音乐回路：追踪并控制大型音乐生成模型中的风格动态

RuijiYu[1], **WenyeMa**[2] **&GusXia**[1]1穆罕默德·本·扎耶德人工智能大学 [2]麦吉尔大学 {ruiji.yu,gus.xia}@mbzuai.ac.ae, wenye.ma@mail.mcgill.ca

arXiv:2611.29475v1 [cs.SD] 22 Nov 2026

摘要

音乐风格是一种贯穿整首乐曲的全局属性，也是现代大型音乐生成模型中用户广泛使用的调控手段。然而，风格如何在这些模型内部表征，以及哪些组件驱动风格，仍在很大程度上尚未得到探索。在语言模型中，回路分析会将某种表征追踪到由少数关键组件构成的稀疏子网络，即注意力头和多层感知机神经元；该方法已被成功用于解释和控制情绪等全局属性。在本研究中，我们利用回路分析探究音乐风格背后的机制。基于 YuE2，我们构建了一个将相同歌词与五种风格配对的数据集，并设计了一个三步框架。首先，阶段定位表明，风格主要形成于乐谱规划阶段。其次，对该阶段各位置进行静态观察进一步揭示，风格并非固定向量，而是跨词元的动态表征。第三，我们因此将各位置的激活差异合并为每种风格对应的一个方向，并追踪驱动该方向的神经元和注意力头，从而得到一个稀疏电路。在不使用任何风格文本的情况下，引导该电路能够成功地将一首乐曲转换为目标风格，其效果可媲美甚至超越风格提示。据我们所知，这是首项在大型音乐生成模型中从电路层级分析和控制音乐风格的研究。音频和乐谱示例见 https://aristo233.github.io/Music_Circuit/。

1 引言

音乐生成模型正在快速发展。[1] 符号模型逐个词元地写入音符事件或 ABC 记谱法等文本乐谱（Huang 等，2019；Yu 等，2022；Yuan 等，2024；Qu 等，2024；Wang 等，2025b），而音频模型则生成离散声学词元，并使用文本控制流派和配器等属性（Dhariwal 等，2020；Agostinelli 等，2023；Copet 等，2023）。在可控音乐生成中，风格是用户控制的最重要属性之一。音乐风格迁移（Dai 等，2018）保留歌词或旋律等内容，同时改变乐曲的整体特征，包括调式、节奏、配器和音色

以及演唱方式。流派是审视音乐风格的主要视角之一。然而，这种控制仍主要依赖文本提示，而我们对风格作为一种全局表征在模型内部如何运作知之甚少。

机制可解释性研究为探究这类问题提供了工具：通过分析模型的内部表征，我们可以找到与特定概念对应的方向，并沿着这些方向直接引导生成（Bereska & Gavves, 2024; Turner et al., 2023; Zou et al., 2023; Rimsky等，2024）。在音乐模型中，已有研究表明，风格同样可以提取为激活空间中的一个方向，并用于引导，例如作为作曲家向量（Jiang et al., 2026），或作为通过稀疏自编码器发现的音乐概念（Singh et al., 2026）。
尽管这些方法在实践中颇为有效，但它们仅作用于残差流中的方向。它们并未解释这种风格方向是如何由特定内部组件形成的。电路方法为回答这些问题提供了一种途径：它们将行为特定计算分解为一个包含重要组件的内部稀疏子网络，即注意力头和多层感知机（Elhage

[1]代码：https://github.com/Aristo233/Music_电路_代码。

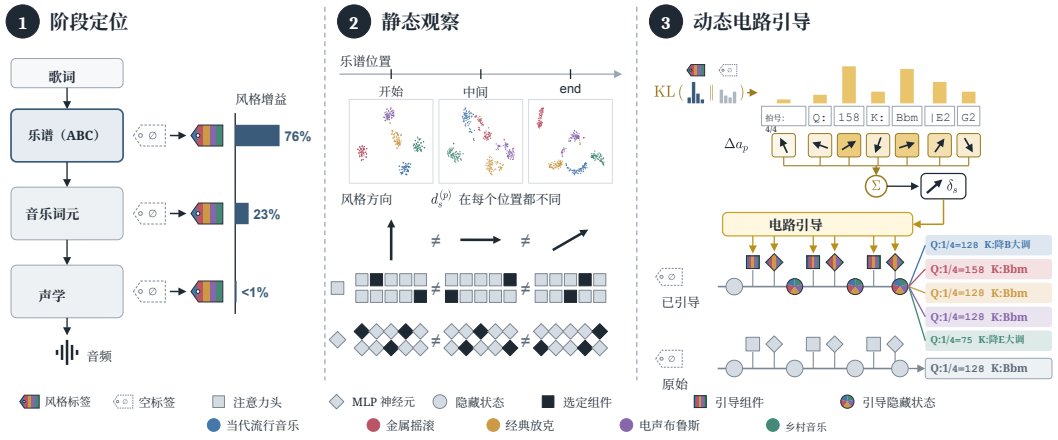


图1：概览。**(1)阶段定位**：每次只在一个生成阶段中将空标签替换为风格标签，结果表明风格增益大多来自乐谱阶段。**(2) 静态观察**：在每个乐谱位置上，风格都可以区分，但风格方向 $d_s^{(p)}$（省略层索引）以及对其贡献最大的注意力头和多层感知机神经元会随位置而变化。**(3)动态电路引导**：用两个标签的下一词元分布之间逐步计算的 KL 散度对其激活差异 Δa_p 加权，并将结果求和，得到每种风格对应的一个方向 δ_s，以引导电路。在不提供任何风格文本的情况下，该电路能以五种风格写出相同的歌词，效果与风格提示相当，且往往更胜一筹；右侧的 ABC记谱标题是针对一条测试歌词生成的真实结果。

等，2021；Wang 等人，2023；Conmy 等人，2023）。此类方法已被用于解释和控制类似于风格的全局属性。关于情绪的研究在模型内部发现了可读的情绪表征，并表明对这些表征进行干预会改变模型的行为（Wang 等人，2025a；Anthropic，2026）。通过干预显著驱动情绪表征的组件，Wang 等人（2025a）在没有任何情绪提示词的情况下，使模型针对同一场景生成带有明显不同情绪的回复。这可以被视为自然语言中的一种风格迁移形式。然而，回路分析在音乐风格研究中的应用仍然非常有限。

因此，我们探究风格在大型音乐生成模型中是如何表征的，并将这一问题分解为三个部分。(1) 作为一种动态的全局概念，风格以何种形式存在于模型的表征中？(2) 哪些内部机制和组件会显著驱动这种表征？(3) 能否利用这些组件有效地控制风格？

为回答这些问题，我们提出了一种基于 YuE2（Yuan等（2025））的风格电路发现与控制方法。YuE2 是一款最先进的开源音乐生成模型。它首先将旋律规划为乐谱，然后生成音乐词元并将其渲染为音频。我们首先构建了一个数据集，其中相同的歌词以五种风格生成，分别为当代流行音乐、金属摇滚、经典放克、电声布鲁斯和乡村音乐（§4.1）。在此基础上，我们设计了一个三步框架（图1；§3）。**(1)阶段定位**。在三个生成阶段分别消融风格标签后发现，大部分风格效应是在乐谱规划阶段形成的。**(2) 静态观察**。我们提取风格方向，并在乐谱和音乐词元的多个位置选取组件，检查它们在生成过程中是否保持稳定。我们发现，每个位置的风格都明显可分，但风格方向及其对应的重要组件会随位置不断变化，彼此几乎没有重叠。因此，我们推测，全局风格并非由固定位置上的静态向量决定，而是由跨词元的动态表征决定。**(3) 动态电路引导**。据此，我们根据风格标签与空标签条件下输出分布之间的 KL 散度为每个步骤加权，并将各位置的激活差异汇总为一个方向，以捕捉动态生成过程中保持不变的风格概念。随后，我们追踪驱动该方向的神经元和注意力头，将它们组合成一个稀疏电路，并通过消融、增强以及引导生成至目标风格来测试其作用。在完全不提供风格文本的情况下，这种电路干预的效果与风格提示相当，且往往更胜一筹：根据 CLAP 的评估，在 15 条测试歌词对应的 150 次生成中，有 133 次命中目标风格，而使用提示词时为 113 次；具有相同组件数量和注入范数的随机电路则停留在空标签基线水平（34；§4.5）。

我们的贡献主要有三方面。**(1) 对音乐风格表征的深入理解**。我们以流派作为理解音乐风格的切入点，这类似于将情绪作为自然语言中的一个概念。我们发现，音乐生成模型中的风格以动态方式存储：它可以在每个位置读出，但承载它的方向和组件会随位置变化。这与音乐在多个层面组织信息的方式相符，例如调式、节奏和配器。**(2)电路层级控制**。我们提出了 KL 加权动态风格方向及其稀疏电路，两者结合后，无需任何风格文本即可在电路层级控制音乐生成。**(3) 机制证据**。我们迈出了音乐风格组件级机制可解释性的第一步。

2 相关工作

2.1 音乐生成模型

符号模型生成音符事件或文本乐谱，例如 ABC 记谱法（Huang 等，2019；Yu 等，2022；Lu 等，2023；Yuan 等，2024；Qu 等，2024；Wang 等，2025b）；音频模型则生成声学词元或潜变量，并使用文本控制流派和配器等属性（Dhariwal 等，2020；Huang 等，2023；Agostinelli 等，2023；Copet 等，2023；Liu 等，2024；Evans 等，2024；Yuan 等，2025）。在这两个层面上，风格都曾通过专用模型进行迁移或控制（Brunner 等，2018；Cífka 等，2020；Wu 等，2024），而在音乐信息检索中，流派是风格的标准标签（Tzanetakis & Cook，2002）。既有可解释性工作探究了音乐模型的内部表征（Castellon 等，2021；Ma 等，2024），通过干预内部激活进行引导（Ma 等，2024；Facchiano 等，2025），并使用稀疏自编码器发现可解释的音乐概念（Singh 等，2026），但这些方法无法追踪写入这些信息的特定组件。

2.2 电路发现与机制可解释性

机制可解释性旨在逆向工程模型计算，将其解析为人类可理解的机制（Bereska & Gavves，2024）。电路发现是一种核心方法，它将 Transformer 分解为从残差流中读取信息并向其中写入信息的注意力头和多层感知机（Elhage 等，2021；Geva 等，2021；Olsson 等，2022），并通过因果干预（Vig 等，2020；Geiger 等，2021；Meng 等，2022；Wang 等，2023；Hanna 等，2023）和自动化方法（Conmy 等，2023；Syed 等，2024；Ferrando & Voita，2024）定位实现特定行为的电路。另一类研究则发现线性概念方向（Nanda 等，2023；Park 等，2024）或稀疏特征（Bricken 等，2023；Cunningham 等，2024），并通过将它们添加到残差流中来引导生成（Turner 等，2023；Zou 等，2023；Rimsky 等，2024）。情绪电路（Wang 等人，2025a）结合了这两种方法：它们提取情绪方向，随后定位并干预沿这些方向写入信息的组件。音乐生成模型中的电路发现仍鲜有探索。

2.3 YuE2

YuE2（Yuan 等人，2025年）[2] 是一种条件音乐生成模型，将符号乐谱规划与音频生成相结合。歌词提供演唱内容，风格文本描述流派、配器和人声特征，可选乐谱则指定旋律、节奏和和弦。YuE2采用混合变换器架构，包含自回归（AR）分支和非自回归（NAR）分支。启用乐谱规划时，AR分支首先生成一份ABC乐谱，然后根据文本和乐谱预测音乐词元。NAR分支利用这些词元和条件上下文，通过流匹配生成声学潜变量。随后，变分自编码器（VAE）解码器将这些潜变量转换为音频。YuE2还支持使用用户提供的乐谱，以及不进行乐谱规划的生成方式。其风格标签会影响全部三个阶段，我们分别称之为乐谱规划（乐谱）阶段、音乐令牌阶段和声学渲染阶段。这与实践中对音乐风格的广义理解一致：音乐风格涵盖旋律动机、和声色彩、配器、演奏技法和人声表现。

[2]YuE2 架构和推理详情：https://huggingface.co/m-a-p/YuE2-3B。

3 方法

我们的方法分为三个步骤（图1）：阶段定位、静态观察和动态电路引导。首先，分阶段标签消融（§3.2）表明，约四分之三的风格是在乐谱阶段（§4.2）形成的，因此后续分析和干预聚焦于这一阶段。其次，单位置分析（§3.3）表明，每个位置都能读出风格，但在不同乐谱位置上，沿这些方向推动的方向和组件各不相同（§4.3）；而且，风格标签只会在少数位置显著改变输出（附录 C.3），没有任何单一位置能够代表整个生成过程。第三，我们汇总不同位置上的激活差异，得到一个动态方向，并根据各步骤的 KL 散度为其加权，据此构建稀疏电路（§3.4）。这些组件具有因果作用（§4.4），即使不提供任何风格文本，其效果也能媲美或超过风格提示（§4.5）。

3.1 问题设定与符号

我们研究 YuE2 针对同一组歌词以五种风格生成的歌曲（§4.1）。方向、组件和干预设置仅在表征分析集 $\mathcal{I}_{\text{ext}}$ （提取方向所用的数据集）上估计；测试集仅用于评估。

YuE2 在乐谱阶段、音乐词元阶段和声学渲染阶段生成歌曲（§2.3），且每个阶段都可以读取风格标签。我们将风格 s 的标签前缀记为 c_s，将空标签前缀记为 $c_\varnothing$。为了提取表征，我们将歌曲的前缀（风格标签和歌词），连同模型已生成的乐谱和音乐词元，一并输入自回归模型进行一次不采样的前向传播，并读取每一层、每个位置的内部状态。比较两个标签时，两次前向传播共享相同的乐谱历史 x，只有前缀不同，因此任何差异都仅来自风格标签。在层 ℓ 和位置 p 处，$h_p^{(\ell)}$ 表示残差流，$a_p^{(\ell)}$ 表示组件激活：对于 MLP 神经元，是进入下投影 $W_{(\ell)\,\text{down}}$ 的门控激活；对于注意力头，是进入输出投影 $W_O^{(\ell)}$ 的头输出。

