arXiv:2302.05393v1 [cs.SD] 10 Feb 2023

GTR-CTRL：基于 Transformer 的吉他导向音乐生成乐器与流派条件控制

Pedro Sarmento[1][0000−0002−4518−0194]、Adarsh Kumar[2],Yu-Hua Chen[3],CJ Carr[4],Zack Zukowski[4],和 Mathieu Barthet[1][0000−0002−9869−1668]

[1] 伦敦玛丽女王大学，英国
{p.p.sarmento,m.barthet}@qmul.ac.uk[2] 印度理工学院卡拉格布尔分校，印度[3] 国立台湾大学，台湾[4] Dadabots，https://dadabots.com/

摘要。 近年来，深度学习技术推动符号音乐生成不断取得进展。该领域的大多数研究聚焦于 MIDI表示，但较少关注使用吉他六线谱（六线谱）进行符号音乐生成，而六线谱可用于编码多种乐器。除节奏和音高外，六线谱还包含有品弦乐器的表现技法和指法信息。本研究使用 DadaGP 数据集进行吉他谱音乐生成；该语料库包含超过 2.6 万首以 GuitarPro 和词元格式表示的歌曲。我们介绍了多种方法，用于对 Transformer-XL 深度学习模型进行条件控制，使其根据指定的配器（inst-CTRL）和流派（genre-CTRL）生成吉他谱（GTR-CTRL）。训练语料库中每首歌曲的开头都附加了专门的控制标记。我们评估了模型在有无条件控制时的表现，并提出乐器出现度量指标，以评估 inst-CTRL 模型对给定配器提示的响应。我们训练了一个 BERT 模型用于下游流派分类，并用它评估 genre-CTRL 模型的生成结果。统计分析表明，经过条件控制与未经过条件控制的模型之间存在显著差异。总体而言，结果显示，与无条件控制的模型相比，GTR-CTRL 方法为吉他导向的符号音乐生成提供了更大的灵活性和控制能力。

关键词： 可控音乐生成 · 深度学习 · 条件控制 · Transformer · 交互式音乐人工智能 · 吉他六线谱

1 引言

音乐生成领域涵盖了多种不同的方法，并在自动化方面不断取得进步：从18世纪作曲家用来创作完整乐曲的音乐骰子游戏 [20], 到早期的计算机音乐创作实例 [10], ，再到后续的探索

基于规则的模型 [14], 、马尔可夫模型 [22], 或遗传算法 [23]等技术。近年来，机器音乐创作领域已通过采用深度学习方法 [3][4]达到最先进水平。

本研究聚焦于利用符号记谱法进行自动音乐生成，这类记谱法以数字形式编码乐谱属性 [25]。我们探讨吉他音乐的生成，此处指旋律与和声编排主要由有品弦乐器（即吉他和贝斯吉他）呈现的音乐。尽管有品乐器是六线谱的核心，本研究提出的生成技术支持多种乐器（如吉他、键盘乐器、贝斯、鼓）。以往的符号音乐生成研究通常依赖采用 MIDI、MusicXML 和 ABC 等格式的数据集 [8]。与此不同，我们使用 DadaGP 数据集，该数据集基于适用于吉他音乐符号记谱的 GuitarPro（GP）[5]格式。对于弦乐器，这种吉他谱格式不仅能指定演奏什么音乐，还能指定如何演奏 [17]。GP 吉他谱格式能够支持吉他音乐的表现性演绎，提供指法和吉他专属技巧等信息，而 MIDI 目前尚不支持这些信息。

[28] 提出了 CTRL，这是一种条件 Transformer [31] 语言模型，可在文本生成任务中控制风格和内容结构。控制代码是附加到文本序列末尾的特殊词元，用于对文本进行分类，也可在推理阶段用作提示，以控制生成特征。受这项工作的启发，我们采用了类似的方法，通过附加特殊控制代码，使用一组乐器和音乐流派来控制吉他谱生成。

本工作的主要贡献包括：（1）GTR-CTRL：用于控制基于 Transformer 的模型的方法。该模型在 DadaGP 数据集上训练，可自动生成吉他谱。GTR-CTRL 的主要目标是在推理时提供更大的灵活性和控制能力。GTR-CTRL 可通过配器（inst-CTRL）和音乐流派（genre-CTRL）进行条件控制；（2）GPBERT：一种使用 DadaGP 数据集将歌曲分类为不同音乐流派的模型，可用于评估生成的音乐是否符合预期的流派。我们希望这项工作能够推动以吉他为核心的符号音乐和人工智能研究。

2 背景

2.1 基于深度学习的符号音乐生成

基于深度学习的符号音乐生成技术可根据所采用的架构进行分类，包括变分自编码器模型 [30], 生成对抗网络 [9],以及源自自然语言处理（NLP）领域的模型，例如循环神经网络（RNN） [19], 长短期记忆网络（LSTM） [29],或 Transformer [31]。Transformer 是一种序列到序列模型，能够学习

[5]可从以下网址获取：https://www.guitar-pro.com/；也可使用以下替代软件：https://sourceforge.net/projects/tuxguitar/、http://www.power-tab.net/ guitar.php。

通过引入自注意力的概念，捕捉给定序列中各元素之间的依赖关系和模式。Huang 等人的 Music Transformer[12]首次将自注意力机制用于生成更长的符号化钢琴音乐序列。其他具有开创性的代表作包括 MuseNet [24], ，该模型采用大型 Transformer 模型 GPT-2，生成不同音乐流派的符号化多乐器音乐；Pop Music Transformer [13],，以 Transformer-XL [7] 为骨干架构，能够生成节奏结构更佳的流行钢琴符号音乐；以及 Compound Word Transformer [11], ，它探索了用于训练的符号音乐分词新方法，并提高了效率。

