arXiv:2512.17946v1 [cs.SD] 15 Dec 2025

让模型学会感受：面向符号音乐情绪识别的调式引导调性注入

夏海英[1,2], **黄忠义**[1,2],**谭玉梅**[1,2*]、**宋书香**[1,2*][1]广西类脑计算与智能芯片重点实验室，电子与信息工程学院，广西师范大学，中国桂林 541004[2]集成电路与微系统重点实验室，广西壮族自治区教育厅，中国桂林 541004 xhy22@gxnu.edu.cn, zoeyhuang@stu.gxnu.edu.cn,tanyumei@gxnu.edu.cn, songshuxiang@gxnu.edu.cn

摘要

音乐情感识别是符号音乐理解（SMER）中的一项关键任务。近期方法通过对大规模预训练模型（例如符号音乐理解领域的基准模型 MIDIBERT）进行微调，将音乐语义映射到情绪标签，已取得了令人鼓舞的结果。尽管这些模型能够有效捕捉分布式音乐语义，但往往忽略调性结构，尤其是音乐调式；根据音乐心理学，音乐调式在情感感知中发挥着关键作用。本文研究了 MIDIBERT 的表征能力，并发现其在捕捉调式与情感之间的关联方面存在局限。为解决这一问题，我们提出调式引导增强（MoGE）策略，将关于调式的心理学洞见融入模型。具体而言，我们首先开展调式增强分析，结果表明 MIDIBERT 无法有效编码情感与调式之间的相关性。基于这一观察，我们进一步识别出 MIDIBERT 中情感相关性最弱的层，并提出调式引导的特征级线性调制注入（MoFi）框架，以注入显式调式特征，从而增强模型在情感表征和推理方面的能力。在 EMOPIA 和 VGMIDI数据集上开展的大量实验表明，我们的调式注入策略显著提升了 SMER 性能，准确率分别达到 75.2% 和 59.1%。这些结果验证了调式引导建模在符号音乐情感识别中的有效性。

扩展版—

https://github.com/ZoeyHuang-paper/MoFi

引言

符号领域中的音乐情感识别在音乐理解中发挥着重要作用。它在音乐生成、音乐心理治疗、音乐推荐等方面具有广泛的应用。

推荐系统和人机交互。与基于音频的音乐情感识别相比，符号音乐提供了明确的结构信息，可以像自然语言中的词语一样进行处理。

由于符号音乐由遵循规则的符号序列构成，自然语言处理（NLP）的进步为符号音乐分析带来了新机遇，使其能够采用语言建模技术。在

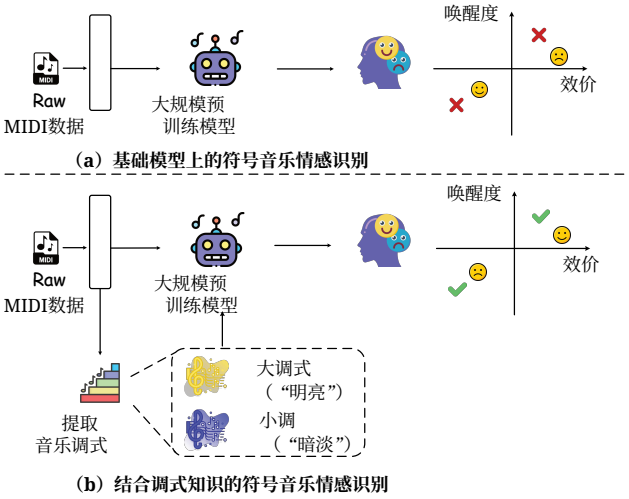


图 1：所提出的 MoFi 框架的整体概览。我们提取调式信息，并将其注入大规模预训练模型。(a) 展示了基础模型在没有明确调式指导的情况下进行微调的过程，并说明了音乐识别的局限性。(b) 展示了具备调式知识的模型。

近年来，基于 Transformer 的预训练模型（Vaswani 等，2017）表现卓越，激发了人们将 Transformer 架构（Huang 等，2019）应用于符号音乐理解任务的大量研究。在大规模 MIDI 音乐数据集上训练的 BERT 模型取得了显著进展，提升了人们对音乐的理解（Zeng 等，2021；Chou 等，2024）。现有方法通过设计多样化的预训练策略并改进符号编码格式，提高了模型捕捉更丰富音乐信息的能力（Zeng 等，2021；Tian 等，2024）。然而，由于符号音乐情感识别数据集规模有限，这些预训练模型在微调过程中往往无法有效习得与情感相关的音乐特征。此外，情感是人类参与音乐体验的主要驱动力，使人们能够体验并回应丰富多样的情感状态。相比之下，计算系统依靠智能算法，通过截然不同的机制处理和理解音乐。

*通讯作者。

尽管当前模型在符号音乐理解方面展现出强大能力，但它们并未捕捉到人类情绪反应背后的内在推理。正如（Li 等，2024）所指出的，为了更好地理解情感，模型需要明确捕捉与人类对情绪触发因素的心理反应相关的特征。

在本研究中，我们旨在使模型以符合人类理解的方式学习音乐感知的关键要素。为此，我们提出调式引导增强（MoGE）策略和调式引导的逐特征线性调制注入（MoFi）框架，用于诊断大规模预训练模型的局限性，并增强其情绪表征能力。我们采用 MIDIBERT（Chou 等，2024）作为骨干模型，它是符号音乐理解领域的基准模型。心理学和音乐理论领域的现有文献揭示了音乐调式与情绪感知之间存在密切联系（Kastner 和 Crowder，1990a年）。具体而言，我们首先进行调式增强分析，以探究 MIDIBERT 在多大程度上已理解音乐调式与情绪表达之间的关系。随后，为了更好地了解 MIDIBERT 的能力，我们逐层探测模型，估计其已编码的情绪信息量。这项分析找出了情绪信息量最低的层，将其选为 MoFi 的目标层。MoFi 框架提取显式的调式特征，并将其注入该层，以提升 MIDIBERT 的情绪表征和推断能力（见图 1）。

