IteraSong：大语言模型引导的网页数字音频工作站中的迭代式符号音乐创作

Di Lu[1,2,3], XinranZhang[1,2,3], Yan Gao[1,2,3],Yuan Zhang[1,2,3],Xiaoqing Wang[1,2,3],Qingwen Zhou[1,2,3], Xiaobing Li[1,2,3],Maosong Sun[1,2,3,4*]

[1]中央音乐学院人工智能音乐与音乐信息技术系 [2]中央音乐学院音乐与脑科学重点实验室，教育部，中国
[3]中央音乐学院人工智能音乐实验室，教育部哲学与社会科学实验室，中国[4]清华大学计算机科学与技术系
ld@ccom.edu.cn，sms@tsinghua.edu.cn

arXiv:2611.29225v1 [cs.SD] 22 Nov 2026

摘要

音乐创作本质上是迭代式的，然而许多符号生成系统仍然主要着眼于生成或重新生成孤立的输出。我们提出 **IteraSong**，这是一个基于网页的交互式数字音频工作站（DAW），用于大语言模型引导的迭代式符号音乐创作。不同于一次性生成系统，IteraSong 支持在持久化的多轨项目上进行生成—试听—编辑—完善循环。用户首先初始化一个带有完整标注的流行音乐草稿，然后通过自然语言指令逐步完善。大语言模型将请求映射到由模型上下文协议（MCP）工具提供的分层编辑原语动作空间，涵盖全局、音轨、段落以及小节／音符范围。该演示支持参会者使用自然语言编辑多轨作品、应用确定性编辑和生成式编辑，并实时撤销/重做更改。

视频— https://youtu.be/XZaHHms5Kl0

引言

符号音乐生成发展迅速，现已能够按照明确的音乐结构生成可控的多音轨作品（von Rütte 等，2023；Pasquier 等，2025）。近期基于大语言模型的方法进一步表明，语言模型能够理解符号音乐，并支持音乐相关推理与生成（Yuan 等，2024；Yu 等，2023）。然而，实际的音乐创作很少一次完成：创作者通常会经过多轮聆听、修改、比较和完善草稿。因此，我们需要能够支持围绕不断演进的音乐项目进行迭代创作的交互式工具，而不是将生成视为创作的最后一步。

现有的交互式符号音乐系统主要聚焦于局部补全或结构化作曲（Malandro 2023, 2024; Hu et al. 2025; He, Xia, and Wang 2025），而基于大语言模型的音乐系统则探索自然语言交互和工具编排（Yuan et al. 2024; Yu et al. 2023）。然而，迭代创作不仅需要孤立的生成或工具调用：每次编辑都应基于当前项目，仅修改预期的音乐内容，并允许修改不满意的变更或将其撤销。鲜有现有系统能将这些能力整合到统一的项目级工作流中，以支持多轮自然语言符号音乐编辑。

我们推出了**IteraSong**，这是一款用于迭代式符号音乐创作的网页数字音频工作站。我们的主要贡献如下：

- 一个以项目为中心的交互系统，通过自然语言支持多轮、项目级的符号音乐编辑。
- 通过兼容 MCP 的工具提供由编辑原语构成的分层动作空间，支持从全局结构到单个音符的精准、限定范围编辑。
- 一个端到端工作流，整合初稿生成、大语言模型引导的编辑和浏览器内播放，专为实时动手交互而设计。

系统设计

图 1 展示了 IteraSong 工作流。首先，轻量级流行音乐生成器生成一个带有完整标签的多音轨初稿，该初稿随后被规范化为 SongProject，并在迭代创作过程中作为持久化的项目状态加以维护。

在之后的每一轮交互中，用户都会提交自然语言编辑请求。大语言模型解读该请求，选择一个或多个编辑原语，并通过兼容 MCP 的工具调用它们（Hou 等，2026）。

成功执行后，更新后的项目会立即在网页数字音频工作站中提供试听。用户可以继续编辑、请求另一个变体，或撤销之前的更改。每次接受的编辑都会成为下一轮交互的起始状态。

初始草稿提供器。 IteraSong 从可编辑的符号音乐草稿开始，而不是从空项目开始。当前原型使用轻量级流行音乐生成器作为可控制的测试平台，同时，该框架也可以接受来自更先进的符号音乐生成模型的草稿。该生成器提供全局属性（例如调性、调式、速度）、音轨的标识与角色、乐段边界、单个音符事件以及和弦进行。由此产生的