尽管吉他谱数量不断增加且十分丰富，迄今针对吉他音乐的符号音乐生成研究仍不多。 [16,2]. [18]通过探索概率模型，提出了一种吉他独奏生成器，可生成吉他谱格式的乐曲，并同时依赖输入和弦和调性序列。关于基于深度学习的吉他谱音乐生成，Chen 等人 [5]提出了一种指弹吉他生成器，该生成器基于 Transformer-XL 模型训练，使用了包含 333 个示例的数据集（未采用 GuitarPro 格式）。

2.2 可控符号音乐生成

尽管深度学习模型在自动符号音乐生成方面取得了令人瞩目的成果，但模型的可解释性和可控性问题依然存在。这推动了对生成过程进行条件控制和引导的方法的研究 [32]。Wang 等人 [32] 提出了一种变分自编码器模型，可生成受和弦结构和风格条件控制的短篇钢琴作品。同样，在 Music FaderNets [30], 中，研究者使用变分自编码器架构生成受节奏和音符密度条件控制的钢琴曲。关于流派条件控制， [15],采用变分自编码器框架，生成巴赫众赞歌或西方民歌曲调风格的 MIDI 曲目。Theme Transformer 与我们的工作密切相关，因为它采用 Transformer 架构 [27] 进行条件式符号音乐生成；这是一种新颖的架构，可生成以音乐动机或主题为条件控制的 MIDI 乐谱 [27]。

3 动机

本工作的主要动机是设计用于创作富有表现力的吉他核心音乐的人工智能技术，从而赋予制作人一定的自主创作空间。本文所提出模型的改编可能促成面向艺术家/乐队的音乐创作工具，使他们能够将生成过程引向不同的创意方向。据我们所知，此前尚无支持多种乐器且可控的吉他核心符号音乐生成模型。我们使用一个无条件模型开展的初步实验表明，通过使用提示可以实现一定程度的控制。例如，为某一乐器组合中的每种乐器定义一个音符，便可以将输出引导至特定的乐器编排。不过，定义这些初始音符通常会迫使乐曲确定其调（这可能是理想的特性，也可能并非如此）。通过使用给定流派的歌曲片段作为提示，生成的音乐很可能符合源音乐流派，但其内容往往会模仿初始提示动机，从而限制创作可能性。为消除这些限制并赋予用户更大的创作自主权，我们研究了有助于对吉他谱音乐生成实现更精细控制的替代令牌表示方法。

对于所需的乐器组合，可以为每种乐器定义一个音符，从而将输出引导至特定的乐器编排。然而，定义这些初始音符通常会迫使乐曲确定其调（这可能是理想的特性，也可能并非如此）。通过使用给定流派的歌曲片段作为提示，生成的音乐很可能符合源音乐流派，但其内容往往会模仿初始提示动机，从而限制创作可能性。为消除这些限制并赋予用户更大的创作自主权，我们研究了有助于对吉他谱音乐生成实现更精细控制的替代令牌表示方法。

4 条件控制实验

4.1 DadaGP 数据集

DadaGP 数据集 [26] 包含 26,181 首歌曲，采用两种表示形式：令牌格式，类似于文本表示；以及 *GuitarPro* 格式，其名称源自用于六线谱编辑和播放的 GuitarPro 软件。借助使用 PyGuitarPro [1], 这一操作 GuitarPro 文件的 Python 库的专用编码/解码工具，这两种文件格式可以相互转换。我们将 DadaGP 用作以吉他为核心的符号音乐数据集，因为它以语法格式表示音符，最适合吉他等有品乐器。

在 DadaGP 的令牌格式中，歌曲以艺术家、降调、速度和开始标记开头；这些头部标记 对于将词元解码为 GuitarPro 格式（反之亦然）至关重要。有音高乐器的音符由 instrument:note:string:fret 形式的词元表示。这种语法不仅适用于弦乐器，因为 string/fret 组合最终会映射为一个 MIDI 音符。它也可用于 DadaGP 支持的任何其他有音高乐器。打击乐器，即架子鼓，使用 drums:note:type 形式的词元表示。为量化音符时值或休止符，使用 wait:ticks 词元。与大多数数字音频工作站（DAW）一样，采用每四分音符 960 个 tick 的分辨率，解码器则根据 tick 值推断音符时值。有关 DadaGP 的更详细说明，请参阅 [26]。

4.2 模型描述

我们使用 Transformer-XL 模型 [7] 作为骨干架构，因为它通过修改位置编码方案并引入循环机制的概念，对原始 Transformer [31]进行了扩展，使其能够利用当前片段之前出现的词元信息。为解决吉他导向的音乐生成问题，我们采用了 Pop Music Transformer 模型 [13] ，该模型使用类似的架构来生成 MIDI 钢琴作品。我们训练了 Transformer-XL 模型配置的三个变体，该配置由 12 个自注意力层和 8 个多注意力头组成：(i) 一个没有任何无条件 模型，无法进行控制

的可能性（除了提示所产生的副作用之外）；(ii) 一个用于控制配器的模型（inst-CTRL）；(iii) 一个用于控制音乐流派的模型（genre-CTRL）。所有模型均训练了 300 个训练轮次，学习率为$1e-4$，批量大小为 8 个样本。模型参数根据先前的实验进行启发式调整。

