# EMelodyGen：基于情感条件的 ABC 记谱法旋律生成与音乐特征模板

Monan Zhou[1], Xiaobing Li[1], Feng Yu[1], Wei Li[2,3*]
[1]Department of Music AI and Information Technology, Central Conservatory of Music, Beijing, China
[2]School of Computer Science and Technology, Fudan University, Shanghai, China
[3]Shanghai Key Laboratory of Intelligent Information Processing, Fudan University, Shanghai, China



**摘要—EMelodyGen 系统专注于通过音乐特征模板控制的 ABC 记谱法的情感旋律生成。由于结构良好且带有情感标签的乐谱稀缺，我们设计了一个模板，该模板通过从小型情感符号音乐数据集和音乐心理学结论中得出的音乐特征与情感标签之间的统计相关性来控制情感旋律的生成。然后，我们使用该模板自动为一个大型、结构良好的乐谱集合标注了粗略的情感标签，将其转换为 ABC 记谱法，并通过数据增强减少了标签不平衡，从而得到了一个名为 Rough4Q 的数据集。我们的系统骨干在 Rough4Q 上进行了预训练，可以达到 99%的 music21 解析率，而由我们模板生成的旋律可以在盲听测试中与情感表达达成 91%的一致性。消融研究进一步验证了模板中特征控制的有效性。可用代码和演示位于这里。**

***Index Terms*—旋律生成，可控音乐生成，ABC 记谱法，情感条件**

## I. 介绍

最近，一些基于情感条件的音乐生成方法如 [1]–[6] 使用 MIDI 数据而不是 ABC 记谱法进行符号音乐生成，并且其中的一些并未严格在 Russell 4Q [7] 情绪标签系统中进行研究，而我们的工作则尝试在由音乐特征模板控制的 Russell 4Q 情绪空间中生成 ABC 记谱法旋律，因此没有直接可比的任务。我们选择使用 ABC 记谱法是因为与乐谱 MIDI 和 XML/MusicXML/MXL 相比，它具有更高的音乐信息密度。在用 ABC 记谱法生成乐谱的领域中，Tomasz Michal Oliwa 之前探索了用于摇滚音乐创作的遗传算法 [8] 而非情感条件下的作曲。最近，如 Tunesformer [9]、abcMLM [10] 和 MelodyT5 [11] 等方法利用基于变换器的语言模型在没有情感条件的情况下生成 ABC 记谱法的音乐。其中，MelodyT5 和 abcMLM 基于变换器编码器-解码器架构。因此，由于其相对较轻量级的特点，我们选择了仅含变换器解码器的 Tunesformer 作为系统的主体。然而，Tunesformer 生成乐谱的有效性高度依赖于训练数据的质量，所以用杂乱无章的乐谱训练出的模型可能会生成无法被 music21[1] 正确解析并最终正确渲染为音频的错误乐谱。

到目前为止，带有情感标签的结构良好的乐谱数据很少。像 EMOPIA [12] 和 VGMIDI [13] 这样的知名数据集仅包含带情感注释的 MIDI 数据，而不是 XML/ABC 格式的乐谱。尽管可以使用 music21 或*移斯 core*[2] 等工具将表演 MIDI 文件转换为 XML/ABC 格式，但生成的曲谱往往杂乱无章，这会降低旋律生成的质量。为了解决这个问题，我们选择了结构良好的乐谱作为基础训练材料，并随后通过控制音乐特征施加了情感模板来生成旋律。这种方法不仅保证了一定水平的生成输出质量，还避开了结构良好乐谱数据集中缺乏情感标签的问题。为了识别与情感控制高度相关的关键音乐特征，我们将带有情感标签注释的 EMOPIA 和 VGMIDI 数据集合并为一个统一的分析数据集。然后从中提取特征并进行音乐特征与情感标签之间的相关性分析。为了避免可能出现的数据分布偏好导致的潜在偏差，我们还参考了音乐心理学文献 [14] 中的先验知识，最终设计出一种专门用于情感控制旋律生成的音乐特征模板。所选特征可以分为两类：可以在输出阶段直接修改的可控特征，如八度、音量等；需要模型通过深度学习来解释的嵌入特征，如音域范围、平均音高（avg pitch）、音高标准差（pitchSD）、旋律上升/下降（direction）、调式等。

设计的模板集成了上述类别中的五个特征，其中两

* Corresponding author.

