TimberAgent：面向可执行音乐效果控制的格拉姆引导检索

何世豪*、夏一涵*、刘芳、王涛涛和张胜利

arXiv:2603.09332v1 [cs.SD] 10 Mar 2026

摘要——**数字音频工作站提供了丰富的效果链，但感知用户意图与底层信号处理参数之间仍存在语义鸿沟。我们研究基于检索的音频效果控制，其输出是可编辑的插件配置，而非最终波形。我们的重点是纹理共振检索（TRR），这是一种基于投影后的中层 Wav2Vec2 激活值的格拉姆矩阵构建的音频表征。该设计保留了与纹理相关的共激活结构。我们在吉他效果基准测试中评估 TRR，其中包含 1,063 个候选预设和 204 个查询。评估采用 Protocol-A，这是一种可避免训练集与测试集数据泄漏的交叉验证方案。我们将 TRR 与 CLAP 及内部检索基线（Wav2Vec-RAG、Text-RAG、FeatureNN-RAG）进行比较，并使用以物理 DSP 参数范围为基础的最小-最大归一化指标。消融研究验证了 TRR 的核心设计选择：投影维度、层选择和投影类型。近重复敏感性分析证实，结果不受简单知识库匹配的影响。TRR 在所评估的方法中取得了最低的归一化参数误差。一项有 26 名参与者参与的多刺激听音研究提供了补充性的感知证据。我们将这些结果视为基准测试证据，表明纹理感知检索有助于实现可编辑的音频效果控制；而更广泛的个性化和真实音频鲁棒性主张，仍不属于本文所呈现的已验证证据范围。**

索引词—**音频效果控制、音频检索、数字音频工作站、多模态音乐制作、可解释音频系统**

I. 引言

神经音频合成面临着保真度 与控制之间的根本性矛盾。近期的高保真生成模型主要基于潜扩散 [1, 2],，能够实现出色的感知质量，但往往如同“黑箱”般运作。它们生成最终波形，其中的内部属性——例如压缩器的起音时间或混响的衰减——彼此纠缠，无法进行事后编辑。相反，可微数字信号处理（DDSP）[3] 提供了可解释的结构化控制，但常常受逆参数估计问题适定性不足的困扰。

这种张力推动了基于检索的音频控制系统的发展，使其能够将感知意图映射为可解释的 DSP 参数。此类系统通过检索相关预设，而非从头生成参数，旨在生成可编辑的输出，并与数字音频工作站工作流程集成。从这一角度看，检索的价值不在于它能以封闭形式求解逆问题，而在于它能将查询置于参数空间中一个合理且可执行的邻域内，供用户继续检查和完善。

关键挑战在于通过检索，将感知描述与参数空间建立依据。标准音频表征（例如 CLAP 风格嵌入或经均值池化的 Wav2Vec2 特征）会将一个音频片段概括为单个向量；虽然这对于粗略语义十分有效，但这种池化可能会丢弃对纹理主导型效果至关重要的时间共激活模式（例如颤音调制或信号相关失真）。

为弥合这些差距，我们提出一种以检索为依据的参数控制框架，并聚焦于其核心检索组件：**纹理共振检索 (TRR)**。TRR 使用深层特征激活的格拉姆矩阵对音频风格建模 [4], ，以显式捕捉定义时间纹理的二阶相关性。我们将这种表征用作纹理感知检索先验：其作用是引导搜索偏向能够保留与效果相关结构的可执行预设，而非声称它是音色或参数空间相似度的通用度量。尽管该框架面向一般的音频效果参数控制，本文的实证评估是在一个留出的吉他效果基准测试上进行的。

具体而言，本文提出了三项有证据支持的贡献。首先，我们将可编辑的音频效果控制表述为一种以检索为基础的预设选择问题，使输出在数字音频工作站工作流程中仍可执行且可检查。其次，我们提出**纹理共振检索（TRR）**，将其作为纹理感知检索先验，并表明二阶统计量（格拉姆矩阵）能够提升当前基准测试中针对纹理主导型效果的检索表现，且在评估的检索基线中（包括 CLAP 检索基线）实现了最强的参数对齐效果。第三，我们报告了一项经过审计的留出评估，其中包含置信区间、配对置换检验以及一项多刺激听音研究。

II. 相关工作

我们的工作处于智能音乐制作、神经音频合成和基于检索的参数控制三者的交叉领域。

智能音乐制作。

将感知描述转化为技术参数这一挑战，最早由 Stables 等人 [5]进行了正式界定。他们的 SAFE 项目展示了在数字音频工作站中实现语义到参数映射的可行性，但未能解决跨乐器、效果链和风格子流派的泛化问题。Schedl 等人 [6] 提供了

S. He、Y. Xia、F. Liu、T. Wang 和 S. Zhang 任职于中国深圳深圳大学电子与信息工程学院，电子邮箱：2410044045@mails.szu.edu.cn；{xiayihan2023、liuf、ttwang、zsl}@szu.edu.cn。S. He 和 Y. Xia 对本工作贡献相同。（通讯作者：F. Liu 和 T. Wang）

一份关于音乐信息检索技术的全面综述，指出语义自动标注是一项关键挑战。

早期关于交互式控制的研究着重于学习直观的合成器映射。Huang 等人 [7]提出利用高斯过程进行主动学习，以发现感知控制旋钮；而较近期的 Brade 等人 [8]推出了 SynthScribe，这是一款用于合成器声音检索的多模态工具。尽管这些工作推动了界面设计的发展，但其主要针对合成参数（例如振荡器波形），而非复杂的音频效果信号链。

音频分析方法。后续研究借助音频内容分析和可微混音系统 [9–11]实现自动化。虽然当已有目标混音或参考音频时，这种方法效果显著，但它默认目标音频已知。因此，它难以支持仅凭概念性请求（“让它更有复古失真感”）进行创意音色设计，而且往往面临**不透明性问题**：用户看得到输出，却缺乏可理解、可编辑的依据。

可微数字信号处理的最新进展推动了音频效果建模的神经网络方法。Peladeau 等人 [12]提出利用自编码器架构盲估计音频效果参数，以解决参数恢复这一逆问题。Liu 等人 [13]提出 DDSP-SFX，该方法利用可微数字信号处理生成声学引导的音效，并将神经网络与传统数字信号处理模块相结合，实现可控合成。Dees 和 O’Leary [14] 探索了用于神经音频合成的生成式潜在空间，弥合了数字信号处理驱动方法与数据驱动方法之间的差距。与此同时，风格迁移方法也被应用于音频效果；Steinmetz 等人 [15] 推出 ST-ITO，通过基于预训练音频表征的推理时优化来控制音频效果。

我们的定位。与以往将参数预测视为回归问题 或纯粹生成问题的工作不同，我们将其构建为一项 检索任务，从知识库中定位相关且先前已验证的预设，并在下游数字音频工作站工作流中保持可编辑性。

神经音频合成。

神经音频合成发展迅速，实现了高保真度的文本到音频生成 [1, 2, 16, 17]。扩散模型已成为主导范式：Huang 等人[18] 提出了提示增强的文本到音频生成，Wang 等人 [19] 借助潜在一致性模型实现了高效的高质量合成，Ghosal 等人 [20] 则探索了指令引导的潜在扩散以提升可控性。然而，这些模型主要输出最终波形，难以在专业工作流中进行编辑。

我们将参数级生成视为对波形级合成的**补充而非替代**：它面向数字音频工作站内部的可控编辑与迭代优化。

这促使我们提出所谓的**参数–波形二元性**：波形级生成具有最大的表现力，但可编辑性有限；而参数级控制则保留了透明性和迭代优化能力，代价是需要领域知识。

与此同时，可微分 DSP（DDSP） [3] 表明，学习模型可以控制可解释的合成参数，而自监督音频表示学习（例如，Wav2Vec 2.0 [21]）可提供强大的声学嵌入。我们的检索锚定表述利用这些表征进行相似度依据，但目标是输出与以数字音频工作站为中心的工作流兼容的可编辑参数。