4.3 乐器条件实验

配器控制标记：在 [28],介绍控制代码后，我们设计了一组词元来表示配器，即乐曲中演奏的乐器。训练前，配器信息会作为头部标记的一部分，附加到每首歌曲的开头。每首歌曲中各乐器的词元会插入到 inst_start 和 inst_end 词元之间。在推理时，模型会被强制按照这些词元指定的配器生成音乐。如第 3 节所述，使用提示指定每种目标乐器的初始音符（例如 distorted0:note:s6:f0），可以在一定程度上通过无条件模型控制配器。不过，这会间接使生成结果倾向于提示中指定的音高（本例中为 E 小调或 E 大调）。我们提出的 inst-CTRL 方法旨在提供更大的灵活性。

乐器编制推理提示：为了评估 inst- CTRL 控制配器的能力，我们为推理阶段设计了三种不同的初始提示：完整提示，描述乐器编制控制代码以及每种所选乐器的一个初始音符；部分提示，包含乐器编制控制代码和一种乐器的音符；以及空提示，仅包含乐器编制控制代码。为了涵盖多样化的乐器组合，我们选取了八种不同的乐器组合，并将其分为由两种、三种和四种乐器构成的组合：贝斯和鼓（b-d）；失真吉他和鼓（dg-d）；失真吉他、贝斯和鼓（dg-b-d）；清音吉他、贝斯和鼓（cg-b-d）；失真吉他、钢琴和鼓（dg-p-d）；两把失真吉他、贝斯和鼓（dg-dg-b-d）；清音吉他、失真吉他、贝斯和鼓（cg-dg-b-d）；失真吉他、钢琴、贝斯和鼓（dg-p-b-d）。为了比较条件模型与无条件模型的结果，无条件情况使用了一个为每种所选乐器提供初始音符的提示。我们为每个模型/提示生成了 1,200 个示例，其中八种乐器组合各包含 150 个示例。我们使用带有$top\text{-}k$截断的温度控制随机采样方法，为每首歌曲设定了 1,024 个生成词元的上限 [28], ，具体方法见 [13]。

4.4 流派条件控制实验

流派控制标记：与配器控制所采用的流程（第 4.3 节）类似，在流派条件控制实验中，我们创建了一组包含音乐流派信息的词元，并将其置于每首

DadaGP 数据集中的歌曲开头。为了剔除歌曲数量较少的流派，我们只选取训练语料库中示例数超过 200 的流派。最终得到约 100 种不同的音乐流派，其中摇滚和金属乐的流派及子流派占主导地位，同时也包括爵士乐、民谣和古典音乐等其他流派。

流派推理提示： 为评估 genre-CTRL 模型，我们针对以下五种音乐流派进行了推理：金属乐、摇滚、朋克、民谣和古典音乐。推理时使用的提示与乐器条件实验中采用的方式类似。完整提示 提供两个小节以及相应的流派标记。我们从训练语料库中随机选取各流派歌曲的前两个小节。部分提示 提供流派标记，但只使用每段两个小节长的片段中的第一个音符。空提示 仅描述流派标记。无条件 情况下的提示指定前两个小节，但省略流派标记。对于五种不同的音乐流派，我们分别针对五种不同的歌曲提示生成 20 首歌曲，总计 500 首。生成标记数量和采样超参数均与乐器条件实验保持一致。

4.5 示例

按照第 4.3 节和第 4.4 节所述的流程，我们从乐器条件模型和流派条件模型生成的歌曲中精选了一些示例。这些未经任何后处理的示例可供试听[6]。

5 评估方法

5.1 总体音高与节奏指标

为了从旋律/和声和节奏两个视角评估生成结果的整体一致性，我们采用了两项指标，即**音级直方图熵 (PCE)** 和 **律动一致性 (GC)** [33]。我们使用 MusPy 库（一种用于符号音乐生成与评估的工具包）计算了采用乐器与流派条件控制方法生成的示例的这些指标 [8]。此外，我们还使用了针对各类条件控制的特定指标，详见以下各节。

PCE 用于计算歌曲的调性程度。当特定音级占主导地位（例如调中的第一级和第五级）时，熵较低；而熵较高则表明调性不稳定。对于 inst-CTRL，我们针对每种选定的乐器组合，计算其中每种旋律乐器（即不包括鼓）的综合音级直方图熵，并对不同乐器组合的结果取平均值。对于 genre-CTRL，则对不同音乐流派的结果取平均值。

[6]目前可通过以下链接获取：https://drive.google.com/drive/folders/1ds5D01YW-8PAkIf-KxbyACOIBEYqk-he?usp=share_link

[8],中所述的律动一致性可以被视为一种衡量节奏性的指标。在节奏明确定义的乐曲中，各小节之间的律动模式保持一致，因此得分较高。相反，对于各小节之间缺乏节奏一致性的歌曲，该指标的得分较低。

5.2 配器指标

如第 4.3 节所述，在推理阶段，我们通过提示模型并测试不同的提示完整度，来考察配器控制代码的影响。为了客观衡量这些效果，我们引入了两个新指标。**提示乐器存在度（PIP）得分**能够反映模型生成控制代码和配器提示中所选乐器的能力。对于某一种乐器，若某小节在整个时长内均为休止符，则该小节为空。对于一首歌曲，PIP 指标表示包含提示乐器的小节数占所有乐器小节总数的百分比，如公式 1 所示：

$$PIP = \frac{\sum_{i \in P} M_i - E_i}{\sum_{i \in A} M_i} \tag{1}$$

其中，A 是该歌曲中出现的**所**有乐器的集合，P 是歌曲中**提**示的乐器集合，因此是A的子集； M_i 表示乐器i 对应的小节总数（包括空小节）， E_i 表示分配给乐器i的空小节总数。为了提高指标的稳健性，我们将至少包含一个音符的小节视为有效小节，即 M_i；我们从中减去空小节 E_i，以避免器乐声部中空小节较多时 PIP 分数虚高。