3.2 阶段定位：逐阶段标签消融

将三个阶段各自的风格标签独立设为 c_s 或 $c_\varnothing$；切换后续阶段的标签时，沿用较早阶段的输出，因此可以分离各阶段对最终音频的影响。我们使用 CLAP（Wu 等人，2023年）差值来衡量音频风格（§4.1）。由于某一阶段标签带来的增益取决于其他阶段是否也带有该标签（乐谱一旦体现了这种风格，音乐令牌标签的作用就会减弱），因此我们根据各阶段的 Shapley 值，将完整标签相对于空标签的差值增益分摊到各阶段（Shapley，1953）；这些值之和恰好等于总增益（附录 A.4）。

3.3 静态观察：单位置风格方向与电路

风格方向。为了将风格与歌曲的其他内容分离，我们取同一组歌词的五个风格版本，并从每个样本中减去其均值隐藏状态，从而去除这些版本共有的内容（主要是歌词），保留因风格而异的内容。对歌词求和并归一化后，得到层 ℓ、位置 p处的风格方向 s：

$$d_s^{(\ell,p)} \propto \sum_{i\in\mathcal{I}_{\text{ext}}} \Big(h_{i,s,p}^{(\ell)} - \frac{1}{5}\sum_{s'=1}^{5} h_{i,s',p}^{(\ell)}\Big), \qquad \big\|d_s^{(\ell,p)}\big\|_2 = 1. \tag{1}$$

组件归因。每个 MLP 神经元通过下投影矩阵的一列将信息写入残差流，而每个注意力头则通过输出投影中对应的列块写入其输出。一个组件对风格方向的贡献取决于其激活值和写入方向：

$$c_{s,j}^{(\ell,p)} = g_{s,j}^{(\ell,p)}\,\big\langle W_{\text{down},:,j}^{(\ell)},\, d_s^{(\ell,p)}\big\rangle, \qquad c_{s,h}^{(\ell,p)} = \big\langle W_{O,h}^{(\ell)}\, z_{s,h}^{(\ell,p)},\, d_s^{(\ell,p)}\big\rangle, \tag{2}$$

其中，$g_s^{(\ell,p)}{}_{,\,j}$ 是神经元 j 在风格为 s的歌曲上的门控激活值均值（未中心化），$z_s^{(\ell,p)}{}_{,\,h}$ 是注意力头 h的输出均值，而 $W_O^{(\ell)}{}_{,\,h}$ 是该注意力头对应的输出投影列块。$d_s^{(\ell,p)}$在子层输出处拟合，即组件写入的位置：注意力头对应注意力之后，神经元对应多层感知机之后。在每一层中，我们按贡献从高到低选取组件。

因果验证。我们通过消融和增强来测试选定组件 J。以相同方式对组件激活进行中心化，并在表征分析集上取均值，可得组件空间中的风格差异 $\bar{\delta}_s^{(\ell)}$。在风格标签下，消融会将选定组件置零；在空标签下，增强会将这一差异加到这些组件上：

$$\text{ablation: } a_J^{(\ell)} \leftarrow 0, \qquad \text{enhancement: } a_J^{(\ell)} \leftarrow a_J^{(\ell)} + \bar{\delta}_{s,J}^{(\ell)}. \tag{3}$$

我们将该影响度量为跨层残差投影在风格方向上的变化总和：读取每个子层之后的残差，并将其投影到根据该子层输出拟合的方向上，

$$\Delta S = \sum_{\ell=1}^{28} \big\langle \widetilde{h}^{(\ell)} - h^{(\ell)},\, d_s^{(\ell)} \big\rangle, \tag{4}$$

其中，$\widetilde{h}$ 是干预后的残差（省略位置索引），并与每层数量相同的随机组件进行比较；对于增强实验，还使用相同的注入范数。

跨位置比较。我们在多个乐谱位置分别提取方向并选择组件，然后通过方向之间的余弦相似度和选定组件的杰卡德重叠度来比较各位置。如果风格是静态向量，那么这些结果应在不同位置之间保持一致。基于一个位置构建的电路也可作为生成实验中的基线。

3.4 使用 KL 权重进行动态电路引导

位置权重。由于风格表征会随位置变化，我们需要确定每个位置在整首乐曲风格差异中所占的比例。风格标签并不会在每个位置以相同程度改变输出，因此我们通过逐步 KL 散度来衡量其在每个位置的影响

$$\mathrm{KL}_p = \mathrm{KL}\big(p(\cdot \mid c_s, x_{<p}) \,\|\, p(\cdot \mid c_\varnothing, x_{<p})\big). \tag{5}$$

根据链式法则，两个条件在整个生成过程上的分布之间的 KL 散度，等于逐步条件 KL 散度之和的期望：

$$\mathrm{KL}\big(p(x_{1:T} \mid c_s) \,\|\, p(x_{1:T} \mid c_\varnothing)\big) = \mathbb{E}_{x \sim p(\cdot \mid c_s)} \sum_{t=1}^{T} \mathrm{KL}_t. \tag{6}$$

因此，KL 散度给出了整体风格差异在每个步骤上的占比：KL 散度指出风格作用于何处，激活值差异则指出作用方向。

KL 加权方向。在每个歌词组内，我们计算风格标签与空标签之间各组成部分激活值差异的平均值，$\Delta a_{i\,s\,,,p}^{(\ell)} = a_{i\,s\,,,p}^{(\ell)} - a_{i\,\varnothing\,,,p}^{(\ell)}$，再用 KL 权重加权，然后对歌词取平均：

$$\delta_s^{(\ell)} = \frac{1}{|\mathcal{I}_{\text{ext}}|} \sum_{i \in \mathcal{I}_{\text{ext}}} \frac{\sum_p \mathrm{KL}_{i,s,p}\, \Delta a_{i,s,p}^{(\ell)}}{\sum_p \mathrm{KL}_{i,s,p}}. \tag{7}$$

对每个子层之后的残差流应用相同的加权平均并进行归一化，得到残差空间中的单位方向 $u_s^{(\ell)}$。因此，每种风格在整个乐谱阶段都对应一个固定方向：位置特定差异的 KL 加权聚合，即各个重要位置所共有的风格表征部分（附录 B.3）。

电路构建。单位置分数衡量某个组件对风格表征的贡献程度。对于引导，我们关心的是注入量将残差推向何处，因此使用 KL 加权差值 δ_s 为每个组件评分，因为注入的正是这个差值：神经元的分数是其差值分量乘以其写入方向在 $u_s^{(\ell)}$ 上的投影；注意力头的分数则是其差值经输出投影后在 $u_s^{(\ell)}$ 上的投影。两者都取自组件写入的子层输出处（附录 B.3）。正分数意味着注入该组件的差值会在其所在层沿风格方向推动残差，我们只保留此类组件。

因果验证。 我们沿用单位置消融和增强，但现在两种干预都作用于乐谱的每个位置：在风格标签下，消融会将选定组件置零；在空标签下，增强会向这些组件添加 δ_s。残差在 u_s 上的投影变化，按照每句歌词各自的逐步 KL 权重对各个位置取平均（附录 B.3）。这些权重仅用于评估；生成时始终为每种风格注入相同的固定方向。随机对照仅从目标未选中的组件中抽取（附录 B.2）。

电路组装。 每种风格的电路在每个阶段包含 560 个组件（每个子层十个）、392 个 MLP 神经元和 168 个注意力头。在乐谱阶段，我们将所有层放在一起排序，因为 KL 加权分数已经反映了每层的写入量；在音乐词元阶段，预算则根据子层重要性分配到各层（Wang 等人，2025a）（附录 B.1）。每种风格只选择一次组件，并由所有歌词共享。

Ste导向。 生成过程中，在乐谱阶段的每一步，我们都会将 KL 加权差值 ，scal乘以强度 α 后，加到选定组件 J 上，并保持其他所有内容不变：

$$a_J \leftarrow a_J + \alpha\, \delta_{s,J}. \tag{8}$$

由于风格方向会随位置变化，且风格决策分布于整份乐谱之中，因此我们在整个阶段内注入相同的方向，而不仅仅是在阶段开始时注入。音乐令牌阶段同样采用这种方式，将前 128 个步骤中以等权重平均得到的差异注入；所有电路条件和 CAA 均共享这一音乐令牌干预。我们不对声学渲染进行干预，且没有任何阶段接收风格文本。

4 实验

4.1 设置

数据集。 我们将使用 GPT- 6 Astra（OpenAI, 2026年）创作的简短歌词，与五种在节奏、配器和人声方面各具特色的风格提示词配对：当代流行音乐、金属摇滚、经典放克、电布鲁斯和乡村音乐（下文简称流行音乐、金属乐、放克、布鲁斯和乡村音乐）。YuE2 生成 ABC乐谱和音频，我们仅保留五个版本均通过 CLAP 和 Gemini 3.6 Flash（Google DeepMind，2026年）检查的歌词组。69 个歌词组被划分为表征分析集 $\mathcal{I}_{\text{ext}}$（54 个歌词组）和测试集（15 个歌词组；附录 A）。

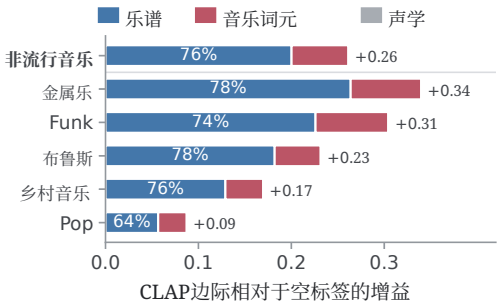


图 2：分阶段风格标签消融（54 条表征分析歌词）：完整标签 CLAP 效果的 Shapley 分解；百分比给出乐谱阶段所占的份额。

评估。 两名自动评审评估风格。CLAP 根据五种风格描述对音频进行评分；差值 是目标风格的分数减去其他风格中最高分，而一次完整的生-成且差值为正的结果即为命中。Gemini 3.6 Flash 在五种风格和一个干扰项中盲选一种风格，选中目标风格即为命中。截断生成结果视为未命中。主观评估作为补充。

基线。 我们将其与空标签基线进行比较；风格提示词，即在三个阶段均使用风格文本；随机电路，其在配方、时序和逐层注入范数方面与动态电路相匹配，但使用与目标不相交的随机组件；以及 CAA（Rimsky等，2024），即一种传统的引导方法，在每个乐谱步骤的一个层中加入风格标签与空标签之间的均值残差差异。CAA 和这些电路共享动态电路的音乐令牌干预，仅在乐谱阶段有所不同。强度 $\alpha = 0.5$ 以及 CAA 的层和强度，是根据表征分析集的调参歌词（附录B和D）确定的。

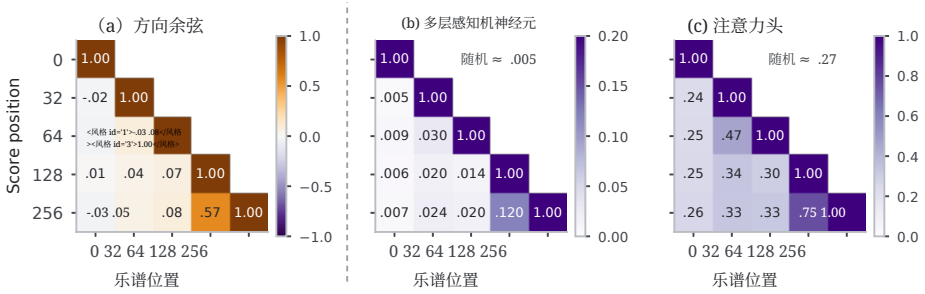


图 3：第 28 层不同乐谱位置上的情况（五种风格的均值）：(a) 方向之间的余弦相似度；(b、c) 所选多层感知机神经元和注意力头的杰卡德重叠度（随机水平分别为 0.005 和 0.27）。

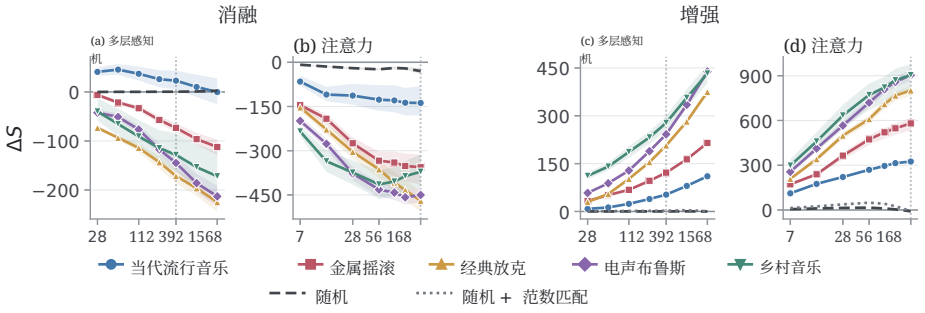


图 4：消融和增强多层感知机神经元及注意力头的效果与编辑组件数量的关系；竖线：生成预算；色带：95% 音程。灰色虚线：每层相同数量的随机组件；灰色点线：使用目标注入范数时的相同结果。

4.2 风格主要在乐谱规划期间形成 ING

分阶段消融显示，乐谱阶段约占完整风格标签效应的四分之三，音乐令牌阶段占四分之一，而声学渲染几乎没有贡献（图2）。另外两项观察结果与此一致：风格标签会改变调性、速度和拍号等全局乐谱设置；仅将空标签下的乐谱替换为风格标签下创作的乐谱，就能使音频明显趋向目标风格（附录 A）。因此，我们在乐谱阶段提取方向并进行干预，同时将音乐令牌阶段作为补充。

4.3 风格在不同乐谱位置上动态编码

经过歌词内中心化后，五种风格在乐谱的开始和结束处，以及音乐词元的起始位置彼此分离；在乐谱中段，隐藏状态几乎无法区分风格，但注意力头输出和多层感知机激活值仍然可以区分（附录 C.1）。在每个位置上，方向都具有可重复性（附录 C.3），而在音乐令牌位置上，其组件具有因果作用（附录 C.2）。然而，不同乐谱步之间几乎没有共享这些方向（图 3，生成令牌数为 0 至 256）：乐谱开头的方向与之后位置的方向几乎正交，其多层感知机神经元与其他位置的神经元几乎没有重叠，而注意力头的重叠度仅处于随机水平；只有较晚的位置彼此明显相似。因此，风格并非固定在某一处的静态向量，而是一种随着乐谱写下而变化的动态表征（音乐令牌方向在第一个位置之后更加一致，见附录 C.3）。