音频表示学习。近年来，音频表示学习取得了长足进展，产生了适用于语义任务的强大嵌入。CLAP [22] 通过对比预训练实现了稳健的文本—音频对齐，而 PaSST [23] 和 PANNs [24]则提供了强大的声学特征，用于分类和标签标注。尽管这些表征在语义理解方面表现出色，但它们主要是为分类和图像描述任务而设计的。我们的工作将音频表征（Wav2Vec2）用于不同目的：捕捉音色纹理以进行参数检索，其中二阶统计量（格拉姆矩阵）可补充一阶语义嵌入。我们使用中层 Wav2Vec2 特征，而非 CLAP 风格的单向量嵌入，因为帧级激活能够用于计算格拉姆矩阵，从而表征纹理。

检索锚定参数控制。

检索增强生成（RAG） [25]为将输出锚定于外部存储库的系统提供了一个概念框架，但本文研究的是一个更狭窄的场景：针对可执行预设的检索锚定参数控制。我们关注的不是开放式文本生成，而是检索可行的参数范例，供用户检查和编辑。这使我们的研究场景有别于波形生成器和以文档为中心的多模态RAG系统。

a) 感知相似性与参数可迁移性：即使是 CLAP 或 PaSST 风格嵌入等强大的通用音频表征，也面临一个根本挑战：感知相似并不一定意味着底层效果参数具有可迁移性。近期的音频语言模型提升了零样本分类和字幕生成能力 [22，26–28]，，但其主要面向离散预测任务，而非连续参数估计。遵循指令的音频语言模型进一步拓展了音频-文本理解 [29, 30]，，但可执行效果链控制仍有其独特性，因为系统必须遵守插件约束和可行的连续控制范围。我们的工作通过采用二阶统计量（格拉姆矩阵）来捕捉与纹理相关的特征，从而弥合这一差距；这些特征与参数空间邻近性更为一致。

音频纹理与语义音频检索。

音频纹理的概念已在风格迁移和合成领域得到广泛研究。Gatys等人[4]为图像提出了基于格拉姆矩阵的风格表示，随后Bruna与Mallat [31]等人将其应用于音频。关于声音纹理感知的奠基性工作 [32] 以及对用于声音合成的音频表征的综述 [33] ，凸显了捕捉二阶统计特征对于音色表征的重要性。与以往主要将格拉姆统计量用于纹理合成或风格迁移的做法不同，我们将其用作检索嵌入，以定位可执行的参数范例，并将检索视为后续约束推理的先验。

在语义音频检索领域，WikiMuTe 数据集 [34]为连接自然语言描

述与音乐音频提供了丰富资源，而检索引导方法[35] 则展示了将检索与生成相结合用于音乐描述生成的价值。检索增强文本生成音乐 [36] 进一步验证了检索增强生成方法在音频任务中的有效性。

我们的 TextureResonance 检索（TRR）基于这些见解，使用格拉姆矩阵通过二阶统计量捕捉音频纹理，从而为参数推断提供更具纹理敏感性的检索先验。

表 I 总结了我们的定位。我们的独特贡献是通过 TRR 实现面向可执行参数控制的纹理感知音频检索。

III. 问题表述

我们将音频效果参数检索形式化为一个受约束的合成问题：给定用户查询和输入信号，生成一个参数向量，使其 (i) 与用户意图在语义上保持一致，且 (ii) 可由实时 DSP 引擎执行。不同于波形生成，参数合成必须遵守严格的插件定义有效性约束，而从感知语言（例如“温暖”“有冲击力”）到高维控制向量的逆映射本质上是病态的。因此，我们的目标并非声称在 Θ 中找到全局最优解，而是生成一个可行的点估计，该估计定义明确、可解释，并以具体的可执行原型为依据。为进行实证评估，我们使用吉他多效果处理链实例化 Φ 和 Θ。

A. 信号处理动态

考虑一个（可能不可微的）数字信号处理渲染算子 $\Phi : \mathcal{X} \times \Theta \to \mathcal{X}$，其中 $\mathcal{X}$ 表示离散时间音频信号的空间，$\Theta \subset \mathbb{R}^d$ 表示 d 维控制参数空间。给定输入信号$\mathbf{x}_{\text{in}} \in \mathcal{X}$，该算子会生成输出 $\mathbf{x}_{\text{out}} = \Phi(\mathbf{x}_{\text{in}}; \theta)$。在实际的数字音频工作站和插件系统中，$\Phi$ 可能包含非线性、条件分支（例如旁通逻辑）和内部状态（例如平滑处理），因此，端到端微分可能不可靠，或不适用于直接在 θ 上进行优化。

参数空间 Θ 并非自由向量空间，而是受结构约束的有界超矩形：

$$\Theta = \left\{ \theta \in \mathbb{R}^d \mid \mathbf{l} \preceq \theta \preceq \mathbf{u}, \quad \mathcal{C}(\theta) = 1 \right\} \tag{1}$$

其中，$\preceq$ 表示逐元素不等式，$\mathbf{l},\mathbf{u}$ 定义物理运行范围（例如，频率界限 $[20, 20\text{k}]$），而$\mathcal{C}: \mathbb{R}^d \to \{0, 1\}$ 表示用于编码参数间依赖关系的二元有效性函数（例如，确保滤波器谐振$Q > 0$）。在实践中，数字信号处理参数空间可能是非凸且部分离散的（例如，开关或模式选择器），因此我们将式 1 用作务实的可执行抽象，而非声称其具有凸性或光滑性。这正是我们依赖检索支撑的合成和确定性有效性约束，而非直接对 Θ进行基于梯度的优化的原因。

B. 目标与约束

用户意图由多模态查询$q = (t,\mathbf{a}_{\text{ref}})$概括，其中包含自然语言描述 $t \in \mathcal{T}$和可选的音频参考 $\mathbf{a}_{\text{ref}} \in \mathcal{X} \cup \{\emptyset\}$。我们的目标是生成与查询一致的有效 参数配置：

$$\tilde{\theta} \in \Theta \quad \text{given } q, \tag{2}$$

其中，有效性由插件特定约束函数 $\mathcal{C}(\theta) = 1$ 和物理边界 $\mathbf{l} \preceq \theta \preceq \mathbf{u}$ （公式 1）定义。我们用 θ^* 表示一个理想（可能不唯一）的解，该解能最佳匹配用户的感知意图；在实践中，我们只能使用基于检索的近似，因此˜旨在获得一个可行的点估计 θ。

a) 操作目标（不涉及概率语义˜）。: θ 的查询一致性无法直接用可微损失衡量，因为目标具有感知性，并且取决于复杂的信号链。因此，我们采用基于检索的操作目标：利用有限预设数据库，在 Θ 中构建可执行的合理邻域，然后在该邻域内进行调整，同时确保可行性。由此得到的系统，其中间产物（检索到的原型、草拟参数、约束修复）在设计上均可审计，这对于强调系统严谨性和可复现性的期刊论文至关重要。

C. 基于检索的方法

由于从感知意图到参数的逆映射存在病态性，且 DSP 有效性约束十分严格，我们采用基于检索的处理流程。令 z 表示从有限数据库中检索到的预设（离散原型）。我们根据 q检索排名前K 的候选项：

$$\theta_{\text{selected}} = \text{Top-}K(q) \tag{3}$$

直观而言，检索提供了一个粗略的寻模态步骤，将查询置于 Θ 的合理邻域中。

a) 设计原则*1*（模态缺失时仍能明确定义合成）。: 由于$\mathbf{a}_{\text{ref}}$ 是可选的，因此当$\mathbf{a}_{\text{ref}} = \emptyset$时，合成流程仍须有明确定义。我们的形式化将 q 视为包含显式空元素的元组，并要求每个下游阶段（检索和投影）在输入模态缺失时都指定确定性的回退行为。