[1]https://pypi.org/project/music21
[2]https://musescore.org

个属于嵌入式特征，只能通过嵌入实现情感控制，因此需要将部分特征标注到训练材料中进行微调。Rough4Q 数据集是通过对结构良好的乐谱集合自动注释这两个特征而构建的，并且通过微调将这两个特征嵌入到骨干模型中。在 Rough4Q 上预训练的骨干模型达到了 99% 的 music21 解析率，模板生成的音乐与人类期望的情感一致性在盲听测试中达到了 91%。此外，还进行了消融实验以验证模板中的这五个控制条件对整体情感表达的有效性。总之，本文的主要贡献如下：

- 我们进行音乐特征与情感之间的整体相关性分析，以设计融合音乐心理学研究成果的情感控制模板。
- 我们提出了一种两阶段方法来避免标签稀缺问题：自动标注结构良好的数据集以控制生成质量，并使用模板进行情绪控制。
- 我们的工作是首次通过 ABC 记谱法进行情绪条件下的旋律生成的探索性研究。

## II. 数据集

四个数据集被创建用于各种实验：处理过的 EMOPIA 和处理过的 VGMIDI 被用来比较在这些数据集上进行微调后的骨干网络的 music21 解析率与结构良好的乐谱之间的差异；分析数据集被用于检查各种音乐特征与情感标签之间的相关性；Rough4Q 被用于通过微调将嵌入特征应用到最终模型中。这四个数据集的数据结构和统计摘要详情在后续的小节中进行了描述。

### *A. 处理后的 EMOPIA & VGMIDI*

确保两个处理后的数据集与预训练骨干所需的输入格式兼容是至关重要的。我们发现用于预训练骨干的数据集中平均的乐句数量约为 20，而预训练骨干支持的最大乐句数为 32。因此，我们将原始的 EMOPIA 和 VGMIDI 数据转换成 XML 乐谱，并过滤掉错误项，将其分割成每段 20 个乐句的小块。每个小块末尾都添加了一个结束标记，以防止模型在遇到没有终止标志的重复旋律时无休止地生成。对于乐谱的结尾部分，如果一段超过 10 个乐句，则进一步划分；否则与前一段合并。这种方法确保了最终得到的乐谱片段不超过 30 个乐句，从而保证所有片段都在骨干支持的最大乐句数限制内，并且平均约为 20 个乐句。随后，我们将分割后的 XML 片段转换成 ABC 记号格式，通过转调到 15 个键进行数据扩增，并提取带有控制码的旋律线，生成最终处理过的 EMOPIA 和 VGMIDI 数据集。两个数据集都具有相同的结构，包括三列：第一列是 Tunesformer 的控制码，第二列是 ABC 字符，第三列包含从原始数据集中继承的 4Q 情绪标签。经过处理后的 EMOPIA 数据共有 21,480 条记录，VGMIDI 数据为 9,315 条记录，并且均按照 10:1 的比例分为训练集和测试集。它们的标签分布见图 1 的第一至第二子图。

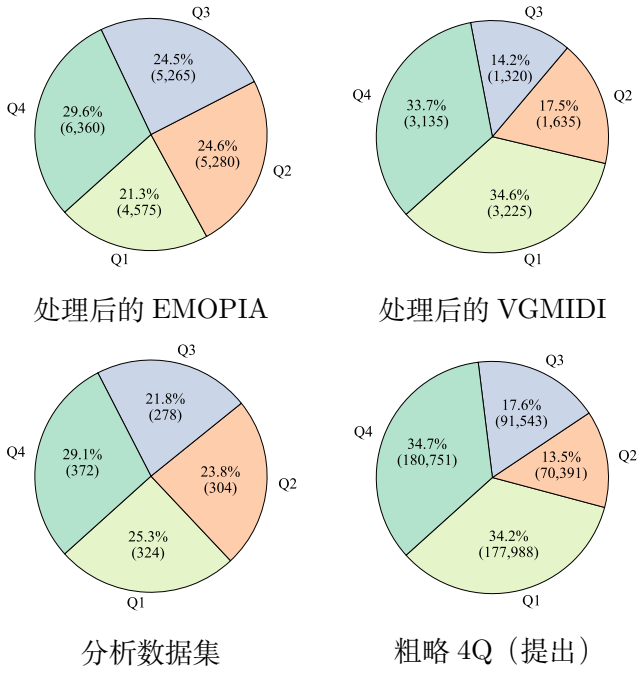


图 1: 饼图展示了处理后的数据集中不同情绪类别所占的比例。

### *B. 分析数据集*

分析数据集源自将原始的 EMOPIA 和 VGMIDI 合并到 Russell 4Q 标签系统中。该数据集包含 11 列：前三列是情感标签，具体为标签（Russell 4Q 情感）、效价（低=0 或高=1）和唤醒度（低=0 或高=1）；剩余的八列表示特征，分别是音调（12 个音调之一：*C, C#, D, 升 E, E, F, F#, G, G#/降 A, A,* 降 *B, B*）、模式（小调=0 或大调=1）、方向（下降=0 或上升=1）、平均音高（八度）、音域范围、音高标准差、速度和 RMS（音量）。