4.4 稀疏电路因果性地控制风格

以 KL 加权方向为参考，仅将 0.23% 的多层感知机神经元置零，就会降低风格投影，并使输出偏离风格标签输出的程度达到移除标签时的约 80%；而注入两者之差，则会在没有标签的情况下推动表征趋向目标风格（图 4）。注意力头的表现类似：注入少数注意力头即可恢复大部分输出差异，注入更多则恢复得更少，这也解释了生成强度为何较小。随机组件即使经过范数匹配，

表1：测试集上两位评判器对各风格的命中数（满分 30 分，其中包含 15 段歌词和两组种子对）；截断生成结果计为未命中。粗体表示该列最佳结果。

	CLAP						Gemini 3.6 Flash					
条件	流行音乐	金属乐	放克	布鲁斯	乡村音乐	总计	流行音乐	金属乐	放克	布鲁斯	乡村音乐	总计
空标签基线	27	0	0	0	3	30	14	1	0	0	15	30
风格提示	29	17	13	**29**	25	113	24	12	15	**30**	**30**	111
随机电路	27	0	0	3	4	34	13	0	0	0	20	33
CAA	26	22	**26**	21	23	118	24	22	**25**	23	26	120
动态电路（我们的）	**30**	**24**	25	27	**27**	**133**	**25**	**23**	22	27	29	**126**

bare也几乎没有影响，只有将注意力头置零会不加区分地损害输出。流行音乐，即默认风格 ，has 一种较弱的方向，其投影在消融后上升（附录 B.2）。

4.5 电路引导无需任何风格文本即可写出目标风格

所有条件都从歌词开始，不借助任何风格文本写出各自的乐谱和音乐；15 段测试歌词和五种风格均使用两组种子对生成，每种条件生成 150 次。

表1给出了两种评判器对每种风格的命中数。两种评判器的结论一致：不使用任何风格文本时，动态电路的命中次数高于风格提示词。差距体现在金属乐和放克上：提示词的命中率仅约为生成次数的一半，未命中的结果大多被听作流行音乐或布鲁斯；而电路的命中率约为 80%。在布鲁斯和乡村音乐上，提示词的命中率接近上限，而电路在布鲁斯上的表现略逊一筹。随机电路的表现与空标签基线相当，因此效果来自选定组件，而非注入本身；单个位置的电路或等权平均方向构成的电路命中率要低得多（附录 D.1），因此 KL 加权聚合至关重要。

与 CAA 相比，电路的优势主要在于稳定性。在每个乐谱步骤都添加相同的固定向量，往往会使生成偏离预期：CAA 在 150 次生成中有 23 次被截断，而电路只有 4 次，运行时间也达到两倍以上（177 秒，而电路为 84 秒）；在完整生成中，CAA 的命中率不低于电路，但若将截断视为未命中，其总计得分在 CLAP 下显著较低，在 Gemini 下也略低。电路的运行时间也长于提示词（66 秒），但随机电路并非如此（58 秒），因此原因并非注入规模。

电路在音乐层面改变了什么。我们考察电路引导了哪些音乐决策：乐谱中设定的调性，以及音乐词元中设定的配器（图5）。数据集中每种风格都有其典型调性（金属乐为 D♯小调，放克和布鲁斯常用 G 小调、D 小调等小调，乡村音乐常用 G 大调、E♭ 大调等大调，流行音乐常用 F 大调和 D 大调），而空标签的调性不随风格变化（附录 D.4）。提示词使几乎所有乐谱都采用典型调性（金属乐中 93% 为 D♯ 小调，布鲁斯中全部为 G 小调）。电路也会朝相同方向调整调性，但幅度较小：一半金属乐谱转为 C♯小调，乡村音乐乐谱则大多转为 E♭大调；而十分之四的放克和布鲁斯乐谱仍保留空标签的调性，且电路从未写出布鲁斯提示词中的 6/8 拍；随机电路则不会改变调性。

使用 CLAP 乐器标签测量的配器，是电路表现得最像提示词的方面：金属乐和布鲁斯采用失真吉他、重型鼓组和击勾贝斯，放克采用不含原声吉他和钢琴的击勾贝斯，乡村音乐采用踏板钢棒吉他和班卓琴。在 13 种以上的乐器中，各种风格下电路与提示词的变化相关性为 0.92–0.98，而随机电路近乎为零的变化相关性仅为 0.46–0.76。因此，电路通过乐谱和配器共同实现了金属乐和乡村音乐，而主要通过配器实现了放克和布鲁斯。

主观评估。听众对使用相同歌词和随机种子生成的三段匿名化片段（基线、提示词和电路）进行风格契合度评分，范围为1至5分，并选出最接近目标风格的一段（图6；附录D.3）。初步结果（11名听众）与自动评审的结论一致，即电路能够生成目标风格：电路和提示词的评分都显著高于基线；在二者之间，听众目前在非流行风格中更偏好提示词，在流行风格中更偏好电路，但两者之间的差异均不显著，而且这两种条件的评分彼此之间都比与基线的评分更为接近。

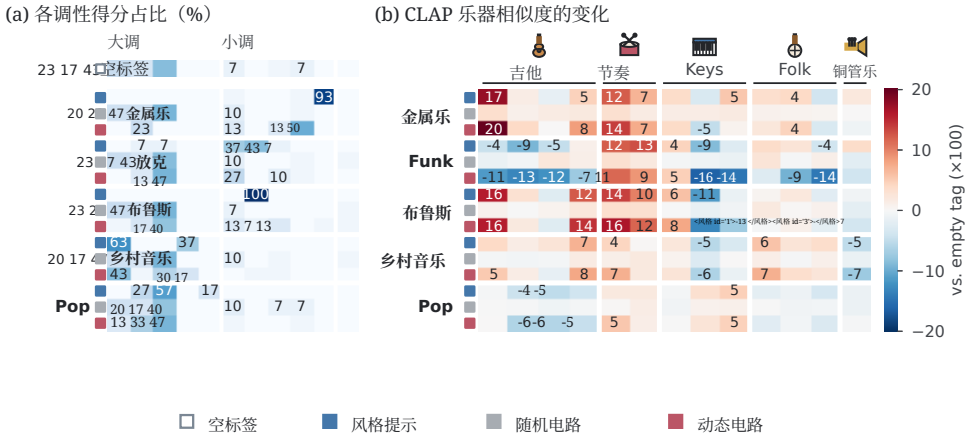


图5：测试集上的调性与配器。(a) 每种调性下每行 30 份乐谱（15 首歌词、两个随机种子）所占的百分比；空标签只有一行。(b) 每种乐器描述对应的 CLAP相似度减去相同歌词和随机种子的空标签音频的 CLAP相似度，取平均后乘以 100。

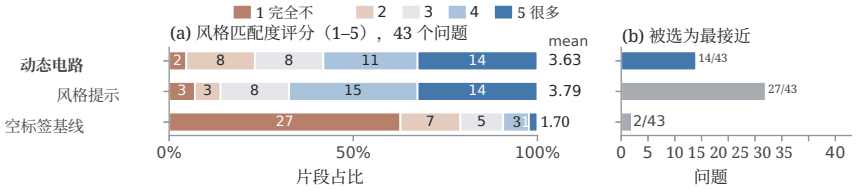


图6：主观评估（初步评估，43 个非流行音乐问题）；柱形表示片段数量。

5 结论

我们在电路层级研究了大型音乐生成模型中的音乐风格。通过YuE2，我们发现风格主要形成于乐谱规划阶段，并以动态方式分布在乐谱的各个词元中：在每个位置都可以读出风格，但承载风格的是会随位置变化的方向和组件。在此发现的基础上，我们根据每一步与基线之间的KL散度进行加权，从而为每种风格获得一个动态方向，并追踪驱动该方向的稀疏电路。这些少量神经元和注意力头在消融和增强实验中都表现出因果作用；直接对其进行引导，在控制风格方面达到了或超过了风格提示的效果，而且无需任何风格文本。该分析已经迁移应用于人声性别属性，并采用了针对属性的聚合方法（附录 E）；未来我们将把它扩展到配器和情绪等更多属性以及其他模型，并追踪信息如何在电路组件之间流动。

参考文献

Andrea Agostinelli、Timo I. Denk、Zalán Borsos、Jesse Engel、Mauro Verzetti、Antoine Caillon、Qingqing Huang、Aren Jansen、Adam Roberts、Marco Tagliasacchi、Matt Sharifi、Neil Zeghidour 和 Christian Frank。《MusicLM：从文本生成音乐》，2023年。网址：https://arxiv. org/abs/2301.11325。Anthropic。《大语言模型中的情绪概念及其作用》。Transformer Cir- cuits Thread，2026年。网址：https://transformer-circuits.pub/2026/emotions/ index.html。Leonard F. Bereska 和 Efstratios Gavves。《人工智能安全的机制可解释性 —— 综述》。《机器学习研究汇刊》，2024年。Trenton Bricken、Adly Templeton、Joshua Batson 等。《迈向单义性：通过字典学习分解语言模型》。*Transformer Circuits Thread*，2023年。网址：https://transformer-circuits.pub/2023/monosemantic-features/index.html。

Gino Brunner、Andres Konrad、Yuyi Wang 和 Roger Wattenhofer。《MIDI-VAE：音乐动态与配器建模及其在风格迁移中的应用》。载于第*19*届国际音乐信息检索学会会议论文集，2018年。Rodrigo Castellon、Chris Donahue 和 Percy Liang。《编码音频语言建模学得了适用于音乐信息检索的表征》。载于第*22*届国际音乐信息检索学会会议论文集，2021年。Ondˇrej Cífka、Umut ¸Sim¸sekli 和 Gaël Richard。《Groove2Groove：利用合成数据监督进行单样本音乐风格迁移》。*IEEE/ACM* 音频、语音与语言处理汇刊，28：2638–2650，2020年。Arthur Conmy、Augustine N. Mavor-Parker、Aengus Lynch、Stefan Heimersheim 和 Adrià Garriga-Alonso。《迈向机制可解释性的自动化电路发现》。载于神经信息处理系统进展，2023年。Jade Copet、Felix Kreuk、Itai Gat、Tal Remez、David Kant、Gabriel Synnaeve、Yossi Adi 和 Alexandre Défossez。《简单且可控的音乐生成》。载于神经信息处理系统进展，第36卷，第47704–47720页，2023年。Hoagy Cunningham、Aidan Ewart、Logan Riggs、Robert Huben 和 Lee Sharkey。《稀疏自编码器在语言模型中发现高度可解释的特征》。载于国际学习表征会议，2024年。Shuqi Dai、Zheng Zhang 和 Gus G. Xia。《音乐风格迁移：一篇立场论文》。载于国际音乐元创作研讨会（*MUME*），2018年。

Prafulla Dhariwal、Heewoo Jun、Christine Payne、Jong Wook Kim、Alec Radford 和 Ilya Sutskever。Jukebox：一种音乐生成模型，2020 年。网址：https://arxiv.org/abs/2005. 00341。

Nelson Elhage、Neel Nanda、Catherine Olsson 等。《Transformer电路的数学框架》。《Transformer Circuits Thread》，2021年。网址：https://transformer-circuits.pub/2021/framework/index.html。Zach Evans、Julian D. Parker、CJ Carr、Zack Zukowski、Josiah Taylor 和 Jordi Pons。《Stable Audio Open》，2024年。网址：https://arxiv.org/abs/2407.14358。Simone Facchiano、Giorgio Strano、Donato Crisostomi、Irene Tallini、Tommaso Mencattini、Fabio Galasso 和 Emanuele Rodolà。《用于可解释引导音乐生成的激活修补》，2025年。网址：https://arxiv.org/abs/2504.04479。Javier Ferrando 和 Elena Voita。《信息流路径：大规模自动解释语言模型》。载于《*2024*年自然语言处理实证方法会议论文集》，第17432–17445页，2024年。Atticus Geiger、Hanson Lu、Thomas Icard 和 Christopher Potts。《神经网络的因果抽象》。载于《神经信息处理系统进展》，2021年。Mor Geva、Roei Schuster、Jonathan Berant 和 Omer Levy。《Transformer前馈层是键值记忆》。载于《*2021*年自然语言处理实证方法会议论文集》，2021年。Google DeepMind。《Gemini 3.6 Flash：模型卡》，2026年7月。网址：https://deepmind.google/models/model-cards/gemini-3-6-flash/。访问日期：2026年9月21日。Michael Hanna、Ollie Liu 和 Alexandre Variengien。《GPT-2 如何计算大于关系？：解读预训练语言模型的数学能力》。载于《神经信息处理系统进展》，2023年。Cheng-Zhi Anna Huang、Ashish Vaswani、Jakob Uszkoreit、Ian Simon、Curtis Hawthorne、Noam Shazeer、Andrew M. Dai、Matthew D. Hoffman、Monica Dinculescu 和 Douglas Eck。《Music Transformer：生成具有长期结构的音乐》。载于《国际学习表征会议》，2019年。

Qingqing Huang、Daniel S. Park、Tao Wang、Timo I. Denk、Andy Ly、Nanxin Chen、Zhengdong Zhang、Zhishuai Zhang、Jiahui Yu、Christian Frank、Jesse Engel、Quoc V. Le、William Chan、Zhifeng Chen 和 Wei Han。《Noise2Music：基于扩散模型的文本条件音乐生成》，2023年。网址：https://arxiv.org/abs/2302.03917。

Xunyi Jiang、Mingyang Yao、Jingyue Huang 和 Julian McAuley。Composer vector：在潜变量空间中进行符号音乐风格引导生成，2026年。链接：https://arxiv.org/abs/2604. 03333。

Haohe Liu、Yi Yuan、Xubo Liu、Xinhao Mei、Qiuqiang Kong、Qiao Tian、Yuping Wang、Wenwu Wang、Yuxuan Wang 和 Mark D. Plumbley。AudioLDM 2：通过自监督预训练学习整体音频生成。*IEEE/ACM* 音频、语音与语言处理汇刊，32：2871–2883，2024年。