IV. 方法论

我们通过**纹理共振检索（TRR）**实现第三节中定义的基于检索的框架，利用二阶纹理统计量检索候选预设。

A. 系统架构

系统通过解耦设计实现算子 $\Phi(\mathbf{x}_{\text{in}}; \theta)$，如图1所示。

该架构由两个主要组件构成。(i) **实时 DSP 引擎**将 Φ 实现为一个由六模块效果链构成的低延迟音频处理图，并执行从知识库中检索到的参数。我们将

表 I 与既有工作的定位比较

Work	输入	输出	依据	可编辑性
SAFE [5]	Text	参数	None	Yes
DDSP [3]	音频	参数	None	Yes
LLM2Fx [37]	Text	参数	None	Yes
AudioLDM [1]	Text	波形	预训练	No
HM-RAG [38]	文本+文档	Text	检索	N/A
基于 TRR 的检索控制（我们的方法）	**文本+音频**	**参数**	**双模态检索**	**Yes**

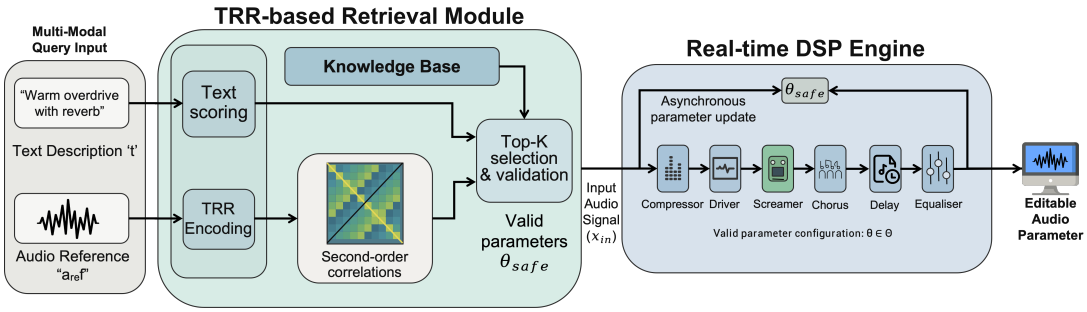


图 1. 检索增强的音效控制流水线的系统概览。实时 DSP 引擎采用六模块串联效果链（如框架所示），其中每个模块都可以独立旁路。检索到的参数在应用前会根据各模块的物理边界（式 1）进行验证。

DSP 引擎使用的参数状态记为 θ_{safe}，可随时间平滑处理，以避免拉链噪声和可闻伪影。**(ii) 检索模块**通过异步执行双模态搜索（文本和音频），以低于音频回调的速率实现映射 $q \to \theta_{\text{safe}}$，并返回保证满足插件约束的有效参数。

a) 设计原则 *2*（运行时隔离）。：在推理时，实时音频路径是 $(\mathbf{x}_{\text{in}},\ \theta_{\text{safe}})$ 经由 $\Phi(\mathbf{x}_{\text{in}};\theta_{\text{safe}})$ 得到的确定性函数。推理代理不会在实时回调中执行，只能在有效性检查后通过异步更新 θ_{safe} 来影响音频。

B. 纹理共振检索（*TRR*）

纹理感知检索的核心挑战在于，为音频模态计算一种能够保留与效果相关动态特征的相似度度量。标准嵌入（例如，均值池化的 Wav2Vec2 特征）会将时间结构压缩成静态向量，丢失对调制和失真等效果至关重要的信息。

为解决这一问题，我们提出**纹理共振检索（TRR）**。基于音频纹理编码在二阶特征统计中的假设 [32],我们计算深层激活的格拉姆矩阵。

令$\mathbf{H}_\ell \in \mathbb{R}^{T_{\text{frm}}\times 768}$ 为预训练 Wav2Vec2 Base 模型第 ℓ 层的帧级激活图。在基准测试实现中，我们使用中层集合 $\mathcal{L} = \{4,5,6\}$，通过冻结的随机投影（采用 Xavier 初始化，未经训练）将每一帧线性投影到 32 维，为每一层计算一个格拉姆矩阵，将得到的 32×32个矩阵取平均，把结果展平为一个 1024 维向量，并

算法 1 基准测试用纹理共振检索（TRR）编码 **要求：** 音频样本 a、Wav2Vec2 模型 $\mathcal{W}$、层集合 $\mathcal{L}=\{4,5,6\}$、冻结的随机线性投影 P：$\mathbb{R}^{768}\to\mathbb{R}^{32}$**确保：** 纹理嵌入 $\mathbf{g}_{\text{norm}}$

1. 预处理：对 a归一化并填充至最小长度。对于每个 $\ell\in\mathcal{L}$，执行 $\mathbf{H}_\ell \leftarrow \mathcal{W}_{\text{layer }\ell}\ a \in \mathbb{R}^{T_{\text{frm}}\times 768}$ ( ) $\mathbf{H}_\ell \leftarrow P$ $(\mathbf{H}_\ell) \in \mathbb{R}^{T_{\text{frm}}\times 32}$ $\widehat{\mathbf{G}}_\ell \leftarrow \frac{1}{T_{\text{frm}}}\widehat{\mathbf{H}}_\ell^{\top}\widehat{\mathbf{H}}_\ell \in \mathbb{R}^{32\times 32}$循环结束$\bar{\mathbf{G}} \leftarrow \frac{1}{|\mathcal{L}|}\sum_{\ell\in\mathcal{L}}\mathbf{G}_{\ell}$ $\mathbf{g}_{\text{norm}} \leftarrow \text{vec}(\mathbf{G})/\|\text{vec}(\mathbf{G})\|_2$返回 $\mathbf{g}_{\text{norm}}$

对其进行 L2 归一化。平均纹理描述符可表示为

$$\bar{\mathbf{G}} = \frac{1}{|\mathcal{L}|}\sum_{\ell\in\mathcal{L}}\frac{1}{T_{\text{frm}}}\widehat{\mathbf{H}}_\ell^{\top}\widehat{\mathbf{H}}_\ell, \tag{4}$$

其中 $\widehat{\mathbf{H}}_\ell$ 表示投影特征图。条目$\bar{\mathbf{G}}^{ij}$ 捕获投影通道 i 与 j 的共激活，而与其绝对时间位置无关。这种时间不变性使 TRR 能够匹配纹理（例如“快速颤音”），即使参考与数据库条目未按移相-对齐。对 vec(G) 进行 L2 归一化等价于对矩阵表示进行 Frobenius 归一化，随后相似度按展平描述符之间的余弦相似度计算。在计算上，格拉姆矩阵的构建在投影之后按 $O(D^2\cdot T_{\text{frm}})$ 扩展，其开销高于均值池化，但对于较短的缓存片段仍然可行。