通过我们的设计，所提出的方法具备了调式知识，能够模拟人类对音乐的感知。此外，我们的模型继承了预训练 MIDIBERT 模型强大的音乐语义理解能力。因此，我们的方法在两个不同规模的数据集 EMOPIA（Hung 等，2021年）和 VGMIDI（Ferreira 和 Whitehead，2019年）上取得了显著成果。

我们的主要贡献如下：

- 我们采用调式引导增强（MoGE）策略，应对捕捉人类情绪反应内在推理机制这一挑战。该策略包含一项有针对性的诊断实验，用于分析模型的局限性。

- 我们发现，大规模预训练模型 MIDIBERT 忽视了调式特征与情绪之间的关系，因此提出了调式引导的逐特征线性调制注入（MoFi）框架。该框架能够将明确的音乐理论先验以细粒度且参数高效的方式融入 MIDIBERT，为弥补其已发现的知识缺口提供了一种有理论依据且可解释的解决方案。

- 我们在 EMOPIA 和 VGMIDI 这两个规模不同的数据集上进行了实验，取得了优于或媲美现有最先进方法的结果。

相关工作

符号音乐情感识别

音乐情感识别（MER）是符号音乐理解中的一项核心任务，要求模型直接根据音乐的创作结构推断高层次的情感状态。

早期的符号音乐情感识别（SMER）方法主要依赖基于音乐理论构建的手工特征（Wu 等，2014年；Malheiro 等，2016年），例如音高直方图、节奏密度和和声复杂度，并结合支持向量机（SVM）等传统分类器（Lin、Chen 和 Yang，2013）。尽管这些方法奠定了重要基础，但它们受限于对大量特征工程的需求，而这一过程不仅耗费人力，而且通常不足以捕捉音乐表现所具有的复杂且长程的时间依赖关系。近年来，随着深度学习的兴起，尤其是循环神经网络（RNN）（Lin 等，2017年）和卷积神经网络（CNN）（Fukushima，1980年）的应用，研究范式发生了显著转变，使得能够从原始符号数据中自动提取特征（Simonetta 等，2019年）。

近年来，随着 MIDIBERT（Chou 等，2024）、MusicBERT（Zeng 等，2021）和 PopMAG（Ren 等，2020）等大规模、基于 Transformer 的预训练模型问世，符号音乐理解领域取得了显著进展。受 BERT（Devlin 等，2019）在自然语言处理领域取得成功的启发，这些模型采用由预训练和微调两个阶段组成的框架。在预训练阶段，它们利用大规模无标注 MIDI语料库和自监督目标（例如掩码词元预测），学习丰富且包含上下文信息的音乐表征。随后，这些表征通过轻量级分类器适配到下游任务，在包括符号音乐情感识别（SMER）在内的多项挑战中取得了最先进的表现。尽管这些模型在实证中表现出色，其内部运作机制仍大多不透明。预训练目标侧重于词元级重建，却没有明确的激励机制促使模型学习抽象的音乐概念，例如调性不变性或调式相关的情感结构。这引出了一个根本问题：这些模型表现出色，是因为它们学到了可泛化的音乐理论知识，还是因为它们依赖于训练数据中的表面相关性？我们的研究旨在通过考察这些模型是否真正内化了音乐理论的基本原则，尤其是对情感识别至关重要的音乐调式，来回答这一问题。

音乐特征与情感

音乐结构与人类情感之间的关系，是音乐学和音乐认知心理学中的一项基本原理。几个世纪以来，理论家和作曲家一直强调音乐调式的情感作用（Bowling 2013）。实证研究一再表明，大调与小调的区别是西方音乐中感知情感效价的主要线索（

Justus、Gabriel 和 Pfaff 2018；Gagnon 和 Peretz 2003）。具体而言，大调音乐始终与快乐、喜悦等积极情绪相关联，反映出较高的情感效价。相较之下，小调音乐则稳定地与悲伤、忧郁等消极情绪相关联，表明其效价较低（Krumhansl 1997）。文献（Kastner 和 Crowder 1990b）表明，积极情绪与大调音乐相关，而小调音乐往往会引发消极情绪。Gerardi 和 Gerken（Gerardi 和 Gerken 1995）、Gregory 等人（Gregory、Worrall 和 Sarge 1996）以及 Dalla Bella 等人（Dalla Bella 等人 2001）也报告了类似发现。尽管这种关联并非绝对，也可能受到文化背景等其他因素的影响，但其统计稳健性已有充分记录，并构成人类听众音乐感知与音乐素养的基本组成部分。

关键在于，调式与情绪之间的关联对调性相对不变。尽管移调会带来细微变化，但听众通常会认为，C大调的乐曲与G大调的乐曲传达着相似的喜悦情绪。传达主要情感特征的是调式的音程结构，而非具体调性（BOYDEN 1994年）。因此，为了减少噪声，我们在四分类任务中仅考虑注入两种调式（大调和小调）。在此基础上，我们提出调式引导增强（MoGE）策略。首先，MoGE考察大规模预训练模型是否捕捉到了调式与情绪的关联。基于这一分析，我们随后利用调式引导的特征级线性调制注入（MoFi）框架，将显式的调式知识注入表现不佳的情绪识别层，从而增强MIDIBERT的识别能力。