对于特征提取，关键、模式、方向、平均音高、音高范围和音高标准差特征是通过 music21 在符号级别直接提取的。然而，从 MIDI 文件中提取的节奏通常默认为 120 BPM，这可能无法反映实际值，我们使用*穆奇分谱*将这些 MIDI 文件以默认钢琴音色渲染成 44.1

表 I: EMOPIA 和 VGMIDI 合并数据中情绪与特征之间的皮尔森相关统计。

| Emotion | Feature | Correlation coefficient | Relevance | P-value | Confidence level |
|---|---|---|---|---|---|
| Valence | Key | +0.0123 | Weak positive | 6.594e-01 | $p \geq 0.05$ insignificant |
| **Valence** | **Mode** | **+0.3850** | **Positive** | **2.018e-46** | $\boldsymbol{p < 0.05}$ **significant** |
| Valence | Tempo | +0.0621 | Weak positive | 2.645e-02 | $p < 0.05$ significant |
| Valence | Direction | +0.0010 | Weak positive | 9.709e-01 | $p \geq 0.05$ insignificant |
| Valence | Avg pitch (guides octave control) | +0.0102 | Weak positive | 7.161e-01 | $p \geq 0.05$ insignificant |
| Valence | Pitch range | -0.0771 | Weak negative | 5.794e-03 | $p < 0.05$ significant |
| Valence | PitchSD | -0.0676 | Weak negative | 1.568e-02 | $p < 0.05$ significant |
| Valence | RMS (guides volume control) | +0.1174 | Weak positive | 2.597e-05 | $p < 0.05$ significant |
| Arousal | Key | -0.0007 | Weak negative | 9.809e-01 | $p \geq 0.05$ insignificant |
| Arousal | Mode | -0.0962 | Weak negative | 5.748e-04 | $p < 0.05$ significant |
| Arousal | Tempo | +0.1579 | Weak positive | 1.382e-08 | $p < 0.05$ significant |
| Arousal | Direction | -0.0958 | Weak negative | 6.013e-04 | $p < 0.05$ significant |
| Arousal | Avg pitch (guides octave control) | -0.1818 | Weak negative | 5.919e-11 | $p < 0.05$ significant |
| **Arousal** | **Pitch range** | **+0.3276** | **Positive** | **2.324e-33** | $\boldsymbol{p < 0.05}$ **significant** |
| **Arousal** | **PitchSD** | **+0.3523** | **Positive** | **1.179e-38** | $\boldsymbol{p < 0.05}$ **significant** |
| **Arousal** | **RMS (guides volume control)** | **+0.3800** | **Positive** | **3.558e-45** | $\boldsymbol{p < 0.05}$ **significant** |

kHz 采样率的 WAV 格式。随后，我们使用了 *librosa*[3] 库来估算更准确且可区分的节奏数据，该库还计算了渲染音频的均方根值。

值得注意的是，从渲染音频中提取后两个特征（节奏和 RMS）的效率较低，相比之下前六个特征更为高效。然而，由于组合数据集仅包含 1,278 首音乐，这些特征的渲染时间对于分析阶段来说是可以接受的。因此，我们构建了分析数据集，并根据 Russell 4Q 分类将其分布展示在图 1 的第三个饼状图中。对于统计相关性分析，我们计算了第二列和第三列（效价和唤醒）与剩余八列（特征）之间的相关性，并得到了表 I。

### C. 粗糙 4Q 数据集

该大规模数据集是基于表 I 中的结论和音乐心理学文献，通过自动标注大量结构良好的乐谱而创建的。该数据集的数据来源在表 II 中详细列出，包括 XML/MXL/MusicXML 格式以及 ABC 记号格式的曲目。过滤掉错误和重复的曲目，并将它们统一转换为 XML 格式后，我们通过 music21 快速提取了两个计算上可管理的特征：音高标准差(pitchSD)和调式(mode)。