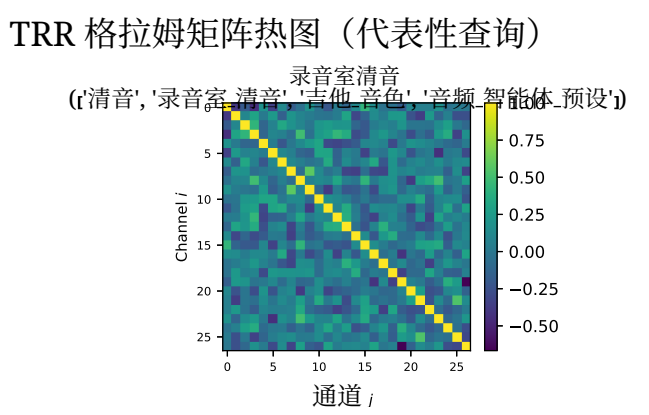


图 2. 代表性查询预设的格拉姆矩阵热图。每个矩阵展示了投影通道间的共激活结构，揭示了不同效果风格的独特纹理模式。

a) 性质与局限：TRR 可视为深层特征的二阶摘要：它保留跨通道相关结构，同时舍弃绝对时间对齐。图 2 展示了多种查询预设的代表性格拉姆矩阵热图，说明 TRR 针对不同效果风格所捕获的不同共激活模式。

这对于效果空间中的检索是可取的，因为感知上的“风格”往往通过重复的调制模式或稳态纹理来表达，而非单一的时间锁定事件。Frobenius 归一化进一步降低了对特征激活全局尺度变化的敏感度，使余弦相似度比较较少受幅度主导。然而，TRR 并非通用表示：通过对时间进行聚合，它有意丢弃细粒度的顺序信息，并可能对瞬态主导效果表征不足。此外，与其他深度表示一样，检索行为可能取决于所选的骨干层以及上游特征提取器。

b) 性质 *3*（对齐不敏感纹理统计）：设$\mathbf{H} \in \mathbb{R}^{C\times T_{\text{frm}}}$为帧级特征矩阵，设$\mathbf{P} \in \mathbb{R}^{T_{\text{frm}}\times T_{\text{frm}}}$为任意帧置换矩阵。则 TRR 格拉姆矩阵满足$\frac{1}{T_{\text{frm}}}\mathbf{H}\mathbf{H}^{\top} = \frac{1}{T_{\text{frm}}}(\mathbf{H}\mathbf{P})(\mathbf{H}\mathbf{P})^{\top}$。因此，TRR 取决于时间聚合共激活，对帧的绝对对齐不敏感，这也使其适合作为面向纹理的检索信号。

C. 复杂度与延迟考量

期刊读者通常不仅关注概念上的正确性，也关注运行约束。我们的设计将 (a)实时执行与 (b) 异步推理分离开来：前者必须保持确定性并具备低延迟，后者可以更慢，且不会在音频回调中执行。在推理侧，TRR 编码通过 $O(C^2T_{\text{frm}})$格拉姆矩阵构造（第 IV-B 节）占据音频检索成本的主要部分，随后执行标准的 top-K 最近邻搜索。检索系统与实时音频路径异步运行，从而确保音频处理延迟不受检索操作影响。

表 II 每次查询的本地检索路径延迟明细（仅缓存）。报告来自原始 211 次查询分析样本池的 633 次运行（211 × 3 次重复）的中位数和第 95 百分位数。

组件	中位数（毫秒）	第 95 百分位数（毫秒）
文本评分	3.027	3.586
音频评分（TRR 余弦相似度）	1.369	1.564
融合	0.110	0.146
投影	0.031	0.039
端到端（本地流水线）	4.538	5.269

a) 延迟测量协议：我们在仅缓存分析设置下，报告本地检索路径可复现的离线延迟分解。测量在一台 Ubuntu 24.04 LTS 服务器（x86-64、英特尔 CPU；Python 3.12）上进行，使用原始的 211 条查询分析样本池，每条查询执行 5 次预热迭代和 3 次重复测量。我们报告所得 633 次计时查询运行的中位数和第95百分位数。这些测量不包括实时网络大语言模型延迟，因此应将其理解为本地检索路径的计时结果，而非端到端云部署预算。

检索系统与实时音频回调异步运行，确保检索操作不会影响音频处理延迟。

D. 多模态融合作为诊断扩展

除核心检索基准之外，该代码库还包含一种实验性融合机制，可在输入质量下降条件下结合文本侧分数与音频侧分数。候选预设 z_i 的融合分数计算如下：

$$s_{\text{fused}}(z_i) = w_{\text{text}} \cdot s_{\text{text}}(z_i) + w_{\text{audio}} \cdot s_{\text{audio}}(z_i), \qquad (5)$$

其中 $w_{\text{text}} + w_{\text{audio}} = 1$ 与权重由质量感知启发式规则确定：若文本查询被判定为模糊（例如检索置信度较低），则 $w_{\text{text}} \to 0$；若音频侧嵌入范数低于阈值，则 $w_{\text{audio}} \to 0$。在当前代码库中，该机制被实现为一种实用的质量感知回退规则，而非单一、清晰的概率模型，因此我们将其结果视为诊断压力测试，而非论文主要经验证的贡献之一。因此，本稿的核心证据链仍是 Protocol-A 中仅检索的基准测试。

V. 实验

我们通过回答一个核心研究问题来验证所提出的框架：对二阶纹理统计量（TRR）进行建模，是否能相较一阶基线改进参数推断?

A. 实验设置

a) 数据集与知识库：我们的完整服务器端语料库包含从吉他效果器预设收集的 1267 条配对记录（音频–文本–参数三元组）。对原始基于歌曲名称的留出列表（211 条测试查询、1056 个知识库条目）的事后审计发现了 16 条跨数据划分共享的已解析音频路径和 48 对跨数据划分的近重复配对

在廉价指纹下。为了从主要客观基准测试中消除精确的音频复用，我们在对留出集进行采样之前，按已解析的本地音频路径对记录重新分组。由此得到的更严格 Protocol-A 划分包含 204 条测试查询和 1063 个知识库条目，在当前审计下跨划分的精确共享已解析音频路径为零。我们将其保留为试点基准测试，因为廉价近重复配对仍然存在，且语料库仍仅限吉他。

b) 来源说明： 数据集 JSON 暴露了 SongName、Style、Feature、Parameters 和 Vectors，但不包含单独且可审计的文本来源字段。因此，我们将 Style 和 Feature 视为代码库中唯一明确可用的文本侧来源，并将存储的 Parameters 字段作为操作层面的真实值。

c) 范围与协议： 除非另有说明，本节的客观结果均采用更严格的按解析音频分组的划分上的 **Protocol-A** （仅检索），且所有检索方法均使用 Top-1 检索。Protocol-C 和延迟概况则基于各自特定协议的数据池进行报告，并明确标注。

d) 近重复敏感性分析： 尽管解析音频划分移除了跨划分的完全重复音频，但知识库中仍存在近重复预设（仅在少数参数值上有所不同的预设）。为量化它们对检索指标的影响，我们通过逐步从知识库中移除近重复项来开展敏感性分析，移除依据为参数空间 L2 距离阈值 $\tau \in \{0.005, 0.01, 0.02, 0.05\}$，随后重新评估所有检索方法。完整的近重复统计数据和各阈值下的结果见补充材料中的近重复敏感性表。在所有测试阈值下，TRR 相较基线方法的相对优势仍然保持；尽管随着知识库缩小，绝对性能平稳下降，但这证实了结果并非由简单的近重复匹配所驱动。

e) 资源与可复现性： 为便于复现，我们提供了一份供审阅的匿名补充材料包，其中包含预处理细节、嵌入计算步骤、检索索引构建、提示模板、评估脚本，以及所有已报告表格所用的数据划分规范。论文录用后，我们将公开经过整理的代码库和制品。

f) 基线： 我们将提出的**TRR**（第 IV-B 节）与四种检索基线进行比较。**Text-RAG** 基于元数据字段执行稀疏词汇检索。**FeatureNN-RAG**使用基准测试包中发布的 Feature NN 向量，因此应将其视为固定的内部稠密检索参考，而非单独重新实现的外部模型系列。**Wav2Vec-RAG** 使用 Wav2Vec2 最后一层 Transformer 层的均值池化激活。若经验证的制品集中提供了缓存的 CLAP 音频嵌入，**CLAP**便使用这些嵌入执行稠密检索。