鉴于模型可能会为配器控制代码中未包含的乐器生成声部，我们实现了一种**未提示乐器存在度（UIP）得分**，用于衡量含有未提示乐器的小节占非空小节总数的百分比，如公式 2 所示：

$$UIP = \frac{\sum_{i \in (A \setminus P)} M_i - E_i}{\sum_{i \in A} M_i - E_i} \tag{2}$$

根据此前实验的观察，当模型为某种乐器只生成少量音符时，DadaGP 中的解码器程序会自动用休止符填充该乐器其余部分，使该乐器的小节数与其他乐器的小节数一致。例如，在一首长达 20 小节的乐曲中，如果某种乐器只有一个小节包含音符信息，那么它就会被分配 19 个空小节。

由于我们关注的是未提供提示的乐器中的活跃音乐内容（即音符），因此在 UIP 表述中舍弃了空小节。

5.3 流派指标

鉴于流派分类存在诸多困难，即使是人类专家也常常无法达成共识 [21]，并且生成数据量极大，我们使用机器分类器评估 genre-CTRL 模型。在符号域中进行音乐流派识别十分复杂；本研究中的网络正是在这一领域输出音乐，因为声学信息能够提供用于识别流派的线索（例如音色）。虽然可以将生成的符号乐谱渲染为音频，但虚拟乐器的选择和音质可能会影响流派识别。 [6] 中的作者介绍了 MIDIBERT，这是一种基于双向 Transformer 编码器表征（BERT）的掩码语言模型，在复调钢琴 MIDI 乐曲上训练，可配置用于下游分类任务。我们采用类似的方法，提出了 GPBERT，这是一种变体模型，首先在 DadaGP 数据集上预训练 50 个训练轮次，随后针对流派分类任务微调 10 个训练轮次。对于后者，我们从第 4.4 节所述的五种流派中各选取了 800 首歌曲，并按 80/10/10 的比例划分训练集、验证集和测试集。该模型的总体测试准确率达到 90.67%，因此我们认为它适用于评估 genre-CTRL 模型。不过，这种评估方式存在局限，因为它使用一个 AI 系统对另一个 AI 系统的输出进行分类，且没有人工反馈。尽管如此，该方法的一项优势是，用于训练 GPBERT 的音乐流派元数据来自 Spotify Web API，其中提供了由 Spotify 策划团队人工制作的流派标签。

6 定量分析与讨论

6.1 音高与节奏指标结果

我们计算了针对乐器（共 1,200 个示例）和流派（共 500 个示例）条件控制实验所生成片段的 **音级直方图熵** 和**律动一致性**。为了对所提出的条件控制方法进行基准评估，我们将训练语料库用作真实值数据（理论上的最佳情况），并将通过随机化某些音乐属性生成的语料库用作基准（理论上的最差情况）。对于 PCE（分别对应 GC），随机化语料库是通过随机改变 DadaGP 歌曲中的音符音高（分别对应节奏）得到的。本节报告统计分析结果，考察音乐来源（4 种不同的提示、真实值和随机）对 PCE 和 GC 因变量的影响，第一类错误率 α 为 .05。图1左侧的图中展示了配器（上）和流派（下）条件控制对应的音级熵缺口箱线图。

我们进行了 Kruskal-Wallis 秩和检验，结果显示音乐来源（4 种提示、真实值和随机）对音级熵有显著影响（$H(5) = 3169.6$, $p < .001$）。我们使用 Wilcoxon 秩和检验进行了两两比较，并在图1中用括号标出显著差异。这些检验表明，对于配器条件控制（左上）和流派条件控制（左下），三个条件模型以及无条件模型产生的音级熵均与真实值和随机条件下的结果存在显著差异。

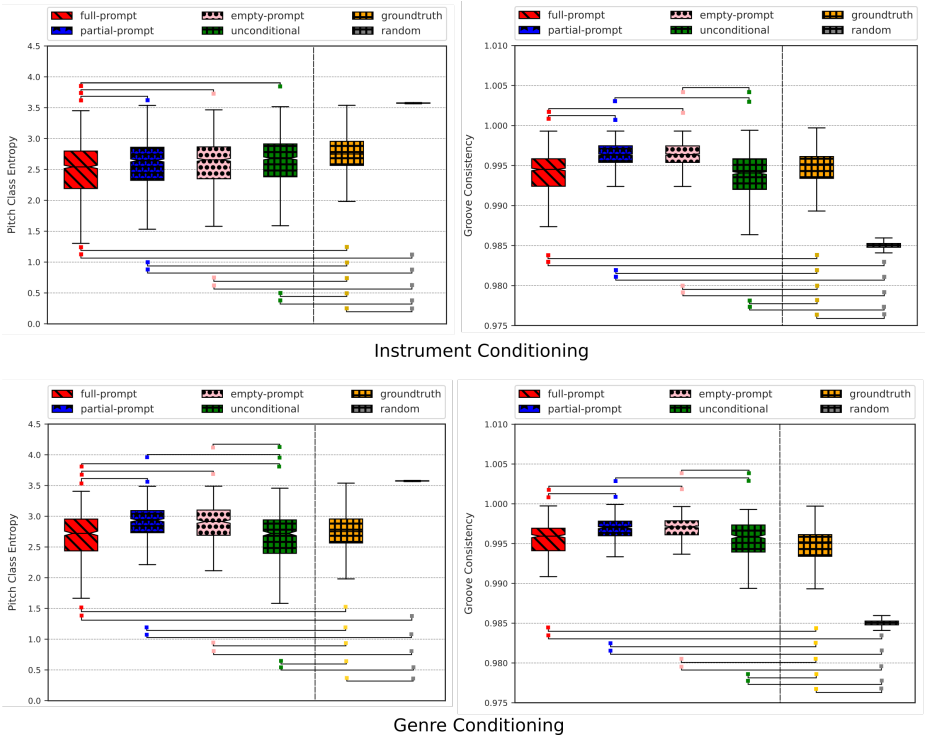


图 1. 乐器（上）和流派（下）条件控制实验中，音级熵（左）和律动一致性（右）的缺口箱线图，并与真实值组和随机组进行比较。括号表示在采用 Bonferroni 校正后的显著性水平 .008（.05/6）时，成对 Wilcoxon 符号秩检验中存在显著差异（$p < .001$）。