为了统一 Rough4Q 和其他数据集之间的标签系统，分别使用 mode 和 pitchSD 作为 valence 和 arousal 的近似值，因为它们与这些情感维度有很强的正相关性。这也导致了一个 4Q 标签系统的形成，这就是将其称为 Rough4Q 的原因之一。原始数据集中存在严重的

[3] https://pypi.org/project/librosa

表 II: 按大小升序排列的 Rough4Q 源数据集比较。

| 源数据集 | 大小 | 原始格式 |
|---|---|---|
| *MIDI-*波 双向流行音乐 [15] | 111 | MusicXML |
| 巴赫众赞歌 [16] | 366 | MXL |
| 诺丁汉 [17] | 1,015 | ABC notation |
| 维基风歌 [18] | 6,394 | MXL |
| 埃森民歌 [19] | 10,369 | ABC notation |
| 爱尔兰人 [9] | 216,281 | XML |

类别不平衡问题，Q2 和 Q3 类别的数据量比其他类别低一个数量级。为了解决这个问题，特别针对 Q2 和 Q3 类别应用了 15 键的数据增强。从表 I 的统计结论来看，键与情感之间几乎没有相关性，这表明转置到 15 个键不太可能显著影响标签分布。最终增强数据集的统计信息呈现在图 1 的最后一个子图中。

## III. 方法论

在相关性分析阶段，所采用的方法包括皮尔逊相关系数（PCC）[20] 和高斯核密度估计（KDE）[21]。前者用于计算特征与情感之间的相关系数，而后者则用于绘制特征上情感标签的分布图。随后的微调阶段讨论了主干网络的细节。

### A. 相关性分析

我们在 Russell 的 4Q 情感标签系统中的分析数据集上进行了统计分析，计算了情感维度与乐谱特征之间的 PCCs，并使用高斯 KDE 图对多尺度特征进行了绘制，使用条形图对二分特征进行了展示。由于 EMOPIA

和 VGMIDI 分别使用 4Q 和二元效价/唤醒 (V/A) 标签。两者都不是连续值，可以按如下方式相互转换：

$$\mathcal{Q}(V,A) = I_{V<0}I_{A\geq 0} + 2I_{V<0}I_{A<0} + 3I_{V\geq 0}I_{A<0} \quad (1)$$

其中 $I$ 是指示函数。因此，我们选择了 V/A 值的正负符号作为两个水平：低和高，其值分别为 0 和 1。

关于音乐特征的选择，我们提取了八个音乐特征：调性、模式、节奏、方向、平均音高、音域、音高标准差和均方根从旋律中。其中，计算平均音高、音高标准差和方向需要进一步解释：假设一个旋律 $M = \{(p_1,d_1),(p_2,d_2),\ldots,(p_n,d_n)\}$ 包含 $n$ 个音符，其中 $p_i$ 和 $d_i$ 分别代表第 $i$ 个音符的音高和时长。这里的时长是通过 music21 提取的，1 表示四分音符的长度。平均音高（记为 $\bar{p}$）是按时间加权后的音高低：

$$\bar{p} = \frac{\sum_{i=1}^{n} p_i d_i}{\sum_{j=1}^{n} d_j} \quad (2)$$

基于 $\bar{p}$，我们进一步计算了按持续时间加权的标准差音高 pitchSD ：

$$pitchSD = \sqrt{\frac{\sum_{i=1}^{n}(p_i-\bar{p})^2 d_i}{\sum_{j=1}^{n} d_j}} \quad (3)$$

对于特征方向，由于它描述的是短语级别的音乐特性而非整个作品，我们通过统计分析上升和下降段落的持续时间来接近这一问题。通过比较这些持续时间，我们确定作品的整体音调方向。具体来说，如果上升段落的总持续时间较长，则该作品被标记为具有上升音调方向。否则，该作品被标记为具有下降音调方向。

我们计算了表 I 中给出的 PCC，并在图 2 中绘制了 V/A 与八个特征之间的分布图表。这些发现指导了后续的数据处理和情感控制实验设计。第 II-C 节提到的 Rough4Q 粗略情绪标注策略也来源于此。

### B. 骨干网络

我们选择了预训练的 Tunesformer 作为骨干网络，该网络专门用于生成 ABC 记谱法。其输入数据格式分为两部分：控制码和 ABC 字符。后者表示传统的 ABC 记谱法音乐，其数据结构在文档 *ABC 音乐记谱法* [4] 中有详细说明。前者由四个标记组成：S（节段数）、B（小节数）、E（编辑距离相似度）和 D（第一段的时长）。

