third

# 世界音乐人工智能大会

Summit on Music Intelligence



## 写歌，正在从创作变成选择吗？

Is Songwriting Becoming Selection Rather Than Creation?

### 徐帆

Xu Fan

Suno公司软件工程师

Software Engineer, Suno Inc.

somi  
2026

中国·北京  
2026.04.25-26



中国人工智能学会  
Chinese Association for Artificial Intelligence



中央音乐学院  
CENTRAL CONSERVATORY  
OF MUSIC



# 写歌：从创作到选择？


SONGWRITING: FROM CREATION TO SELECTION?

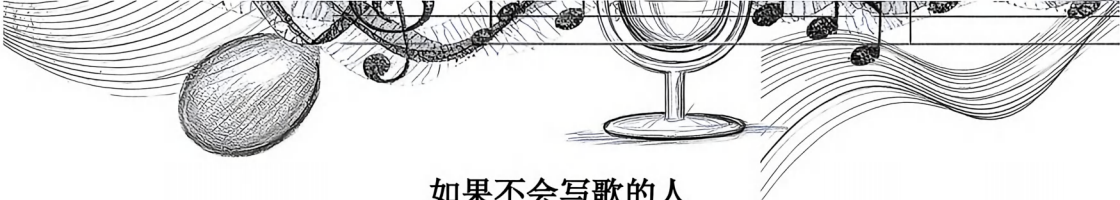
AI音乐生成与创作方式的变化

AI-Generated Music and the Changing Nature of Music Creation

徐帆 (Suno)

Fan Xu (Suno)



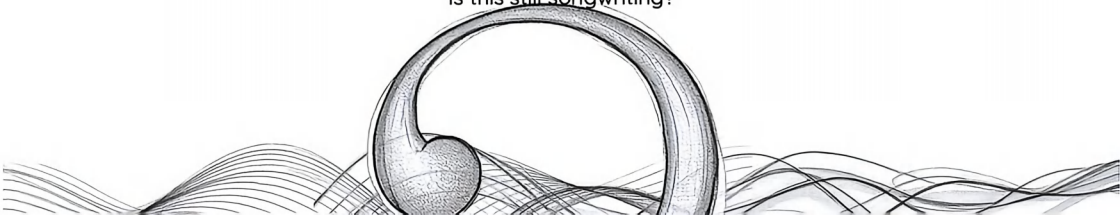


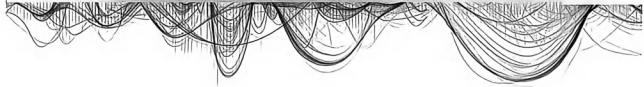
如果不会写歌的人  
也能做出一首歌

IF PEOPLE WHO DON'T KNOW HOW TO WRITE SONGS CAN  
CREATE A SONG

还是创作吗？

Is this still songwriting?





# 现在AI可以做什么？

WHAT CAN AI DO TODAY?

文字 → 歌曲的快速生成

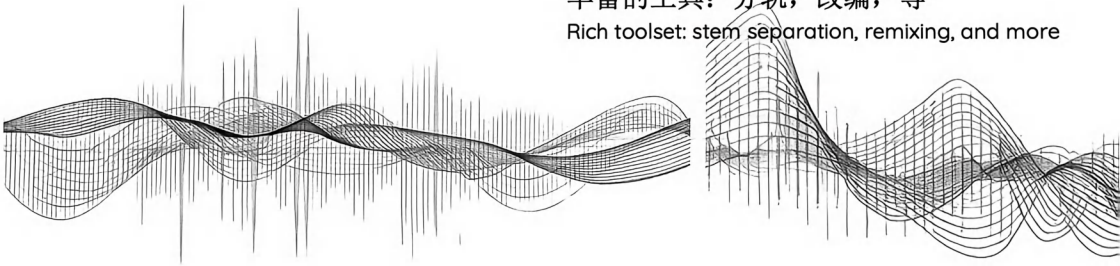
Fast generation from text to full songs


个性化的曲风，人声

Personalized styles and vocals

丰富的工具：分轨，改编，等

Rich toolset: stem separation, remixing, and more



The background features several hand-drawn sketches. In the top left, there's a sketch of a building's corner. In the top center, musical notes are scattered along a curved line. On the right side, a vertical sketch of a smartphone is visible. In the bottom left, there's a sketch of a fan-like structure. In the bottom right, there's a sketch of a bar chart with two bars of different heights. The overall style is artistic and sketchy.

# AI音乐已经进入现实世界

AI MUSIC HAS ENTERED THE REAL WORLD

随着AI歌曲在短视频平台和音乐流媒体快速传播，AI生成的音乐内容正在以惊人的速度获得关注和分享。

As AI songs rapidly spread across short-form video platforms and music streaming services, they are gaining attention and shares at an unprecedented pace.

AI生成音乐已进入各大榜单，AI音乐应用下载量快速增长，越来越多的普通用户开始使用AI工具创作属于自己的音乐作品。

AI music is now appearing on major charts, and downloads of AI music apps are growing quickly. More and more everyday users are beginning to use AI tools to create their own music.




# 创作的改变

## CHANGES IN CREATION

生成和选择的快速迭代，加快了歌曲demo的创作周期，也降低了难度  
Rapid iteration between generation and selection accelerates the demo creation cycle and lowers the barrier to entry

创作入门级的音乐变得更加容易，大量个人用户成为音乐创作者，主要为个人或短视频写歌

Entry-level music creation is becoming much easier, enabling a large number of individuals to become creators—primarily making music for personal use or short-form video content





# 创作的改变

CHANGES IN CREATION

经典的音乐市场受到冲击

Disruption to the Traditional Music Market

- 快速的创作节奏，冲击了初级音乐人的工作

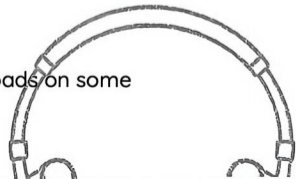
Faster creation cycles are putting pressure on entry-level musicians

- 同时对音乐制作人的品味和市场嗅觉提出了更高的要求

At the same time, higher demands are being placed on producers' taste and market intuition

- 音乐分发平台显示AI音乐占上传量约44%（保守估计）

AI-generated music accounts for approximately 44% of uploads on some distribution platforms (conservative estimate)



# 听众的改变

## CHANGES IN LISTENER BEHAVIOR

为个人而做的歌，用户更加关注歌词本身的意义。

As more music is created for individuals, listeners are placing greater emphasis on the meaning of lyrics

音乐流媒体平台AI音乐播放只占1-3%（美国）

On music streaming platforms, AI-generated music currently accounts for only about 1-3% of total plays (U.S.)

短视频平台AI音乐逐渐增加（快速实验，版权需求）

On short-form video platforms, AI music is steadily increasing—driven by rapid experimentation and demand for flexible, low-friction content



# 音乐在教育中的改变

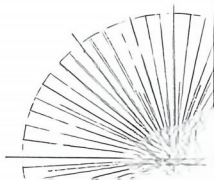
CHANGES IN MUSIC IN EDUCATION

Suno进入学生音乐课堂

Suno is being introduced into music classrooms

歌曲作为老师帮助学生记忆知识的手段

Songs are used as a tool for teachers to help students memorize and learn through music



创作→选择?

CREATION→SELECTION?

创作>选择

CREATION>SELECTION

第三届  
The Third

# 世界音乐人工智能大会

Summit on Music Intelligence



## 推动开源音乐生成的边界

Pushing the Boundaries of Open-Source Music Generation

龚俊民

Gong Junmin

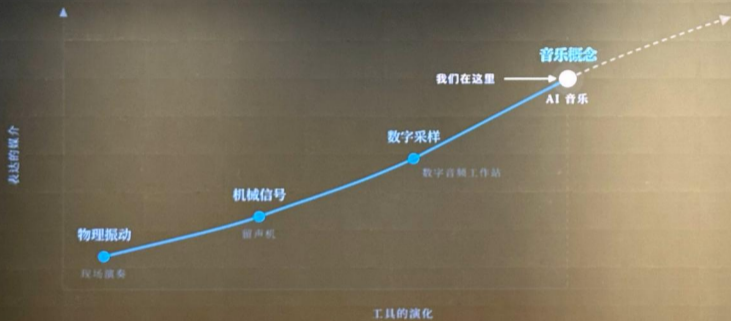
ACE Studio 合伙人

Partner, ACE Studio

# 推动开源音乐生成的边界

PUSHING THE BOUNDARIES OF OPEN-SOURCE MUSIC GENERATION

ACE Studio · ACE-Step · 龚俊民



AI 降低创作门槛，真正有音乐素养的人反而更重要了。

这个行业已经有了非常优秀的开创者们。

致敬。

我们想补上另一块拼图。

# 我们是 ACE Studio。

做了 5 年音乐人工具——从歌声合成，到一整套创作生态。



AI Singer



AI Instruments



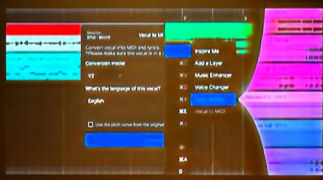
Music Enhancer



Add Layer



Music Composer



More Tools

# ACE-STEP

一个开源的音乐生成模型

## MODEL

**4B 参数 · v1.5 XL**

Diffusion + 规划 LM

## LICENSE

**MIT**

开源 · 商用免费

## RUNTIME

**RTX 3090 · 16 GB**

消费级 GPU 即可本地运行

## INPUT

**文字 · 标签 · 参考音频**

多模态提示

## OUTPUT

**完整歌曲**


含人声 · 多语言

## WHERE

**GitHub · HuggingFace**

社区微调生态

CO-BUILT BY

ACE Studio  StepFun

3 - 4 人 · 主体研发 · 两年

## 三代架构, 一个向上的梯度

### 1.0

概念验证

文 · 音

DiT · 2D mel latent

基于 tags 生成

抽卡随机  
低辨音质

### 1.5 · XL

基本可用

文 · 音 · 音

**LN** + DiT · 1D VAE latent

基于详细描述生成

音频规划生成

上下文感知

分轨生成

专业知识  
高辨音质

### 2.0

世界一流  
专业做深

文 · 音 · 音

乐谱 / 合辑 / 规划 · 音

MUSIC AGENTS

世界知识

上下文编辑

studio 音质

# 1.0

2025 · 05

概念验证  
暴露问题

## 不是语音合成

NOT SPEECH SYNTHESIS

### 更像文生艺术画

CLOSER TO TEXT-TO-ART

人声是乐器  
不是必须

### 主观

SUBJECTIVE

没有统一指标  
专业评估昂贵

### 不要用粗 tags

一对多，太依赖生成能力

### 不要用 mel

压缩过头，不利于高保真重建

### 不要做人声伴奏分离

分类先验不对

# 1.5

2026 · B2 → B4

基本可用  
逐项修正

## 用语言模型桥接意图

LANGUAGE MODELS BRIDGE INTENT

1.0 给出经验判断, 1.5 把它们变成实践和解法。

### 先救音质

FIX FIDELITY

waveform VAE

### 细粒度标注

FINER LABELS

style / instrument  
mood / form

### 先思考规划

PLAN FIRST

language models  
bridge intent

### THREE DON'TS

#### 不要用纯自回归

数据依赖强 · prefix bias 重

#### 不要让扩散只做渲染

drop 70% semantic codes

#### 不要让歌词主导风格

补纯音乐数据 · 独立 dropout

# 从全局描述，到音乐蓝图

FROM GLOBAL DESCRIPTION TO MUSIC BLUEPRINT

简单标签

中文民谣流行

男声

温暖

怀旧

木吉他

1.5 · 详细描述

这是一首温暖怀旧的中文民谣流行歌，整体带一点复古感。前奏由指弹木吉他和口琴进入，主歌里曼陀林在人声句间点缀，手风琴与低音提琴持续托住温暖底色。到了后半段，刷镲鼓轻轻加入，把情绪从讲述感推向更开阔的抒情。男声用温暖略沙哑的中音区演唱，整体像一间小酒馆里自然收录下来的现场。



# 2.0

NOW

世界一流  
专业做深

## 世界一流，专业做深

WORLD CLASS, GO DEEPER

### 音质

QUALITY

高频 / 动态  
真实乐器纹理

### 可控

CONTROL

局部编辑  
结构规划

### 理解

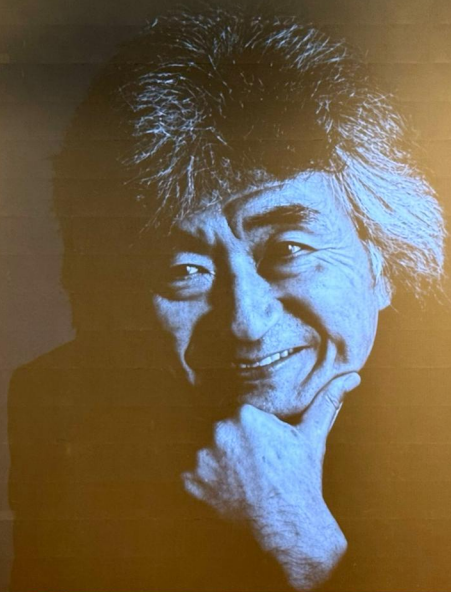
UNDERSTANDING

术语 / 语境  
音乐世界知识

真正进入专业工作流



艺术是以  
艺术家为中心的。





AI 是乐器。  
让它真正响起来的是艺术家。



第三届  
The Third

# 世界音乐人工智能大会

Summit on Music Intelligence



## 从一杯奶茶到音乐创作和消费的Agent

From a Cup of Milk Tea to Agents for Music Creation and Consumption

### 姜涛

Jiang Tao

自由量级公司 技术专家

Technical Expert, Ziyou Liangji Inc.