然而，与随机条件相比，模型生成的样例的音级熵更接近真实值。完整提示 的音级熵与其他提示显著不同，其中位数最低（略低于真实值的中位数）。这意味着，平均而言，完整提示 条件产生的调性稳定性最高。然而，音级熵较低并不足以确保音乐质量，因为较少的

调性变化也可能是由生成音乐中的重复结构等因素造成的。平均而言，真实样例包含 84.6±57.0小节，而生成样例平均包含 34.4 ± 24.6 小节。我们推测，与训练语料库相比，调性稳定性略高是因为生成片段的长度更短；片段较短往往意味着乐曲会保持在给定的调内，且不会发生或只发生少量和声转调，而这会提高音级熵得分。我们强调，由于音乐具有主观性，很难精确定义音级熵方面何谓“更好”的结果。较低的音级熵不一定意味着样例听起来更好，而是总体上表明音乐更倾向于保持在给定的调性之内。音级熵差异对听众音乐预期的影响，还应通过感知测试进一步评估。

我们进行了 Kruskal-Wallis 秩和检验，以评估音乐来源对律动一致性（GC）的影响（图1右侧图表）。结果显示影响极其显著（$H(5) = 3666.9$, $p < .001$）。在乐器条件控制实验中（右上图），部分提示 和 空提示 配置的GC中位数略高于真实值，且差异具有统计显著性，表明其节奏稳定性更高。较长的音乐片段（例如真实语料库中的片段）往往会出现更多节奏变化。生成片段的节奏更为稳定，可能是因为片段时长较短，也可能是因为其中存在高度重复的结构。流派条件控制实验也可以得出类似结论。

6.2 配器指标结果

对于配器指标，由于其含义不言自明，因此未与真实值和随机样例进行比较。

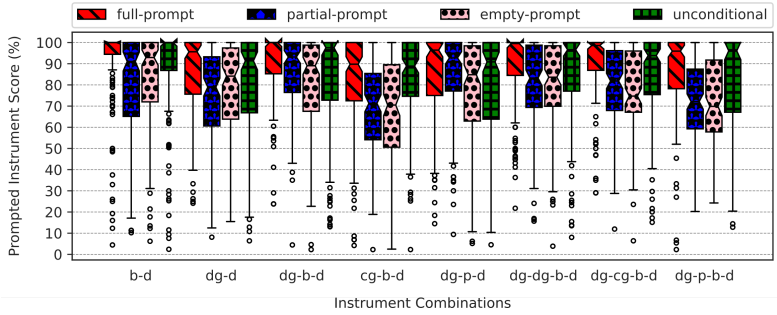


图2。 各乐器组合对应提示乐器得分的缺口箱线图，分别展示完整提示 （红色）、部分提示 （蓝色）、空提示 （粉色）和 无条件 （绿色）样例（数值越高表示性能越好）。

图2展示了**提示乐器得分**的缺口箱线图。所有提示配置的PIP分数中位数均超过70%。

完整提示在每种乐器组合中都取得了最佳的PIP中位数（78.2%，更接近100%）。对于每种配器，无条件模型的表现都优于部分提示和空提示，但$dg\text{-}p\text{-}b\text{-}d$ （失真吉他、钢琴、贝斯和鼓）除外。我们认为，这是由于数据集中缺少采用这种配器的歌曲，导致训练示例较少（参见图4(h) [26]中DadaGP的乐器分布）。Kruskal-Wallis 秩和检验显示，提示配置对PIP分数具有显著影响（$H(3) = 304.93$， $p < .001$）。表1报告了三个条件模型与无条件模型的平均PIP分数及其显著差异（采用带Bonferroni 校正的Wilcoxon 秩和检验）。结果表明，完整提示取得了显著优于无条件模型的结果。

表 1. 完整提示、部分提示、空提示和无条件条件下的提示乐器得分与未提示乐器分数均值。星号表示根据成对 Wilcoxon 秩和检验并采用 Bonferroni 校正后，与无条件模型相比存在显著差异（$p < .001$ 适用于所有情况）。最佳结果以**粗体**显示。

	完整提示	部分提示	空提示	无条件
提示乐器得分	**87.54*** **	78.23***	77.26***	83.62
未提示乐器分数	4.09	4.19	**4.01**	5.2