[4]https://trillian.mit.edu/ jc/music/abc/doc/ABC.html

对于预训练过程，我们考虑一个由配对数据集 $S$ 组成的得分数据集，其中每个配对为 $(x,y)$，$x$ 是输入音乐记谱，而 $y$ 是目标音乐记谱。每个得分表示为一系列小节补丁，每个小节补丁进一步分解为一系列字符 $\{(b_1^1,b_2^1,\ldots,b_n^1),(b_1^2,b_2^2,\ldots,b_n^2)\ldots,(b_1^m,b_2^m,\ldots,b_n^m)\}$。主干模型基于输入的记谱和先前生成的标记以自回归方式训练来预测目标记谱中的每个字符标记。形式上，预训练的目标是最小化目标序列中所有标记的交叉熵（CE）损失：

$$\mathcal{L}_{CE}(\theta) = -\sum_{(x,y)\in S}\sum_{i=1}^{m}\sum_{j=1}^{n} log p_\theta(b_j^i | x, b_{<j}^{<i}) \quad (4)$$

其中 $p_\theta$ 表示由 $\theta$ 参数化的骨干的概率分布函数，用于预测正确的字符，而 $b_j^i$ 表示第 $i$ 个条形补丁中第 $j$ 个字符，在第 $y$ 分数的第 $i$ 条补丁左侧第 $j$ 个字符之前的字符用 $b_{<j}^{<i}$ 表示。

在此基础上，我们引入了两个嵌入特征和三个可控特征以实现情感条件生成。微调过程中使用的损失函数与预训练过程中使用的一致。将主干结构与情感控制模块相结合的整体系统架构如图 3 所示。

## IV. 实验

### A. music21 解析速率

music21 解析率是指模型生成的乐谱中能被 music21 成功无误解析的比例。该指标有助于识别并过滤可能引发渲染失败、影响生成质量的错误乐谱。我们使用处理过的 EMOPIA、处理过的 VGMIDI 和 Rough4Q 数据集对骨干进行了微调，这些数据集具有相同的数据结构，包含三列：控制码、ABC 字符和 4Q 标签。鉴于 transformer 解码器的注意力机制是向右的，我们将 4Q 标签解析成字符串形式并合并到控制码的左侧。

我们在 Linux 环境中使用单个 H800 GPU 对这三个数据集进行了微调，批量大小为 1。一旦在微调期间的评估损失低于预训练模型观察到的最小评估损失，训练就提前停止，并保存表现出最佳性能的模型权重。这种方法有助于缓解过拟合，防止模型生成与训练集中旋律过于相似的旋律。我们使用上述获得的三个微调模型进行推理，从每个模型中生成 100 个 ABC 记谱法。然后计算这些乐谱的 music21 解析率，并在表 III 中给出。由于这三个数据集的大小不一致，使用了不同的采样率来控制用于微调的数据量以控制变量。