SOMI  
2026

中国·北京  
2026.04.25-26



中国人工智能学会  
Chinese Association for Artificial Intelligence



中央音乐学院  
CENTRAL CONSERVATORY  
OF MUSIC

# 事件1 千问订奶茶，打破了 Agent 次元壁

奶茶店都崩溃了！  
千问推出一句话订奶茶，  
有商家单量暴增至10倍

奶茶爆单中



我想喝一杯黑糖珍珠  
鲜奶茶，少糖



下单成功



Agent打破次元壁

“我想喝一杯黑糖珍珠  
鲜奶茶，少糖”



连接门店




下单支付



奶茶送达





# 事件2: AI音乐诞生, 音乐创作被重构

 suno



2023.12


 **文生歌曲**  
一键生成完整歌曲


 **社区爆发**  
百万用户快速增长

 udio



2024.04


 **高品质生成**  
音质与结构大幅提升


 **个性化定制**  
可控风格与歌词创作


 produce.ai



2024.08


 **专业级工具**  
编曲、混音全流程支持


 **AI协作**  
音乐人共创新范式

 mureka



2024.11


 **多语言支持**  
覆盖全球多种语言


 **风格多样化**  
多种曲风自由探索

 音潮  
hitto.ai



2025 → 2026

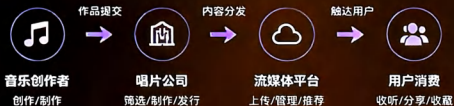
 **2025.7 WAIC主题曲**  
发布2.5版本

 **2026.2**  
发布3.0版本

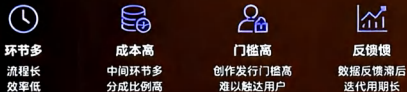
# 音乐生产-消费的循环有了更畅通的路径

## 过去

音乐生成和消费的循环链条比较长，要经过多个环节



## 传统链条的挑战



## 现在

音乐生成和消费可以在一个平台内完成，路径更短，效率更高



## "vibe coding + AI音乐作品" 创建百花齐放的个性化APP



## 路径更短

从创作到消费  
一步直达

## 效率更高

实时生成、分发  
即时反馈



## 门槛更低

人人可创作  
人人可发布



## 生态更活跃

个性化应用百花齐放  
音乐生态繁荣发展

# 猜想：流媒体平台的压力不是来自同行，而是来自Agent

传统流媒体平台面临多重压力

- 🎵 内容同质化
- 👥 用户注意力稀缺
- 💰 商业模式固化
- 📈 数据价值受限

被动推荐·千人一面·用户流失

未来，用户的音乐消费、娱乐和社交也通过Agent来完成



更加个性化、创意的音乐消费体验



量身定制



实时生成



无限风格



情感共鸣

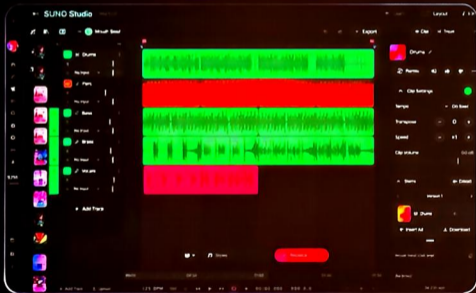
# 化繁为简

# 丰俭由人



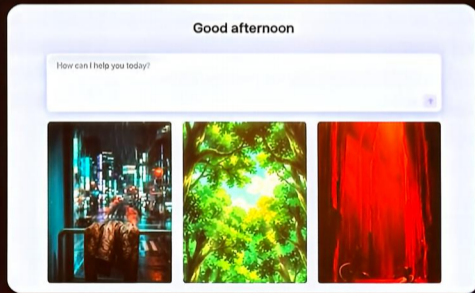
以 suno studio 为代表  
AI 创作工具在向 DAW 考虑

面向专业音乐创作者



以 hitto.ai 为代表  
的 Agent 在向对话形式考虑

面向广大的用户



# 用hitto.ai 通过对话创作音乐



今天上海下雨了，天气闷闷的，我想要创作一首歌让自己开心起来



雨天需要一首能让心情飞起来的歌！问你两个问题：

1. 你想要什么风格？比如轻快的流行、活泼的电子、温暖的民谣，或者你有其他偏好？
2. 歌词语言用中文还是英文？

轻快的流行，中文，加入一个温馨的爱情小故事



好的！让我来为你创作一首轻快流行的中文情歌 🎵

# 用 hitto.ai 通过对话创作音乐



这首歌写好啦，你看看满意吗？

## 半把伞的晴天

轻快流行

[Verse 1]

滴滴答答 外滩的钟声敲响  
乌云跑来 遮住了魔都太阳  
我站在街角 咖啡店的玻璃窗  
没带伞的尴尬 让人有点小慌张

[Pre-Chorus]

突然一个转身 撞进你的目光  
你撑着透明雨伞 笑得像颗薄荷糖  
“要不要一起走？” 你的声音多明朗  
这阴郁的雨天 瞬间被你点亮

[Chorus]

我们就这样走在 武康路的街头  
雨滴在伞面上 弹奏轻快的节奏  
你的肩膀微微 向我这边靠拢  
心跳扑通扑通 连风都变温柔  
上海的雨天 原来也可以这么甜  
多希望这条路 永远走不到终点

## 从灵感到成品，只需一句话的距离

🎵 故事线是：雨天忘带伞，街角偶遇一个撑着透明伞的人，  
两人共伞走过武康路，在雨声和心跳中悄悄靠近 🌧️

喜欢这个故事吗？如果满意的话，我就帮你生成音乐！

还不错，但是希望故事更成熟一点

🎵 好的，让我把故事改得更更有质感，少一点青涩，  
多一点成熟的情感温度~

改好啦，你看看这个版本：

## 武康路的半把伞

轻快流行

[Verse 1]

黄梅天的雨 停停下下  
武康路的梧桐 被洗刷得发青  
推开咖啡馆的门 你收起黑色的伞  
相视一笑 抖落肩上的水花

[Pre-Chorus]

两杯热拿铁 在雾里升华  
聊着工作 聊着日常的喧哗  
不用像二十岁那样 试探着表达  
懂你的沉默 胜过千言万语啊

[Chorus]

上海的雨天 滴滴答答  
我们在屋檐下 步伐不急不慢地走着  
你把伞倾斜 淋湿了左肩的衣襟  
这份克制的温柔 我全都偷偷记下  
成熟的爱情 像一杯温热的茶  
不沸腾 却暖到心里去啦

# 用 hitto.ai 通过对话创作MV



Please help me create an MV for the song "武康路的半把伞"

Let me grab the song first! Got it! Now choose which type of MV you'd like to make.

Premium MV  
Auto-generated visuals from music, 180 credits

Lip-Sync MV  
Lip-sync video from your photos, 100 credits

I want to create a Premium MV.

武康路的半把伞

00:00 03:19

00:36

Duration: 00:30 00:36 - 01:06

Landscape  Portrait

Pick your favorite 30-second clip from the waveform above, choose portrait or landscape, and confirm — I'll get the MV rolling right away!

Selected clip: 36.5s - 66.5s, Landscape. Generate MV now.




# 用 hitto.ai 通过对话创作MV



只需一句话， hitto.ai 帮你把灵感变成精美的音乐视频

武康路的半把伞



Choose actions

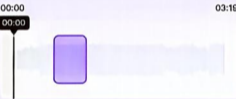
- Healing & Warm
- Energetic & Confident
- Melancholy & Sentimental
- Cool & Edgy
- Dreamy & Ethereal

3 selected / 3

- Taking a deep breath and exhaling slowly
- Confident smirk and sharp gaze
- Eyes filled with mist and longing

I want to create a Lip-Sync MV.

00:00 03:19



Duration: 00:30 00:38 - 01:08

Landscape  Portrait



# hitto.ai的未来



用对话，重新定义音乐的创作与体验



## 通过对话方式 支持更加精细的音乐编辑

用自然语言精准控制音乐的每一个细节，  
让创意表达更高效、更自由。

把副歌部分的鼓点更强烈一些，  
加入电子合成器音色

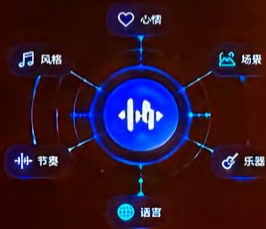


已为你优化副歌部分的鼓点，  
并加入电子合成器音色



## 成为个性化 音乐创作和消费的入口

理解你的喜好与情绪，随时为你创作，  
推荐和生成专属的音乐内容。



## 与用户建立信任 成为专属的音乐人

长期陪伴，深度理解，与你共同成长，  
成为你最懂你的音乐伙伴。

我想要一首温暖治愈的歌，  
送给未来的自己。



好的，这首歌为你而作。

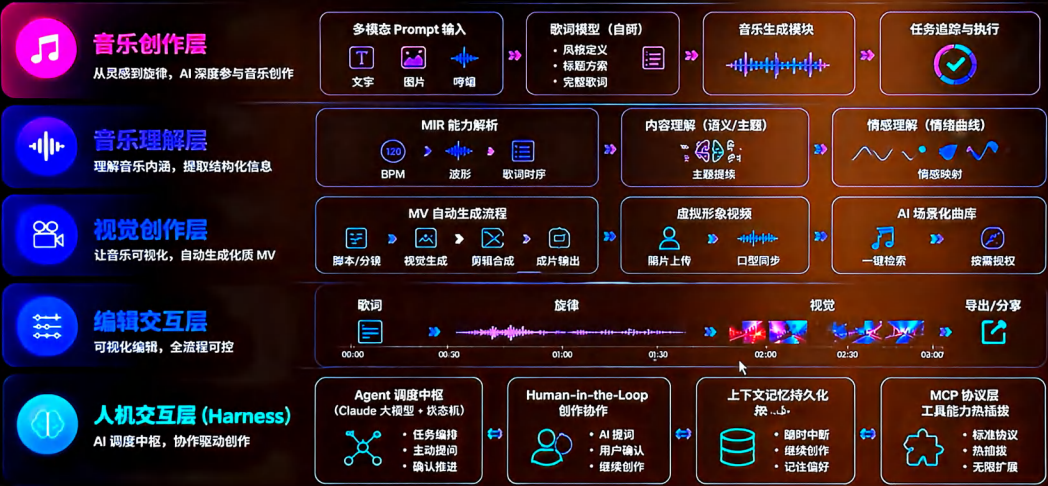


这就是我想要的感觉，  
谢谢你一直懂我。

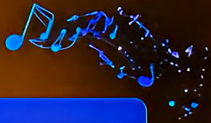


我会一直在这里，陪你创作属于你的音乐。❤️

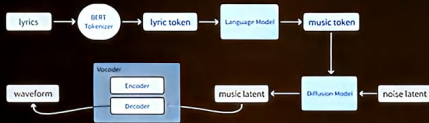
# hitto.ai的架构设计



# 音潮3.0模型国内领先



## 模型能力

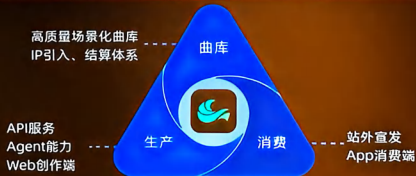


模型：自主0-1的大语言模型+DiT 扩散模型

特点：

- 商业级音乐质量，多语种、多风格支持
- 支持IP音色定制
- 具备续写、二次编辑能力
- 秒级生成
- 音乐专家知识库推理
- 多模态Prompt输入
- 细粒度音乐指令遵循能力

## 产品生态



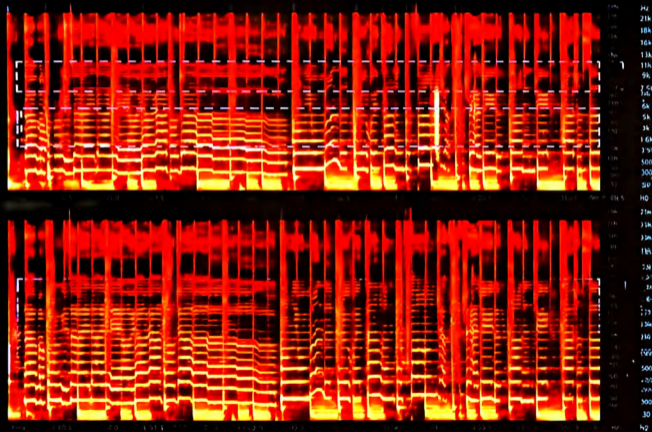
To C

- Web 创作工作站：多轨编辑 + AI 伴奏人声，专业级输出
- 移动端社区 App：自由“玩音乐”，音乐话题玩法社区
- 创作者经济闭环：作品自动入库 → 平台分发 → 播放/授权分成