B. 指标

我们通过五项操作性指标评估参数空间中的对齐情况，这些指标基于展平后的数值参数叶节点计算。具体而言，对于嵌套参数字典，我们递归地将数值叶节点展平为稀疏向量，并以点分隔键作为索引（例如 DelayOn.Mix），同时忽略非数值叶节点（例如字符串）。除非另有说明，我们取预测结果与真实值所含键的并集，并在计算指标时将缺失的数值键视为零。需要注意的是，我们不进行逐参数归一化；因此，L2 范数可能会被少数取值范围较大的参数主导，应仅在各自的协议内进行解读，不宜跨协议比较。

*a) L2*误差（均方根误差）： 给定并集键集合中大小为 d 的展平向量 $v(\theta_{\text{gt}})$ 和 $v(\theta)$，我们报告均方根误差

$$\text{L2} = \sqrt{\frac{1}{d}\sum_{i=1}^{d}\left(v(\theta_{\text{gt}})^{(i)} - v(\tilde{\theta})^{(i)}\right)^2}. \tag{6}$$

b) 准确率*@0.1*： 我们报告绝对误差在固定容差 $\tau = 0.1$ 范围内的展平数值参数所占比例：

$$\text{Acc@0.1} = \frac{1}{d}\sum_{i=1}^{d}\mathbb{I}\left(\left|v(\theta_{\text{gt}})^{(i)} - v(\tilde{\theta})^{(i)}\right| \le 0.1\right). \tag{7}$$

c) 召回率（活跃数值参数）： 若满足 $\left|v(\theta_{\text{gt}})^{(i)}\right| > 0.05$, 则将真实数值参数定义为活跃 。召回率是指预测值落在同一容差 $\tau = 0.1$ 范围内的活跃真实维度所占比例。

d) 余弦相似度： 我们基于并集键集合上的展平向量计算余弦相似度：

$$\text{Cosine} = \frac{v(\theta_{\text{gt}}) \cdot v(\tilde{\theta})}{\|v(\theta_{\text{gt}})\|_2 \, \|v(\tilde{\theta})\|_2}. \tag{8}$$

e) 模块一致性： 我们将根据以 On 结尾的键推断出的活跃模块集合之间的杰卡德相似度作为模块级一致性：

$$\text{Module} = \frac{|M_{\text{gt}} \cap M_{\text{pred}}|}{|M_{\text{gt}} \cup M_{\text{pred}}|}. \tag{9}$$

f) 归一化*L2*（归一化*L2*）： 为确保不同物理范围参数的误差具有可比性，我们还报告对每个参数叶节点进行最小-最大归一化后计算的L2误差，归一化至 $[0, 1]$ ，所用范围为插件源代码中定义的DSP物理范围（例如，以毫秒计的延迟时间、以分贝计的增益）。这可防止范围较大的参数主导总体误差。

C. Protocol-A：检索对比评估

目的。 在更严格的防泄漏划分下，将 TRR 与用于可执行参数控制的检索基线进行基准测试。

协议。 本小节中的所有结果均遵循 Protocol-A（仅检索），基于按解析音频分组的划分（N_{test} =204， $N_{\text{kb}} = 1063$），并对所有方法采用 Top-1 检索。由于当前经验证的制品集中缺少三个查询侧缓存嵌入，CLAP 行使用了 201 个查询。

表III报告了当前检索基线之间的主要比较结果。

表III PROTOCOL-A：在按解析音频分组的划分上的直接检索对比（$N_{\text{TEST}} = 204$, $N_{\text{KB}} = 1063$）。L2越低越好；其余所有指标越高越好。TRR在每一项所报告指标上均取得最佳均值表现；CLAP行使用201条共享查询以及缓存CLAP嵌入。

方法	n	L2 ↓	归一化L2 ↓	准确率@0.1 ↑	召回率 ↑	余弦 ↑	模块 ↑
TRR（本文）	204	**8.0467**	**–**	**0.5169**	**0.4595**	**0.8247**	**0.8051**
检索基线							
Wav2Vec-RAG	204	23.8197	–	0.4251	0.3505	0.5292	0.7043
Text-RAG	204	24.1705	–	0.4717	0.3967	0.4943	0.7456
FeatureNN-RAG	204	19.2613	–	0.4198	0.3600	0.5479	0.7134
CLAP	201	22.3102	–	0.4492	0.3971	0.5725	0.7240

结果。在更严格的数据划分中，TRR 在表III所有完全配对的行中均表现最佳。与 Wav2Vec-RAG 相比，TRR 将均值 L2 降低了 15.7730，自助法 95% 置信区间为[12.2807, 19.3414]，余弦相似度提高了 0.2956 [0.2305, 0.3630]；经过 Holm 校正后，五项配对比较在每项指标上仍均具有显著性（每项指标均为$p_{\text{Holm}} = 0.001$ ），且效应量较大（所有指标的 Cohen's $d > 0.8$ ）。在 201 个共享查询上，与 CLAP 相比，TRR 将均值 L2 降低了 14.1460 [10.3041, 18.0624] ，余弦相似度提高了 0.2503[0.1833, 0.3183],；同样，经过 Holm 校正后结果仍具有显著性，且效应量较大（补充材料）。

逐查询视角使这一结果更为具体。与 Wav2Vec-RAG 相比，TRR 在 204 个配对查询中的 125 个查询上改善了 L2，40 个查询上表现较差，39 个查询持平。然而，有符号改进中位数仅为 0.1238，远小于 15.7730 的均值改进；与 CLAP 相比也呈现出相同的长尾模式（中位数为 0.0658，均值为 14.1460）。因此，表III中的总体增益并不意味着整个基准测试都发生了统一的重新排序。相反，增益集中在一部分困难案例中：在这些案例里，二阶结构改变了检索邻域，避免了语义上看似合理、但参数不匹配的预设；与此同时，仍有不可忽视的一部分案例未能奏效。因此，TRR 是在该基准测试中评估过的最强检索先验，但我们不声称它在当前协议之外具有普适性。

a) 说明性案例研究：为提供具体直观的理解，考虑一个“布鲁斯独奏”查询，其真实预设启用了轻度过载（DriverOn.Distortion=0.35、DriverOn.Tone=0.6）和混响（ReverbOn.Mix=0.25）。TRR 检索到一个驱动和混响设置相近的预设（Distortion=0.38、Tone=0.55、Mix=0.28），对于这些启用模块，其逐参数归一化误差低于 0.05。相比之下，Wav2Vec-RAG 检索到一个启用 theScreamermodule 的高增益金属风格预设，导致驱动参数上的归一化误差超过 0.4。该案例说明，二阶纹理统计量能够区分风格相似但参数不同的效果配置，而一阶嵌入会将它们混为一谈。

D. 模态退化下的诊断压力测试

目的。考察在一种模态退化时，融合启发式方法能否恢复性能。

协议。为了与仓库现有诊断产物保持连续性，我们在原始的 211 条查询池上测试两种退化场景，而非更严格的 Protocol-A 划分：（1）模糊文本：将准确描述替换为通用查询“温暖吉他音色”；（2）合成音频劣化：扰动音频侧检索表示。我们将单模态系统（仅文本、仅 TRR）与融合启发式方法进行比较。这些实验有助于压力测试，但不能替代完全贴近现实的真实音频劣化基准测试。

表 IV 协议 C 边界案例精选（$N = 211$）。当前融合启发式方法主要充当回退规则：在单模态退化时，它会依赖完好的模态；但在出现明确的模态冲突时，性能会大幅下降。