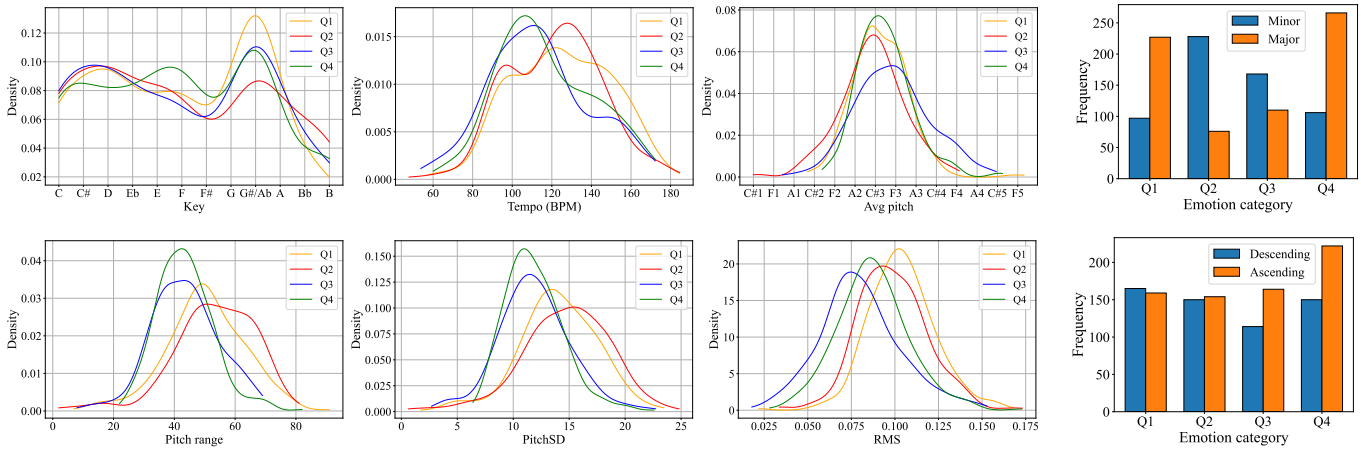


图 2: 高斯 KDE 图表用于显示 Russell 4Q 情绪在六个与音乐相关的特征上的分布：调式、节奏、平均音高、音高范围、音高标准差和 RMS（左侧的六个子图）；条形图显示了在不同模式和方向上 Russell 4Q 情绪频率的分布（右侧的两个子图）。

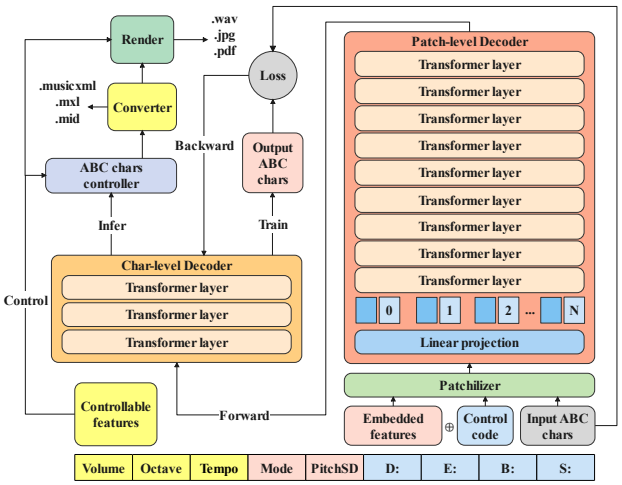


图 3: 整体系统架构包括主干的训练和推理分支，其下部分概述了当前使用的音乐特性。

表 III: 经过处理的 EMOPIA、处理过的 VGMIDI 和 Rough4Q 数据集微调的主干网络输出中的 music21 解析率比较。

| 处理过的数据集 | 情绪 opia | VGMIDI | 粗略 4Q |
|---|---|---|---|
| Sampling rate (%) | 24.58 | 56.68 | 1.01 |
| music21 parsing rate (%) | 28 | 75 | 99 |

旋律的质量是一个很难定义的指标，需要进行广泛的主观测试以尽量减少偏差从而获得可靠的结果。music21 的解析率作为可计算且客观的度量标准，是质量的一个必要条件。因此，本实验可以在一定程度上反映生成结果的质量。

### *B. 消融研究*

基于表 III 的结果，发现通过处理后的 EMOPIA 和 VGMIDI 进行微调的模型在旋律生成方面音乐 21 解析率不令人满意。因此，在后续实验中，我们使用了用 Rough4Q 微调的主干网络进行进一步研究。基于表 I 和图 2 的统计分析、音乐心理学结论的指导以及我们的手动听觉尝试，我们选择了以下五个情感控制特征：模式、节奏、音高标准差、音量（RMS）和八度（平均音高），来设计情感条件生成模板，在这个模板中，模式和音高标准差通过嵌入进行管理，其余则在主干的输出阶段进行控制。这五个特征的情感控制模板如下：

- **Q1** 主要模式，高音 SD，随机节奏从 160-184 BPM（大致在急板 – 活泼的之间），八度不变，音量增加 5dB；
- **Q2**: 次要模式，高音 SD，从 184-228BPM（大约在预备齐全 – 急板之间）随机节拍，音调降低两个八度，音量增加 10dB；
- **Q3**: 次要模式，低音调 SD，从 40-69BPM 的随机节奏（大致在广幅的 – 如歌的慢板之间），音调降低一个八度，音量不变；

- **Q4** 主要模式，低音 SD，从 40-69BPM 的随机节奏，八度和音量不发生变化。

使用此模板，我们为每个情绪类别生成了 25 个旋律，共计 100 首。在盲听条件下，我们首先让三名音乐爱好者聆听这些作品，并根据他们感知到的 4Q 情感对其进行标记。为了最小化主观偏见，我们采用了三分之二策略：如果至少两位听众将某一首曲子识别为某种特定情绪，则该情绪被认为是真实的；如果三位听众提供了三种不同的回答，我们将随机解决分歧并替换这三位听众重新进行测试。我们重复了这一过程直到所有分歧得到解决。在获取所有来自听众的标签后，我们将它们与提示中的情感条件进行了比较以检验情感控制模板的有效性。当前模板的情感生成准确率为 91%。此外，我们在该模板内的五个控制条件下进行了同样的消融实验，所有结果见表 IV 和图 4。这里的消融是指关闭模板中的指定特征的控制。