To B

- AI 场景化曲库：一键检索、按需授权
- Agent 级创作接口：歌词 → 旋律 → 混音全流程生成，按调用量计费
- 数据回流 & 分成：使用数据实时反馈，版权收益自动结算

# 音潮3.0模型国内领先



清晰丰富的  
人声“甜点”频段



清晰的  
乐器泛音和厚实中频

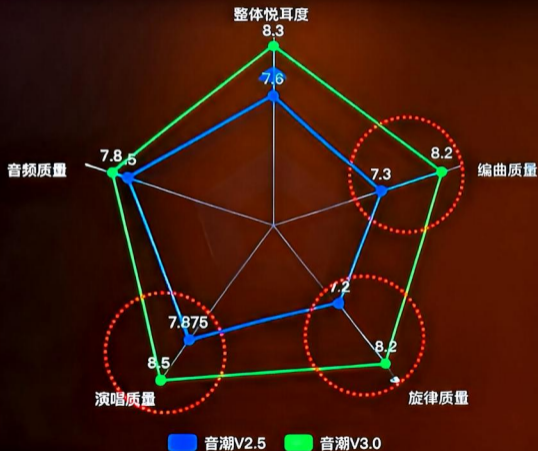


多样化的  
音乐风格表达



超强的  
旋律悦耳度和记忆性

# 音潮3.0模型国内领先



## 演唱质量的大幅度提升

更加细腻的演唱技巧，更加充沛的演唱表达  
让每个想用音乐表达的人的故事更加生动



## 编曲多样性的显著提升

更多样化的音乐风格让表达变得更加丰富  
更加合理的配器，更好的音乐整体性



## 旋律悦耳度和记忆性的明显提升

几乎每一条旋律都有让值得记忆的闪光点  
更加流畅的旋律线条，更加起伏的音符张力  
不仅仅是好听，让人每个人的表达更有“个性”

第三届  
The Third

# 世界音乐人工智能大会

Summit on Music Intelligence



## AI 赋能音乐教育

AI-Empowered Music Education

## 刘晓光

Liu Xiaoguang

DeepMusic首席执行官

DeepMusic CEO

SOMI  
2026

中国·北京  
04.25-26



中国人工智能学会  
Chinese Association for Artificial Intelligence



中央音乐学院  
CENTRAL CONSERVATORY  
OF MUSIC

# 对音乐科技的思考：应当解决什么问题？

场景1：教节奏型

2/4拍：以四分音符为一拍，每小节有两拍

4/4拍：以四分音符为一拍，每小节有四拍

3/4拍：以四分音符为一拍，每小节有三拍

生活中的3/4拍歌曲  
祝你生日快乐



现象：大多数人唱生日快乐歌拍手节奏是不对的。

流行音乐中的3/4拍歌曲  
发现爱



问题：3/4拍是不是都很欢快？

音乐知识举例太困难了，音乐为什么不可以通过乐理知识来检索？

# 对音乐科技的思考

场景1：教节奏型

2/4拍：以四分音符为一拍，每小节有两拍

4/4拍：以四分音符为一拍，每小节有四拍

3/4拍：以四分音符为一拍，每小节有三拍

听起来很悲伤的3/4拍流行音乐  
丁香花



问题：用什么方法可以让一首歌悲  
伤？

歌词、旋律、和声、乐器、音色等都在影响音乐的情感。怎么定性？怎么定量？

# 对音乐科技的思考

场景1：教节奏型

6/8拍：以八分音符为一拍，每小节有六拍

12/8拍：以八分音符为一拍，每小节有十二拍

被强制输入的12/8拍歌曲

别看我只是一只羊



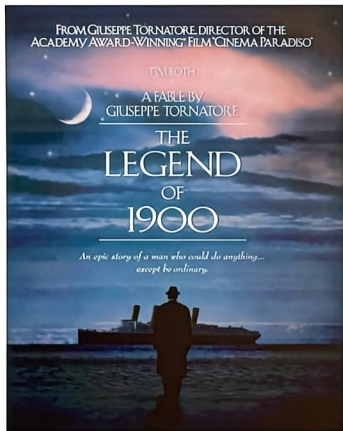
现象：小朋友学习12/8拍非常轻松

通过被强行输入的音乐作为样例学习音乐知识，是很好的方式。

# 对音乐科技的思考



## 场景2：学习和声



## Play Love (Ennio Morricone)

海上钢琴师 (1998)



电影音乐中的重属导和弦/升四级减和弦

# 对音乐科技的思考

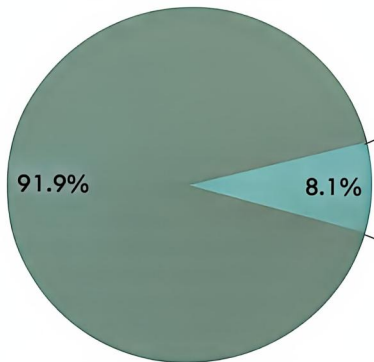
音乐知识举例太困难了，音乐为什么不可以通过乐理知识来检索？

歌词、旋律、和声、乐器、音色等都在影响音乐的情感。数据结构怎么统一？怎么定性？怎么定量？

通过被强行输入的音乐作为样例学习音乐知识，是很好的方式。

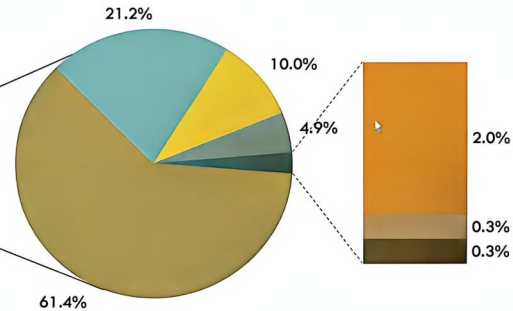
# 检索 / 和弦分析

## #4 级和弦用法知识库



■ 使用 #4 级和弦   ■ 未使用 #4 级和弦

## #4 级和弦去向分布



■ 下属功能   ■ 大调属功能   ■ 主功能   ■ 小调属功能  
■ 重属功能   ■ 五度圈   ■ 离调用法

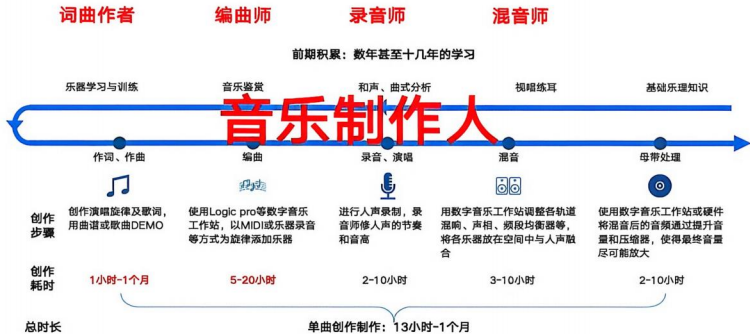
## 升四级减和弦作乐句终止功能用例检索

歌曲名称	段落	功能	局部和弦连接	起止位置
说散就散 	副歌	乐句终止	Gm7 – Em7b5 – Cm7	3:18 – 3:22
世界末日 	副歌	乐句终止	Em/D – C#m7b5 – Am	1:38 – 1:40
祝福 	主歌	乐句终止	G/D – C#m7b5 – C(add2)	1:00 – 1:04
大鱼 	预副歌	乐句终止	Bbm7(add4) – Gm7b5 – Ebm7	1:29 – 1:31
匿名的好友 	副歌	乐句终止	D/A – G#m7b5 – G	3:43 – 3:47
囚鸟 	副歌	乐句终止	Bb/F – Em7b5 – Cm7	3:45 – 3:50
背叛 	副歌	乐句终止	D#m/F# – E#m7b5 – C#m7	1:29 – 1:33
彩虹 	副歌	乐句终止	Dm/C – Bm7b5 – Gm7	3:35 – 3:40

# 音乐AI能如何影响 workflow

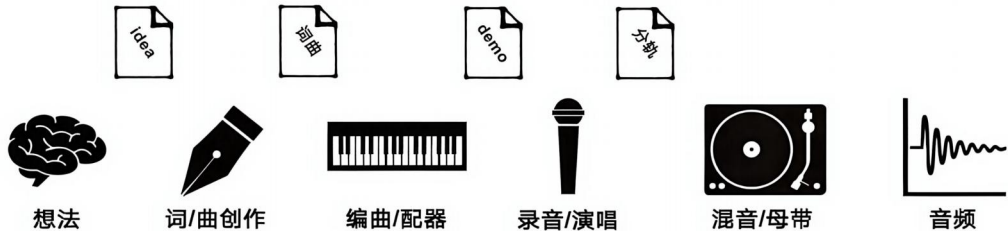
创作所需的音乐基础需要经过漫长的学习积累

音乐创作过程专业性要求极高，耗时极长



# 音乐工作流的变化

音乐创编：



LLM

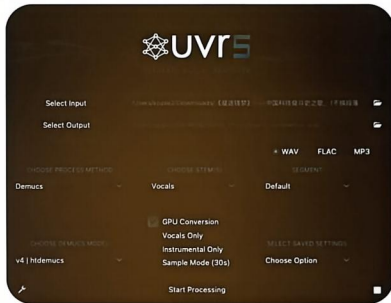
音乐模型

# 音乐工作流的变化

翻唱:



Remix:



音频操作变得很重要

将音频分离成:

Vocal, Bass, Drum, Others  
<https://ultimatevocalremover.com/>



音频

# 音乐AI产品在音乐制作流程的定位

## 春节序曲

Allegro con fuoco

李焕之作曲

笛子 Piccolo  
长笛 Flauti I  
双簧管 Oboi  
单簧管 Clarineti (Bb)  
大管 Fagotti

纯音乐

音乐创作

总谱  
功能简谱

音乐制作

歌曲

和弦

4M7 57 3/5# 6m7 5 6m7

↑ 51 11 1 62 ↑ 2 12 21 76 ↑ 6 0 0 0 | 0 0 1 2 |

後 悔 也 草 重 故 事 結 尾 分 手

(小鼓第4拍)Strings In

PN(1)

5 6m7 3m7 旋律+歌词

31 23 31 55 ↑ 5 53 32 1 | 16 71 16 33 ↑ 3 37 76 5 |

應 該 體 面 誰 都 不 要 說 抱 歉 何 來 虧 欠 我 敢 給 就 敢 心 碎

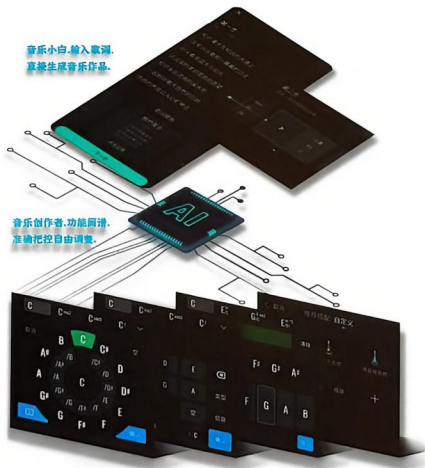
4 3m7 6m7 67/1# 2m7 3/7 6m7 57

16 71 15 65 ↑ 55 31 12 36 ↑ 63 33 32 13 ↑ 3 - 1 2 |

鏡 頭 前 面 是 從 前 的 我 們 在 喝 彩 流 著 淚 聲 嘶 力 竭 離 開

开始没有清晰的创作制作之分

# 和弦派：跨平台、一站式、低门槛AI音乐工作站



## ■精准编辑歌曲信息

编辑简谱，让AI精准演唱输入的歌词和旋律。支持多个声部。  
编辑和弦并选择风格，让AI像“电子琴”一样自动伴奏。

## ■支持所有设备

支持平板、手机、电脑、教学电视。  
苹果、Windows、安卓全平台支持。

## ■多种格式导出

支持mp3、wav格式导出音频，支持midi导出进行进一步编辑。  
可以导出简谱和五线谱。支持工程文件分享。



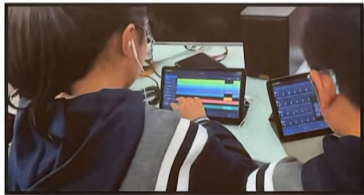
武汉市教研活动：小学音乐课《乃哟乃》  
学生分组用和弦派实践，用do、mi、sol三个音对作品进行改编