初步研究

在将调式相关知识注入预训练符号音乐模型之前，首要问题是：MIDIBERT是否已经学会了音乐调式所蕴含的情绪含义？回答这一问题对于判断是否需要这些理论特征，以及应将其整合到模型架构中的什么位置，至关重要。

为实证探究这一问题，我们采用 EMOPIA（Hung 等，2021年）作为初步研究数据集。EMOPIA 是一个广泛用于符号音乐情感识别任务的数据集，其中包含标注了 Russell 四象限模型（Russell，1980年）情绪标签的 MIDI文件。基于这一设置，我们开展了以下两项初步研究。

基于数据增强的诊断研究 为评估 MIDIBERT 是否已内化调式与情绪之间的乐理关联，我们设计了一项基于保持调式的移调的诊断实验。具体而言，我们通过在单个八度范围内进行保持调式的移调，对初步研究数据集进行增强，并评估 MIDIBERT 在原始数据和增强数据上的表现。若增强后模型表现显著提升，则表明模型未能捕捉音乐调式的情感特征，暗示其对调式相关知识的整合不足。反之，若模型在两种条件下的表现相当，则说明这种调式与情绪之间的关联可能已隐式编码在模型学习到的表征中。

逐层表征分析 在本研究中，我们冻结预训练 MIDIBERT 模型的所有参数，仅保留分类前的自注意力层和 softmax 分类头可训练。我们的目标是确定哪些层保留了与情感相关的信息，以及哪些层未能有效编码此类知识。该分析有助于为后续注入与调式相关的先验知识选择目标层。

数据增强的影响

我们通过将每段音乐片段随机上移或下移若干个半音，对 EMOPIA 进行保持调式不变的数据增强。移调范围限制在一个八度内，并统一应用于片段中的所有音符。由于片段中的所有音符一同移调，其内部音程关系保持不变，从而保留了潜在的音乐调式及其相关情感特质。这种增强方法利用音乐理论先验知识合成新的样本，在保留与情绪相关结构的同时，为训练集提供了一种简单而有效的扩充方式，并注入任务特定的归纳偏置。

我们遵循 MIDIBERT（Chou 等，2024）提供的官方数据划分，将训练集、验证集和测试集的比例设为 8:1:1，并确保同一首 Song 的所有片段都被分配到同一子集中，以防止数据泄漏。为减少训练方差的影响，我们报告使用不同随机种子进行五次运行的平均表现。

表 1 展示了 MIDIBERT 在原始数据集与增强数据集上训练后的分类准确率。结果显示，使用增强数据时，识别准确率显著提升，这表明 MIDIBERT 尚未充分内化与音乐调式相关的情感规律。

数据集	**原始**	**增强**
准确率	0.675	0.723

表 1：使用和不使用数据增强时在 EMOPIA 数据集上的准确率。

MIDIBERT 在符号音乐情感识别任务上的逐层分析

图2展示了在未经过数据增强的原始数据集上，MIDIBERT 各层在 SMER 任务中的表现。微调期间，仅更新了一个用于权重平均的自注意力模块和分类头，其余参数均被冻结。因此，各层之间的表现差异

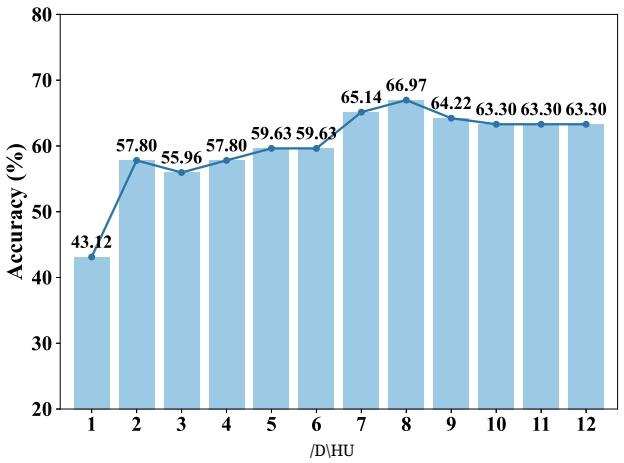


图2：MIDIBERT 在 EMOPIA 数据集上各层的表现。

主要反映了它们各自在 SMER 任务中提取与情感相关特征方面的贡献。中间层的准确率最高，而较低层和较高层的效果均有所下降。这一模式与自然语言处理领域的既有发现一致：中间的 Transformer 层往往编码出语义信息最丰富的表征（Clark等（2019）；Hewitt和Manning（2019））。

进一步分析表明，较低层的表现始终最弱，这说明它们编码的情感特征相关信息相对较少。相比之下，中间层的表现显著更强，表明它们可能在捕捉符号音乐中的情感模式方面发挥着重要作用。较高层的表现也不尽如人意，这可能是因为它们专门用于模型最初的预训练目标，即基于音乐上下文信息进行掩码词元预测，而非以情感理解为核心的任务。

总之，数据增强研究表明，MIDIBERT 尚未有效捕捉符号音乐情感识别中的音乐调式信息。逐层性能分析进一步显示，较低层对情感表征的贡献最小，说明这些层最需要改进。综合来看，这些发现指明了一个很有前景的优化方向：将与调式相关的先验知识注入 MIDIBERT 的最低层（第一层），以增强其情感建模能力。下一节将介绍我们实现这一策略的方法。