Peiling Lu、Xin Xu、Chenfei Kang、Botao Yu、Chengyi Xing、Xu Tan 和 Jiang Bian。MuseC-oco：从文本生成符号音乐，2023年。链接：https://arxiv.org/abs/2306. 00110。

Wenye Ma、Xinyue Li 和 Gus Xia。音乐大语言模型是否学会了符号概念？一项使用探测和干预的试点研究。载于*Audio Imagination*：*NeurIPS 2024 Workshop* 人工智能驱动的语音、音乐与声音生成，2024年。

Kevin Meng、David Bau、Alex J. Andonian 和 Yonatan Belinkov。在 GPT 中定位和编辑事实关联。载于神经信息处理系统进展，2022年。

Neel Nanda、Andrew Lee 和 Martin Wattenberg。自监督序列模型世界模型中涌现的线性表征。载于第六届 *BlackboxNLP* 研讨会论文集：自然语言处理神经网络的分析与解读，2023年。

CatherineOlsson、Nelson Elhage、Neel Nanda 等。上下文学习与归纳注意力头。*Transformer* 电路专题，2022年。URL https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html。

OpenAI。GPT-6 Astra：新一代智能。https://openai.com/index/ gpt-6-astra/，2026年。访问于2026年9月。

Kiho Park、Yo Joong Choe 和 Victor Veitch。线性表征假说与大语言模型的几何。在 国际机器学习大会，2024年。Xingwei Qu、Yuelin Bai、Yinghao Ma、Ziya Zhou 等。MuPT：生成式符号音乐预训练变换器，2024年。URL：https://arxiv.org/abs/2404.06393。

NinaRimsky、Nick Gabrieli、Julian Schulz、Meg Tong、Evan Hubinger 和 Alexander Turner。通过对比激活添加引导 Llama 2。在 计算语言学协会第*62*届年会论文集，2024年。

Lloyd S. Shapley。n人博弈的一个值。载于 Harold W. Kuhn 和 Albert W. Tucker（编），博弈论的贡献 *II*，第 307–317 页。普林斯顿大学出版社，1953年。Nikhil Singh、Manuel Cherep 和 Pattie Maes。大型生成音乐模型中可解释概念的发现与引导。载于国际学习表征会议，2026年。Aaquib Syed、Can Rager 和 Arthur Conmy。归因修补优于自动化电路发现。载于第七届 *BlackboxNLP*研讨会论文集：分析与解释自然语言处理神经网络，第 407–416 页，2024年。Alexander Matt Turner、Lisa Thiergart、Gavin Leech、David Udell、Juan J. Vazquez、Ulisse Mini 和 Monte MacDiarmid。运用激活工程引导语言模型，2023年。网址：https://arxiv.org/abs/2308.10248。George Tzanetakis 和 Perry Cook。音频信号的音乐流派分类。*IEEE*语音与音频处理汇刊，10(5):293–302，2002年。

Laurens van der Maaten 和 Geoffrey Hinton。使用 t-SNE 可视化数据。《机器学习研究期刊》，9:2579–2605，2008年。

Jesse Vig、Sebastian Gehrmann、Yonatan Belinkov、Sharon Qian、Daniel Nevo、Yaron Singer 和 Stuart Shieber。使用因果中介分析研究语言模型中的性别偏见。载于神经信息处理系统进展，2020年。

Chenxi Wang、Yixuan Zhang、Ruiji Yu、Yufei Zheng、Lang Gao、Zirui Song、Zixiang Xu、Gus Xia、Huishuai Zhang、Dongyan Zhao 和 XiuyingChen。大语言模型是否“有感受”？情绪电路的发现与控制，2025a。网址：https://arxiv.org/abs/2510.11328。

Kevin Wang、Alexandre Variengien、Arthur Conmy、Buck Shlegeris 和 Jacob Steinhardt。野外的可解释性：GPT-2 small 中用于间接宾语识别的电路。载于国际学习表征会议，2023年。

Yashan Wang、Shangda Wu、Jianhuai Hu、Xingjian Du、Yueqi Peng、Yongxin Huang、Shuai Fan、Xiaobing Li、Feng Yu 和 Maosong Sun。NotaGen：通过大语言模型训练范式提升符号音乐生成的音乐表现力。载于第三十四届国际人工智能联合会议论文集，第10207–10215页，2025b。

Shangda Wu、Zhancheng Guo、Ruibin Yuan、Junyan Jiang、Seungheon Doh、Gus Xia、Juhan Nam、Xiaobing Li、Feng Yu 和 Maosong Sun。CLaMP 3：面向未对齐模态和未见语言的通用音乐信息检索。载于计算语言学协会 *ACL 2025* 研究成果，第 2605–2625 页，奥地利维也纳，2025年。计算语言学协会。doi：10.18653/v1/2025.findings-acl.133。网址：https://aclanthology.org/ 2025.findings-acl.133/。

Shih-Lun Wu、Chris Donahue、Shinji Watanabe 和 Nicholas J.Bryan。Music ControlNet：用于音乐生成的多种时变控制。*IEEE/ACM* 音频、语音与语言处理汇刊，32：2692–2703，2024年。

Yusong Wu、KeChen、Tianyu Zhang、YuchenHui、Taylor Berg-Kirkpatrick 和 Shlomo Dubnov。通过特征融合和关键词到字幕增强进行大规模对比式语言音频预训练。载于*IEEE* 国际声学、语音与信号处理会议（*ICASSP*），2023年。

Botao Yu、Peiling Lu、Rui Wang、Wei Hu、Xu Tan、Wei Ye、Shikun Zhang、Tao Qin 和 Tie-Yan Liu。Museformer：采用细粒度与粗粒度注意力的 Transformer，用于音乐生成。载于神经信息处理系统进展，第 35 卷，第 1376–1388 页，2022年。

Ruibin Yuan、Hanfeng Lin、Yi Wang、Zeyue Tian、Shangda Wu、Tianhao Shen、Ge Zhang、Yuhang Wu、Cong Liu、Ziya Zhou 等。ChatMusician：利用大语言模型从内在层面理解和生成音乐。载于 计算语言学协会 *ACL 2024* 论文集，第 6252– 6271 页，2024年。

Ruibin Yuan、Hanfeng Lin、Shuyue Guo、Ge Zhang、Jiahao Pan、Yongyi Zang、Haohe Liu、Yiming Liang、Wenye Ma、Xingjian Du、Xinrun Du、Zhen Ye、Tianyu Zheng、Zhengxuan Jiang、Yinghao Ma、Minghao Liu、Zeyue Tian、Ziya Zhou、Liumeng Xue、Xingwei Qu、Yizhi Li、Shangda Wu、Tianhao Shen、Ziyang Ma、Jun Zhan、Chunhui Wang、Yatian Wang、Xiaowei Chi、Xinyue Zhang、Zhenzhu Yang、Xiangzhou Wang、Shansong Liu、Lingrui Mei、Peng Li、Junjie Wang、Jianwei Yu、Guojian Pang、Xu Li、Zihao Wang、Xiaohuan Zhou、Lijun Yu、Emmanouil Benetos、Yong Chen、Chenghua Lin、Xie Chen、Gus Xia、Zhaoxiang Zhang、Chao Zhang、Wenhu Chen、Xinyu Zhou、Xipeng Qiu、Roger Dannenberg、Jiaheng Liu、Jian Yang、Wenhao Huang、Wei Xue、Xu Tan 和 Yike Guo。《YuE：扩展开放基础模型以实现长篇音乐生成》。*arXiv* 预印本 *arXiv:2503.08638*，2025年。网址：https://arxiv.org/abs/2503.08638。

Andy Zou、Long Phan、Sarah Chen、JamesCampbell、Phillip Guo 等。《表征工程：一种自上而下的人工智能透明度方法》，2023年。网址：https://arxiv.org/abs/2310. 01405。

数据构建与生成协议

我们构建了一个包含不同风格匹配歌词的音乐数据集，用于研究 YuE2 中的风格表征。我们首先使用 GPT-6 Astra 生成 32 段基础歌词文本，然后为每个歌词类别创作四种叙事变体，得到 128 段由四行组成的英语歌词文本。每段文本都配有当代流行音乐、金属摇滚、经典放克、电声布鲁斯和乡村音乐的提示词。启用乐谱规划后，YuE2 会先根据歌词和风格文本生成 ABC乐谱，再结合相同的歌词和风格文本以及生成的乐谱，生成音乐词元和音频。每组样本的歌词完全相同，而乐谱则针对每种风格独立生成。生成过程会持续到音频自然结束，并保留完整的 48千赫立体声输出。
接下来，我们验证生成的样本是否符合其目标风格。对于 CLAP，我们汇总覆盖整首音轨的各个 10 秒窗口的得分，并按覆盖范围加权。目标风格必须在候选描述中排名第一；候选描述包括五种风格，以及一个干扰项——中国传统风格，且其相似度间隔必须至少比得分最高的其他选项高 0.03。对于初步筛选后缺失的歌词与风格组合，我们会生成额外的 ABC乐谱和完整音频轨道，并保留第一个通过 CLAP阈值的候选样本。由此得到 640 个候选样本。这一筛选标准比用于评估引导的命中标准（§4.1）更严格；后者仅比较五种风格描述，并要求差值为正，但不设阈值。

接下来，我们使用 Gemini 3.6 Flash（Google DeepMind，2026年）对 640 条完整音频轨道逐一分类，从五种风格和同一个干扰项中进行选择。我们使用匿名文件名，并将风格选项的顺序随机打乱。共有 569 个样本通过 Gemini 验证和原始 CLAP阈值筛选。为便于使用相同歌词比较不同风格，我们仅保留五种风格版本均通过筛选的完整歌词组。最终数据集包含 69 个歌词组，每种风格各有 69 首歌曲，总计 345 首歌曲。

A.1 数据构建示例

图7展示了一个当代流行音乐样本，该样本根据歌词和风格文本生成，启用了乐谱规划，且未输入乐谱。

A.2 不同风格的乐谱头部分布

在写下第一个音符之前，ABC乐谱会在乐谱头部确定其拍号、速度和调性。表2比较了不同风格下的这些设置，并以相同歌词在空标签下创作的乐谱作为参考。

表2：345 个保留样本以及相同 69 首歌词对应的空标签乐谱的 ABC 头部字段。小调表示处于小调的乐谱所占比例；拍号列统计各拍号下的乐谱数量；速度以每分钟四分音符数表示。

组别	n	小调	4/4	2/4	6/8	拍	速度、中位数（四分位距）	最常见调性
当代流行音乐	69	9%	56	13	0	100	（90–118）	F 大调（14）、D 大调（13）、A（9）
金属摇滚	69	100%	40	29	0	90	（87–118）	升D小调（51）、Em（6）、D 小调（4）
经典放克	69	96%	66	3	0	105	（100–118）	G 小调（21）、D 小调（20）、Em（8）
电声布鲁斯	69	100%	34	0	35	86	（85–88）	G 小调（59）、D 小调（8）、升D小调（2）
乡村音乐	69	1%	57	12	0	80	（78–85）	G 大调（20）、降B大调（15）、降E大调（14）
空标签	69	54%	56	11	0	93	（85–120）	G 小调（11）、D 小调（9）、D 大调（8）

金属摇滚、电声布鲁斯和经典放克几乎总是采用小调，当代流行音乐和乡村音乐几乎总是采用大调，而只有电声布鲁斯使用 6/8 拍。仅凭这三个设置，就能识别出 77.1% 的风格（随机水平为 20%）。

A.3 仅交换乐谱

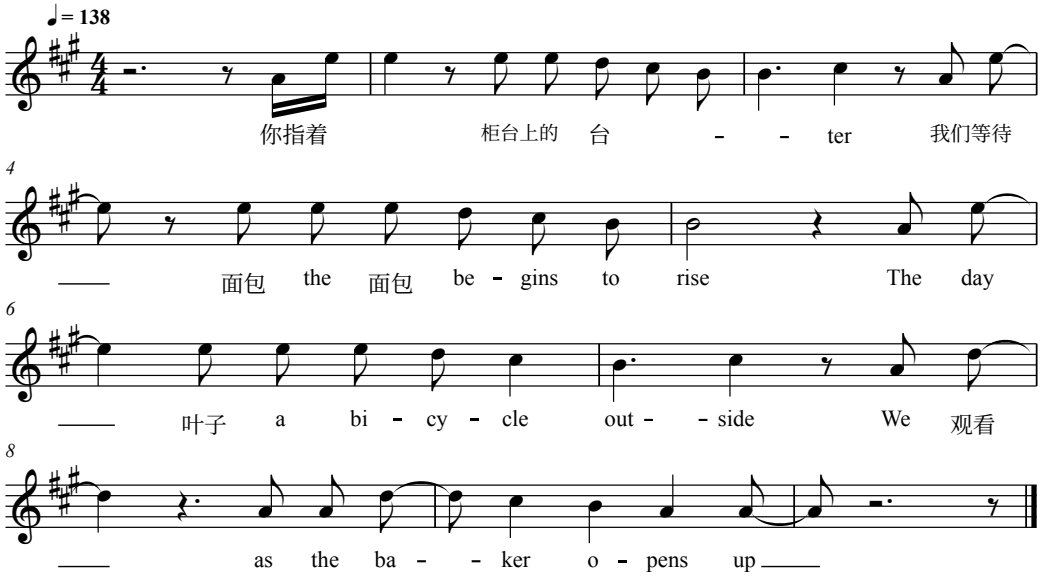
对于54个表征分析歌词中的每一个以及每种风格，我们生成两个样本，两者唯一的区别在于乐谱是否写在风格标签下；所有后续阶段都使用空标签（表3）。

YuE2 输入 当代流行音乐

先生成不含和弦符号、仅包含旋律的 ABC记谱转录，再根据给定条件生成带有编解码器标记的音乐。

输入歌词是下面展示的四行诗节。

生成的人声旋律 夜间面包店



音频：30.12 秒 CLAP边际：0.259 Gemini：当代流行音乐

图7：一个数据构建示例，展示了生成的 ABC 中第 1–10 小节的人声部分。歌词衬底仅供示意，并非经过测量的音频对齐。CLAP 和 Gemini 3.6 Flash 用于验证风格；这些歌词的五个版本均通过筛选并予以保留。