杭州市滨江区教研活动：《军民团结一家亲》  
运用和弦派备课，将作品转化为多个风格鉴赏



长沙市麓山国际实验小学：  
学生分组将语文课学习的诗词，在和弦派中创编音乐作品



上海市复旦二附中：  
学生分组，运用和弦派的AI乐器改编老师原创作品的配器



人大附中

## 歌曲写作与AI作曲应用

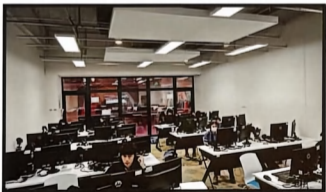
- 课程上，学生运用和弦派进行作品的伴奏创作
- 学生自带手机或平板



北师大实验中学

## Music Creation

- 运用和弦派为主要软件，进行音乐创编教学
- 运用学校的安卓平板
- 已开设第二学年



稻香湖学校

## AI音乐制作

- 运用和弦派教学音乐乐理知识
- 通过和弦派精准控制音乐中的元素，生成音频参考，配合SUNO创编
- MIDI教室Windows电脑



首师大附中

## AI音乐创编

- 运用和弦派为主要软件，进行音乐创编教学
- 通过和弦派精准控制音乐中的元素，生成音频参考，配合SUNO创编
- 运用学校的安卓平板
- 已开设第二学年

60

如愿  
66

80

告白气球  
90

110

120  
逆战

永不失联的爱  
41

梦底  
56

多远都要在一起  
68

悬溺  
72

小幸运  
79

万有引力  
90

想你时风起  
42

淘汰  
60

不知所措  
64

大眠  
68

大鱼  
70

那些年  
80

Blank Space  
96

Love Story  
120

你还要我怎样  
48

搁浅  
61

专属天使  
65

枫  
67

年轮  
73

月牙湾  
87

乌梅子酱  
85

有何不可  
100

阳光宅男  
135

夏天的风  
57

一千年以后  
62

跳楼机  
64

孤勇者  
65

爱错  
73

浪漫手机  
85

Blueming  
105

Apt  
148

Always Online  
60

我好像在哪见过你  
63

好感伤啊

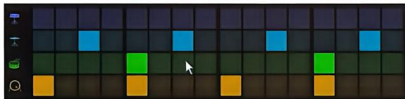
我那么的爱你你却不爱我  
感谢你无私的爱

那就这样吧  
还是美好的

甜甜的爱情  
美好的生活

勇敢冲冲冲

House



Techno



摇滚流行



四拍三分



8beat  
抒情流行



16beat  
流行



Dubstep/Trap  
/Futurebass

8beat  
欢快/动感流行

UK Garage  
/Jersey Club

Drum&Bass





# DeepMusic AI 生成能力亦得到音乐大厂的验证与高度认可

## AI 编曲



- 覆盖4000余首头部歌曲
- 为全民K歌提供“花样唱”及“AI曲风”功能，累计使用3亿次，是音乐AIGC的最大规模落地

## 曲谱生成



- 支持全民K歌自动修音/弹唱等录唱功能
- 生成简谱、钢琴演奏/弹唱谱、吉他弹唱谱等，已上线酷狗音乐及QQ音乐

## 功能音乐



- 与TME长音频合作孵化哄睡AI角色，上线半年，AIGC睡眠音乐已超千万播放量

个都呢，大家都平时  
大家有什

第三届  
The Third

# 世界音乐人工智能大会

Summit on Music Intelligence

panies, this is all uh,

世界音乐人工智能大会  
Summit on Music Intelligence

CHALLENGE



SOMI  
2026

中  
2026



主办单位 Hosts



人工智能学会  
Society for Artificial Intelligence



中央音乐学院  
CENTRAL CONSERVATORY OF MUSIC





# MSU-Bench

Musical Score Understanding Benchmark:

Evaluating LLMs' Comprehension of Complete Musical Scores

音乐全谱理解基准：大模型对完整乐谱理解能力的评测与分析

---

## AUTHORS

Congren Dai, et al.

戴球人

## INSTITUTION

Central Conservatory of Music · Imperial College London · Tsinghua University

中央音乐学院 · 伦敦帝国理工学院 · 清华大学





CONTENTS · 目录

# Table of Contents

## 演讲结构

**ABOUT ME · 关于我**

导师介绍 · 教育背景 · 实习经历 · 学术发表

**BACKGROUND · 研究背景**

乐谱理解的重要性 · 幻觉问题动机

**MSU-BENCH · 基准测试**

整体概览 · 四层次框架 · 基准特点 · 案例分析

**EXPERIMENTS · 实验结果**

实验设置 · 零样本评估 · 层次成功率 · LoRA微调

**CONCLUSION · 总结**

主要发现 · 贡献与未来方向



## BACKGROUND · 研究背景

# Why Musical Score Understanding Matters

为什么乐谱理解如此重要？

### Musical Scores

乐谱是音乐的书面语言，编码了音高、节奏、力度、表情、和声与结构等多层信息。理解完整乐谱需要跨越局部符号识别与整体音乐结构推理。

Pitch · Rhythm · Harmony · Form

### LLMs & VLMs

大语言模型 (LLMs) 和视觉语言模型 (VLMs) 在自然语言与多模态任务中取得了显著进展，但其对音乐符号和完整乐谱的理解能力仍缺乏系统研究。

GPT · Claude · Gemini · Grok · Qwen

### Research Gap

现有音乐 AI 的基准评测多聚焦于音频、短片段或选择题形式，较少系统考察模型对完整乐谱的视觉理解、符号推理与开放式问答能力。

→ MSU-Bench fills this gap

# The Hallucination Problem in Score Reading

谱阅读中的幻觉问题

(a) Hallucination.

Pictures at an Exhibition

Modest Mussorgsky



Which articulation is used in bar 7?



I "saw" **staccato** markings in bar 7.

(b) Ideal scenario.

Pictures at an Exhibition

Modest Mussorgsky



Which articulation is used in bar 7?



I saw **tenuto** markings in bar 7.

MSU-BENCH · 基准测试

# MSU-Bench at a Glance

基准测试整体概览

150

Musical Scores

巴赫、贝多芬、肖邦、德彪西等作曲家的完整乐谱，涵盖多种风格和体裁

1,800

QA Pairs

人工策划的生成式问答对，每首乐谱12个问题，覆盖四个递进层次

2

Modalities

文本模态（ABC记谱法）与视觉模态（PDF乐谱），支持跨模态对比评估

15+

SOTA Models Evaluated  
评估模型数量

4

Comprehension Levels  
理解层次

Zero-shot

+ Fine-tuning Evaluation  
评估设置

Open Source

Publicly Available  
开源可用



# Four Progressive Levels of Musical Comprehension

四个递进层次的音乐理解框架





## MSU-BENCH · 基准测试

# What makes MSU-Bench distinctive?

与现有基准相比，MSU-Bench 新在哪里？

### Textual + visual

同时支持 ABC textual QA 与 PDF visual QA

### Full scores

面向完整乐谱，而不是短片段或局部截图

### Generative QA

开放式问答，更贴近真实音乐分析

### Human-curated

人工标注与专家校验，可靠性更高

### Homophony-aware

显式覆盖多声部 / 同音性场景

### Trainable benchmark

支持 zero-shot 与 LoRA fine-tuning 评测



# Illustration of Multi-level Understanding in MSU-Bench

以穆索尔斯基《图画展览会》为例的多层次理解示例

(a) Raw score. Pictures at an Exhibition Level 1:  
Modest Mussorgsky

*♩ = 112*  
*Promenade* Level 4: Melodic motif Level 3:  
G minor

(b) ABC notation's Metadata.

```
X:1
T:Pictures at an Exhibition
C:Modest Mussorgsky
L:1/4
Q:1/4=112
M:5/4
K:Bb
V:1 treble nm="Piano" snm="Pno."
V:2 bass
```

(c) Musical content in ABC notation.

```
V:1
""\n"fr""Promenade" |tenuto:G |tenuto:F |tenuto:D |tenuto:F/F |tenuto:D || %1
[M:6/4] |tenuto:c/r/ |tenuto:d |tenuto:B |tenuto:G |tenuto:F || %2
[M:5/4] |B,,D,G |A,,C,F |B,,D,B |((C-A-c)/|C,A,F)/ |F,A,D || %3
[M:6/4] |((C-A-c)/|C,A,F)/ |F,B,D |D,G,B |E,G,C |G,,C,G |A,,C,F || %4
[M:5/4] |tenuto:F |tenuto:G |tenuto:D |tenuto:F/G/ |tenuto:C || %5
V:2
z5 |[M:6/4] z6 |[M:5/4] |G,,G |A,,F |G,,G |F,,F |D,,D || %3
[M:6/4] |F,,F |B,,B |G,,G |C,,C |E,,E |F,,F |[M:5/4] z5 || %5
```

(d) Sample questions for Levels 1, 2, 3, and 4.

Level 1: Q: Who is the composer? A: Modest Mussorgsky.	Level 2: Q: In bar 5, what is lowest note? A: Middle C (C4).	Level 3: Q: In bar 3, is there a G minor chord? A: Yes.	Level 4: Q: In which bar does the melodic motif appear? A: In bars 1 and 2.
--	--	---	---



## EXPERIMENTS · 实验结果

# Models & Evaluation Protocol

## 评估模型与实验协议

- ChatGPT-5
- Gemini 2.5 Pro
- Qwen3-Max
- ChatGPT-5-mini
- Gemini 2.5 Flash
- Qwen3-32B
- Claude Opus 4
- Grok 4
- Qwen2.5-VL-72B
- Claude Sonnet 4
- Llama 4 Maverick
- + more models...

①

### Zero-shot Evaluation

在不提供任何示例的情况下，直接评估模型对乐谱问题的回答能力。每首乐谱评估12个问题。

②

### Fine-tuning (LoRA)

使用低秩自适应 (LoRA) 对模型进行微调，验证领域适配是否有效提升乐谱理解性能。

③

### LSR Metric · 层次成功率

引入层次成功率 (Level-wise Success Rate) 衡量模型在多层次问题上的连贯推理能力。



EXPERIMENTS · 实验结果

# Zero-shot Performance & Modality Gap

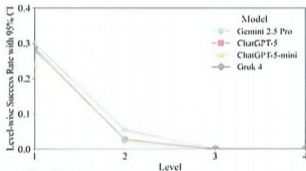
零样本性能与模态差距

Model	Level 1	Level 2	Level 3	Level 4	Overall
<b>TEXTUAL QA (ABC Notation) · 文本模态</b>					
<b>Gemini 2.5 Pro</b>	<b>65.33</b>	<b>56.00</b>	<b>38.67</b>	<b>37.78</b>	<b>49.44%</b>
ChatGPT-5	62.00	50.22	38.44	38.44	47.28%
ChatGPT-5-mini	59.11	43.56	31.33	<b>40.89</b>	43.72%
Grok 4	62.00	40.00	31.11	37.11	42.61%
Claude Opus 4	57.11	36.89	35.56	35.56	41.28%
<b>VISUAL QA (PDF Score) · 视觉模态</b>					
<b>Claude Opus 4</b>	<b>25.33</b>	<b>21.78</b>	<b>30.44</b>	<b>19.33</b>	<b>24.22%</b>
Gemini 2.5 Pro	22.00	22.44	29.11	20.00	23.39%
Claude Sonnet 4	27.11	16.44	27.33	18.44	22.33%

EXPERIMENTS · 实验结果

# Level-wise Success Rate (LSR)

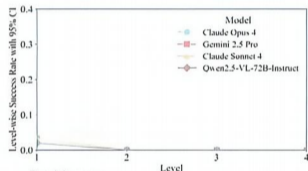
层次成功率：性能随难度急剧下滑



Remaining scores

Model	Level 1	Level 2	Level 3	Level 4
○	4.5	0	0	0
□	4.2	4	0	0
△	3.6	3	0	0
◇	4.3	4	0	0
·	1	2	3	4

(a) Textual QA.



Remaining scores

Model	Level 1	Level 2	Level 3	Level 4
○	1	0	0	0
□	5	0	0	0
△	5	0	0	0
◇	3	0	0	0
·	1	2	3	4

(b) Visual QA.

● Textual QA: L1 ~50% → L2 ~9% → L3&4 ≈ 0%

● Visual QA: L1 ~5% → L2+ ≈ 0%

● Chain-of-level reasoning collapse

## EXPERIMENTS · 实验结果

# LoRA Fine-tuning Substantially Improves Performance

## LoRA微调显著提升乐谱理解性能

### Low-Rank Adaptation (LoRA)

参数高效的微调方法，仅训练少量可更新参数，即可实现面向乐谱理解任务的领域适配，同时尽量保留模型原有的通用知识与推理能力。

### Structured Input Advantage · 结构化输入优势

相比 PDF 乐谱，ABC 记谱法提供了更明确的音高、节奏、声部与小节信息，能够减少视觉解析噪声，从而提升模型在乐谱问答任务中的表现。

### General Knowledge Preserved · 通用知识保留

MMLU 结果显示，LoRA 微调后模型在通用基准上的性能基本保持稳定，说明领域适配并未以明显牺牲通用能力为代价。

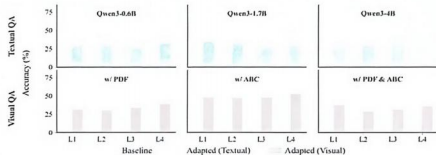


Table 6: Evaluation of models conducted before and after LoRA on MMLU. Qwen2.5-VL-3B-Instruct is adapted using LoRA across the three input modalities described in Section 5.1.