方法论

在本节中，我们将音乐调式的先验知识融入 MIDIBERT 模型，以增强其情感表征能力。根据上一节的研究发现，我们将知识整合的目标设定为模型的第一层。鉴于每个注意力层都会计算复合词（CP）之间的成对关系

词元（Hsiao等（2021）），且自然语言处理领域的先前研究表明，第一层呈现出宽泛且缺乏区分性的注意力模式（Clark等（2019）；Rogers、Kovaleva和Rumshisky（2021）），因此我们着重利用与调性相关的信息调节第一层中的注意力计算。下文将介绍所提出方法的具体细节。

MIDIBERT与复合词

MIDIBERT（Chou 等，2024）是一种基于 Transformer 的预训练基准模型，用于符号音乐理解。该模型使用大规模 MIDI 文件语料进行训练，以学习丰富的结构和上下文模式，并生成高质量的音乐表征。我们选择 MIDIBERT 作为骨干编码器，因为它具备强大的表征能力和出色的泛化能力。这些特质使其适用于各种符号音乐理解和生成任务。它已被广泛用作符号音乐领域的标准基础模型。

与自然语言处理中的词语类似，符号音乐序列（MIDI 文件）首先被转换为一系列离散的 CP 事件，然后作为输入传入 MIDIBERT。该编码方案将每个时间步出现的多种音乐属性（如小节、位置、音高和时值）合并为一个词元。每个 CP 词元由四个不同的子事件构成，它们共同编码音乐结构的时间（水平）和和声（垂直）维度。附录中提供了 CP 词的详细构造方法。

- 小节事件标记新小节的开始，并传达有关节拍边界的信息。
- 位置事件指定音符在小节中的相对位置，并体现节奏结构。
- 音高事件表示音符的绝对音高，包括音级、八度和变音记号。
- 时值事件表示音符的时值，并有助于保持音乐流动的连贯性。

MIDIBERT 的核心是一个标准的 Transformer 编码器堆栈，采用了 BERT 在自然语言处理中确立的架构范式。每个 Transformer 编码器层由一个多头自注意力机制和一个前馈网络组成。

MIDIBERT 中的多头注意力 该机制使模型在编码每个词元时，能够关注序列中的所有其他复合词嵌入，并根据上下文相关性分配权重。给定输入嵌入序列 $\mathbf{E} \in \mathbb{R}^{L \times d_{\text{model}}}$，其中 L 是序列长度，d_{model} 是隐藏维度，多头自注意力层首先将输入投影为查询 (Q)、键 (K)和值 (V) 矩阵。对于每个注意力头，这些投影通过可学习的参数矩阵$\mathbf{W}_i^Q$、$\mathbf{W}_i^K$和$\mathbf{W}_i^V$计算得出：

$$\text{多头}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{拼接}(\text{头}_1, \cdots, \text{头}_h)\,\mathbf{W}^{\mathbf{O}}$$

$$\text{head}_i = \text{Attention}\left(\mathbf{Q}\mathbf{W}_i^{\mathbf{Q}}, \mathbf{K}\mathbf{W}_i^{\mathbf{K}}, \mathbf{V}\mathbf{W}_i^{\mathbf{V}}\right) \quad (1)$$