场景	参考分支	融合（本文方法）	解读
模糊文本	仅文本：12.6517	**0.2970**	回退到音频
噪声音频	仅 TRR：33.8549	**1.5148**	回退到文本
冲突	仅 TRR：0.2970	4.5259	失效模式

E. TRR 消融研究

我们对 TRR 的三项核心设计选择进行消融，以了解它们对 Protocol-A 下检索质量的影响。

a) 投影维度：线性投影 $P : \mathbb{R}^{768} \to \mathbb{R}^d$ 控制格拉姆矩阵的维度，因此决定了容量—效率权衡。针对 $d \in \{32, 64, 128, 256\}$ 的消融实验d=32证实，性能在不同维度下保持稳定；经基准测试的配置（d=32）实现了最佳权衡，而更大的维度在当前包含 1063 项的知识库上没有带来收益。

b) 层选择：对六种层组合进行消融后确认，均衡的中间层集合 $\mathcal{L} = \{4, 5, 6\}$能够提供稳健的性能。单层（{4}、{5}、{6}）和两两组合的结果相近，这表明在中间层范围内，TRR 对具体的层选择并不敏感。使用更接近输入的层会捕获过多声学细节，而更高层则会丢失与纹理相关的时间结构。

c) 投影类型：对随机投影（Xavier 初始化、冻结）与 PCA 拟合投影的消融比较证实，在当前基准测试中，两种方法的结果相当。这验证了采用冻结随机投影这一实用选择的合理性：它无需训练数据，也避免了依赖知识库的拟合。

结果。在模糊文本条件下，仅文本性能急剧下降（L2=12.65），而融合启发式方法完全回退到音频（$w_{\text{audio}} = 1.0$），与仅 TRR 持平（L2=0.30）。在合成音频劣化条件下，仅 TRR 性能急剧下降（L2=33.85），而融合启发式方法回退到文本（w_{text} =1.0），与仅文本持平（L2=1.51）。在这两种设置下，当前融合规则的主要作用是切换到未受损的模态；现有证据不足以支持其能实现更稳健的跨模态整合这一更强结论。该启发式方法在明确的模态冲突下显现出局限性：融合结果升至 L2=4.53，而仅 TRR 仍为 0.30。

虽然 Protocol-C 会在合成退化条件下评估融合机制，但并不评估

表 V 试验 1（风格匹配）：条件层面统计（被试内均值）。

条件	Mean	中位数	Std
参考	86.85	92.50	16.70
清音爵士	76.19	81.00	20.80
移相	75.38	80.50	20.88
布鲁斯独奏	72.96	69.50	21.19
环境吉他	72.35	73.50	18.45
镶边	69.12	74.00	23.08
现代金属	65.58	70.00	30.22
合唱	64.92	66.50	23.12

检索到的预设相对于真实值的感知质量。为通过主观证据补充客观检索指标，我们遵循既定的多刺激协议开展了感知聆听研究。

F. 感知听音测试（带隐藏参考的多刺激测试）

协议。我们邀请 26 名参与者进行带隐藏参考的多刺激听力测试（18 名女性、8 名男性；25 名参与者自报年龄，均值为 22.2 岁，范围为 18–30 岁；自报音乐经验：12 名参与者有 $\geq$3 年器乐训练经历，8 名有非正式聆听经验，6 名未接受过正规训练），共在 10 次试验中收集了 910 条评分。每次试验均使用连续的 0–100 滑块（分数越高越好），以随机顺序呈现刺激，并将真实预设渲染结果作为隐藏参考（参考）；我们未纳入 ITU-R BS.1534 要求的明确低质量锚点，因此不将本研究称为标准 MUSHRA 测试。试验 1 评估七种带标签目标条件下的风格匹配，试验2–5 比较我们的基于 TRR 的系统与手动参数调节，试验6–10 则比较基于 TRR 的系统与波形生成基线（MusicGen）。统计分析基于参与者的原始评分，采用重复测量非参数检验（Friedman 总体检验；事后 Wilcoxon 符号秩检验并进行 Holm 校正）。仓库日志保留了参与者 ID 和评分，但未记录设备元数据、音乐专长或手动调节的来源及时间预算；因此，听力研究可作为检索基准测试的补充，但不应被视为对工作流试验公平性至关重要的评估，也不应被视为对手稿其他部分讨论的每项系统扩展的直接验证。

1) 试验 *1*：风格匹配：**数据**。208 个评分（26 名参与者 $\times$ 8 条件）。总体均值为 72.92（中位数 77.50，标准差 22.72）。

2) 试验*2–5*：吉他独奏比较：试验2–5对比了我们的基于 TRR 的系统与人工参数调优在四种吉他独奏条件下的表现。基于 TRR 的系统的受试者内均值为 71.55，而人工调优为 51.72；经 Holm 校正后，差异仍具有显著性（$p_{\mathrm{Holm}} = 1.5 \times 10^{-4}$）。不过，由于日志未记录人工基线操作人员、其专业水平或时间预算，这些试验应被视为辅助性的

表 VI 试验6–10（相似度）：条件层面统计（被试内均值）。

条件	Mean	中位数	Std
参考	89.02	91.50	10.04
基于 TRR 的系统	42.31	41.90	19.30
MusicGen	41.29	43.90	19.58

感知上下文。每次试验的完整统计数据、分布和条件明细见补充材料。

3) 试验*6–10*：与参考的相似度：**数据**。共收到 390 份评分（26 位参与者 $\times$ 3 个条件 $\times$ 5 次试验）。总体均值为 57.54（中位数 59.00，标准差 32.24）。基于 TRR 的系统与 MusicGen 的相似度得分相当：被试内均值分别为 42.31（基于 TRR 的系统）和 41.29（MusicGen），而参考条件的得分为 89.02。Friedman 总体检验仍然显著（$Q = 36.2308$，置换检验 $p = 5\times10^{-5}$），但经 Holm 校正后，TRR 与 MusicGen 的成对比较不显著（$p_{\mathrm{Holm}} = 0.7797$），因此我们将这部分结果解读为两者相当，而非前者更优的证据。

a) 解读：这些听力测试结果补充了参数空间指标，但不应将其视为对任何单一客观指标的一对一验证。相反，这些结果表明，检索锚定控制能够在数字音频工作站工作流中生成感知上合理且可编辑的起点。由于人工基线溯源信息不完整，试验2–5 最好解读为补充性的感知上下文，而非决定性的人机优化基准；试验6–10 同样表明，在这项相似度任务中，系统与 MusicGen 表现相当，而非优于后者。

VI. 讨论

A. 当前证据的范围

最有力的证据来自 Protocol-A 检索基准。听力研究提供了补充性的主观背景，而 Protocol-C 则为融合启发式方法提供了边界条件诊断。个性化和记忆组件仅得到初步伪用户模拟的支持，因此不纳入本文主要的已验证结论。

B. 对音频效果控制的启示

结果表明，基于检索的音频控制存在一种实用模式：通过检索真实预设来提供可行性约束和风格先验，然后开放检索到的配置供用户直接编辑。在所评估的方法中，TRR 实现了最强的参数对齐，同时保留了数字音频工作站工作流中的可编辑性。至于同样的优势是否适用于其他音源、效果族和制作阶段，则需要专门评估。