Models	STEM	Humanities	Social Sciences	Other Subjects
Qwen3-4B	72.63	81.14	63.21	74.61
w/ LoRA	71.09 <sup>+(-0.36)</sup>	81.51 <sup>+(-0.10)</sup>	63.51 <sup>+(-0.30)</sup>	75.11 <sup>+(-0.50)</sup>
Qwen2.5-VL-3B-Instruct	60.60	75.63	58.72	69.65
w/ PDF	60.90 <sup>+(-0.30)</sup>	75.66 <sup>+(-0.03)</sup>	58.15 <sup>+(-0.27)</sup>	69.55 <sup>+(-0.15)</sup>
w/ ABC	60.17 <sup>+(-0.13)</sup>	75.79 <sup>+(-0.16)</sup>	58.13 <sup>+(-0.59)</sup>	69.55 <sup>+(-0.15)</sup>
w/ PDF&ABC	60.50 <sup>+(-0.10)</sup>	75.87 <sup>+(-0.22)</sup>	58.28 <sup>+(-0.44)</sup>	69.55 <sup>+(-0.15)</sup>



## CONCLUSION · 总结

# Key Takeaways

## 论文核心发现总结

- 1 Pronounced Textual-Visual Modality Gap**  
文本模态与视觉模态之间存在显著差距：基于 ABC 记谱法的问答表现明显优于直接读取 PDF 乐谱，说明视觉乐谱解析与小节定位仍是主要瓶颈。
- 2 Fragile Level-wise Success Rates**  
层次成功率 (LSR) 随层级推进迅速下降，表明模型虽然能回答部分单独问题，但难以在同一乐谱中持续保持跨层级正确性。
- 3 LoRA Fine-tuning is Highly Effective**  
LoRA 微调能显著提升模型在乐谱理解任务上的表现；同时，MMLU 结果显示模型的通用知识基本保持稳定。
- 4 MSU-Bench as a Rigorous Foundation**  
MSU-Bench 为 AI 音乐学、符号音乐理解与多模态乐谱推理提供了严格、可复现的评估基础。



## CONTRIBUTIONS · 主要贡献

- Introduce MSU-Bench, a human-annotated benchmark for complete musical score understanding  
提出 MSU-Bench，一个面向完整乐谱理解的人工标注基准
- Establish a dual-modality evaluation setting with ABC notation and PDF scores, covering 1,800 generative QA samples  
构建文本 ABC 与视觉 PDF 双模态评估设置，覆盖 1,800 个生成式问答样本
- Reveal key limitations of current SOTA models in hallucination, bar localisation, modality gaps, and hierarchical reasoning  
揭示当前 SOTA 模型在幻觉、小节定位、模态差距与层次推理上的局限
- Introduce Level-wise Success Rate (LSR) to measure continuous correctness across hierarchical levels  
引入层次成功率 (LSR)，用于衡量模型跨层级连续正确的能力
- Validate the effectiveness of LoRA fine-tuning while showing that general knowledge is largely preserved  
验证 LoRA 微调在乐谱理解任务中的有效性，并显示通用知识基本保留

## FUTURE WORK · 未来方向

- Expand the dataset to cover broader styles, regions, historical periods, and notation traditions  
扩展数据集规模，覆盖更多风格、地域、时期与记谱传统
- Explore multimodal architectures and pre-training strategies tailored to musical score understanding  
探索面向乐谱理解的多模态架构与预训练策略
- Improve visual score parsing, especially bar localisation, voice identification, and long-score understanding  
提升视觉乐谱解析能力，尤其是小节定位、声部识别与长谱面理解
- Extend evaluation to downstream tasks such as performance guidance, score transcription, and music analysis assistance  
将评估扩展至表演指导、乐谱转录、音乐分析辅助等下游任务
- Explore joint multimodal understanding across audio, MIDI, and visual scores  
探索音频、MIDI 与视觉乐谱的联合多模态理解

# Q & A

感谢聆听 · Thank You

---



arXiv Paper



GitHub Repo



Google Scholar



中央音乐学院  
CENTRAL CONSERVATORY  
OF MUSIC

# 音频驱动的弦乐演奏动作生成

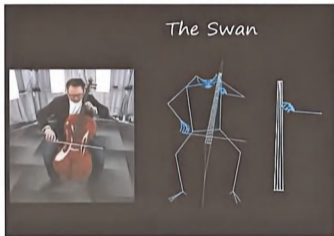
ELGAR: Expressive Cello Performance Motion Generation for Audio Rendition

中央音乐学院 | 清华大学

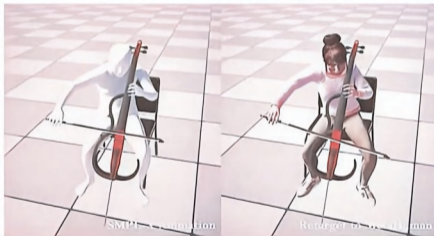
丘治平

## 从感知到具身：音乐演奏的智能演进

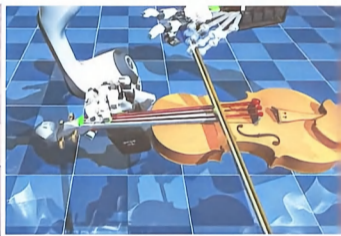
### 弦乐演奏动作捕捉



### 弦乐演奏动作生成



### 弦乐机器人



- **感知层**：将演奏家精巧的技法数字化记录下来，为后续的智能建模提供多模态教科书
- **认知层**：代表了AI对音乐信息的理解能力，即根据音乐思考如何运弓与按弦的逻辑思考能力
- **执行层**：音乐具身智能的终极形态，在复杂物理约束下验证音乐交互的无限可能

## 通用动作生成模型的发展趋势与局限性

- **输入模态的高度泛化:** 支持文本、音频和图片等
  - **控制精度的精细化:** 实现精细化控制 (手-物交互等)
  - **垂类任务的适配:** 需要垂直领域知识的约束与学习
- ∴ 乐器演奏 = 丰富的听觉意境 + 生动的视觉表达
- ∴ 生动的视觉表达 ≡ 演奏家的艺术表现力
- ∴ 乐器演奏 (Performance) ≠ 简单的动作 (Movements)

```
def get_movements_of_musician (text, audio, ...)  
def get_expressive_performance_of_musicians (audio)
```

Add prideful emotions  
when playing the violin

Lower body moves  
with rhythm of guitar

MotionRefit



Text: Play Guitar  
Music: Anti-Hero

Text: Play Drums  
Music: Anti-Hero

LMM



## 现有演奏动作生成研究的局限与挑战

### ■ 监督学习

- 动作精度不足：难以生成精细的手指演奏动作
- 缺乏交互感知：忽视了人与乐器之间的交互约束

### ■ 强化学习

- 输入依赖性强：受限于符号化音乐作为输入
- 协同难度大：难以实现全身多肢体的复杂控制

当前学术界工作多集中于**局部动作生成**（通常仅限躯干或手部），忽略了乐器演奏作为一种复杂具身行为的整体协调性。

我们的研究目标：从原始音频出发，实现**全身演奏动作生成**

- **精细化控制**：实现高精度的指法与弓法建模
- **深度具身交互**：生成符合物理逻辑的真实器乐交互过程

### Supervised Learning



Shlizerman et al. 2018



Kao and Su 2020

### Reinforcement Learning



Wang et al. 2024



Xu and Wang 2024

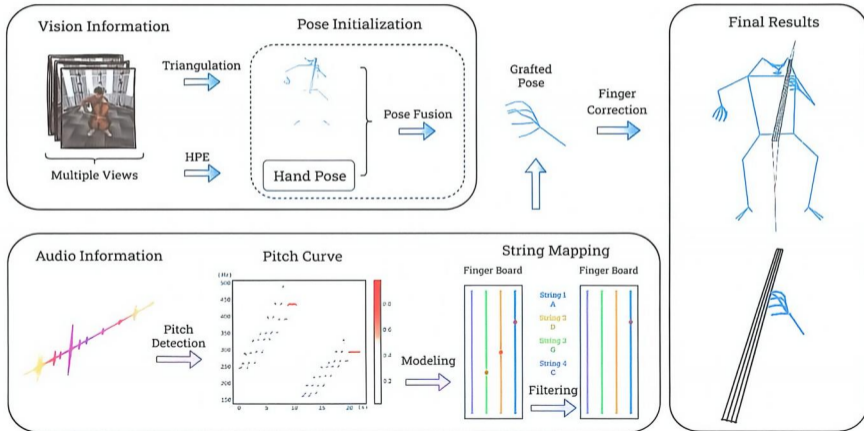


## 乐器演奏相关数据集

Dataset	Instrument	Pieces	Duration	Camera Views	Mocap Annotation
<i>Marker / Sensor Based Dataset</i>					
TELEMI [Volpe et al. 2017]	Violin	41	2.4 h	3 + 13 (infrared)	Body, Instrument, Bow
QUARTET [Papiotis et al. 2016]	String quartet	30	0.5 h	1 + 26 (infrared)	Body, Instrument, Bow
MMG [Perez-Carrillo et al. 2016]	Guitar	10	0.17 h	N/A	Body, Hands, Instrument
EEP [Marchini et al. 2014]	String quartet	23	N/A	0 (wired EMF)	Bow
Bowstroke [Young and Deshmane 2007]	Violin	N/A	N/A	1	Bow
<i>Markerless Dataset</i>					
CCOM-HuQin [Zhang et al. 2022]	HuQin	N/A	1.29 h	3	N/A
URMP [Li et al. 2018]	Multi-instrument	44	1.3 h	1	N/A
C4S [Bazzica et al. 2017]	Clarinet	54	4.5 h	1	N/A
ENST-Drums [Gillet and Richard 2006]	Drum kit	N/A	3.75 h	2	N/A

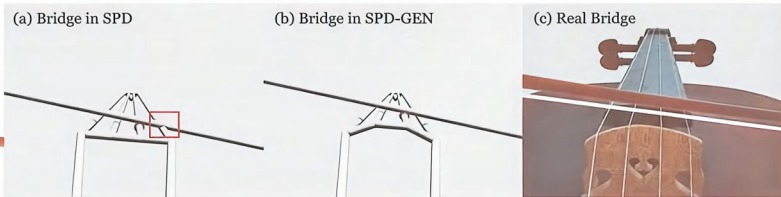
- 数据来源匮乏，数据规模较小，采集时长较短
- 采集的数据标注成本高，间接导致标注的粒度粗，缺少精细的标注
- 音乐演奏数据往往存在版权问题，部分数据集未开放使用

## 音视频融合的无标记弦乐演奏动作捕捉

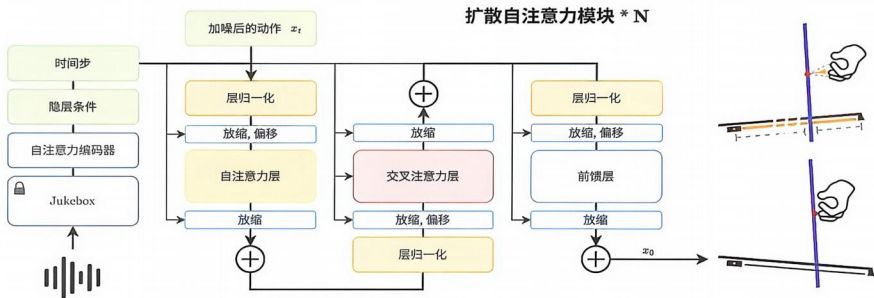


## 数据归一化与运动学校准

- 构建统一的规范化空间：**如同让同一个人使用同一把乐器将所有曲目演奏一遍**
  - 对人体和大提琴进行归一化，消除不同表演者与乐器间的几何差异
  - 使用SMPL-X模型对人体进行拟合，统一躯干形态、全局朝向及位移
  - 保持演奏交互语义：根据原始数据的演奏位置进行逆向动力学还原
- 物理几何矫正：重构琴桥的拱形形状，确保弓-弦接触存在可行解
- 先验约束：将弓根 (Frog) 定位于关键手部关节 (图中红色高亮部位) 的中点



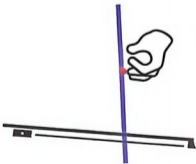
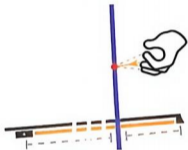
## 模型架构设计



模型以音乐编码表征、加噪的全身大提琴演奏动作以及加噪时间步作为降噪模型DiT的输入，以降噪后的全身大提琴演奏动作作为训练模型的输出，通过交互接触损失函数（图右上）来进一步迭代优化。



## 损失函数设计



交互接触损失函数包含手交互接触损失函数 $\mathcal{L}_{hand}$ 和弓交互接触损失函数 $\mathcal{L}_{bow}$

$$\mathcal{L}_{hand} = \mathbb{1}_{note} \|\widehat{d}_{cp} \odot I_{f_0}\|_2^2 + \mathbb{1}_{others} \|(\widehat{d}_{cp} - d_{cp}) \odot I_{f_0}\|_2^2 \quad (1)$$

$$\mathcal{L}_{bow} = \|\widehat{d}_{l_s, l_b} \odot I_{f_0}\|_2^2 + \|(\widehat{d}_{p, l_s} - d_{p, l_s}) \odot I_{f_0}\|_2^2 \quad (2)$$