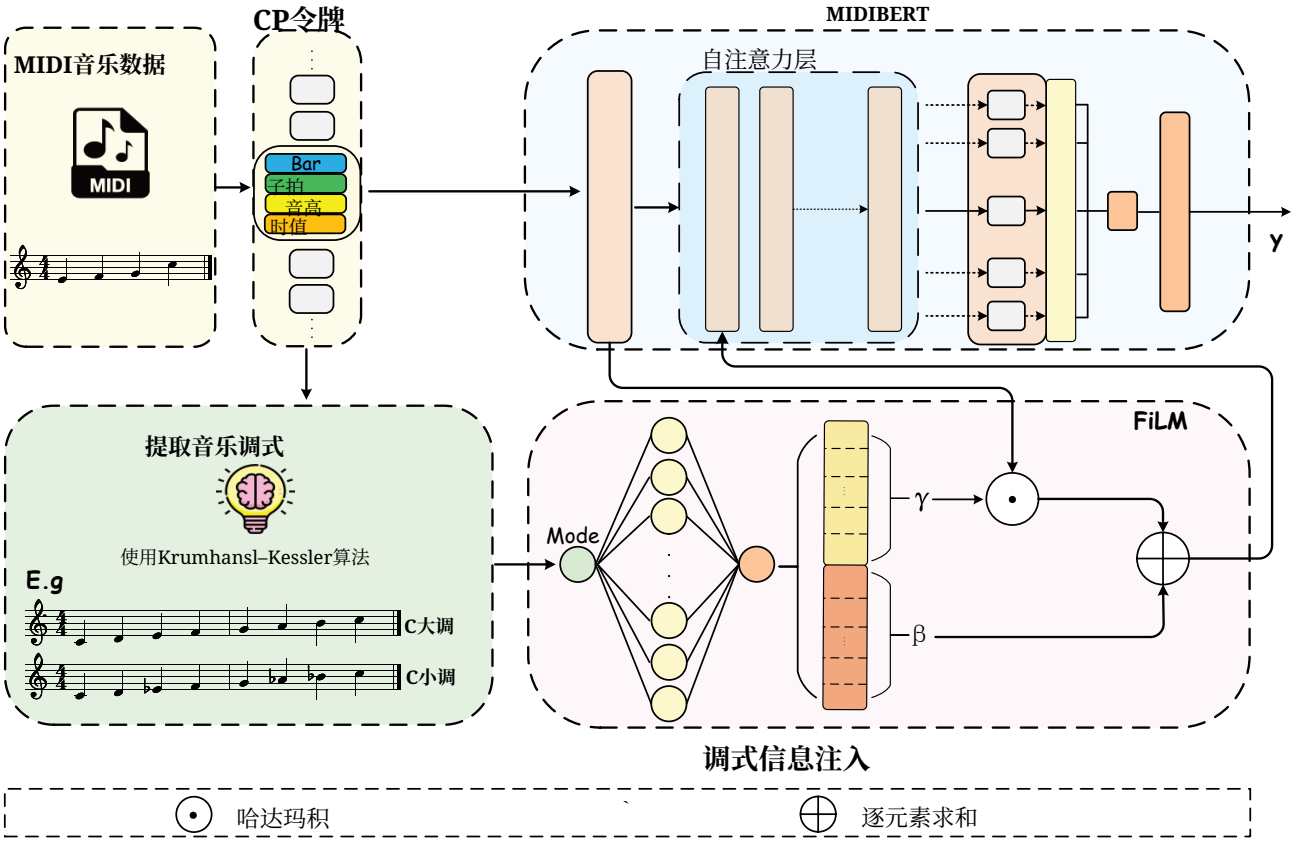


图 3：MoFi 的架构示意图。我们的方法首先使用 Krumhansl–Kessler算法提取 CP词元的调式知识（大调/小调），然后将其传递给 FiLM（特征级线性调制）方法。注入的信息被整合到大规模预训练模型 MIDIBERT 的第一层中，从而增强其捕捉符号音乐情感细微差异的能力。调式知识的注入使模型能够更好地适应情感特征，同时保留预训练 MIDIBERT 框架的优势。

其中，$\mathbf{W}_i^Q$ 、$\mathbf{W}_i^K$和$\mathbf{W}_i^V$是可学习的投影矩阵，用于将输入嵌入映射到较低维的子空间。所有注意力头的输出会先拼接，再使用矩阵$\mathbf{W}^O$投影回原始维度 d_{model} 。每个注意力头均采用缩放点积注意力机制计算：

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V \qquad (2)$$

其中， d_{k} 是键向量的维度，softmax函数用于对序列中的注意力分数进行归一化。该机制使模型能够捕捉符号音乐中的长程依赖和复杂的结构模式，而这些对于理解高层次的音乐语义至关重要。

前馈网络 自注意力机制的输出会传入一个两层的位置前馈网络，该网络在每个序列位置独立进行非线性变换。该组件进一步优化表示，并增强模型捕捉复杂音乐模式的能力。

MIDIBERT 通过在大型 MIDI文件语料库上采用掩码音乐建模目标，以自监督方式进行预训练。这种训练范式使模型能够学习符号音乐丰富且具备上下文感知能力的表示。在这一预训练基础之上，我们融入音乐调式的外部知识，以增强模型识别音乐情感内容的能力。

调式知识提取

我们的调式引导的逐特征线性调制注入（MoFi）框架的一个关键步骤，是从符号音乐数据中提取准确且可迁移的音乐调式先验知识。这些知识构成了后续注入模型的基础。在西方音乐理论中，大调式通常与明亮感、愉悦和喜悦等积极情感状态相关联，而小调式则更常与忧郁、悲伤和沉思等消极或内省情绪相关联（Horn和Huron，2015年）。调式与情感之间这一广为确立的关联，是符号音乐情感识别的重要先验。

为捕捉关键调式线索并减少数据噪声，我们采用二元分类方案，仅区分大调式和小调式。尽管还存在多利亚调式和利底亚调式等其他调式类型，但它们在符号音乐语料库中相对少见，且呈现出的情感特征一致性较低。这种简化保留了与调式相关的最显著情感差异，同时提升了数据平衡和模型泛化能力。此外，由于大调式往往传达更明亮的情感，而小调式则与较阴暗的感受相关，因此在四分类任务中，我们采用二元调式分类（大调式与小调式），而不使用全部24个调性（以主音定义），以减少噪声。

我们采用 Krumhansl–Kessler（K-K）算法（Krumhansl 1997），以自动且符合音乐理论的方式从符号化 MIDI 数据中提取调式信息。该算法是音乐信息检索（MIR）领域广泛认可的技术，植根于认知心理学，能够紧密模拟人类的调性识别感知过程。我们使用 music21 工具包实现该方法。每首乐曲提取出的调式先验都编码为一个独热向量，用以表示大调或小调。

通过 FiLM 注入调式

为了将提取出的调式先验融入 MIDIB-ERT 模型，我们采用特征级线性调制（FiLM）（Perez 等人，2018年）。FiLM 是一种最初为计算机视觉领域开发的条件化机制。它对神经网络的中间特征应用仿射变换，使模型能够根据外部输入动态调制其内部特征表示。与注意力机制等其他注入方法不同，FiLM 灵活性更高且参数效率更优，因此特别适合将有针对性的领域知识整合到预训练架构中。