因此，SongProject 会公开明确的音乐对象，以便在网页数字音频工作站中进行可视化，并在后续编辑中进行操作。

*通讯作者。版权所有 © 2027，人工智能促进协会 (www.aaai.org)。保留所有权利。

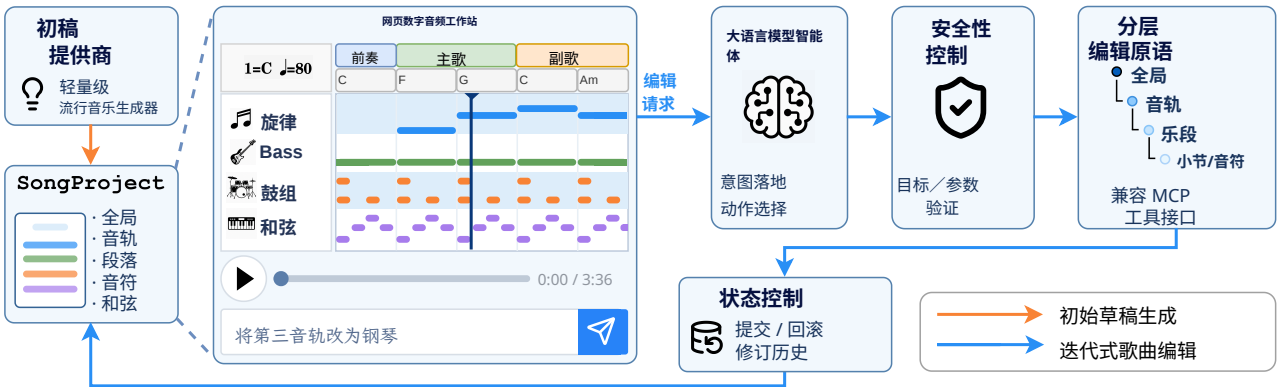


图 1：IteraSong 在网页数字音频工作站中进行初始草稿生成和迭代式歌曲编辑的整体架构。

大语言模型智能体。 大语言模型并不直接操作原始的符号化歌曲数据，而是充当意图解释器和动作选择器。收到自然语言请求后，它会识别预期的音乐目标和操作，然后调用相应的兼容 MCP 工具。例如，“将第三音轨改为钢琴”会映射到确定性原语集_instrument，而“重新生成副歌 2 中的贝斯”则会映射到乐段范围的生成式原语 regenerate_section。

安全控制。 在执行任何突变之前，安全模块会检查请求的目标是否存在且明确无歧义、参数是否有效，以及请求的操作是否受支持。未通过这些检查的请求将被拒绝，且不会修改项目。

分层编辑原语。 IteraSong 提供一组有界的音乐领域原语，按全局、音轨、乐段和小节/音符级别组织。动作空间包括观察原语和 DAW 控制原语，以及确定性和作用范围明确的生成式编辑原语。观察原语用于获取项目、音轨、结构、和弦及音符信息；DAW 控制则提供播放、停止、定位、循环、静音和独奏操作。确定性和生成式原语总结于表 1。

级别	确定性	生成式
全局	设置_速度，设置_调性_调式，歌曲_移调	更改_开头_策略，更改_结尾_策略
音轨	设置_乐器，添加_音轨，清除音轨_	重新生成_音轨
乐段	设置_和弦，清除_乐段，移调_乐段	重新生成_乐段，重新生成_乐段和声编配_
小节/音符	添加_音符，删除_音符，移调_音符	重新生成_音符

表 1：分层编辑原语。

状态控制。 状态控制负责管理交互轮次中不断演进的项目。每次符号突变都会以事务方式应用于当前项目修订版本：成功的编辑会创建新修订版本和历史记录，而失败的操作则会使先前状态保持不变。此机制还支持撤销和回滚，使后续交互轮次能够从有效且明确选定的项目状态继续进行。

演示场景

演示展示的是一个连续的创作会话，而非孤立的工具调用。用户首先只需点击一次，即可生成一首带有段落、和弦和乐器轨道标签的多轨流行歌曲。试听初稿后，用户可以连续发出自然语言请求，更换乐器、重新生成副歌 2 中的伴奏、请求另一种变体、撤销/重做编辑，或修正单个音符。每次接受的编辑都会以可视化形式呈现并供用户试听，同时保留同一项目以供后续请求使用。用户也可以通过修订历史查看先前版本。