在式 (1) - (2) 中,  $\mathbb{1}_{note}$  表示手指是演奏音符的手指,  $\mathbb{1}_{others}$  表示手指不是演奏音符的手指,  $\widehat{d}_{cp}$  表示预测的指尖到接触点的距离,  $d_{cp}$  表示指尖到接触点的真实距离,  $I_{f_0}$  是指示函数, 当检测到音高 (即基频 $f_0$ ) 时将被激活,  $\widehat{d}_{l_s, l_b}$  表示演奏弦与生成弓之间的距离,  $\widehat{d}_{p, l_s}$  表示预测的琴弓端点与演奏弦之间的距离,  $d_{p, l_s}$  表示琴弓端点与演奏弦之间的真实距离。



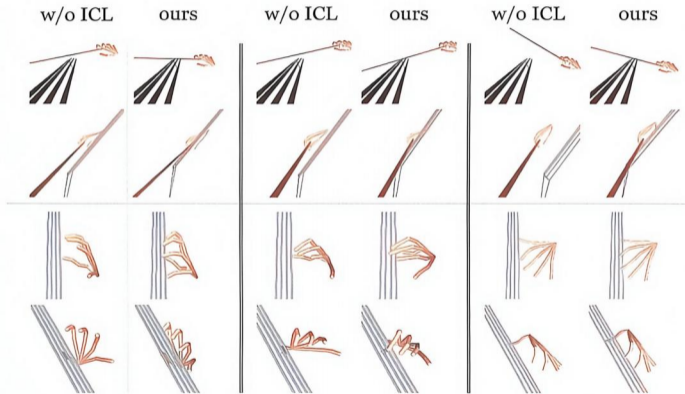
## 定性结果分析

### ■ 缺乏物理感知的演奏动作

- 模型仅学习整体动作分布，忽略了弦乐演奏中重要的接触点
- 手指与琴弦发生脱离/穿模，运弓位置有着明显的空间差异

### ■ 解决方案：交互接触损失函数

- 精准对齐于理论演奏位置
- 维持稳定的接触状态
- 具备明显的几何约束



引入交互接触损失函数 (ICL) 的前后对比

## 定量结果分析

### ■ 传统通用动作生成指标的局限性

- 物理约束导致分布偏移：引入ICL强行纠正了生成动作的物理偏差，但是偏离了训练集的数据分布
- 小规模数据集的随机性：在SPD-GEN这种样本稀疏的数据集中，训练集与测试集的分布差异存在噪声

### ■ 我们提出专为弦乐动作生成设计的量化指标

- FCD (Finger-Contact Distance): 衡量手指触弦的空间贴合精度
- BSD (Bow-String Distance): 衡量弓毛与琴弦的物理接触精度
- Bowing Scores (BF1 & BCS): 衡量运弓的节奏一致性与轨迹相似度

Loss Configuration	FCD ↓	BSD ↓	BF1 ↑	BCS ↑
w/o ICL	18.64	25.20	0.4332	0.6965
w/ HICL only	<b>14.56</b>	23.98	0.4082	0.6646
w/ both HICL and BICL	15.60	<b>5.40</b>	<b>0.4721</b>	<b>0.7515</b>

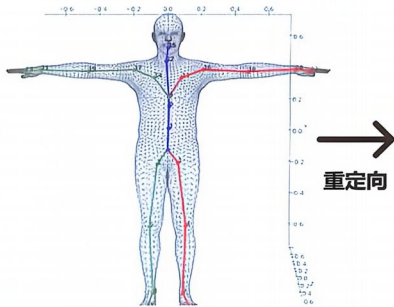
## 演奏动作的跨角色重定向

### ■ 骨骼结构差异

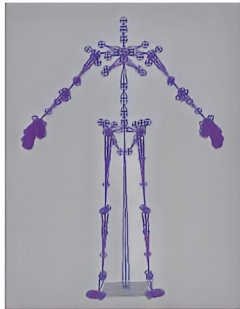
- 学术模型：骨骼层级**简化**，适合深度学习训练
- 工业资产：具有**复杂**的异构拓扑、关节坐标系和肢体比例

### ■ 技术关键：动作重定向

- 实现动作数据在**保持语义**的情况下进行自然迁移
- 确保物理约束不被打破，不发生动作的形变/穿模



SMPL-X 骨架结构



MetaHuman 骨架结构

## 演奏动作视频生成

### ■ 语义引导的视频生成

- 交互门槛低
- 空间想象力强
- 语义模糊性






### ■ 结构引导的视频生成

- 如ControlNet (Zhang, et al), 能够根据草图/2d关键点序列作为确定性的空间提示词, 实现像素级的生成轨迹约束

### ■ Text to Video Leaderboard (No Audio) Artificial Analysis

Added to the leaderboard in the last month:

HappyHorse-1.0, Dreamina Seedance 2.0 720p

Category: All		Current models	All models	No Audio	With Audio	All	Open weights	Global Leaderboard	Personal Leaderboard
Rank	Range	Creator	Model	ELO	95% CI	Samples	Released		
1	1	 Alibaba-ATH	HappyHorse-1.0	1,365	-9/9	15,808	Apr 2026		
2	2	 ByteDance Seed	Dreamina Seedance 2.0 720p	1,270	-8/8	8,704	Mar 2026		
3	3	 KlingAI	Kling 3.0 1080p (Pro)	1,246	-9/9	5,808	Feb 2026		
4	4-6	 Skywork AI	SkyReels V4	1,236	-9/9	6,564	Mar 2026		
5	4-6	 xAI	grok-Imagine-video	1,232	-8/8	6,247	Jan 2026		

somi



中央音乐学院  
CENTRAL CONSERVATORY  
OF MUSIC

# Video Echoed in Music: Semantic, Temporal, and Rhythmic Alignment for Video-to-Music Generation

Xinyi Tong<sup>\*</sup>, Yiran Zhu, Jishang Chen, Chunru Zhan, Tianle Wang, Sirui Zhang, Nian Liu,  
Tiezheng Ge, Duo Xu, Xin Jin, Feng Yu, Song-Chun Zhu<sup>†</sup>

Central Conservatory of Music, Beijing Institute for General Artificial Intelligence, Alibaba Group, Peking University

 Xinyi Tong

2026.04.25

## :: Background

- ▶ Music, akin to video, evokes sensory perception and emotional responses, thereby playing a critical role for **enhancing the overall audiovisual experience** in video edits.
- ▶ However, the use of music in media production often raises **complex copyright concerns**. The process of manually composing original music is not only **highly resource-intensive** but also **time-consuming**, which presents a significant challenge for content creators and producers.



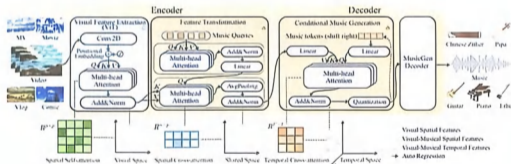
**Intelligent & Immersive  
Audiovisual Experience**



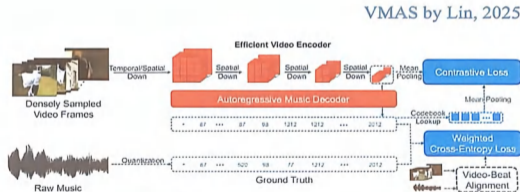
**Manual & Costly Challenges**

# :: Background and Motivation

- ▶ Video-to-Music (V2M) seeks to automatically generate appropriate background music for videos. V2M generation presents a promising solution with broad applications in film, advertising, gaming, etc.
- ▶ But currently, it suffers from **two critical limitations**:
  - incomplete video details, leading to weak semantic alignment
  - inadequate temporal and rhythmic correspondence, particularly in achieving beat synchronization



GVMGen by Zuo, 2025



VMAS by Lin, 2025

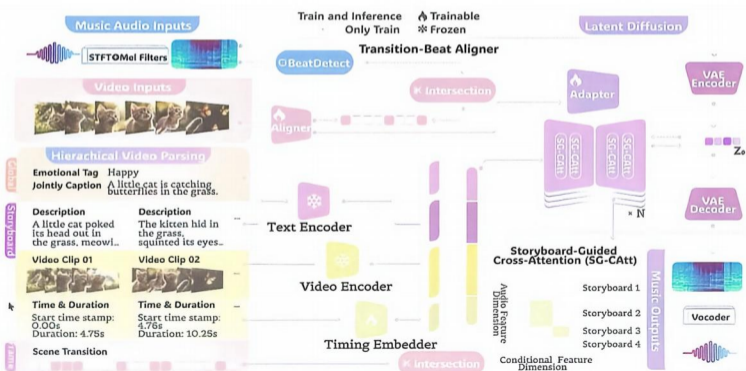
[1] Lin, Y.-B.; Tian, Y.; Yang, L.; Bertasius, G.; and Wang, H. 2024. VMAS: Video-to-Music Generation via Semantic Alignment in Web Music Videos. *EEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2025.

[2] Zuo, H.; You, W.; et al. GVMGen: A General Video-to-Music Generation Model With Hierarchical Attentions. *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 39. No. 21. 2025.

# Methods

## How to achieve the semantic, temporal, and rhythmic alignment for Video-to-Music Generation?

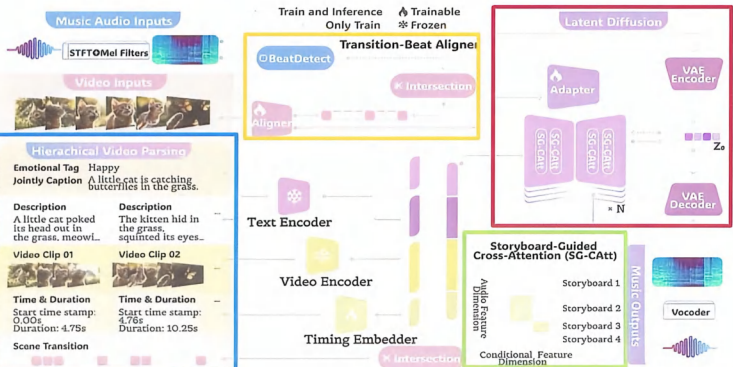
We employ a diffusion architecture grounded in video constraints, enabling the model to comprehend video content like a human composer. The generated music maintains both global thematic consistency and precise local storytelling, while realizing millisecond-level audio-visual synchronization.



# Methods

## How to achieve the semantic, temporal, and rhythmic alignment for Video-to-Music Generation?

We employ a diffusion architecture grounded in video constraints, enabling the model to comprehend video content like a human composer. The generated music maintains both global thematic consistency and precise local storytelling, while realizing millisecond-level audio-visual synchronization.



❑ **Core Generation Architecture:** Multimodal-Conditioned Latent Music Diffusion

❑ **Music Conductor:** Hierarchical Video Parsing

❑ **Multimodal Alignment:** Storyboard-Guided Cross-Attention (SG-CAtt)

❑ **Rhythmic Synchronization:** Frame-Level Transition-Beat Aligner and Adapter (TB-As)

# ::: Methods

## ① Method

**Latent Diffusion Model** as the backbone to perform the diffusion process in latent space, utilizing a VAE to encode audio Mel-spectrograms into latent vectors

### Music Conductor: Hierarchical Video Parsing

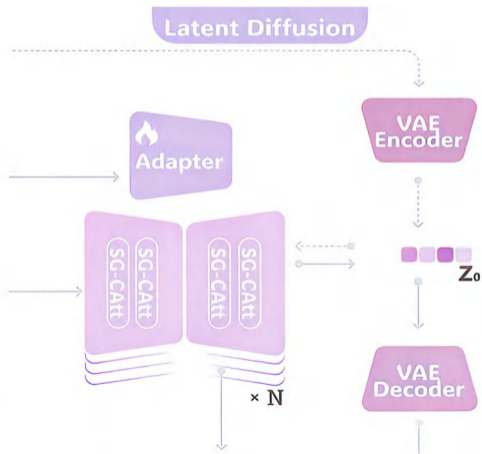
- Global Level: Extracts the overall emotional tone and core thematic description
- Storyboard Level: Adopting a "film script" perspective, the video is segmented into coherent shot groups
- Frame Level: Precisely capture the frame-level timestamps of visual cuts (scene changes)

### Storyboard-Guided Cross-Attention (SG-CAtt)

- SG-CAtt integrates global themes with local storyboard details, ensuring consistent global tonality while achieving precise local narrative synchronization

### Transition-Beat Aligner and Adapter (TB-As)

- Aligner: predicts the intersection timestamps of transitions and beats to determine which visual cuts require accentuation
- Adapter: modulates the music for accents or variations at transition moments, achieving frame-level synchronization



# ::: Methods

## ① Method

**Latent Diffusion Model** as the backbone to perform the diffusion process in latent space, utilizing a VAE to encode audio Mel-spectrograms into latent vectors

### Music Conductor: Hierarchical Video Parsing

- **Global Level:** Extracts the overall emotional tone and core thematic description
- **Storyboard Level:** Adopting a "film script" perspective, the video is segmented into coherent shot groups
- **Frame Level:** Precisely capture the frame-level timestamps of visual cuts (scene changes)

### Storyboard-Guided Cross-Attention (SG-CAtt)

- SG-CAtt integrates global themes with local storyboard details, ensuring consistent global tonality while achieving precise local narrative synchronization

### Transition-Beat Aligner and Adapter (TB-As)

- **Aligner:** predicts the intersection timestamps of transitions and beats to determine which visual cuts require accentuation
- **Adapter:** modulates the music for accents or variations at transition moments, achieving frame-level synchronization

## Hierarchical Video Parsing

Global

**Emotional Tag** Happy  
**Jointly Caption** A little cat is catching butterflies in the grass.

Storyboard

Description	Description
A little cat poked its head out in the grass, meowi...	The kitten hid in the grass, squinted its eyes...