FiLM 的核心思想是利用条件向量c （表示调式知识）生成两个调制参数：缩放因子 γ 和平移因子 β。这些参数通过逐元素仿射变换应用于序列特征表示 x。具体而言，参数生成网络 f_{cond} 将条件向量 c 映射到参数空间：

$$[\gamma, \beta] = f_{\text{cond}}(c) \tag{3}$$

随后，FiLM 操作将这些参数应用于输入特征表示 x，如下所示：

$$\text{FiLM}(x, c) = \gamma \odot x + \beta \tag{4}$$

其中，$\odot$ 表示哈达玛积。在这一表述中，条件信息 c 以细粒度的方式调制 x 的激活模式，从而根据先验知识引导模型强调或减弱特定的特征维度。

如前述分析所示，MIDIBERT 中的第一个 Transformer 层编码了最低层级的与调式相关的情感信息。在我们提出的 MoFi 架构中，我们在复合词嵌入层与第一个 Transformer 编码器层之间引入了一个基于 FiLM 的条件化模块。该集成使输入表示在进行自注意力计算之前融入对调式感知的条件信息，从而使所有后续表示都受到这一关键音乐理论先验的影响。如图 3 所示。

实验

设置

数据集 在我们的实验中，我们在两个广泛使用的基准数据集上评估我们的方法：小规模的 VG-MIDI 和相对大规模的基于 MIDI 的音乐情感数据集 EMOPIA。

EMOPIA EMOPIA 数据集（Hung 等，2021）是一个全面收录流行钢琴音乐片段的数据集，旨在支持符号与音频领域的情感识别。该数据集包含 1,087 个片段，取自 387 首不同的歌曲，每个片段均标注了情绪标签。情绪采用 Russell 四象限环形模型（Russell，1980）进行分类，该模型定义了四个情感象限（Q1–Q4）（见表 2）。该数据集已广泛应用于音乐情感分类。

原始音频录音通过线上公开可用的元数据收集。相应的 MIDI 文件通过高保真钢琴转录模型（Kong 等，2021）进行转录生成。数据集创建者随机抽取了部分节选片段进行人工检查，并确认音高、力度和时值均得到了准确保留。由于经处理的环境音效会对转录质量产生负面影响，因此最终数据集中排除了带有此类音效的曲目。

分类体系	描述
快乐（Q1）	高效价高唤醒（HVHA）
悲伤（Q2）	低效价高唤醒（LVHA）
平静（Q3）	低效价低唤醒（LVLA）
愤怒（Q4）	高效价低唤醒（HVLA）

表2：Russell四象限模型的分类体系。

VGMIDI VGMIDI数据集（Ferreira 和 Whitehead，2019年）包含 MIDI 格式的电子游戏音乐曲目。其中包括 200 首带有相应情绪标注的 MIDI 乐曲、 97 个对齐的音频版本，以及另外 3,850 首未标注乐曲。本研究仅使用带标注的 MIDI 曲目。每段带标注的片段均由 30 名标注者依据情感环形模型进行评分，该模型由效价–唤醒度（VA）维度定义。为确保与 EMOPIA 保持一致，VGMIDI 中的 VA 标注被映射为 EMOPIA 使用的 Russell 四象限分类方案（见表 2）。数据集划分为训练集（80%）、验证集（10%）和测试集（10%）。表 4 报告了 EMOPIA 和 VGMIDI 的其他统计信息。

模型	EMOPIA		VGMIDI	
	准确率	**F1**	**准确率**	**F1**
支持向量机（Lin、Chen 和 Yang，2013年）	0.477	0.476	0.451	0.377
LSTM-Attn（Hung 等，2021年）	0.647	0.563	0.417	0.260
MIDIGPT（Ferreira、Lelis 和 Whitehead，2020年）	0.587	0.572	0.538	0.505
MT-MIDIGPT（Qiu、Chen 和 Zhang，2022年）	0.625	0.611	0.585	0.509
MT-MIDIBERT（Qiu、Chen 和 Zhang，2022年）	0.676	0.664	0.498	0.453
BiLMA（Xiao 等，2024年）	0.708	0.631	0.572	0.478
MIDIBERT-Piano（基线）（Chou 等，2024）	0.634	0.628	0.473	0.432
Ours	**0.752**	**0.751**	**0.591**	**0.587**

表3：现有 MIDI 模型与我们的方法的比较。

INFO	**EMOPIA**	**VGMIDI**
MIDI 数量	1087	200
训练集-验证集-测试集划分	8:1:1	8:1:1
来源	YouTube	电子游戏原声带
音乐类型	各种	各种
单首时值	约 30 秒	无

表 4：EMOPIA 和 VGMIDI 概览。

训练详情 我们采用 MIDIBERT 作为骨干模型，并沿用其原始架构（Chou 等，2024）。该模型包含 12 个 Transformer 层，每层有 12 个注意力头，隐藏层大小为 768，总参数量为 1.11 亿。我们使用随 MIDIBERT 论文（Chou 等，2024）一同发布的预训练检查点初始化模型。

对于 SMER 任务，两个数据集均按照 8:1:1 的比例划分为训练集、验证集和测试集。对于 EMOPIA，我们采用与 MIDIBERT（Chou 等，2024）相同的划分方式。EMOPIA 和 VGMIDI 的批次大小分别设为 16 和 8。模型在单张 NVIDIA GeForce RTX 3090 GPU 上最多微调 20 个周期，总训练时间不到 30 分钟。当验证集性能连续三个周期没有提升时，便提前停止训练。

训练开始时，FiLM 层初始化为缩放因子 $\gamma = 1$ 和平移因子 $\beta = 0$，从而有效保留 MIDIBERT 中蕴含的预训练知识。这种初始化有助于保持训练稳定性，使模型能够从预训练的 MIDIBERT 嵌入开始，逐渐融入与调式相关的信息，避免随机初始化的参数造成干扰。