图 3 展示了未提示乐器分数的缺口箱线图。

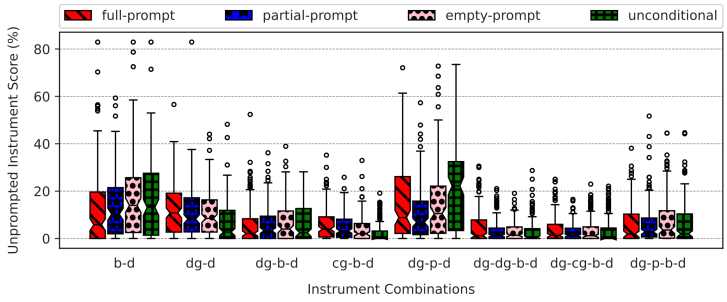


图 3. 各乐器组合在完整提示 （红色）、部分提示 （蓝色）、空提示 （粉色）和无条件（绿色）情况下的未提示乐器分数缺口箱线图（数值越低表示表现越好）。

随着乐器数量较少的组合逐渐过渡到乐器数量较多的配置，避免出现未在提示中指定的乐器（UIP分数较低）似乎有了逐步改善。这可能与 DadaGP 中包含的两种或三种乐器的歌曲数量少于包含四种或更多乐器的歌曲有关。然而，$dg\text{-}p\text{-}d$ 和$dg\text{-}p\text{-}b\text{-}d$ 配置的乐器

组合总体上获得了最差的 UIP 分数，这可能与 DadaGP 中这两种配置分别只有 4 个和 25 个示例有关（相比之下，其余六种选定乐器组合的示例总数为 4,930，范围为 [98-3,058]）。我们进行了 Kruskal-Wallis 秩和检验，以评估乐器组合对 UIP 的影响，并按乐器数量分组（分别为 2、3 和 4 种乐器）。乐器组合中的乐器数量对 UIP分数有显著影响（$H(2) = 352.6$, $p < .001$）。采用 Bonferroni 校正的成对 Wilcoxon 秩和检验显示，三组乐器组合之间存在显著差异（$p < .001$）。包含 2、3 和 4 种乐器的组合，其平均 UIP 分别为 6.9%、4.35% 和 2.69%。

6.3 流派指标结果

GPBERT 的流派分类分数见表 2，其中报告了每种流派和提示下生成的 100 个示例的平均 softmax 分数。

表 2. GPBERT 的流派分类 softmax 分数，分别对应完整提示 (F- P)、部分提示 (P-P)、空提示 (E-P)和无条件 设置，并以金属乐、朋克、摇滚、古典音乐 和民谣为条件。每种流派在各提示下的最高结果以**粗体**显示。

		流派分类分数				
		金属乐	朋克	摇滚	古典音乐	民谣
Metal	F-P	**0.6680**	0.0925	0.2028	0.0263	0.0103
	P-P	0.2217	0.1929	**0.4537**	0.1162	0.0155
	E-P	0.2509	0.1412	**0.4322**	0.1577	0.0180
	UNC		0.0675	0.1697	0.0404	0.0035
Punk	F-P	0.0041	**0.8126**	0.1727	0.0012	0.0094
	P-P	0.1264	0.3406	**0.4807**	0.0434	0.0088
	E-P	0.0997	0.2176	**0.5043**	0.1342	0.0441
	UNC	0.0290	**0.6954**	0.2464	0.0113	0.0179
Rock	F-P	0.0236	0.3325	**0.6403**	0.0021	0.0015
	P-P	0.1069	0.1661	**0.6401**	0.0757	0.0112
	E-P	0.0863	0.0608	**0.5763**	0.1994	0.0772
	UNC	0.0266	0.2207	**0.7487**	0.0030	0.0010
Classical	F-P	0.0338	0.0128	0.2293	**0.6753**	0.0488
	P-P	0.0300	0.0126	0.1883	**0.7035**	0.0657
	E-P	0.0877	0.0644	0.3175	**0.4767**	0.0538
	UNC	0.0415	0.0227	0.2638	**0.5940**	0.0780
Folk	F-P	0.0692	0.0451	**0.4338**	0.2751	0.1768
	P-P	0.0751	0.0533	0.2904	**0.4323**	0.1489
	E-P	0.0822	0.1141	0.3192	**0.3900**	0.0946
	UNC	0.0885	0.0430	**0.4849**	0.2901	0.0934

无条件 和完整提示 在金属乐、朋克 和摇滚上的流派分类结果最佳。虽然 无条件 模型在摇滚 和金属乐上的表现更好，但完整提示 在朋克上的得分更高。值得注意的是，GPBERT 经常混淆一些人类听众也难以区分的流派（例如，将金属乐/朋克 判为

摇滚）。总体而言，民谣 流派的结果常被归类为摇滚 （完整提示 和无条件），或古典音乐 （部分提示 和空提示）。对于部分提示 和空提示，当目标流派为金属乐 或朋克时，生成的乐曲往往会被归类为摇滚。我们认为，这是由于 DadaGP 中摇滚/金属乐 偏向较强，这两种流派的曲目占比超过 50%（见 [26] 中的图 3）。

7 主观分析与讨论

我们对精心挑选的生成样例进行了主观分析。在流派控制方面，尤其是在空提示 和部分提示 的情况下，我们发现模型难以强制生成指定的多乐器配置，经常只生成一个或两个器乐声部。这一行为会影响 GPBERT 的流派分类分数，因为正确分类需要多乐器作品。为解决这一问题，一个值得尝试的方法是将 inst-CTRL 和 genre-CTRL 合并为一个模型，同时考虑配器和流派的可控性。