### Video Clip 01



### Video Clip 02



### Time & Duration

Start time stamp: 0.00s  
Duration: 4.75s

### Time & Duration

Start time stamp: 4.76s  
Duration: 10.25s

Frame

### Scene Transition



## ① Method

**Latent Diffusion Model** as the backbone to perform the diffusion process in latent space, utilizing a VAE to encode audio Mel-spectrograms into latent vectors

### Music Conductor: Hierarchical Video Parsing

- Global Level: Extracts the overall emotional tone and core thematic description
- Storyboard Level: Adopting a "film script" perspective, the video is segmented into coherent shot groups
- Frame Level: Precisely capture the frame-level timestamps of visual cuts (scene changes)

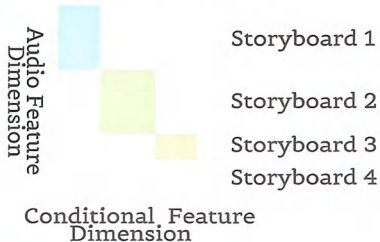
### Storyboard-Guided Cross-Attention (SG-CAtt)

- SG-CAtt integrates global themes with local storyboard details, ensuring consistent global tonality while achieving precise local narrative synchronization

### Transition-Beat Aligner and Adapter (TB-As)

- Aligner: predicts the intersection timestamps of transitions and beats to determine which visual cuts require accentuation
- Adapter: modulates the music for accents or variations at transition moments, achieving frame-level synchronization

## Storyboard-Guided Cross-Attention (SG-CAtt)



# ::: Methods

## ① Method

**Latent Diffusion Model** as the backbone to perform the diffusion process in latent space, utilizing a VAE to encode audio Mel-spectrograms into latent vectors

### Music Conductor: Hierarchical Video Parsing

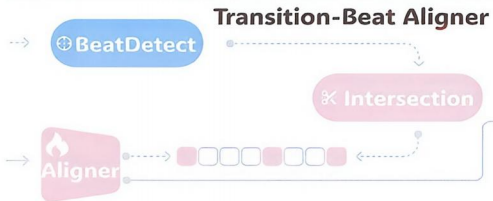
- Global Level: Extracts the overall emotional tone and core thematic description
- Storyboard Level: Adopting a "film script" perspective, the video is segmented into coherent shot groups
- Frame Level: Precisely capture the frame-level timestamps of visual cuts (scene changes)

### Storyboard-Guided Cross-Attention (SG-CAtt)

- SG-CAtt integrates global themes with local storyboard details, ensuring consistent global tonality while achieving precise local narrative synchronization

### Transition-Beat Aligner and Adapter (TB-As)

- **Aligner**: predicts the intersection timestamps of transitions and beats to determine which visual cuts require accentuation
- **Adapter**: modulates the music for accents or variations at transition moments, achieving frame-level synchronization



# :: Methods

## ① Method

**Latent Diffusion Model** as the backbone to perform the diffusion process in latent space, utilizing a VAE to encode audio Mel-spectrograms into latent vectors

### Music Conductor: Hierarchical Video Parsing

- Global Level: Extracts the overall emotional tone and core thematic description
- Storyboard Level: Adopting a "film script" perspective, the video is segmented into coherent shot groups
- Frame Level: Precisely capture the frame-level timestamps of visual cuts (scene changes)

### Storyboard-Guided Cross-Attention (SG-CAtt)

- SG-CAtt integrates global themes with local storyboard details, ensuring consistent global tonality while achieving precise local narrative synchronization

### Transition-Beat Aligner and Adapter (TB-As)

- Aligner: predicts the intersection timestamps of transitions and beats to determine which visual cuts require accentuation
- Adapter: modulates the music for accents or variations at transition moments, achieving frame-level synchronization

## ② Data and Training

Constructed **TB-Match**, a high-quality dataset with strict rhythmic synchronization

### Data: TB-Match Dataset

- ✓ Data Source: Curated ~18,000 rhythmically consistent samples from e-commerce ads and short videos.
- ✓ Cleaning Process: Employed a dual-stage framework (algorithmic + manual) involving 20 music conservatory students to ensure high aesthetic quality.
- ✓ Significance: Establishes a high-standard benchmark for fine-grained audio-visual alignment research.

### Training

- ✓ Stage 1: Pre-train VAE and Aligner to ensure audio reconstruction and rhythmic extraction quality.
- ✓ Stage 2: Train the backbone and SG-CAtt (with frozen visual encoders) to focus on semantic and temporal understanding.
- ✓ Stage 3: Jointly fine-tune the Adapter and backbone to refine precise rhythmic synchronization.

	Au.	Vd.	IS $\uparrow$	FAD $\downarrow$	KLD $\downarrow$	CLAP $\uparrow$	LB $\uparrow$	tw-CLAP $\uparrow$	tw-LB $\uparrow$	B $_{Iol\uparrow}$	TB $_{Iol\uparrow}$
GroundTruth			-	-	-	0.247	0.928	0.252	0.932	1.000	0.559
CMT	×	✓	1.131	7.151	5.540	0.109	0.728	0.113	0.775	0.254	0.213
Diff-BGM	×	✓	1.173	6.940	4.870	0.112	0.781	0.109	0.792	0.227	0.261
M <sup>2</sup> UGen	✓	×	1.211	5.902	3.350	0.158	0.892	0.163	0.893	0.307	0.331
VidMuse	✓	✓	1.206	7.437	4.210	0.102	0.704	0.103	0.718	0.335	0.352
GVMGen	✓	✓	1.227	6.137	3.210	0.212	0.899	0.219	0.917	0.465	0.357
Ours	✓	✓	<b>1.263</b>	<b>4.043</b>	<b>3.160</b>	<b>0.244</b>	<b>0.930</b>	<b>0.249</b>	<b>0.935</b>	<b>0.594</b>	<b>0.364</b>

## ① Quantitative

**Quantitative evaluation** demonstrates VeM consistently outperforms 5 SOTA baselines across all 9 metrics.

- **Audio Quality:** VeM achieves the lowest FAD and highest IS, indicating superior audio fidelity compared to both MIDI-based and audio-based methods.
- **Semantic Alignment:** The highest CLAP and LB scores show the hierarchical parsing effectively captures global and local semantics.
- **Rhythmic Precision:** On the proposed fine-grained metrics, VeM surpasses the best baseline, validating the effectiveness in synchronizing visual cuts with music beats.

## ② Qualitative

**Qualitative evaluation** with 50 participants, including 30 experts (musicians and film professionals), reveal a strong preference for our method.

- **Preference Rate:** VeM achieves a Top-1 preference rate in both expert and non-expert groups, significantly outperforming the second-best model.
- **MOS Scores:** VeM ranks highest for both Music Quality (MOS-Q) and Video-Music Alignment (MOS-A), demonstrating that our generated soundtracks are perceived as both musically pleasing and narratively coherent by human evaluators.

# :: Results

	Preference Rate		Preference Score			
	Top-1		MOS-Q		MOS-A	
	Expert	Non-expert	Expert	Non-expert	Expert	Non-expert
CMT(Di et al. 2021)	3.625%	2.000%	5.622 <sub>+0.213</sub>	6.139 <sub>+0.329</sub>	4.680 <sub>+0.247</sub>	4.924 <sub>+0.189</sub>
Diff-BGM(Li et al. 2024c)	2.250%	2.125%	5.406 <sub>+0.185</sub>	5.935 <sub>+0.314</sub>	4.387 <sub>+0.243</sub>	4.530 <sub>+0.212</sub>
M <sup>2</sup> UGen(Liu et al. 2023)	5.375%	5.125%	5.340 <sub>+0.162</sub>	5.863 <sub>+0.307</sub>	5.814 <sub>+0.221</sub>	6.127 <sub>+0.205</sub>
VidMuse(Tian et al. 2024)	4.250%	2.750%	4.767 <sub>+0.234</sub>	4.992 <sub>+0.128</sub>	5.467 <sub>+0.229</sub>	5.270 <sub>+0.210</sub>
GVMGen(Zuo et al. 2025)	11.125%	10.125%	5.418 <sub>+0.223</sub>	5.693 <sub>+0.262</sub>	6.467 <sub>+0.197</sub>	6.374 <sub>+0.251</sub>
Ours	<b>73.375%</b>	<b>77.875%</b>	<b>6.892</b> <sub>+0.173</sub>	<b>7.537</b> <sub>+0.195</sub>	<b>7.341</b> <sub>+0.174</sub>	<b>7.852</b> <sub>+0.260</sub>

## ① Quantitative

**Quantitative evaluation** demonstrates VeM consistently outperforms 5 SOTA baselines across all 9 metrics.

- **Audio Quality:** VeM achieves the lowest FAD and highest IS, indicating superior audio fidelity compared to both MIDI-based and audio-based methods.
- **Semantic Alignment:** The highest CLAP and LB scores show the hierarchical parsing effectively captures global and local semantics.
- **Rhythmic Precision:** On the proposed fine-grained metrics, VeM surpasses the best baseline, validating the effectiveness in synchronizing visual cuts with music beats.

## ② Qualitative

**Qualitative evaluation** with 50 participants, including 30 experts (musicians and film professionals), reveal a strong preference for our method.

- **Preference Rate:** VeM achieves a Top-1 preference rate in both expert and non-expert groups, significantly outperforming the second-best model.
- **MOS Scores:** VeM ranks highest for both Music Quality (MOS-Q) and Video-Music Alignment (MOS-A), demonstrating that our generated soundtracks are perceived as both musically pleasing and narratively coherent by human evaluators.

# :: Results

## ■ Universality evaluation

Universality evaluation on external domains shows that VeM consistently maintains high performance.

	SymMV		Sora		Others	
	LB $\uparrow$	TB $_{Tot}$ $\uparrow$	LB $\uparrow$	TB $_{Tot}$ $\uparrow$	LB $\uparrow$	TB $_{Tot}$ $\uparrow$
CMT	0.912	0.314	0.758	0.671	0.578	0.337
Diff	0.643	0.253	0.898	0.667	0.589	0.325
M <sup>2</sup> U	0.925	0.296	1.029	0.725	0.885	0.332
Vid	0.787	0.312	0.982	0.785	0.670	0.400
GVM	0.910	0.260	1.084	0.814	0.887	0.391
Ours	<b>0.989</b>	<b>0.331</b>	<b>1.106</b>	<b>0.829</b>	<b>0.895</b>	<b>0.453</b>

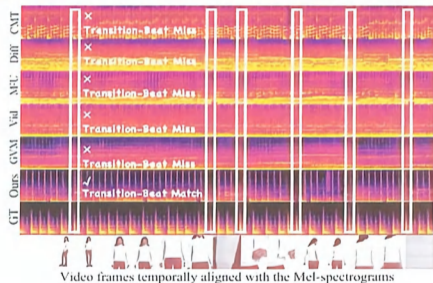
## ■ Ablation Study

Ablation studies confirm the individual contributions of each component.

HVP-Cond	SG-CAtt	TB-As	IS $\uparrow$	FAD $\downarrow$	KLD $\downarrow$	CLAP $\uparrow$	LB $\uparrow$	tw-CLAP $\uparrow$	tw-LB $\uparrow$	B $_{Tot}$ $\uparrow$	TB $_{Tot}$ $\uparrow$
×	×	×	0.823	6.692	4.714	0.180	0.624	0.188	0.625	0.221	0.197
×	×	✓	0.772	7.217	5.097	0.172	0.639	0.181	0.643	0.433	0.283
✓	✓	×	1.191	4.382	3.608	0.231	0.890	0.236	0.882	0.403	0.265
✓	×	×	1.140	5.712	3.869	0.218	0.735	0.227	0.742	0.383	0.220
✓	✓	✓	<b>1.263</b>	<b>4.043</b>	<b>3.160</b>	<b>0.244</b>	<b>0.930</b>	<b>0.249</b>	<b>0.935</b>	<b>0.594</b>	<b>0.364</b>

## ■ Visual Comparisons

Mel-spectrogram Visualization highlights the rhythmic precision of VeM that coincides perfectly with music beats.



# :: Contributions

**The main contributions are claimed as follows:**

- ▶ **Hierarchical Video Parsing as Conductor:** We propose a novel strategy that utilizes hierarchical video parsing (encompassing Global, Storyboard, and Frame levels) as a "music conductor" to comprehensively orchestrate multimodal constraints for generation.
- ▶ **Unified Alignment Framework:** We design a diffusion-based architecture that integrates a Storyboard-Guided Cross-Attention (SG-CAtt) mechanism to ensure semantic-temporal coherence, and a Transition-Beat Aligner and Adapter (TB-As) to strictly synchronize visual scene transitions with musical beats.
- ▶ **Dataset and Metrics:** We introduce TB-Match, a dataset comprising 18k video-music pairs with high rhythmic correlation (sourced from ads and short-form videos), and propose novel metrics, specifically tailored to evaluate fine-grained rhythmic precision. Extensive experiments demonstrate the superiority.