结果分析

表3总结了所提出方法在符号音乐情感识别（SMER）任务上的整体表现，包括准确率和宏平均F1。结果与先前研究报告的发现（Xiao 等，2024）一致。作为比较，我们还报告了传统支持向量机（SVM）（Lin、Chen 和 Yang，2013）基线以及若干先前的符号音乐情感识别模型。值得注意的是，我们的调式引导特征级线性调制注入（MoFi）框架在准确率方面实现了在 EMOPIA 和 VGMIDI数据集上均提升11.8%，相较于基线，F1分数分别提升12.3%和15.5%。这些提升仅通过最小程度的架构修改即可实现。

消融研究

表 5 展示了完整模型在有无模式注入情况下的对比消融研究。我们观察到，移除基于 FiLM 的模式注入模块会导致性能明显下降。这一结果表明，引入显式的调式信息有助于提升模型识别音乐情感的能力。通过在表示学习的早期阶段注入这一音乐理论先验，模型获得了原始架构未能捕捉的调式感知能力。

模型	EMOPIA		VGMIDI	
	准确率	**F1**	**准确率**	**F1**
完整模型	0.752	0.751	0.591	0.587
不含调式知识	0.716	0.715	0.500	0.365

表 5：通过 FiLM 注入调式的消融研究。

结论

本文提出调式引导增强（MoGE），作为一种分析策略，用于评估 MIDIBERT 捕捉音乐调式音乐理论知识的能力，并揭示其局限性。在此分析基础上，我们提出调式引导的逐特征线性调制注入（MoFi），这是一种轻量而有效的框架，旨在将调式相关先验注入 MIDIBERT。我们的方法并不主要依赖资源密集型预训练，而是利用音乐调式信息调制输入表示，并引导模型的注意力。通过将符号音乐建模与音乐理论先验相结合，MoFi 框架表明，即使是基于领域知识的轻量级引导，也能带来显著提升。这些发现鼓励人们进一步探索人工智能音乐领域中可解释且以理论为导向的方法，并有望使情感识别以外的广泛任务受益。

致谢

本研究部分得到中国国家自然科学基金项目（编号 62366006 和 62366005）、广西壮族自治区教育厅人工智能与信息处理重点实验室（河池学院）项目（编号 2024GXZDSY011）、广西精密导航技术与应用重点实验室研究项目（编号 DHKL2417）、广西自然科学基金青年项目（编号 2025JJB170055），以及广西师范大学国家自然科学基金2025年联合培养项目的资助。

参考文献

Bowling, D. L. 2013年. 旋律中音乐调式情感特征的声乐基础. *Frontiers in Psychology*, 4: 464. BOYDEN, D. 1994年. 基于《对位法教程》. CARL FISCHER. ISBN 9780825827648. Chou, Y.-H.; Chen, I.-C.; Ching, J.; Chang, C.-J.; and Yang, Y.-H. 2024年. MIDIBERT-Piano：面向符号音乐分类任务的大规模预训练. 创意音乐系统期刊, 8(1). Clark, K.; Khandelwal, U.; Levy, O.; and Manning, C. D. 2019年. BERT关注什么？对BERT注意力的分析. In Linzen, T.; Chrupala, G.; Belinkov, Y.; and Hupkes, D., eds., *2019*年*ACL*研讨会论文集 *BlackboxNLP*：自然语言处理神经网络的分析与解读，*BlackboxNLP@ACL 2019*，意大利佛罗伦萨，*2019*年*8*月*1*日, 276–286. Association for Computational Linguistics.

Dalla Bella, S.; Peretz, I.; Rousseau, L.; and Gosselin, N. 2001年. 音乐中速度与调式情感值的发展研究。*Cognition*, 80(3): B1–B10.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019年. BERT：用于语言理解的深度双向Transformer预训练。载于 Burstein, J.; Doran, C.; and Solorio, T. 编，*2019*年北美计算语言学协会会议论文集：人类语言技术，第*1*卷（长篇与短篇论文），4171–4186。明尼阿波利斯，明尼苏达州：计算语言学协会。

Ferreira, L.; Lelis, L.; and Whitehead, J. 2020年. 面向桌面角色扮演游戏的计算机生成音乐。载于*AAAI*人工智能与互动数字娱乐会议论文集，第16卷，59–65。

Ferreira, L. N.; 和 Whitehead, J. 2019. 学习创作带有情感的音乐。*ISMIR'19*。

Fukushima, K. 1980年。Neocognitron：一种自组织神经网络模型，用于实现不受位置偏移影响的模式识别机制。《生物控制论》，36(4)：193–202。

Gagnon, L.; and Peretz, I. 2003年. 调式和速度对等音调旋律"快乐—悲伤"判断的相对贡献。*Cognitionandemotion*,17(1):25–40.

Gerardi, G. M.; and Gerken, L. 1995年. 情感反应对调式和旋律轮廓的反应发展。《音乐感知》, 12(3): 279–290.

Gregory, A. H.; Worrall, L.; 和 Sarge, A. 1996. 幼儿对音乐的情绪反应的发展。《动机与情绪》，20(4)：341–348。

Hewitt, J.; 和 Manning, C. D. 2019. 用于在词表示中发现句法的结构探针。载于《*2019*年北美计算语言学协会会议论文集：人类语言技术》，第*1*卷（长篇与短篇论文），4129–4138。Horn, K.; 和 Huron, D. 2015. 关于1750—1900年大调与小调的使用变化。《在线音乐理论》，21(1)：1–11。Hsiao, W.-Y.; Liu, J.-Y.; Yeh, Y.-C.; 和 Yang, Y.-H. 2021年。复合词变换器：在动态有向超图上学习创作完整歌曲音乐。载于《*AAAI*人工智能大会论文集，第35卷，178–186。