表3：仅替换乐谱。两组使用相同的所有随机种子；音乐令牌和声学阶段始终接收空标签，且不进行干预。命中表示目标风格的 CLAP 分数最高；差值为目标分数减去其他分数中的最高值。

风格	CLAP 命中数		差值变化
	空标签分数	风格标签分数	
当代流行音乐	48/54	53/54	+0.050
金属摇滚	0/54	33/54	+0.305
经典放克	0/54	9/54	+0.228
电声布鲁斯	2/54	34/54	+0.179
乡村音乐	4/54	39/54	+0.141
非流行风格	6/216	115/216	+0.213

仅替换乐谱，就能将非流行音乐的命中数从 6 提高到 216 首中的 115 首，约为通过提示词引导两个自回归阶段所获得差值增益的 81%（附录 A.4）。符号编码器 CLaMP 3（Wu 等人，2025年）在评判乐谱本身时总体上也得出一致结论，但对各个风格的判断并不一致，因此我们仅将其作为佐证。

A.4 分阶段标签消融

三个阶段中的每个阶段都可以保留风格标签或使用空标签，从而形成八种条件。某个阶段的主要效应，是仅切换该阶段时，在其他两个阶段的所有设置下所产生的平均影响。沙普利值同样是该阶段增益在其他两个阶段设置下的加权平均，但权重不同：其他两个阶段均不携带或均携带

标签的情况各获得 1/3 的权重，而恰好其中一个阶段携带标签的两种情况各获得 1/6 的权重。这些权重使三个贡献之和恰好等于总增益。

表 4：分阶段标签消融中各阶段的影响（54 首表征分析歌词，四种非流行风格，一个种子对）。

阶段	主要效应（差值）	沙普利份额
乐谱	+0.200	76%
音乐词元	+0.061	23%
声学渲染	+0.001	0%
完整标签，所有阶段	+0.262	100%

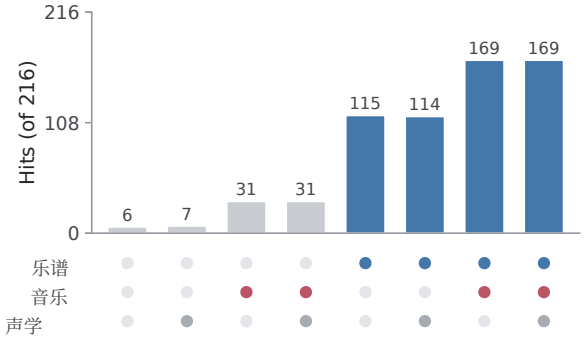


图8：分阶段消融的八种条件下，四种非流行风格的命中数（满分为 216）；下方的圆点标示保留风格标签的阶段。

两个自回归阶段略有重叠：乐谱使用风格标签后，在音乐令牌阶段添加该标签所带来的增益约少三分之一，而声学渲染与两者都没有交互。空提示词下，当代流行音乐已经有命中（54 首中命中 48 首），因此未纳入。

B 实现细节

B.1 子层重要性

局部写入贡献较大的层，对最终表征的影响不一定也大。因此，我们每次只沿风格方向对一个子层添加扰动，重新计算其余层，并将最终表征沿风格方向的变化量除以扰动幅度，作为该子层的重要性。在乐谱阶段，我们使用 KL 加权方向：在每个乐谱位置都添加扰动，并使用 KL 权重对各位置的读出值取平均。我们还记录这种扰动会消除多少由风格标签引起的输出 KL。

两种读出值都在较低层和中间层最大，大致位于第 3 至第 12 层，并随深度增加而下降；在第 20 层之后，扰动单个子层消除的输出 KL 不到十分之一。相比之下，通过对所有层的写入分数进行排序而选出的 MLP 神经元大多位于中间层和较高层，其中约四分之一位于最后一层。这两个量衡量的是不同方面：前者表示沿某个方向推动整个子层所产生的下游影响，后者表示单个组件写入的量。

预算分配。电路预算为子层数量的十倍，即每个阶段 560 个组件，其中多层感知机神经元与注意力头的比例为 7:3。音乐令牌阶段分两步将预算分配给各层。首先，为每个子层分配最低配额，使较低层的隐藏状态在整个生成过程中始终对注意力可见，因此仍能编码风格线索。其次，按照子层重要性比例分配剩余预算，优先考虑对最终风格表征影响更大的子层（Wang 等人，2025a）。在每个子层内，我们选择得分最高的组件。在乐谱阶段，KL 加权分数根据实际注入量计算，且已反映各层的写入量，因此我们取消逐层配额，将所有层放在一起排名。按照子层重要性分配乐谱阶段的预算或许能进一步改善电路，这留待未来工作探讨。

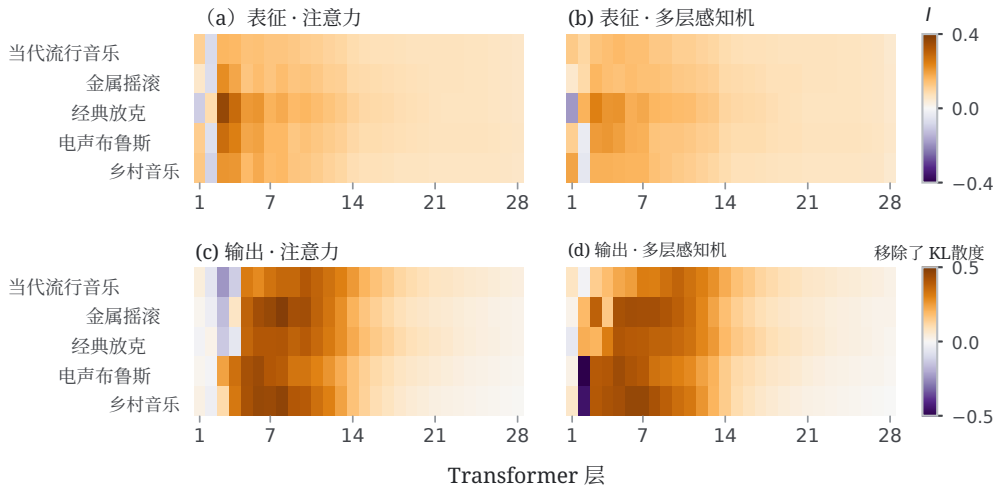


图9：乐谱阶段的子层重要性（54 条表征分析歌词）。上：最终表征的变化。下：移除的输出 KL 散度比例。左：注意力子层；右：多层感知机子层。颜色刻度以上限为最大的正值；少数更多的负值单元格会达到饱和。

音乐令牌阶段根据音乐开始时的子层重要性分配逐层配额（图 10，底行）：每层至少分配 4 个神经元和 2 个注意力头，剩余预算按照重要性的正值部分按比例分配，并采用最大余数舍入；任何一层获得的额外配额不得超过该组件类型预算的 40%。

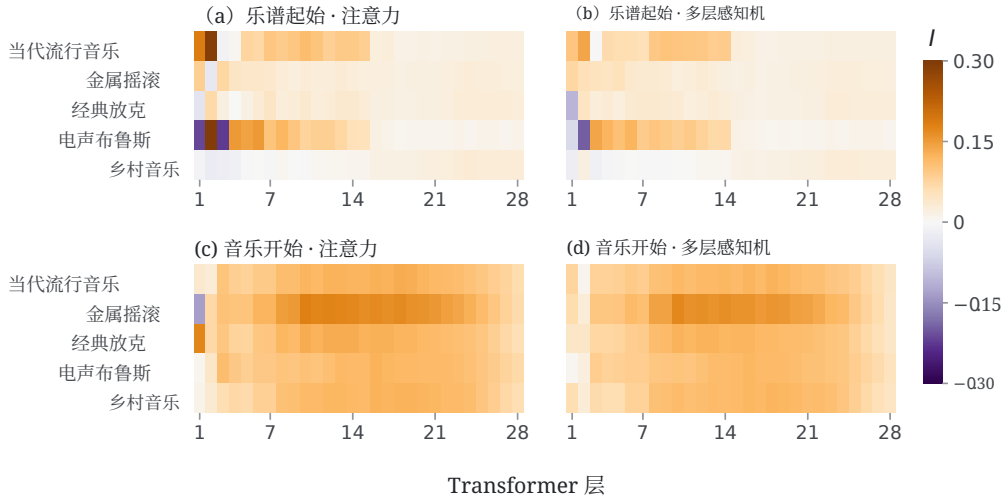


图10：单位置方向的子层重要性。上：乐谱起始标记（乐谱开始）；下：音乐开始。左：注意力子层；右：多层感知机子层。正值促进目标风格，负值则与之相悖。

B.2 随机组件控制

随机对照在每层抽取与目标相同数量的组件，并在三个随机种子上取平均（附录 C.2 中的单位置实验使用十个随机种子）。每层只有 16 个注意力头，而目标在大多数层中会选择其中约三分之一，因此允许重叠的随机抽样会与目标共享许多注意力头。因此，我们仅从目标未选择的组件中抽取；如果某层剩余组件过少，不足的数量将转移到

同样包含目标组件的最近层。对于增强，第二个随机对照会将注入差异重新缩放至目标的逐层范数，从而排除注入大小本身作为解释。

图11展示了表示分析集上的风格投影（上）以及测试集上的输出分布变化（下）。对于消融，下排给出干预后的输出与风格标签输出之间的 KL 散度，除以空标签输出与风格标签输出之间的 KL 散度；对于增强，则给出 1 减去相同比值，即恢复比例。

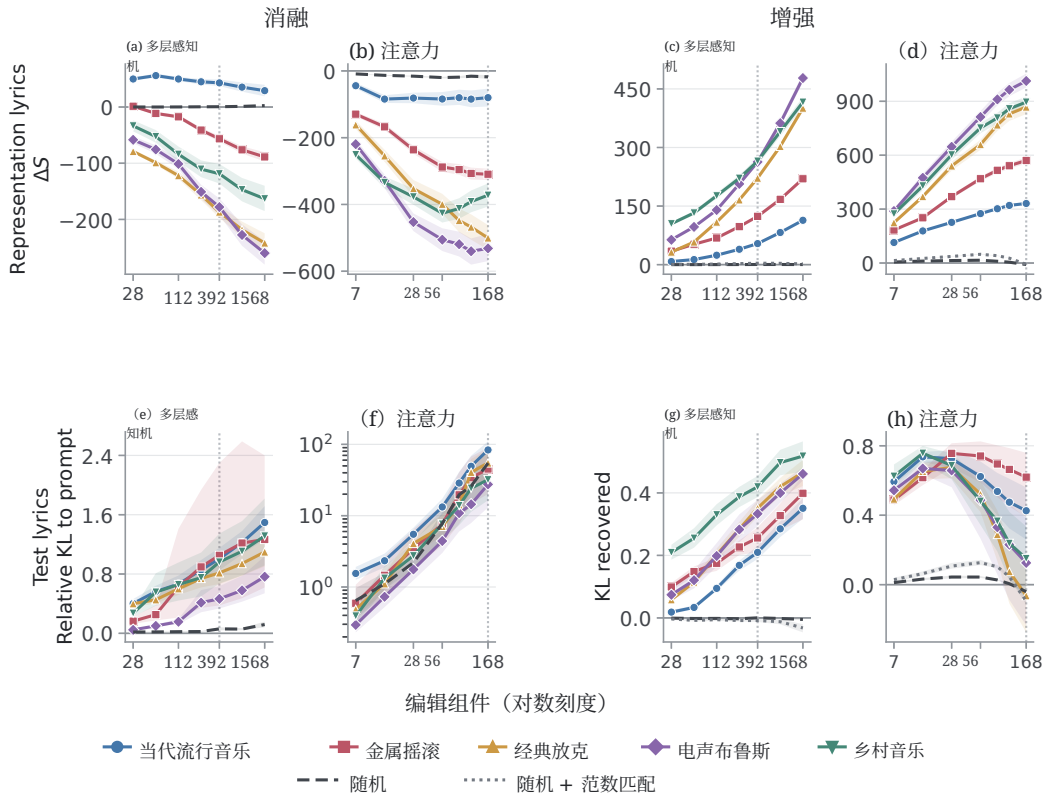


图11：以 KL 加权方向为参考的组件干预。上排（a–d）：54 条表示分析歌词上的风格投影。下排（e–h）：测试集上的输出分布变化。各列与图 4 相同：多层感知机消融、注意力头消融、多层感知机增强和注意力头增强；灰色曲线表示随机对照。

多层感知机结果在两个数据集上相符，而随机神经元几乎没有影响。如果对各位置等权平均，而不是按 KL 权重加权，目标组件仍明显优于随机组件，但多层感知机消融的效果会小得多，这表明将这些神经元置零主要在 KL 值较高的位置发挥作用。在风格投影方面，注意力头也远强于随机组件（图 b 和 d）。输出层面消融是个例外：将同样数量的随机注意力头置零也会显著改变输出（图 f，对数刻度），因此，置零注意力头造成的输出影响并非特异性的。对于增强，恢复比例在 14 至 28 个注意力头时最高，之后下降，而随机注意力头在所有预算下都接近于零（图 h）。

B.3 KL 加权方向和组件选择

KL 权重和方向仅使用表征分析集。对于每个歌词组，一首在空标签下写成的乐谱充当共享历史；我们读取该乐谱在空标签以及五个风格标签中的每个乐谱位置处的下一个词元分布和组件激活。在每个歌词组内，KL 散度值被归一化为权重，将激活值差异乘以这些权重后求和，最后对歌词取平均。同样的

残差流的加权平均得到 $\rho_s^{(\ell)}$和 $u_s^{(\ell)} = \rho_s^{(\ell)}/\|\rho_s^{(\ell)}\|_2$：对于神经元，在多层感知机之后提取；对于注意力头，在注意力之后提取。组件乐谱为

$$c_{s,j}^{(\ell)} = \delta_{s,j}^{(\ell)} \left\langle W_{\text{down},:,j}^{(\ell)},\, u_s^{(\ell)} \right\rangle, \qquad c_{s,h}^{(\ell)} = \left\langle W_{O,h}^{(\ell)}\, \delta_{s,h}^{(\ell)},\, u_s^{(\ell)} \right\rangle, \tag{9}$$