表 IV: 通过比较人类盲听生成旋律和情感提示的情感表现，并进行消融对比来评估生成模型的性能。

| 消融研究 | 准确率。(%) | F1 分数 (%) | 精度 (%) | 召回率 (%) |
|---|---|---|---|---|
| Tempo | **66.0** | 64.9 | **64.8** | **66.0** |
| PitchSD | 67.0 | **64.8** | 65.6 | 67.0 |
| Mode | 71.0 | 70.8 | 71.3 | 71.0 |
| Octave | 72.0 | 71.2 | 74.0 | 72.0 |
| Volume | 86.0 | 85.9 | 87.1 | 86.0 |
| - | 91.0 | 90.9 | 91.6 | 91.0 |

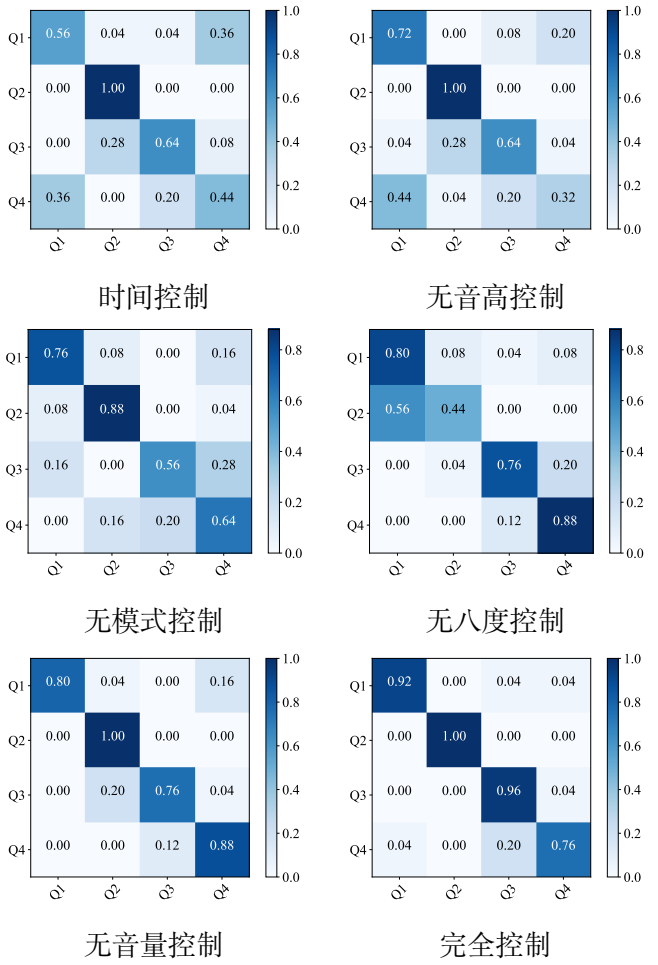


图 4: 人类盲听生成旋律的情绪与情绪提示在全控制和消融选项下的混淆矩阵，其中纵轴代表情绪提示，横轴代表参与者标记的情绪。

## V. 结论

从先前的实验结果来看，直接将 EMOPIA 和 VGMIDI 数据集转换为 ABC 记谱法以对 Tunesformer 进行微调无法保证在旋律生成上的高 music21 解析率。相比之下，在 Rough4Q 上通过音乐特征模板控制对其进行微调可以成为一种更可靠的情绪条件化旋律生成方法。诸如调式、节奏、音高 SD 和 RMS 等关键特性大致符合音乐心理学的研究成果，而平均音高并不能完全匹配情绪判断的复杂性。盲听测试验证了我们的模板对上述五个特性的控制效果，表明我们这种方法虽然不是一种纯粹端到端的情绪嵌入方法，仍能实现约 91%的情感生成表现。消融研究进一步证实了模板中这五个特征的控制可以显著影响生成性能，在其中节奏、音高 SD 和调式发挥着尤为重要的作用。

## 致谢

本工作得到了国家社会科学基金（21ZD19）和国家自然科学基金专项计划（T2341003）的支持。

## 参考文献

[1] Shulei Ji and Xinyu Yang, “Muser: Musical element-based regularization for generating symbolic music with emotion,” in *Proceedings of the AAAI Conference on Artificial Intelligence*. 2024, pp. 12821–12829, AAAI Press.