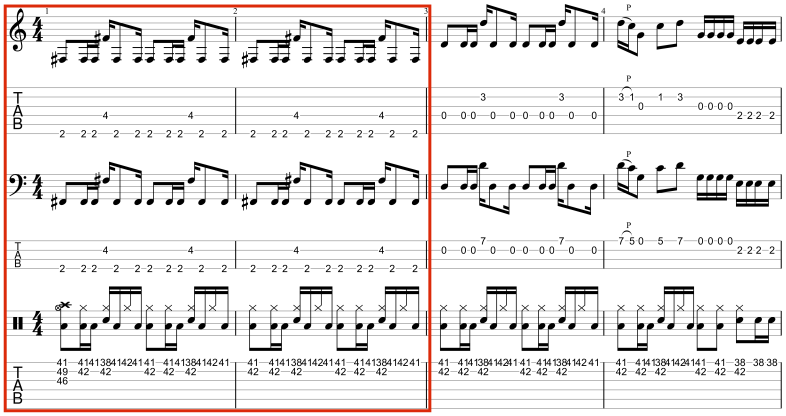


图4. id-139 中的前四小节（提示取自齐柏林飞艇的《移民之歌》）。提示以红色突出显示（可见失真吉他、贝斯和鼓）。

尽管这并非设计目标，但我们发现，一些采用 完整提示流派条件控制的模型可以有效地用作动机／歌曲续写器。例如，在歌曲 id-139[7],中，模型保持了初始动机的整体节奏结构，同时通过不同的和弦进行对其进行调整（见图4）。在 id-67 中（提示来自交响 X 《谎言之海》的前两个小节），两把吉他分别承担主奏与节奏功能，在吉他独奏乐句和和弦结构之间交替演奏。此外，

[7] 带下划线的歌曲 ID 已添加超链接，方便收听。

id-459（提示来自约翰·帕赫贝尔《D大调卡农》的前两个小节）以其不同寻常的和声探索脱颖而出，在保持初始构思特征的同时，id-103（见图5）展示了模型生成符合提示的调与风格的新颖吉他乐句的能力。

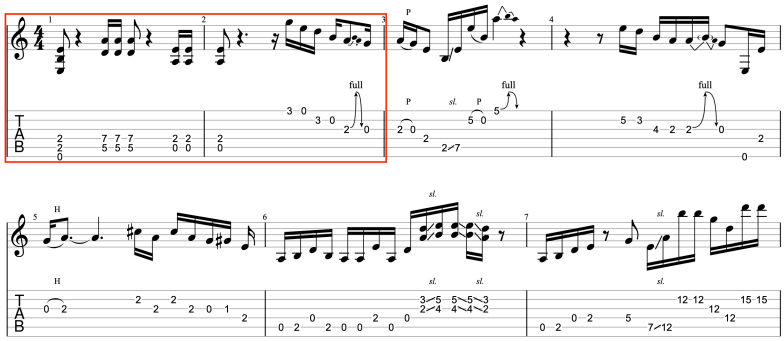


图 5. id-103 吉他音轨的开头几个小节（提示取自 AC/DC 的《Back In Black》）。用作提示的小节以红色标出（仅显示失真吉他）。

8 结论与未来工作

本文介绍了针对吉他音乐生成任务对两个基于 Transformer 的模型 inst-CTRL 和 genre-CTRL 进行条件控制。我们评估了这两个模型相对于无条件模型的表现，并探讨了推理时采用三种不同提示策略的效果。结果表明，两个模型在各自的条件控制任务上均表现良好：inst-CTRL 中的条件提示在所有指标上的表现都优于无条件情形；而对于 genre-CTRL，完整提示在朋克、古典音乐 和民谣 流派上的表现优于无条件模型。我们认为，无条件模型在无条件的摇滚 和金属乐流派上得分最高，是因为数据集中摇滚和金属乐占比较大，使无条件 模型能够很好地生成这些流派的歌曲。尽管结果略逊一筹，空提示 模型无需使用特定的乐器音符提示模型（这些音符往往会控制乐曲的调式或旋律特征），因此灵活性更高，音符和调的选择也完全不受影响。

在未来的工作中，我们计划结合这两个模型，以同时控制配器和流派。我们还打算探索 GPBERT 在不同下游任务中的应用，例如音乐修复以及艺术家/作曲家分类。一个有趣的研究方向是开发一种能够按照特定吉他演奏者的风格生成吉他段落的模型，再次使用 GPBERT 的变体对其输出进行分类，并辅以专家吉他演奏者的聆听测试