其中， $\delta_s^{(\ell)}{}_{,j}$ 和 $\delta_{s\,h}^{(\ell)}{}_{,}$是神经元 j 和注意力头 h上 KL 加权差值的分量，而$W_O^{(\ell)}{}_{,h}$ 是该注意力头输出投影的列块。多层感知机神经元和注意力头分别排名，并保留得分最高的 392 个神经元和 168 个注意力头。

在因果验证中，消融将 $a_J \leftarrow 0$ 设为$a_J \leftarrow a_J + \delta_{s,J}$，增强也将其设为$a_J \leftarrow a_J + \delta_{s,J}$，且两者都作用于每个乐谱位置。读出结果为

$$\Delta S_{i,s} = \sum_p w_{i,s,p} \sum_{\ell=1}^{28} \left\langle \widetilde{h}_{i,p}^{(\ell)} - h_{i,p}^{(\ell)},\, u_s^{(\ell)} \right\rangle, \qquad w_{i,s,p} = \frac{\mathrm{KL}_{i,s,p}}{\sum_{p'} \mathrm{KL}_{i,s,p'}}, \tag{10}$$

其中， $\widetilde{h}$ 是干预后的残差，在每个子层之后读取，并投影到该子层输出的 $u_s^{(\ell)}$上。在输出端，我们还记录干预后的输出与风格标签输出之间的 KL散度，并以空标签输出与风格标签输出之间的 KL散度为参照。

B.4 注入强度

KL 加权方向使用原始差值，因此必须限制注入强度（表 5）。

表 5：四首表征分析歌词和四种非流行风格下的注入强度。在$\alpha = 2.0$ 中，所有乐谱标题栏均已损坏，因此其命中数无法作为风格控制来解读。

强度 α	CLAP 命中数	完整度	ABC 词元均值	音频均值（秒）
0.5	14/16	16/16	697	83
1.0	13/16	13/16	1378	143
2.0	11/16	13/16	915	175

从 $\alpha = 1.0$开始，乐谱会变得长得多，有些还会被截断，因此我们使用 $\alpha = 0.5$。每个阶段和风格的预算为 392 个多层感知机神经元（占所有神经元的 0.23%）和 168 个注意力头。

音乐令牌阶段。 KL 加权的位置依赖聚合仅应用于乐谱阶段，因为 §4.2 的阶段定位表明，乐谱阶段承载了大部分风格效应，因此值得进行详细建模；音乐令牌阶段贡献约四分之一，我们以较简单的方式进行引导：对前 128 个位置取等权均值，强度设为 1.0（经典放克和电声布鲁斯为 1.5）。所有电路条件和 CAA 共用这一音乐令牌干预，因此表1中的条件仅在乐谱阶段有所不同。在整个阶段注入，而非仅在最初几个步骤注入，遵循与乐谱阶段相同的理由：风格决策分布在整个序列中。

C 扩展分析

C.1 单位置风格方向

本节使用全部 345 个样本和 t-SNE（van der Maaten和Hinton，2008），展示第 3.3 节中单位置方向的歌词内中心化效果、组件激活的风格分离，以及更多层上的表征。每个自回归阶段包含三个位置：第一个位置、中间 64 个词元的均值，以及结束标记。

中心化后，五种风格在乐谱的开始和结束处，以及音乐词元的开始处都清晰分离；测试样本也落入各自风格对应的簇中。而在乐谱中间，即使经过中心化，各种风格仍难以区分。

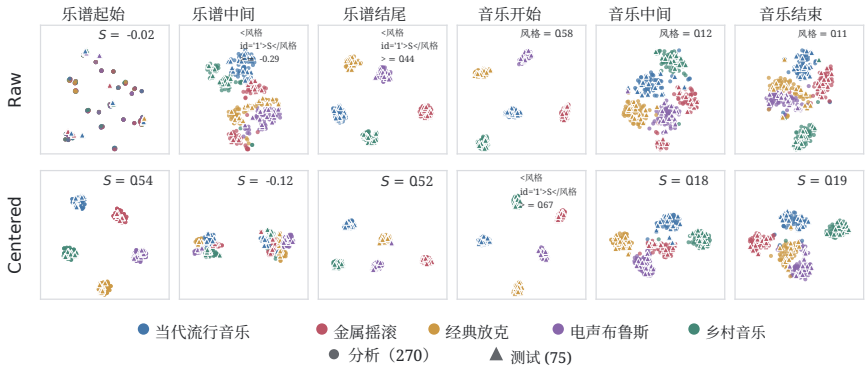


图12：第14层六个位置的隐藏状态（t-SNE）。每个阶段分别取其第一个位置、中间64个词元的均值及其结束标记。上图：原始隐藏状态；下图：经歌词内中心化后的风格表征。圆形表示表征分析样本，三角形表示测试样本；S 为原始空间中的余弦轮廓系数。

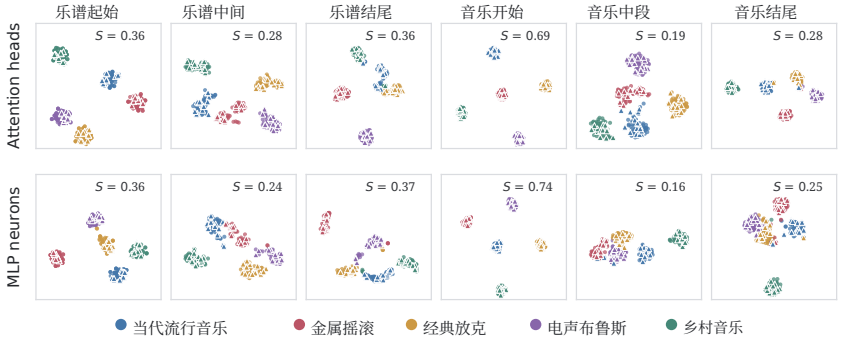


图13：第14层六个位置上五种风格的注意力头输出（上）和 MLP 神经元激活（下）（t-SNE，经歌词内中心化）。标记和 S 与图12相同。

电路所在的两种组件激活也能区分这些风格。在乐谱中间，隐藏状态几乎无法区分风格（$S = -0.12$），而注意力头输出和多层感知机神经元激活仍能做到（$S = 0.28$ 和 0.24）。

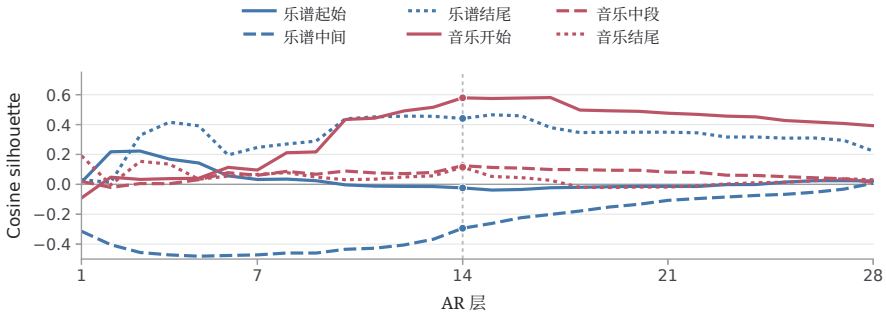


图14：六个位置上不同层的原始隐藏状态的风格区分度（余弦轮廓系数）。蓝色：乐谱阶段；红色：音乐令牌阶段。实线、虚线和点线分别表示开始、中间和结束。

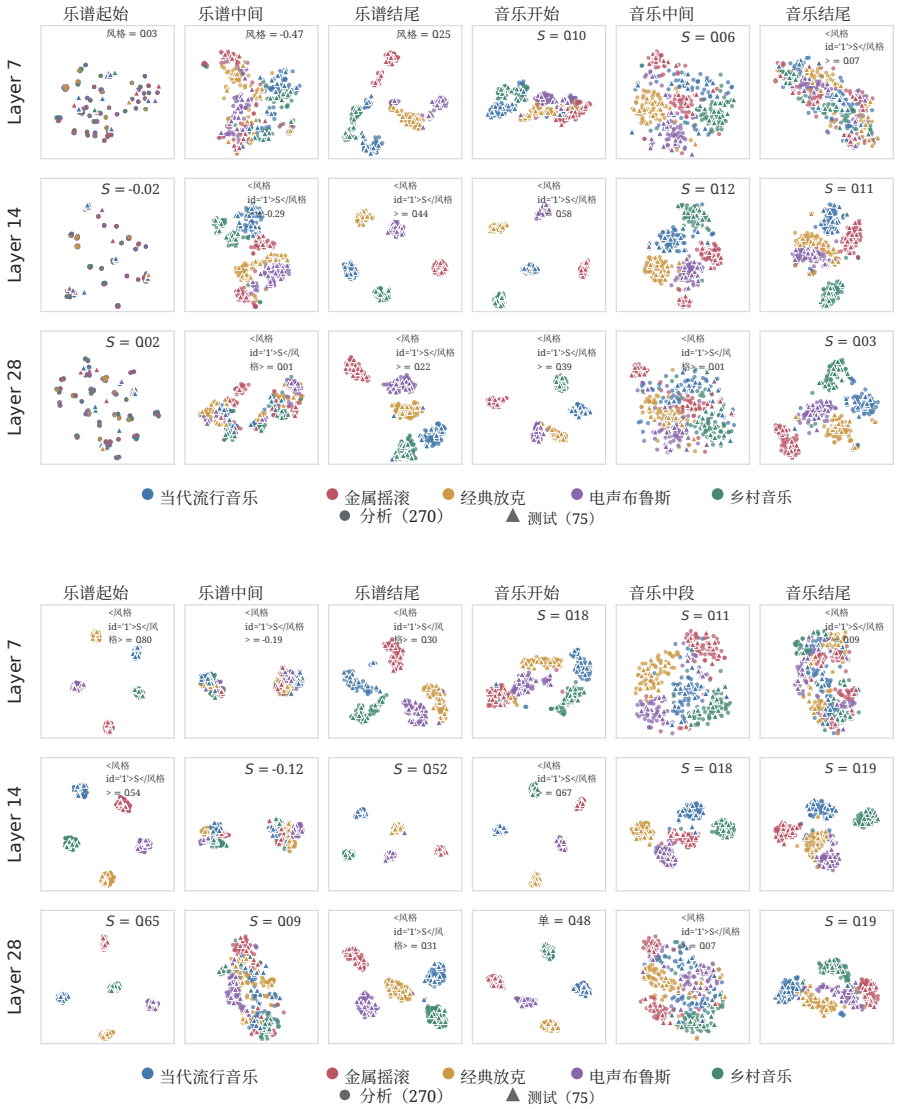


图15：第7层、第14层和第28层中六个位置的隐藏状态。上：原始隐藏状态，以及 S 原始空间中的余弦轮廓系数。下：经过歌词内中心化后的相同结果。

C.2 单位置电路与因果验证

本节报告了§3.3中定义的音乐令牌阶段三个位置上的单位置电路；相关定义见§3.3。图16至图16使用音乐词元开始处（音乐开始）的位置方向，并在另外两个位置重复实验：音乐中段，即音乐序列的中间64个词元；以及音乐结尾，即音乐结束标记之后的位置。横轴表示每层编辑组件的数量。

在这三个位置，移除少量高贡献组件会明显削弱风格投影，而注入差异则会在没有风格提示的情况下推动表征趋向目标风格；这两种效果都明显强于数量和注入范数相同的随机组件。由此可见，单个位置的风格集中在少量组件中。音乐中段与音乐开始一致，而音乐结尾由少数组件主导，并且在不同风格之间变化更大。

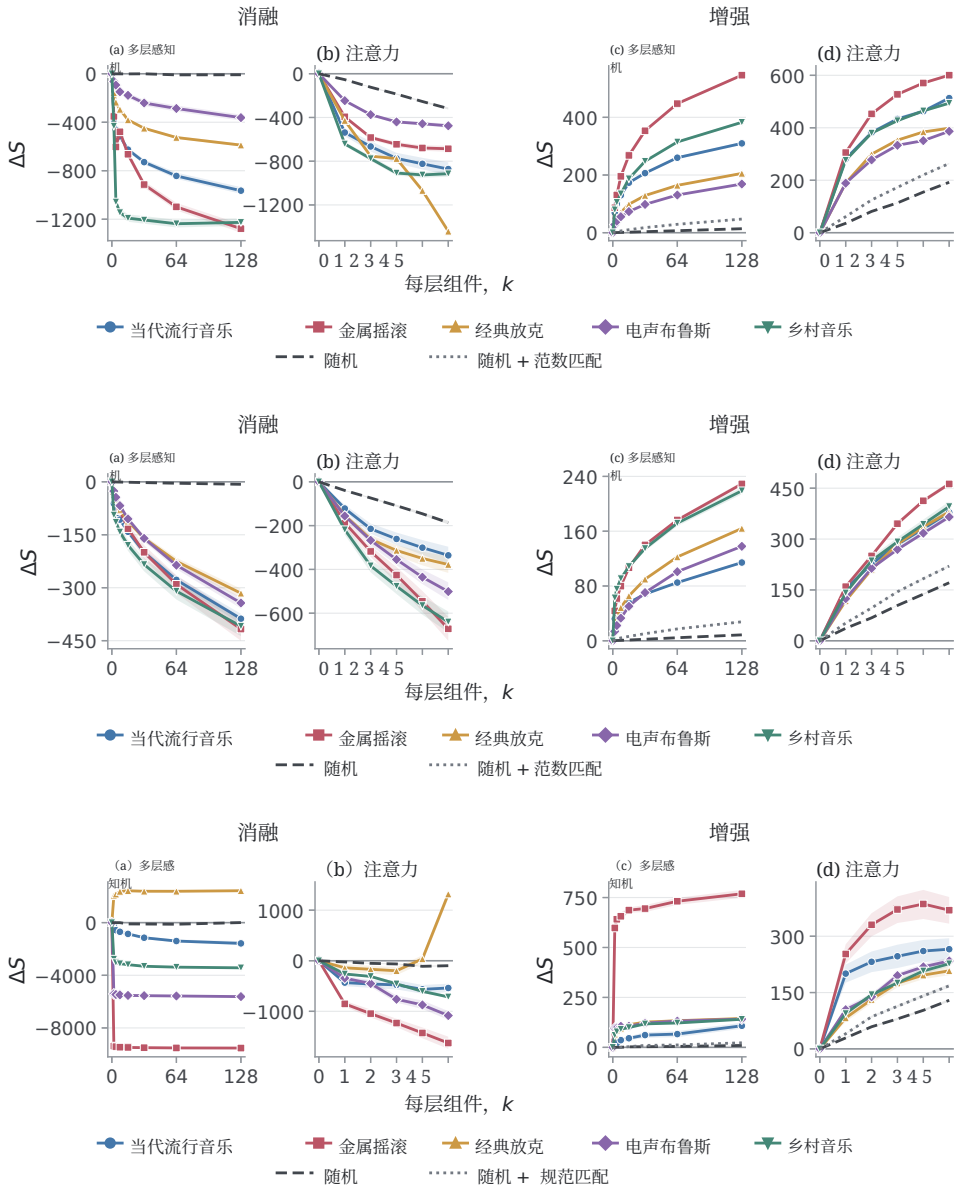


图16：在音乐开始（上）、音乐中段（中）和音乐结尾（下）进行的组件干预。每行从左到右依次为：MLP 消融、注意力头消融、MLP 增强和注意力头增强。

C.3 跨位置稳定性

我们在两个自回归阶段中分别生成 0、32、64、128 和 256 个词元后所处的位置提取风格方向，并在每个位置独立选择组件（表6）。

在乐谱阶段，只有第 128 步和第 256 步彼此接近；在音乐令牌阶段，开始之后所有位置的方向和注意力头都相当一致。KL散度也高度集中：在 54 首表征分析歌词和五种风格中，KL散度最高的 5% 位置平均承载了总 KL散度 的 59%（中位数为 59%，最低为 40%）。对其中六首歌词进行令牌类型计数后发现，在前 128 步等权平均的权重中，近一半落在几乎不受风格影响的标头格式文本上。