[2] Lijun Zheng and Chenglong Li, “Real-time emotion-based piano music generation using generative adversarial network (gan),” *IEEE Access*, 2024.

[3] Jingyue Huang, Ke Chen, and Yi-Hsuan Yang, “Emotion-driven piano music generation via two-stage disentanglement and functional representation,” in *Proceedings of the 25th International Society for Music Information Retrieval Conference, ISMIR 2024*, 2024.

[4] Serkan Sulun, Matthew EP Davies, and Paula Viana, "Symbolic music generation conditioned on continuous-valued emotions," *IEEE Access*, vol. 10, pp. 44617–44626, 2022.

[5] Lucas N Ferreira, Lili Mou, Jim Whitehead, and Levi HS Lelis, "Controlling perceived emotion in symbolic music generation with monte carlo tree search," in *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*. 2022, pp. 163–170, AAAI Press.

[6] Jacek Grekow and Teodora Dimitrova-Grekow, "Monophonic music generation with a given emotion using conditional variational autoencoder," *IEEE Access*, vol. 9, pp. 129088–129101, 2021.

[7] James A Russell, "A circumplex model of affect journal of personality and social psychology 39," *Í6I-I78*, 1980.

[8] Tomasz Michal Oliwa, "Genetic algorithms and the abc music notation language for rock music composition," in *Proceedings of the 10th Annual Conference on Genetic and Evolutionary Computation*, New York, NY, USA, 2008, GECCO '08, p. 1603 – 1610, Association for Computing Machinery.

[9] Shangda Wu, Xiaobing Li, Feng Yu, and Maosong Sun, "Tunesformer: Forming irish tunes with control codes by bar patching," in *HCMIR@ISMIR*, 2023.

[10] Luca Casini, Nicolas Jonason, and Bob LT Sturm, "Investigating the viability of masked language modeling for symbolic music generation in abc-notation," in *International Conference on Computational Intelligence in Music, Sound, Art and Design (Part of EvoStar)*. Springer, 2024, pp. 84–96.

[11] Shangda Wu, Yashan Wang, Xiaobing Li, Feng Yu, and Maosong Sun, "Melodyt5: A unified score-to-score transformer for symbolic music processing," in *Proceedings of the 25th International Society for Music Information Retrieval Conference, ISMIR 2024*, 2024.

[12] Hsiao-Tzu Hung, Joann Ching, Seungheon Doh, Nabin Kim, Juhan Nam, and Yi-Hsuan Yang, "Emopia: A multi-modal pop piano dataset for emotion recognition and emotion-based music generation," in *International Society for Music Information Retrieval Conference*, 2021.

[13] Lucas Ferreira and Jim Whitehead, "Learning to generate music with sentiment," in *Proceedings of the 20th International Society for Music Information Retrieval Conference, ISMIR 2019, Delft, The Netherlands, November 4-8, 2019*, 2019, pp. 384–390.

[14] Klaus R Scherer and Eduardo Coutinho, "How music creates emotion: A multifactorial process approach," *The emotional power of music, multidisciplinary perspectives on musical arousal, expression, and social control*, pp. 121–145, 2013.

[15] Zhaorui Liu and Zijin Li, "Music data sharing platform for computational musicology research (ccmusic dataset)," https://doi.org/10.5281/zenodo.5676893, nov 2021.

[16] Shangda Wu, Xiaobing Li, and Maosong Sun, "Chord-conditioned melody harmonization with controllable harmonicity," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[17] Eric Foxley and Seymour Shlien, "The nottingham music database," https://ifdo.ca/ seymour/nottingham/nottingham.html, October 2011.

[18] Federico Simonetta, "Enhanced wikifonia leadsheet dataset," https://doi.org/10.5281/zenodo.1476555, nov 2018.

[19] Helmut Scaffrath, Damien Sagrillo, Ewa Dahlig-Turek, and Seymour Shlien, "Essen folk song database," https://ifdo.ca/ seymour/runabc/esac/esacdatabase.html, October 2013.

[20] Karl Pearson, "Ii. mathematical contributions to the theory of evolution. ii. skew variation in homogeneous material," *Proceedings of the Royal Society of London*, vol. 57, no. 340-346, pp. 257–260, 1985.

[21] Emanuel Parzen, "On estimation of a probability density function and mode," *The annals of mathematical statistics*, vol. 33, no. 3, pp. 1065–1076, 1962.