实现与验证

实现。 IteraSong 使用基于规则的流行音乐生成器，进行初始草稿生成和生成式编辑。网页数字音频工作站使用 Node.js 后端实现符号化可视化、音符级交互和自然语言聊天。浏览器端的 MIDI 播放采用编译为 WebAssembly (WASM) 的 FluidSynth。Qwen-Max 作为大语言模型智能体，通过兼容 MCP 的工具调用观察、DAW 控制、确定性和生成式基本操作。

验证。 我们在五个 SongProject 和四种工作流中，对 20 场十轮会话（共 200 个回合）进行了评估，涵盖中文和英语请求。使用 Qwen-Max 和实际的 MCP 执行路径，IteraSong 的任务完成率达到 90.63%（145/160），安全拒绝率达到 100%（40/40），状态一致性达到 100%（200/200）。

结论

IteraSong 通过结构化项目状态、分层编辑原语以及网页数字音频工作站中的自然语言交互，实现迭代式符号音乐创作。当前原型专注于流行音乐，并采用基于规则的生成后端。感知质量和用户体验仍有待后续评估。未来工作还将扩展动作空间，并集成更强大的生成式后端。

致谢

本研究获得国家自然科学基金专项项目（项目编号：T2341003）、北京市属高校高精尖学科建设项目以及中央音乐学院一般科研项目（项目编号：25YB14）的资助。感谢各位同事就本研究提出的宝贵讨论和反馈。

参考文献

He, Q.; Xia, G.; and Wang, Z. 2025. TOMI：面向具有完整歌曲结构的多轨作曲的音乐构思转换与组织。载于第*26*届国际音乐信息检索学会会议论文集，337–345。Hou, X.; Zhao, Y.; Wang, S.; and Wang, H. 2026. 模型上下文协议（MCP）：现状、安全威胁与未来研究方向。《ACM软件工程与方法学汇刊》，35(10)。Hu, Z.; Liu, Y.; Chen, G.; and Yu, B. X. 2025. 与我共同作曲：用于符号音乐补全的协作式音乐修补模型。载于第三十九届*AAAI*人工智能会议、第三十七届人工智能创新应用会议及第十五届人工智能教育进展研讨会论文集，AAAI’25/IAAI’25/EAAI’25，1327–1335。AAAI出版社。ISBN 978-1-57735-897-8。Malandro, M. E. 2023. 作曲助手：用于多轨MIDI补全的交互式Transformer。载于第*24*届国际音乐信息检索学会会议论文集，327–334。Malandro, M. E. 2024. 作曲助手2：具有细粒度用户控制的交互式多轨MIDI补全。载于第*25*届国际音乐信息检索学会会议论文集，438–445。Pasquier, P.; Ens, J.; Fradet, N.; Triana, P.; Rizzotti, D.; Rolland, J.-B.; and Safi, M. 2025. MIDI-GPT：用于计算机辅助多轨音乐创作的可控生成模型。载于第三十九届*AAAI*人工智能会议、第三十七届人工智能创新应用会议及第十五届人工智能教育进展研讨会论文集，AAAI’25/IAAI’25/EAAI’25，1474–1482。AAAI出版社。ISBN 978-1-57735-897-8。von Rütte, D.; Biggio, L.; Kilcher, Y.; and Hofmann, T. 2023. FIGARO：使用学习特征和专家特征的可控音乐生成。载于第十一届国际学习表征会议，1–18。Yu, D.; Song, K.; Lu, P.; He, T.; Tan, X.; Ye, W.; Zhang, S.; and Bian, J. 2023. MusicAgent：使用大语言模型进行音乐理解与生成的人工智能智能体。载于*2023*年自然语言处理实证方法会议：系统演示论文集，246–255。计算语言学协会。Yuan, R.; Lin, H.; Wang, Y.; Tian, Z.; Wu, S.; Shen, T.; Zhang, G.; Wu, Y.; Liu, C.; Zhou, Z.; Xue, L.; Ma, Z.; Liu, Q.; Zheng, T.; Li, Y.; Ma, Y.; Liang, Y.; Chi, X.; Liu, R; Wang, Z.; Lin, C.; Liu, Q.; Jiang, T.; Huang, W.; Chen, W.; Fu, J.; Benetos, E.; Xia, G.; Dannenberg, R.; Xue, W.;Kang, S.; and Guo, Y. 2024. ChatMusician：使用大语言模型内在地理解和生成音乐。载于计算语言学协会会议论文集：*ACL 2024*，6252–6271。计算语言学协会。