表6：风格读出与组件选择的稳定性（第 28 层）。读出是一种基于余弦相似度的原型风格分类器，在表征分析集上拟合；“高于随机水平”表示其测试准确率优于家族一致的标签置换零假设（经 Holm 校正的 $p < 0.05$）。一首表征分析歌词的乐谱短于 256 个词元，且没有第 256 步，因此所有位置的方向均使用其余 53 首；风格读出在 13 首测试歌词上进行评估。

比较	结果	
在相同位置训练和测试的风格读出 tion	在 80 个案例中的 79 个案例里高于随机水平	
风格读出跨位置迁移	在 320 个案例中的 228 个案例里高于随机水平；21.5%	when
对歌词进行重采样后重新选择多层感知机神经元（乐谱起始）	在乐谱起始处训练，并在第 256 步进行测试 杰卡德 0.80–0.89	

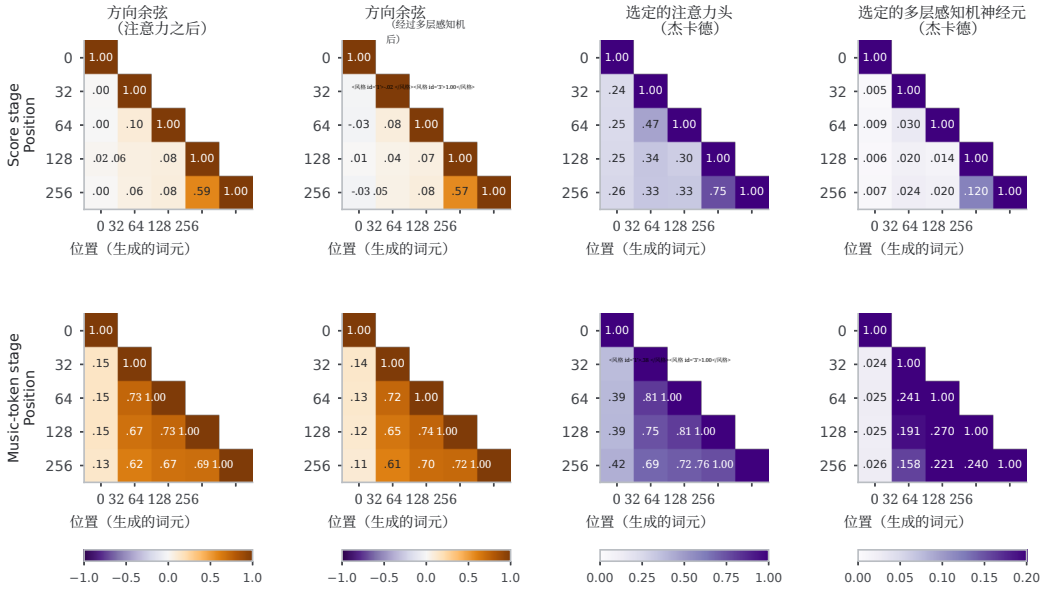


图17：两个自回归阶段中各位置之间的方向余弦和分量重叠（上：乐谱；下：音乐词元），解读方式同图 3；数值为五种风格的均值。

C.4 引导结果的稳健性

配对比较。 在测试集上，按歌词、风格和随机种子进行配对后，该电路的 CLAP边际与 CAA 相当，并超过其他每一种条件（歌词族自助法区间均高于零）。

乐谱与失效模式。 CLaMP 3 评判器将电路生成的金属摇滚和乡村音乐乐谱判定为目标风格的频率，与将提示词生成的乐谱判定为目标风格的频率大致相当，但很少将放克和布鲁斯乐谱判定为目标风格。即使在提示词条件下，它也很少识别出布鲁斯乐谱。风格提示未命中时，CLAP 大多将其识别为当代流行音乐，也就是模型的默认风格；电路未命中时，CLAP 大多将其识别为电声布鲁斯。

长度。 CLAP 对完整曲目中的 10秒窗口取均值，因此我们重新计算前30秒和中间20秒的命中数（表 7）。

在裁剪长度相同的情况下，电路在前30秒与提示词表现相当，并在中间20秒领先：提示词的风格集中在歌曲开头，而电路则能让风格贯穿整首歌曲，因此它的优势并非来自长度。

表7：固定长度裁剪片段上的 CLAP 命中数（测试集，每个条件生成150次）。

条件	音频均值（秒）	完整曲目	前30秒	中间20秒
空标签基线	53	30	30	30
风格提示（所有阶段）	66	113	123	93
单位置电路	64	73	73	62
均值方向电路	72	90	91	89
随机电路	58	34	37	28
动态电路	84	133	122	126

D 大调 风格引导的补充评估

D.1 各风格结果与更多基线

表 8 给出了测试集上所有条件下各风格的 CLAP 命中数。两个提示词写入相同的乐谱和音乐词元，仅在声学渲染方面有所不同，因此其命中数相同；正文保留了全阶段提示词。

表 8：使用两组种子对的 15 条测试歌词的 CLAP 命中数。

条件	Pop	金属乐	Funk	布鲁斯	乡村音乐	总计
空标签基线	27/30	0/30	0/30	0/30	3/30	30/150
风格提示（AR阶段）	29/30	17/30	13/30	29/30	25/30	113/150
风格提示（所有阶段）	29/30	17/30	13/30	29/30	25/30	113/150
单位置电路	30/30	5/30	2/30	15/30	21/30	73/150
均值方向电路	30/30	10/30	13/30	17/30	20/30	90/150
随机电路	27/30	0/30	0/30	3/30	4/30	34/150
CAA	26/30	22/30	26/30	21/30	23/30	118/150
动态电路	30/30	24/30	25/30	27/30	27/30	133/150

单位置电路采用 §3.3 中的方向和组件，在第一个乐谱位置（乐谱起始）注入，并在每个乐谱步都进行注入；均值方向电路则对前 128 个乐谱步的方向进行等权平均。两者与动态电路采用相同的组件预算、强度和音乐令牌干预。它们在金属摇滚和经典放克上几乎不起作用，总体命中次数也少得多，因此使用 KL 权重聚合方向至关重要。CAA 仅在经典放克上与该电路效果相当，而且多数未命中都是截断造成的。

传统引导。除 CAA（Rimsky等，2024）外，我们还测试了两种不使用稀疏电路的引导方法：像 CAA 一样注入 KL 加权残差方向，以及将 KL 加权差值稠密注入所有多层感知机神经元和注意力头。两种方法都将风格文本留空，只改变分数阶段干预，并在与电路相同的表征分析歌词上选择层和强度。两种方法均在 70 个测试生成结果上运行，表9也展示了 CAA 在这些测试中的结果。

表9：对70个测试生成结果进行传统引导（14首歌词和五种风格，第一组种子对 ）.

条件	CLAP	Gemini	截断	长度（秒）
风格提示（所有阶段）	56	54	0	61
CAA	52	54	12	185
KL 加权残差方向	60	50	1	65
稠密注入	56	51	4	91
动态电路	65	58	0	73

这三种方法都能在没有风格文本的情况下引导风格，但命中率低于动态电路，而且 CAA 和稠密注入经常导致生成偏离。KL 加权残差方向使用与电路相同的位置权重，但将注入对象从稀疏组件改为残差流；在 CLAP 评判下，它与电路接近，而在 Gemini 评判下得分较低。

D.2 Gemini 3.6 Flash 作为第二评审

Gemini 3.6 Flash（Google DeepMind，2026年）按照用于构建数据集的协议（附录 A）对每个测试集生成结果进行盲评：完整音频以匿名文件名上传，Gemini 在五种风格和一个干扰项“中国传统音乐”中进行选择；每个样本的选项顺序都会随机打乱，温度参数设为 0。空标签基线针对每个歌词和随机种子生成一段音频，并根据每个目标风格进行评分（表 10）。

表 10：Gemini 3.6 Flash 在 15 条测试歌词上的命中数（使用两对种子对）。

条件	Pop	金属乐	Funk	布鲁斯	乡村音乐	总计
空标签基线	14/30	1/30	0/30	0/30	15/30	30/150
风格提示（所有阶段）	24/30	12/30	15/30	30/30	30/30	111/150
单位置电路	25/30	13/30	7/30	9/30	30/30	84/150
均值方向电路	14/30	13/30	11/30	15/30	26/30	79/150
随机电路	13/30	0/30	0/30	0/30	20/30	33/150
CAA	24/30	22/30	25/30	23/30	26/30	120/150
动态电路	25/30	23/30	22/30	27/30	29/30	126/150

在歌词、风格和随机种子均相同的情况下，动态电路在两位评判器下都胜过风格提示（精确麦克内马尔检验；CLAP：不一致配对为 32 对比 12 对，$p = 0.004$；Gemini：31 对比 16 对，p =0.04）。在 CLAP 下，它胜过 CAA（28 对比 13 对，$p = 0.03$），但在 Gemini 下优势不显著（25 对比 19 对，$p = 0.45$）。由于 Gemini 也被用于筛选数据集，因此它并不是完全独立的评判器。

D.3 主观评估

我们从测试生成结果中抽取空标签基线、风格提示和动态电路生成的 20 秒片段，并进行响度归一化和匿名化。在每个问题中，听众会看到目标风格，为同一歌词和随机种子的三个片段在风格契合度方面进行评分，分数从 1（完全不符合）到 5（非常符合），并选出最接近的片段；各条件的呈现顺序会轮换。在问题开始前，听众会听到取自风格提示生成结果的各风格简短示例，这在一定程度上反而有利于风格提示。如果听众对金属摇滚、经典放克和电声布鲁斯的基线平均评分达到 4 分或更高，而基线几乎从不匹配这些风格，则会被排除（该规则预先固定），且只有完成的问题才会计入。

表11给出了初步结果（11位听众、完成了54个问题，其中43个涉及非流行风格）。对于非流行音乐问题，电路和提示词的评分都远高于基线（威尔科克森符号秩检验，$p < 0.001$ 各项均如此）；两者相比，19个问题中提示词评分更高，13个问题中电路评分更高，另有11个平局（$p = 0.64$）；提示词也更常被选为最接近的选项（27次，对比14次；二项 $p = 0.06$）。在流行音乐方面，电路的评分最高。目前的样本还不足以区分电路和提示词。

表11：均值风格匹配度评分（1–5），以及各条件被选为最接近选项的非流行音乐问题数量。

条件	Pop	金属乐	Funk	布鲁斯	乡村音乐	非流行音乐	选定
空标签基线	3.64	1.09	1.45	1.70	2.55	1.70	2/43
风格提示	4.00	3.82	4.09	3.40	3.82	3.79	27/43
动态电路	4.36	3.73	3.18	3.70	3.91	3.63	14/43

D.4 乐谱示例与配器测量

图18展示了一段测试歌词在三种条件下的声乐旋律线：空标签写成 C大调，风格提示转为 D♯m，而电路则写出更窄、更流畅的 C♯m 旋律。

空标签乐谱的调性。图5的空标签行汇集了测试集的两组种子对，其调性分布由随机种子而非歌词决定：乐谱标题栏在最初几个词元内进行采样，在使用相同随机种子的情况下，几乎每首歌词都会得到相同的调性（第一组种子对中，14份乐谱有12份为C大调；第二组中为E♭ 和B♭；使用第三个随机种子生成的附录A.2中的69份空标签乐谱则分布在G小调、D小调