Huang, C. A.; Vaswani, A.; Uszkoreit, J.; Simon, I.; Hawthorne, C.; Shazeer, N.; Dai, A. M.; Hoffman, M. D.; Dinculescu,M.; 和 Eck,D.2019. 音乐变换器：生成具有长期结构的音乐。载于第*7*届国际学习表征会议，*ICLR 2019*，新奥尔良，路易斯安那州，美国，*2019*年*5*月*6*日至*9*日。OpenReview.net。

Hung, H.; Ching, J.; Doh, S.; Kim, N.; Nam, J.; 和 Yang, Y. 2021年。EMOPIA：用于情感识别和基于情感的音乐生成的多模态流行钢琴数据集。载于 Lee, J. H.; Lerch, A.; Duan, Z.; Nam, J.; Rao, P.; van Kranenburg, P.; 和 Srinivasamurthy, A. 编，第*22*届国际音乐信息检索学会会议论文集，*ISMIR 2021*，线上，*2021*年*11*月*7*日至*12*日，318–325。

Justus, T.; Gabriel, L.; and Pfaff, A. 2018年. 音乐中的形式与意义：重新审视大调与小调的情感特征。《听觉感知与认知》, 1(3-4): 229–247. Kastner, M. P.; and Crowder, R. G. 1990a年. 大调/小调区别的感知：IV. 幼儿的情绪内涵。《音乐感知》, 8(2): 189–201. Kastner, M. P.; and Crowder, R. G. 1990b. 大调/小调区别的感知：IV. 幼儿的情绪内涵。《音乐感知》，8(2): 189–201. Kong, Q.; Li, B.; Song, X.; Wan, Y.; and Wang, Y. 2021年. 通过回归起始与结束时间，实现带踏板的高分辨率钢琴转录。*IEEE/ACM*音频、语音与语言处理汇刊, 29: 3707–3717. Krumhansl, C. L. 1997年. 音乐情绪与心理生理学的探索性研究。《加拿大实验心理学杂志》*/Revue canadienne de psychologie exp´erimentale*, 51(4): 336. Li, J.; Lin, Z.; Wang, L.; Si, Q.; Cao, Y.; Yu, M.; Fu, P.; Wang, W.; and Zhou, J. 2024年. 大声思考：对话中的情绪推断解释。*arXiv*预印本 *arXiv:2406.04758*。

Lin, Y.; Chen, X.; and Yang, D. 2013年. 基于MIDI的音乐情感识别探索。载于*ISMIR*, 221–226. Lin, Z.; Feng, M.; dos Santos, C. N.; Yu, M.; Xiang, B.; Zhou, B.; and Bengio, Y. 2017年. 结构化自注意力句子嵌入。载于国际学习表征会议。Malheiro, R.; Panda, R.; Gomes, P.; and Paiva, R. P. 2016年. 用于音乐歌词分类与回归的情感相关特征。*IEEE*情感计算汇刊, 9(2): 240–254. Perez, E.; Strub, F.; De Vries, H.; Dumoulin, V.; and Courville, A. 2018年. FiLM：通过通用条件层进行视觉推理。载于*AAAI*人工智能大会论文集, 第32卷。Qiu, J.; Chen, C.; and Zhang, T. 2022年. 一种用于符号音乐情感识别的新型多任务学习方法。*arXiv*预印本 *arXiv:2201.05782*。Ren, Y.; He, J.; Tan, X.; Qin, T.; Zhao, Z.; and Liu, T.-Y. 2020年. PopMAG：流行音乐伴奏生成。载于第*28*届*ACM*多媒体国际会议论文集, MM ’20, 1198–1206. 美国纽约：计算机协会。ISBN 9781450379885. Rogers, A.; Kovaleva, O.; and Rumshisky, A. 2021年. BERT学入门：关于BERT工作原理的已知结论。计算语言学协会汇刊, 8: 842–866. Russell, J. A. 1980年. 情感环形模型。人格与社会心理学杂志, 39(6): 1161. Simonetta, F.; Chac´on, C. E. C.; Ntalampiras, S.; Widmer, G.; et al. 2019年. 一种用于符号乐谱旋律线识别的卷积方法。载于*ISMIR 2019*, 1–8. ISMIR。Tian, J.; Li, Z.; Li, J.; and Wang, P. 2024年. N元语法无监督复合化与特征注入，以更好地理解符号音乐。载于*AAAI*人工智能大会论文集, 第38卷, 15364–15372. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L. u.; and Polosukhin, I. 2017年. 注意力就是一切。载于Guyon, I.; Luxburg, U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; and Garnett, R. 编，神经信息处理系统进展, 第30卷。Curran Associates, Inc. Wu, B.; Zhong, E.; Horner, A.; and Yang, Q. 2014年. 通过多标签、多层、多实例、多视图学习进行音乐情感识别。载于第*22*届*ACM*多媒体国际会议论文集, 117–126. Xiao, Y.; Ruan, H.; Zhao, X.; Jin, P.; Tian, L.; Wei, Z.; Cai, X.; Wang, Y.; and Liu, L. 2024年. 一种用于音乐情感识别的高效双模态融合框架。*IEEE*情感计算汇刊。Zeng, M.; Tan, X.; Wang, R.; Ju, Z.; Qin, T.; and Liu, T.-Y. 2021年. MusicBERT：基于大规模预训练的符号音乐理解。载于Zong, C.; Xia, F.; Li, W.; and Navigli, R. 编，计算语言学协会论文集：*ACL-IJCNLP 2021*, 791–800. 线上：计算语言学协会。