C. 重审纹理假说

在所报告的留出设置（表 III）中，TRR 相较于纯文本检索改善了参数空间对齐，

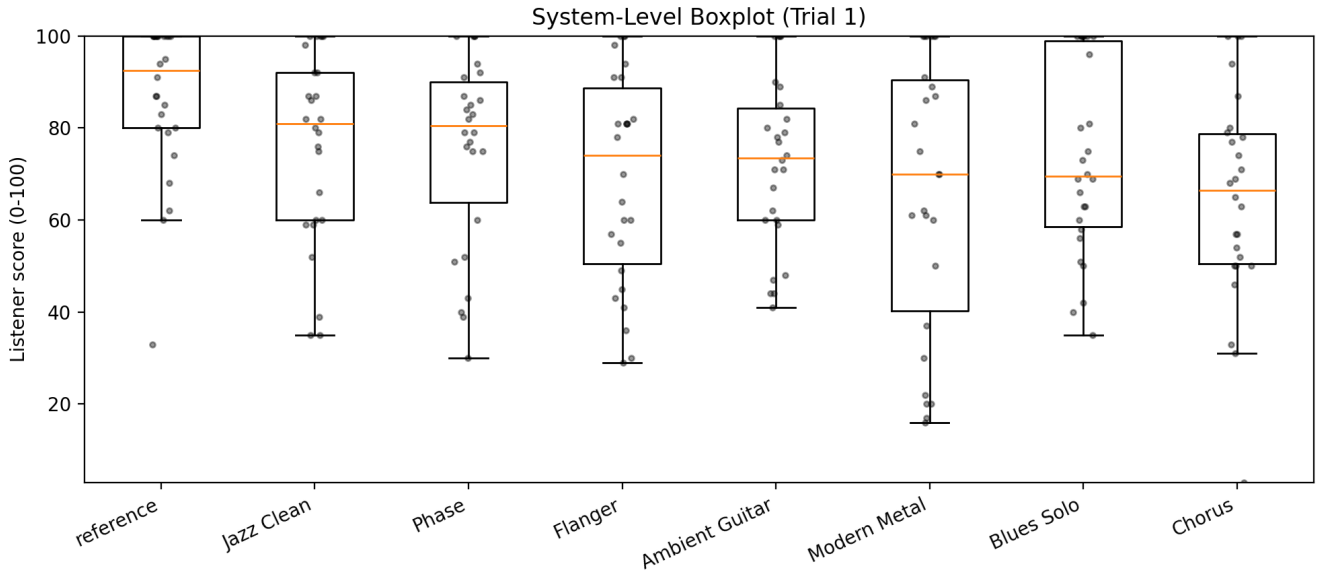


图 3. 试验 1：风格匹配的条件级分数分布（箱线图）。

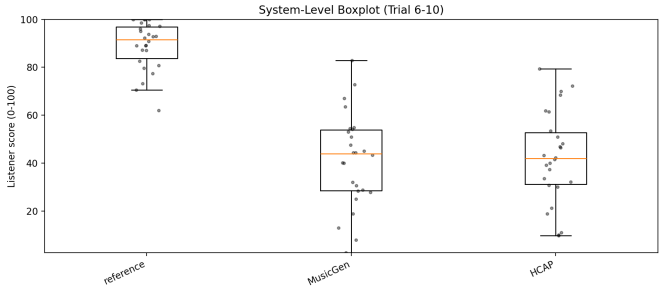


图 4. 试验6–10：条件级分数分布。

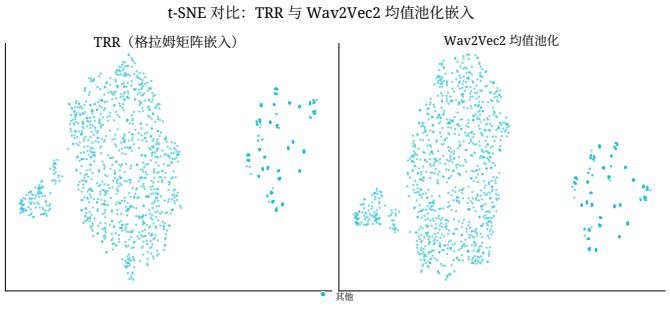


图5. 在完整数据集上对TRR（格拉姆矩阵）与Wav2Vec2均值池化嵌入进行的t-SNE比较，按风格标签着色。TRR表现出更紧密的风格内聚类，支持其作为纹理感知检索先验的效用。

这支持了以下假设：二阶相关结构可作为检索的纹理描述符。图 5 提供了视觉证据：与经过均值池化的 Wav2Vec2 嵌入相比，TRR 嵌入的 t-SNE投影按风格标签聚类得更紧密，表明基于格拉姆矩阵的表征能够捕捉与风格相关的结构全局平均所丢弃的信息。由此得出的逐查询分析使这一论断更为明确：由于许多查询在不同方法之间打成平局，二阶结构不应被理解为对基准测试进行统一的重新排序。相反，它似乎在较小的一部分情形中最为重要；在这些情形中，粗粒度的语义相似度不足，检索决策依赖于特定于纹理的共激活模式。我们并不声称TRR在参数空间中定义了全局校准的距离；最好将其理解为一种检索先验，有利于对纹理敏感的控制。

一个开放方向是，类似的二阶表示能否迁移到其他音频效果场景，包括更多乐器、效果族以及混音阶段任务。

D. 伦理考量

当前的检索驱动系统旨在增强而非取代用户的创造力。风险包括向常见示例邻域同质化，以及对建议设置的过度依赖。在未来的部署中，界面应强调透明度（例如，展示检索到的参考）和用户自主权（例如，直接可编辑性以及选择退出检索）。更一般地说，任何基于现有录音或预设构建的模型或检索语料库，都应考虑版权、许可和数据来源方面的约束。

七、局限性

我们的证据受到六项边界条件的限制，这些条件限定了我们主张的适用范围。

评估范围。该基准测试仅涵盖使用合成音频查询（$N = 204$）的吉他效果。真实音频的鲁棒性、跨乐器迁移以及更大规模的评估尚未得到验证。更广泛的结论需要通过现场录音、多样化乐器和扩展数据集进行实验来支持。

基线完整性。 虽然 CLAP 提供了一个外部基线，但尚未纳入更强的替代方案（PaSST、PANNs、直接参数估计）。我们声称“在已评估的方法中最强”，其范围有意窄于“同类最佳”。

数据集限制。 解析音频分组去除了完全重复项，但仍存在近似重复项。数据集缺少单独的文本来源溯源字段，限制了对语义泄漏的审计。

参数与感知差距。 音色具有多对一的特性：不同参数可能听起来相似。客观指标（L2、余弦相似度）可能无法与感知质量完全吻合。需要开展受控听力研究，以校准这种关系。

鲁棒性证据有限。 Protocol-C 为融合启发式方法提供边界条件诊断，而非一般性的跨模态鲁棒性证据。基于记忆的个性化仍处于初步阶段，且未经审计。

听力研究的不足。日志缺少设备元数据、参与者专业程度标签和人工基线溯源。试验2–5不应被解读为涉及公平性的人机对比。

这些限制使我们的结论仅限于较窄的发现：在此基准测试中，TRR 可为纹理主导的效果提供有用的检索先验；要支持普适指标、大规模部署和广泛的个性化应用，还需要更多证据。

VIII. 结论

本文探讨了纹理感知的音频表征能否改进检索驱动的音频效果控制这一问题。我们提出了一种检索驱动的可编辑音频效果控制方法，并在我们经审核的留出基准测试的已报告比较中，将纹理共振检索（TRR）确定为最强的检索表征。在更严格的已解析音频分组划分（表III）的客观指标上，TRR总体上优于一阶检索基线（包括可用的CLAP比较子集）改善了参数对齐；我们将这一优势解释为更有用的检索先验的证据，而非通用距离度量或逆参数估计闭式解的证据。