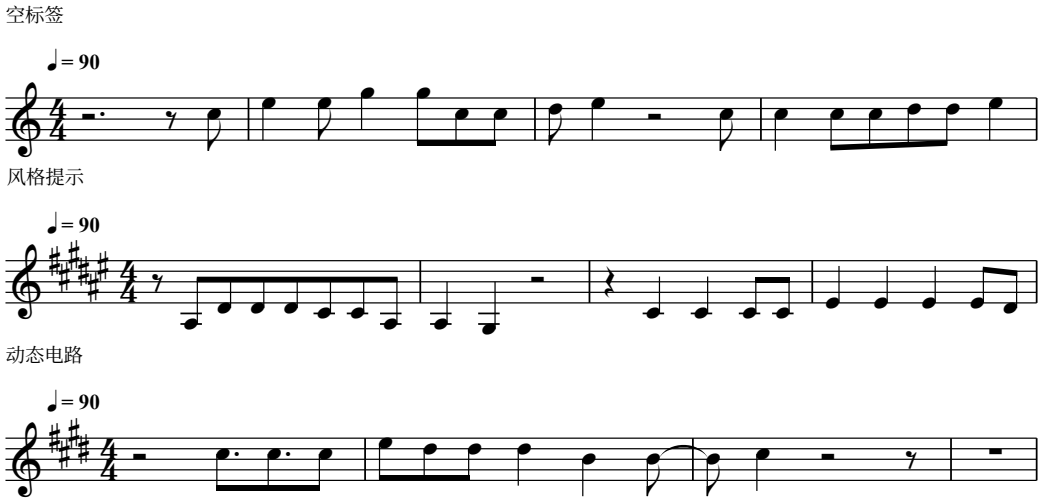


图18：三种条件下，一段具有目标风格金属摇滚的测试歌词的声乐旋律线，声乐进入后的前四小节。

和D大调之间）。因此，空标签没有风格典型调性，而风格提示和电路在两组种子对下都会将乐谱置于该风格的典型调性中。

乐谱未包含任何乐器，因此配器在后续阶段确定。使用 CLAP模型进行风格分析时，我们针对 13 条乐器描述（“此音频包含 {instrument}。”；表12）分别对每段测试音频进行评分，并减去同一歌词和随机种子对应的空标签音频的得分。

表12：配器分析中的乐器标签。

类别	标签	描述中的乐器	典型于
吉他	失真	失真电吉他	金属摇滚，电声布鲁斯
	清音电吉他	清音电吉他	经典放克，电声布鲁斯
	声学	原声吉他	乡村音乐，当代流行音乐
	踏板钢棒	踏板钢棒吉他	乡村音乐
节奏	架子鼓	响亮的架子鼓	金属摇滚、经典放克
	击勾贝斯	击勾贝斯	经典放克
Keys	管风琴	一架哈蒙德风琴	电声布鲁斯
	钢琴	钢琴	当代流行音乐
	合成器	合成器	当代流行音乐
Folk	班卓琴	班卓琴	乡村音乐
	小提琴	小提琴	乡村音乐
	口琴	口琴	电声布鲁斯，乡村音乐
铜管乐	铜管乐	一组带有号角的铜管乐器	经典放克

空标签音频以原声吉他和钢琴为主。CLAP 将电声布鲁斯听成了由失真吉他和鼓演奏的、最接近金属摇滚的音乐，这与它混淆这两种风格的情况相符。

D.5 内容与旋律保留

每种条件都从歌词出发，重新写一份乐谱，因此只有歌词是固定的。我们通过 CLaMP 3 乐谱特征，以及声乐旋律线的音程和节奏相似度，将每份乐谱与相同歌词和随机种子对应的空标签乐谱进行比较；同时以另一组种子对（重采样）对应的空标签乐谱作为自然变化的参考（表13）。

电路和提示词都停留在重采样层面：二者都从歌词出发，写出一段新的旋律。与空标签乐谱更接近的条件，命中目标风格的频率较低，这与乐谱是风格的主要承载者这一观点一致（附录A.3）。引导固定乐谱留待未来工作。

表13：每个条件的乐谱与相同歌词和随机种子的空标签分数之间的接近程度（测试集）。

条件	CLaMP 3余弦相似度	歌词优势	间隔相似度	节奏相似度
空标签重采样（参考）	0.553	0.027	0.310	0.383
风格提示	0.528	0.008	0.322	0.423
动态电路	0.561	0.022	0.323	0.436
均值方向电路	0.678	0.099	0.397	0.508
单位置电路	0.645	0.064	0.361	0.499
随机电路	0.747	0.148	0.483	0.586

E 声乐性别扩展

为测试针对流派发现的电路图景是否适用于另一种风格属性，我们对人声性别重复进行分析，将“男性人声”和“女性人声”描述词用作两个条件；性别描述词起到风格标签的作用，而空描述词则对应空标签。我们使用来自 32 个歌词族的 128 段简短歌词，按歌词族划分为 96 段表征分析歌词和 32 段评估歌词。方向、组件选择和干预设置仅根据表征分析划分进行估计。与流派实验不同，每段歌词都使用一份中性的 ABC乐谱，并在两个条件下逐字节完全相同地重复使用。因此，该实验隔离了乐谱固定后人声性别条件如何发挥作用。为进行表征比较，两次前向传播还会重放相同的已生成音乐令牌历史记录，且仅性别描述词不同；结果取男性生成和女性生成历史记录两种分支的平均值。

我们得出三项发现。在乐谱固定的情况下，剩余的人声性别效应几乎完全由音乐令牌阶段产生，而非声学渲染。在音乐令牌阶段中，音乐开始位置与后续位置差异显著，后续位置则稳定得多。最后，选定组件无需性别文本即可引导生成的人声性别，但对前 128 个位置取平均并不比 Music-Start 电路效果更好。

E.1 乐谱之后的人声性别控制主要是自回归的

我们按照 §3.2 中的分阶段消融方法，分别在音乐令牌阶段和声学渲染中切换性别描述词，以分离乐谱之后两种人声性别控制来源。在来自 24 个歌词族的 42 组配对中，切换音乐令牌使 CLAP 女性减男性差值平均变化 $+0.123$ ，而切换声学渲染条件仅为 $+0.0018$ 。所有 42 组配对中的音乐令牌效应均为正，且约为声学渲染效应的 68 倍；声学渲染效应在 42 组中有 41 组为正。两者的交互效应接近于零（$+0.0006$）。因此，在乐谱固定后，几乎所有剩余的人声性别效应都在音乐令牌阶段形成。本设计未测试乐谱规划本身的贡献。

E.2 音乐开始后人声性别表征趋于稳定

接下来，我们考察随着音乐令牌生成的推进，同一人声性别表征是否会持续存在。在位置 $p \in \{0, 32, 64, 128, 256\}$，其中 $p = 0$ 是音乐开始，我们根据共享历史对比 $h(c_F,\ x_{<p}) - h(c_M,\ x_{<p})$，估算女性减男性的方向，并独立选出沿该方向写入信息的多层感知机神经元和注意力头。我们通过方向余弦相似度和组件杰卡德重叠度比较不同位置，这与图 3 中用于流派的指标相同。两种历史分支得到的方向高度一致（音乐开始后余弦相似度为 0。92–0。99），因此结果并不取决于重放哪段生成历史。我们报告第 8–28 层的结果，因为在这些层中，对比结果并非主要由描述符令牌的浅层词汇痕迹所主导。

结果显示，从音乐开始到后期更稳定的阶段存在明显过渡（图19）。音乐开始方向与位置32–256的余弦相似度仅为0.36–0.42，而后期各位置之间的成对相似度则升至0.76–0.93。组件重叠也发生了相同变化。音乐开始与后期位置之间的MLP神经元Jaccard重叠仅为0.02–0.03，而后期各位置之间为0.15– 0.36，相比之下，随机水平约为0.01。注意力头重叠从0.40–0.42上升至0.66–0.85，而随机水平约为0.35。在留出歌词中，人声性别在每个位置仍然可被识别，其中音乐开始时的区分度最强。前128个位置的等

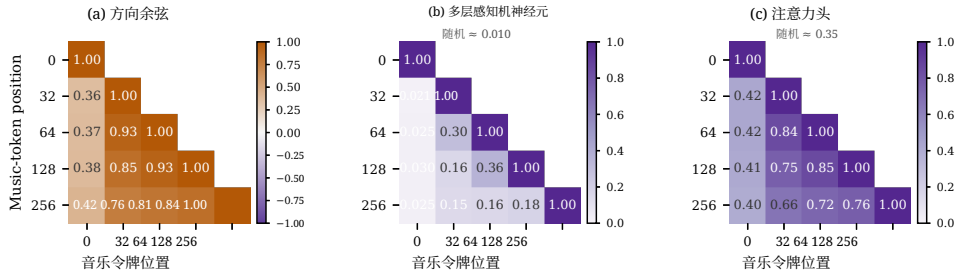


图19：**不同音乐令牌位置上的人声性别表征**。第 8–28 层中，女性减男性的共享历史对比。(a) MLP 之后残差方向之间的余弦相似度；(b) 所选 MLP 神经元的杰卡德重叠度；(c) 所选注意力头的杰卡德重叠度。音乐开始阶段与众不同，而位置 32–256 共享着稳定得多的表征。

权均值仍与位置32–128高度一致（0.93–0.97），并且与位置256也高度一致（0.84–0.86）。

因此，人声性别并非在整个音乐令牌阶段都由一个固定方向表征：音乐开始阶段与众不同，但生成开始后，表征会变得稳定得多。

E.3 人声性别电路的因果效应

Music Start 之后更高的稳定性促使我们构建一个聚合早期音乐令牌生成过程中信息的电路。我们对位置 0–127 上的共享历史分量差异和残差方向取平均，并使用 §3.4 中的音乐令牌分量分数和生成预算，构建一个前128个电路：包含 392 个多层感知机神经元和 168 个注意力头。由于该对比是二元的，一个电路通过相反的符号同时服务于两个目标，$\delta_F = (F - M)/2$ 和 $\delta_M = -\delta_F$。

我们在留出歌词的每个位置（从 0 到 127）测试选定组件。在不含性别描述词的前缀下，增强会添加 δ_F 或 δ_M；消融则在相应的性别前缀下将相同组件置零。随机对照与每层的组件数量相匹配，增强实验还包括一个与目标每层注入范数相匹配的随机对照。

增强产生了清晰的分量级因果效应（图20）。在生成预算下，168 个选定的注意力头为两个目标都产生了约为 56 的与目标对齐的 ΔS，而随机注意力头约为 4，范数匹配的随机注意力头为 7。5。392 个选定的多层感知机神经元产生了约为 5。2 的与目标对齐的偏移，而随机神经元仍接近于零。所有留出单元都朝预期方向移动。

消融的可解释性较低。将选定的注意力头置零后，在两种前缀下，结果都大幅朝女性方向变化，而不是对当前目标产生对称性削弱；此外，MLP 消融只有在男性前缀下才呈现预期方向。因此，我们将增强而非消融视为清晰的因果证据，表明选定组件能够写入声乐性别方向。

E.4 电路引导在没有性别文本的情况下改变声乐性别

最后，我们测试该电路是否会改变生成音频，而非仅改变内部表征。对于 32 段留出歌词，所有电路条件都使用固定的 ABC 乐谱，且不提供性别文本；声学渲染不受干预，并且在每个音乐令牌步骤注入组件差值。首 128 个变体的强度根据 $\{0.5, 1.0, 1.5\}$ 中的四段表征分析歌词选定，并在评估前固定。

我们比较前128个电路与单位置 Music-Start 电路。为了区分注入差异取均值的效果与重新选择组件的效果，我们还纳入了一种混合电路：保留 Music-Start 组件，但注入前128个平均差异。匹配随机电路使用相同的每层组件数量和注入范数。

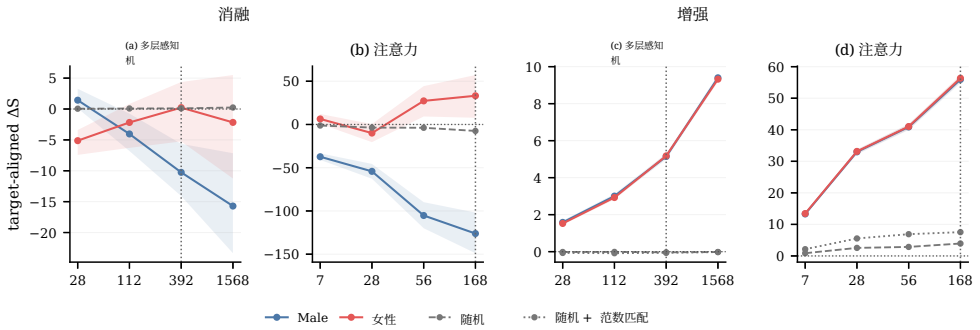


图20：**对前 128 个位置的人声性别电路进行因果干预**。从左到右依次为：对位置 0–127 进行 MLP 消融、注意力头消融、MLP 增强和注意力头增强。纵轴为目标对齐的 ΔS（男性效果的符号反转，因此正值始终指向目标）。竖线标示生成预算；色带为按歌词类别进行自助抽样得到的 95% 音程。灰色虚线为随机对照，灰色点线为增强实验的范数匹配随机对照。增强具有特异性，并且在不同目标间近乎对称；而注意力头消融则不具特异性。

表 14：**不使用性别文本的引导**。在每个目标的 32 条留出歌词上计算 CLAP目标命中（总计 64 次带目标标签的评估），并计算相对于空描述符基线的女性减男性差值的配对目标方向偏移。偏移为正表示朝请求的目标移动。

条件	男性命中数	女性命中数	目标方向总偏移	
空描述符基线	11	21	32	0
性别提示词	13	28	41	+0.045
Music-Start 电路	13	25	38	+0.024
前128个电路	12	22	34	+0.010
Music-Start 组件与前 128 个组件的差异	16	25	41	+0.025
匹配随机电路	10	21	31	−0.003

所有目标电路都会使最终音频朝所请求的性别方向移动，而匹配随机干预则不会（表 14）。Music-Start 电路带来的目标方向 CLAP 边际偏移为 $+0.024$，混合电路为 $+0.025$；相比之下，匹配随机组件为 -0.003。前128个电路也有效（$+0.010$），但效果弱于 Music-Start 电路；直接使用性别提示词时为 $+0.045$。目标命中数也呈现相同的排序：Music-Start 电路为 38，前128个电路为 34，而匹配随机组件为 31。

因此，回路分析也适用于人声性别，但最佳干预方式因属性而异。与流派评分阶段结果不同，对前 128 个音乐令牌位置取平均，并未优于单个位置上的 Music-Start 电路，尽管后续表征之间高度对齐。

局限性。 CLAP 是用于评估人声性别的自动代理指标；我们未报告人工或 Gemini 评判结果。引导在 32 条留出歌词上进行评估，且仅使用一个随机种子；所选前128个电路的强度处于测试范围的上限。注意力头消融缺乏特异性，不应将其解读为干净的逆向干预。最后，该方向是根据规范表述“男性人声”和“女性人声”定义的。释义实验显示结果高度依赖措辞，因此我们不声称存在与措辞无关的抽象性别轴。