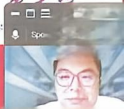
# CLaMP 3

ACL 2025  
VIENNA

## Universal Music Information Retrieval Across Unaligned Modalities and Unseen Languages

Shangda Wu<sup>1</sup> Zhancheng Guo<sup>1</sup> Ruibin Yuan<sup>2</sup> Junyan Jiang<sup>3,4</sup> Seungheon Doh<sup>5</sup>  
Gus Xia<sup>3,4</sup> Juhan Nam<sup>5</sup> Xiaobing Li<sup>1</sup> Feng Yu<sup>1</sup> Maosong Sun<sup>1,6</sup>

- <sup>1</sup> Central Conservatory of Music
- <sup>2</sup> Hong Kong University of Science and Technology
- <sup>3</sup> New York University Shanghai
- <sup>4</sup> Mohamed bin Zayed University of Artificial Intelligence
- <sup>5</sup> Korea Advanced Institute of Science and Technology
- <sup>6</sup> Tsinghua University



# What is Music Information Retrieval (MIR)?



**Goal**

Automatic analysis and retrieval of music

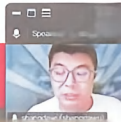
Classification

Joy Anger Sadness Calmness

Retrieval

One of the better known nocturnes, this piece has a rhythmic freedom that came to characterize Chopin's later work.

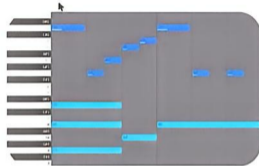
Recommendation



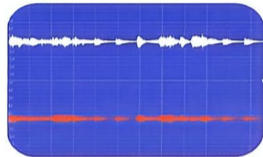
# Why Multimodal MIR?



**Sheet Music  
(ABC Notation)**  
*Composition & Analysis*



**Performance Signals  
(MIDI)**  
*Digital Music Editing*



**Audio  
(Waveform)**  
*Listening & Distribution*



# What is Missing in MIR Systems?

- **Modality Limitation**  
Focused solely on audio; neglects sheet music and performance signals
- **Language Restriction**  
Supports only English queries; lacks multilingual capability
- **Limited Semantic Depth**  
Training data often contain only surface-level tags (e.g., genre, instrumentation), failing to capture deeper musical semantics
- **Examples of Existing MIR Systems**  
MusCALL, MuLan, CLAP, TTMR



# CLaMP 3: A Universal MIR Framework

## Modalities

Supports all major music modalities

→ Text bridges modalities for **emergent alignment**

## Languages

Trained on 27, generalizes to **100 languages**

→ Robust cross-lingual retrieval

## Training Data

M4-RAG (**2.31M music-text pairs**)

→ Web-retrieved, LLM-annotated, multilingual, multi-modal

## Performance

**SOTA** on multimodal & multilingual MIR

→ Best across all modalities and languages

Performance Signal



Sheet Music



Audio Recording



# Core Components

## Multilingual Text Encoder

Uses **XLM-R base** pre-trained on 2.5TB multilingual corpus (100 languages)

→ 12-layer transformer, 768 hidden

→ Strong cross-lingual generalization to unseen languages

## Symbolic Music Encoder

Uses **M3**: self-supervised model for ABC (sheet music) & MIDI (performance signals)

→ Encoded with **Interleaved ABC** and **MTF (MIDI Text Format)**

→ One bar (ABC) or one message (MIDI) = one patch

→ 12-layer transformer, 768 hidden, up to 512 patches / 32k characters

## Audio Music Encoder

12-layer transformer trained from scratch

→ Uses **frozen MERT-v1-95M** for audio features

→ One **5s audio clip** = one embedding (averaged across layers & time)

→ Up to **128 clips (640s)** per input



# Multi-Stage Alignment Strategy

**Naïve approach:** Align all modalities at once

**Problem:** Conflicting distributions caused unstable and ineffective alignment

**Our solution:** **Text as a bridge** between modalities (inspired by *ImageBind*):

- **Stage 1:** Align text ↔ one music modality (text trainable)
- **Stage 2:** Align text ↔ another modality, freeze text encoder
- **Stage 3:** Unfreeze text, refine alignment with new modality
- **Stage 4:** Re-freeze text, re-align with the first modality



**Why this matters:**

- Prevents **representation drift**
- Reduces **interference** between music modalities

# M4-RAG Curation

## Data Sources:

- WebMusicText (1.4M ABC)
- Million MIDI Dataset (1.5M MIDI)
- Web audio (1.8M tracks, ~160K hours)

## Pipeline:

1. *Title Filtering* – remove entries without valid song titles
2. *Web Search* – search top 10 Google results
3. *RAG* – generate annotations with Qwen2.5-72B
4. *Quality Filtering* – remove low-quality or incomplete entries
5. *Postprocessing* – clean fields, verify language with fastText
6. *Multilingual Translation* – translate long-form text into a second language

Table 1: Metadata overview for M4-RAG, grouped into basic information, annotations, and translations. In *Annotations*, *Region* and *Language* are written in English; other fields follow the *Language* specification.

Category	Field	Content	Avg Bytes
<i>Basic</i>	<i>Title</i>	Music Title	20.04
	<i>Artists</i>	Artist names	21.97
	<i>Region</i>	Country of origin	20.69
<i>Annotations</i>	<i>Language</i>	Document language	7.02
	<i>Genres</i>	Genre list	21.83
	<i>Tags</i>	Keywords/playlists	51.91
	<i>Background</i>	Background context	531.79
	<i>Analysis</i>	Musical analysis	770.29
	<i>Description</i>	General description	591.86
	<i>Scene</i>	Scene depiction	750.92
<i>Translations</i>	<i>Language</i>	Translation language	6.38
	<i>Background</i>	Translated background	819.76
	<i>Analysis</i>	Translated analysis	1130.47
	<i>Description</i>	Translated description	888.86
	<i>Scene</i>	Translated scene	1077.07



# M4-RAG Statistics

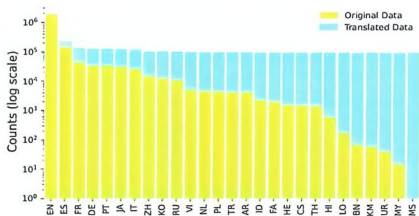


Figure 3: Language distribution of original and translated entries in M4-RAG, covering 27 languages.

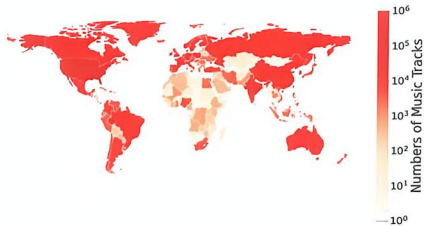


Figure 4: Country-wise distribution of music tracks in M4-RAG, spanning 194 countries.

**2.31M music–text pairs in total**

→ *ABC–text*: 0.58M (from WebMusicText)

→ *MIDI–text*: 0.17M (from Million MIDI Dataset)

→ *Audio–text*: 1.56M (from web sources)

# English Text-to-Music Retrieval

Table 2: Results for English text-to-music retrieval on several benchmarks: WikiMT and MidiCaps have 1,010 pairs, Song Describer Dataset (SDD) has 706 audio and 1,106 captions, and MusicCaps-Remake (MC-R) contains 2,777 pairs. MC-R prevents data leakage by using full-length audio and rewritten captions from AudioSet's evaluation set.

Model	Symbolic Benchmarks		WikiMT-X (Sheet Music)			
	WikiMT	MidiCaps	Background	Analysis	Description	Scene
CLaMP	0.2561	0.1236	0.2122	0.1345	0.0306	0.0426
CLaMP 2	0.3438	0.2695	0.3024	0.2374	0.0418	0.0838
CLaMP 3 <sup>c2</sup> <sub>sa</sub>	<b>0.4498</b>	<b>0.2826</b>	<b>0.4028</b>	<b>0.3382</b>	0.0835	<b>0.1512</b>
CLaMP 3 <sub>saas</sub>	0.3555	0.1798	0.3301	0.2758	<b>0.1274</b>	0.1500

Model	Audio Benchmarks		WikiMT-X (Audio)			
	SDD	MC-R	Background	Analysis	Description	Scene
CLAP	0.1310	0.0657	0.0598	0.0429	0.0318	0.0218
TTMR++	0.1437	<b>0.1248</b>	0.1119	0.0833	0.0584	0.0301
CLaMP 3 <sup>c2</sup> <sub>sa</sub>	0.1612	0.0959	0.1180	0.1206	0.0639	0.0619
CLaMP 3 <sub>saas</sub>	<b>0.1985</b>	0.1177	<b>0.2017</b>	<b>0.1711</b>	<b>0.0988</b>	<b>0.0963</b>

- CLaMP 3 outperforms all baselines on both **symbolic** and **audio retrieval**
- Strong performance even on **abstract texts** (e.g., Scene, Description in WikiMT-X)
- Gains attributed to **multi-stage training** and **diverse M4-RAG annotations**



# Multilingual Text-to-Music Retrieval

Table 3: Results for multilingual text-to-music retrieval on translated WikiMT-X background annotations. Languages marked with asterisks were not included in the M4-RAG training data. The BLEU scores below each language are calculated by back-translating the text with the SeamlessM4T model and comparing it to the original English text.

Model	ru	fr	es	ar	zh	fi*	el*	ta*	kk*	am*
<b>ABC Notation</b>										
<i>CLaMP 2</i>	0.2668	0.2968	0.2934	0.2298	0.1646	0.2795	0.2410	0.0915	0.2543	0.1237
<i>CLaMP 3<sup>ca</sup></i>	<b>0.3614</b>	<b>0.3949</b>	<b>0.3921</b>	<b>0.3155</b>	<b>0.2373</b>	<b>0.3524</b>	<b>0.3226</b>	<b>0.1415</b>	<b>0.3397</b>	<b>0.1871</b>
<i>CLaMP 3<sub>saas</sub></i>	0.2918	0.3214	0.3239	0.2789	0.2358	0.2919	0.2681	0.1246	0.2703	0.1139
<b>MIDI</b>										
<i>CLaMP 2</i>	0.1271	0.1414	0.1452	0.1113	0.0749	0.1438	0.1087	0.0466	0.1079	0.0616
<i>CLaMP 3<sup>ca</sup></i>	<b>0.1921</b>	<b>0.2101</b>	<b>0.2137</b>	<b>0.1681</b>	<b>0.1316</b>	<b>0.2019</b>	<b>0.1702</b>	<b>0.0804</b>	<b>0.1765</b>	<b>0.1039</b>
<i>CLaMP 3<sub>saas</sub></i>	0.1165	0.1319	0.1330	0.1141	0.0937	0.1245	0.1143	0.0601	0.1104	0.0544
<b>Audio</b>										
<i>CLaMP 3<sup>ca</sup></i>	0.1068	0.1150	0.1202	0.0981	0.0877	0.1112	0.1014	0.0720	0.1005	<b>0.0681</b>
<i>CLaMP 3<sub>saas</sub></i>	<b>0.1788</b>	<b>0.1980</b>	<b>0.1962</b>	<b>0.1665</b>	<b>0.1459</b>	<b>0.1770</b>	<b>0.1736</b>	<b>0.0945</b>	<b>0.1561</b>	0.0675

- Translated WikiMT-X background into **10 languages**:
  - 5 included in M4-RAG, 5 not covered (diverse scripts, regions, families)
- CLaMP 3 demonstrates **robust cross-language generalization**
  - Across ABC notation, MIDI, and audio retrieval tasks

# Emergent Cross-Modal Retrieval

- CLaMP 3 supports retrieval across symbolic and audio modalities — **no explicit alignment** required
- Outperforms CLaMP 2, which is limited to symbolic-only retrieval
- All directions outperform random baseline → showing **true cross-modal generalization**
- Room remains to close the performance gap in **symbolic ↔ audio retrieval**

Table 4: Results for emergent cross-modal retrieval on WikiMT-X pairings across different musical modalities. **S**: Sheet Music (ABC notation), **P**: Performance Signals (MIDI, converted from ABC), **A**: Audio recordings.

Model	S→P	S→A	P→S	P→A	A→S	A→P
CLaMP 2	<b>0.5138</b>	-	0.4480	-	-	-
CLaMP 3 <sub>sa</sub> <sup>c2</sup>	0.4547	0.0543	<b>0.5293</b>	0.0313	<b>0.0492</b>	<b>0.0383</b>
CLaMP 3 <sub>saas</sub>	0.3262	<b>0.0578</b>	0.3146	<b>0.0397</b>	0.0410	0.0303

# Conclusions

- CLaMP 3 unifies sheet music, MIDI, audio, and multilingual text in a **shared representation space**
- Shows robust generalization across modalities and languages, including **unaligned and unseen** cases
- Introduces M4-RAG (**2.31M music–text pairs**, 27 languages, 194 countries)
- Sets a new standard for **multimodal and multilingual MIR**, enabling future research in retrieval and generation



# Thank You!



**GitHub Code**



**arXiv Paper**



**M4-RAG Dataset**