附录

多轮评估详情

测试项目与会话设计。评估使用了来自共享生成流水线的 5 个符号流行音乐项目，涵盖 11–12 条音轨、14–16 个乐段和 6,241–8,536 个音符事件。每个项目均通过 4 种十回合工作流进行测试：轨道优先、段落优先、音符优先和混合编辑。最终的 20 个会话中，10 个使用中文，10 个使用英语。每个会话包含 1 次观察、7 个变更请求（包括编辑和撤销/重做），以及 2 个无效请求。Qwen-Max 通过 stdio MCP 客户端—服务器路径执行这些请求。整个会话期间项目状态均得到保留，未进行选择性重试或重置。

每会话结果。表 2 通过展示每个项目–工作流组合的完成次数，扩展了正文中的汇总验证结果。所有观察请求（20/20）和无效请求拒绝（40/40）均成功，而未成功的请求集中在 3 个会话中。

			成功 / 总计		
项目	工作流	回合	Obs.	Mut.	Rej.
项目 1	1	10	1/1	7/7	2/2
	2	10	1/1	7/7	2/2
	3	10	1/1	7/7	2/2
	4	10	1/1	7/7	2/2
项目 2	1	10	1/1	**1/7**	2/2
	2	10	1/1	7/7	2/2
	3	10	1/1	7/7	2/2
	4	10	1/1	7/7	2/2
项目 3	1	10	1/1	7/7	2/2
	2	10	1/1	**2/7**	2/2
	3	10	1/1	7/7	2/2
	4	10	1/1	**3/7**	2/2
项目4	1	10	1/1	7/7	2/2
	2	10	1/1	7/7	2/2
	3	10	1/1	7/7	2/2
	4	10	1/1	7/7	2/2
项目5	1	10	1/1	7/7	2/2
	2	10	1/1	7/7	2/2
	3	10	1/1	7/7	2/2
	4	10	1/1	7/7	2/2
总计（20 个会话）		**200**	**20/20**	**125/140**	**40/40**

表 2：多轮编辑评估的单会话结果。每一行代表在单个 SongProject 上进行的一次十轮会话。工作流 1–4 分别对应轨道优先、段落优先、音符优先和混合编辑。Obs.、Mut. 和 Rej. 分别表示观察请求、突变请求和无效请求拒绝。

具有代表性的十轮会话。 表 3 总结了一次录制的英文段落优先会话。它展示了对同一项目进行的连续编辑、撤销后的修订版本变化，以及请求被拒绝后版本保持不变的情况。为便于阅读，请求内容已作简化。

回合	请求	观察到的效果	Rev.
1	检查结构（只读）。	-	1
2	将副歌 2 及其和弦升高两个半音	316 音符 and 和弦已更改。	2
3	Undo	恢复了原始的音符和和声。	3
4	在副歌 2 中重新生成第3轨	108 个替换音符及其节奏已更改。	4
5	在第2轨上，将音高为33的音符 位于第2小节，第1拍改为34	1个音符已更改。	5
6	升高副歌99两个半音 已拒绝	不进行突变。	5
7	将第2轨升高 127 个半音	已拒绝 不进行突变。	5
8	设置速度为 100 BPM	已提交。	6
9	将第2轨升高一个半音	811 音符 转-置。	7
10	撤销音轨移调	已恢复先前的音乐内容。	8

表 3：一次录制的段落优先会话。修订内容显示在每个回合之后。

评分与失败分析。 评分要求每个回合都达到预期状态，且不得修改授权范围以外的内容。依赖回合还必须满足其前置条件。15 个未完成的突变回合中，3 个属于主要目标绑定错误，12 个属于下游依赖失败。在 3 个中文再生成请求中，智能体提供了乐段_8，而不是正确解析出的乐段_9，因此验证在突变发生前就拒绝了这些不匹配项。

B 网页数字音频工作站界面

图 2 展示了编辑过程中不断演变的 SongProject 如何呈现。音轨、乐段边界、和弦和音符始终清晰可见，同时还显示智能体的请求和执行轨迹，用户因此可以在试听当前编曲时检查项目的变化。

C 范围与局限性

本次功能评估基于固定的开发工作流开展。由于各项目共用同一生成后端，且每次会话中的回合彼此依赖，因此结果仅具描述性，不能证明其具有普遍性。。评估所用版本采用了限定范围再生成适配器。此外，本次评估未考察可听回放、感知音乐质量，也未评估专业音乐人的易用性。



图 2：多轮编辑会话中的 IteraSong Web 数字音频工作站。编曲窗口显示符号化项目内容，智能体面板则显示自然语言请求和 MCP 工具执行。