。最后，我们设想与愿意参与人类与 AI 共同创作乐曲的艺术家/乐队开展合作。

致谢

本研究得到 EPSRC UKRI 人工智能与音乐博士培训中心的支持（资助编号：EP/S022694/1）。

参考文献

1. Abalumov, S.：PyGuitarPro（2014），https://github.com/Perlence/PyGuitarPro，最后访问时间：2022年11月3日 2. Barthet, M.、Anglade, A.、Fazekas, G.、Kolozali, S.、Macrae, R.：音乐学习的音乐推荐：Hotttabs，多媒体吉他导师。载于：音乐推荐与发现研讨会。第7–13页。美国伊利诺伊州芝加哥（2011）3. Briot, J.P.、Hadjeres, G.、Pachet, F.D.：音乐生成的深度学习技术。计算合成与创意系统丛书，Springer（2019）4. Carnovalini, F.、Rodà, A.：计算创造力与音乐生成系统：现状简介。Frontiers in AI **3** （2020）5. Chen, Y.H.、Huang, Y.H.、Hsiao, W.Y.、Yang, Y.H.：利用 Transformer 和律动建模进行吉他谱自动创作。载于：第21届国际音乐信息检索学会会议论文集。第756–763页（2020）6. Chou, Y.H.、Chen, I.C.、Chang, C.J.、Ching, J.、Yang, Y.H.：MidiBERT-Piano：面向符号音乐理解的大规模预训练。技术报告（2021）7. Dai, Z.、Yang, Z.、Yang, Y.、Carbonell, J.、Le, Q.V.、Salakhutdinov, R.：Transformer-XL：超越固定长度上下文的注意力语言模型。载于：第57届计算语言学协会年会论文集。第2978–2989页。意大利佛罗伦萨（2019）8. Dong, H.W.、Chen, K.、McAuley, J.、Berg-Kirkpatrick, T.：MusPY：用于符号音乐生成的工具包。载于：第21届国际音乐信息检索学会会议论文集。第101–108页。加拿大蒙特利尔（2020）9. Dong, H.W.、Yang, Y.H.：利用带二值神经元的卷积生成对抗网络进行复调音乐生成。载于：第19届国际音乐信息检索学会会议论文集。第190–198页。法国巴黎（2018）10. Hiller, L.A.、Isaacson, L.M.：实验音乐；使用电子计算机进行作曲。Greenwood Publishing Group Inc.，美国（1979）11. Hsiao, W.Y.、Liu, J.Y.、Yeh, Y.C.、Yang, Y.H.：Compound Word Transformer：在动态有向超图上学习创作完整歌曲。载于：AAAI人工智能会议论文集。第178–187页（2021）12. Huang, C.Z.A.、Vaswani, A.、Uszkoreit, J.、Shazeer, N.、Simon, I.、Hawthorne, C.、Dai, A.M.、Hoffman, M.D.、Dinculescu, M.、Eck, D.：Music Transformer：利用长期结构生成音乐。载于：第7届国际学习表征会议论文集。美国路易斯安那州新奥尔良（2019）13. Huang, Y.S.、Yang, Y.H.：Pop Music Transformer：富有表现力的流行钢琴作品的基于节拍的建模与生成。载于：第28届ACM国际多媒体会议论文集。第1180–1188页。美国华盛顿州西雅图（2020）14. Johnson-Laird, P.N.：爵士乐演奏者如何即兴演奏。载于：音乐感知，第19卷（3），第415–442页。加利福尼亚大学出版社（2002）

15. Lim, Y.Q.、Chan, C.S.、Loo, F.Y.：风格条件音乐生成。载于：电气电子工程师学会国际多媒体与博览会会议论文集。英国伦敦（2020）16. Macrae, R.、Dixon, S.：吉他六线谱挖掘、分析与排序。载于：第12届国际音乐信息检索学会会议论文集。第453–459页。美国佛罗里达州迈阿密（2011）17. Magnusson, T.：声音书写：材料、符号与信号铭刻技术。布卢姆斯伯里学术出版社（2019）18. McVicar, M.、Fukayama, S.、Goto, M.：AutoLeadGuitar：在六线谱空间中自动生成吉他独奏乐句。国际信号处理会议论文集。第599–604页（2014）19. Meade, N.、Barreyre, N.、Lowe, S.C.、Oore, S.：探索利用人类可解释的控制进行生成音乐系统的条件控制。技术报告（2019）20. Nierhaus, G.：算法作曲：自动音乐生成的范式。施普林格维也纳（2009）21. Oramas, S.、Barbieri, F.、Nieto, O.、Serra, X.：用于音乐流派分类的多模态深度学习。国际音乐信息检索学会汇刊1（1），第4–21页（2018）22. Pachet, F.：The Continuator：风格化音乐交互。载于：国际计算机音乐会议。第333–341页。瑞典哥德堡（2002）23. Papadopoulos, G.、Wiggins, G.：用于生成爵士旋律的遗传算法。载于：人类与人工信息处理：芬兰人工智能会议。第7–9页。芬兰于韦斯屈莱（1998）24. Payne, C.：Musenet（2019），https://openai.com/blog/musenet，最后访问时间：2022年6月12日 25. Raffel, C.、Ellis, D.P.W.：从MIDI文件中提取真实标注信息：MIDIfesto。载于：第17届国际音乐信息检索学会会议论文集。第796–803页。美国纽约市（2016）26. Sarmento, P.、Kumar, A.、Carr, C.、Zukowski, Z.、Barthet, M.、Yang, Y.H.：DadaGP：用于序列模型的标记化 GuitarPro 乐曲数据集。载于：第22届国际音乐信息检索学会会议论文集。第610–618页（2021）27. Shih, Y.J.、Wu, S.L.、Zalkow, F.、Müller, M.、Yang, Y.H.：Theme Transformer：利用主题条件变换器进行符号音乐生成。技术报告（2021）28. Shirish Keskar, N.、Mccann, B.、Varshney, L.R.、Xiong, C.、Socher, R.、Research, S.：CTRL：用于可控生成的条件 Transformer 语言模型。技术报告（2019）29. Sturm, B.L.、Santos, J.F.、Ben-Tal, O.、Korshunova, I.：使用深度学习进行音乐转录建模与作曲。载于：第一届音乐创造力计算机模拟会议论文集（2016）30. Tan, H.H.、Herremans, D.：Music Fader Nets：基于高层特征、通过低层特征建模实现的可控音乐生成。载于：第21届国际音乐信息检索学会会议论文集。第109–116页。加拿大蒙特利尔（2020）31. Vaswani, A.、Shazeer, N.、Parmar, N.、Uszkoreit, J.、Jones, L.、Gomez, A.N.、Kaiser, Ł.、Polosukhin, I.：注意力就是一切。载于：第31届神经信息处理系统会议论文集。美国加利福尼亚州长滩（2017）32. Wang, Z.、Wang, D.、Zhang, Y.、Xia, G.：学习用于可控复调音乐生成的可解释表征。载于：第21届国际音乐信息检索学会会议论文集。第662–669页。加拿大蒙特利尔（2020）33. Wu, S.L.、Yang, Y.H.：前线上的爵士变换器：通过定量指标探索人工智能作曲的不足。载于：第21届国际音乐信息检索学会会议论文集。第142–149页。加拿大蒙特利尔（2020）