综合来看，经审核的客观基准测试为一项范围较窄但具有实际意义的结论提供了主要支持：对可执行预设进行检索可为数字音频工作站工作流提供可编辑的起点，而二阶纹理结构在粗粒度语义相似度不足的一部分困难案例中改进了该检索。听力研究提供了补充性的主观背景，而Protocol-C仅阐明了当前融合启发式方法的回退行为与失效模式。

展望未来，最重要的后续工作包括评估更强的现代基线，在真实音频劣化条件下验证鲁棒性，并测试相同的检索优势能否从吉他效果扩展到更广泛的制作参数空间，例如混音和母带处理工作流。

总体而言，现有证据表明，具备纹理感知能力的检索可作为一种由基准测试支持的、颇具前景的基础，用于实现可编辑的音频效果控制，同时，将关于普适性、稳健性和个性化的更广泛论断留待未来工作进一步探讨。

参考文献

[1] H. Liu、Z. Chen、Y. Yuan、X. Mei、X. Liu、D. Mandic、W. Wang 和 M. D. Plumbley，“AudioLDM：基于潜在扩散模型的文本到音频生成”，载于第*40*届国际机器学习会议，2023年。[2]A. Agostinelli、T. I. Denk、Z. Borsos、J. Engel、M. Verzetti、A. Caillon、Q. Huang、A. Jansen、A. Roberts、M. Tagliasacchi 等，“MusicLM：从文本生成音乐”，*arXiv* 预印本 *arXiv:2301.11325*，2023年。[3] J. Engel、L. Hantrakul、C. Gu 和 A. Roberts，“DDSP：可微数字信号处理”，载于国际学习表征会议，2020年。[4] L. A. Gatys、A. S. Ecker 和 M. Bethge，“一种艺术风格的神经算法”，*arXiv* 预印本 *arXiv:1508.06576*，2015年。[5]R. Stables、S. Enderby、B. De Man、G. Fazekas 和 J. D. Reiss，“SAFE：语义音频描述符提取与检索系统”，载于第*15*届国际音乐信息检索学会会议论文集，最新成果与演示环节，2014年。 [在线]。网址：https://ismir2014.ismir.net/ LBD/LBD15.pdf[6] M. Schedl、E. G´omez 和 J. Urbano，“音乐信息检索：最新进展与应用”，信息检索基础与趋势，第8卷，第2–3期，127–261页，2014年。[7] C.-Z. A. Huang、D. Duvenaud 和 K. Z. Gajos，“使用高斯过程主动学习合成器的直观控制旋钮”，载于*2014*年国际智能用户界面会议论文集（*IUI ’14*），2014年，第115–124页。[8]S. Brade、A. McLeod、M. Blanco、N. J. Bryan 和 G. Fazekas，“SynthScribe：用于合成器声音检索与探索的深度多模态工具”，载于第*29*届国际智能用户界面会议论文集（*IUI ’24*），2024年。[9]M. A. Mart´ınez Ram´ırez、D. Stoller、S. I. Mimilakis 和 J. D. Reiss，“自动混音的深度学习”，ISMIR2022教程，2022年。 [在线]。网址：https://dl4am.github.io/tutorial/landing-page.html [10] C. J. Steinmetz、J. Pons、S. Pascual 和 J. Serr`a，“借助神经音频效果的可微混音控制台实现自动多轨混音”，载于*2021*年*IEEE*国际声学、语音与信号处理会议（*ICASSP*），2021年，第316–320页。[11]E. Moliner 等，“使用效果嵌入生成模型实现自动音乐混音”，*arXiv* 预印本 *arXiv:2511.08040*，2025年。 [在线]。网址：https://arxiv.org/abs/2511.08040[12] C. Peladeau 等，“采用自编码器方法和可微数字信号处理进行音频效果盲估计”，载于*2024*年*IEEE*国际声学、语音与信号处理会议（*ICASSP*），2024年。

[13]X. Liu 等，“DDSP-SFX：利用可微数字信号处理进行声学引导的音效生成”，载于第*27*届国际数字音频效果会议（*DAFx*），2024年。[在线]。网址：https://dafx.de/paper-archive/2024/ papers/DAFx24 paper 51.pdf[14]A. Dees 和 S. O’Leary，“用于音频纹理神经合成的生成式潜在空间”，载于第*28*届国际数字音频效果会议（*DAFx*），2025年。[在线]。网址：https://dafx25.dii.univpm.it/wp-content/uploads/2025/09/DAFx25 paper 56.pdf[15] C. J. Steinmetz等，“ST-ITO：利用推理时优化控制风格迁移中的音频效果”，*arXiv* 预印本 *arXiv:2410.21233*，2024年。 [在线]。网址：https://arxiv.org/abs/2410.21233[16] H. Liu、Y. Yuan、X. Liu、X. Mei、Q. Kong、Q. Tian、Y. Wang、M. D. Plumbley 和 W. Wang，“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，2024年。 [在线]。网址：https://doi.org/10.1109/TASLP.2024.3399607[17] A. Vyas、B. Shi、M. Le、A. Tjandra、Y.-C. Wu、B. Shi、M. Guo、W.-N. Hsu、A. Conneau 和 M. Auli，“Audiobox：利用自然语言提示实现统一音频生成”，*arXiv* 预印本 *arXiv:2312.15821*，2023年。[18]R. Huang、J. Huang、D. Yang、Y. Ren、L. Liu、M. Li、Z. Ye、J. Liu、X. Yin 和 Z. Zhao，“Make-an-audio：利用提示增强型扩散模型实现文本到音频生成”，载于第*40*届国际机器学习会议，2023年，第13963–13982页。 [在线]。网址：https: // proceedings.mlr.press/v202/huang23i.html[19] H. Wang、R. Ye、J. Huang、J. Liu、Z. Zhao 和 Y. Ren，“AudioLCM：使用潜在一致性模型进行文本到音频生成”，*arXiv* 预印本 *arXiv:2406.00356*，2024年。[在线]。网址：https://arxiv.org/abs/2406.00356[20]D. Ghosal、N. Majumder、A. Mehrish 和 S. Poria，“使用指令微调大语言模型和潜在扩散模型进行文本到音频生成”，*arXiv* 预印本 *arXiv:2304.13731*，2023 年。 [在线]。网址：https://arxiv.org/abs/2304.13731[21]A. Baevski、H. Zhou、A. Mohamed 和 M. Auli，“wav2vec 2.0：用于语音表征自监督学习的框架”，载于神经信息处理系统进展，第33卷，2020年，第12 449–12 460页。[22] B. Elizalde、S. Deshmukh、M. Al Ismail 和 H. Wang，“CLAP：通过自然语言监督学习音频概念”，载于*IEEE*国际声学、语音与信号处理会议（*ICASSP*），2023年，第1–5页。[23] K. Koutini、J. Schl¨uter、H. Eghbal-zadeh 和 G. Widmer，“利用Patchout高效训练音频变换器”，载于*Interspeech 2022*论文集，2022年，第2753–2757页。[24] Q. Kong、Y. Cao、T. Iqbal、Y. Wang、W. Wang 和 M. D. Plumbley，“PANNs：大规模预训练音频神经网络——用于音频模式识别的神经网络”，《IEEE/ACM 音频、语音与语言处理汇刊》，第 28 卷，第 2880–2894 页，2020年。[25] P. Lewis、E. Perez、A. Piktus、F. Petroni、V. Karpukhin、N. Goyal、H. K¨uttler、M. Lewis、W.-t. Yih、T. Rockt¨aschel、S. Riedel 和 D. Kiela，《面向知识密集型自然语言处理任务的检索增强生成》，载于《神经信息处理系统进展》，第 33 卷，2020年，第 9459–9474 页。[26]S. Ghosh 等，《Reclap：通过描述声音提升零样本音频分类》，arXiv 预印本 arXiv:2409.09213，2024年。 [在线]。获取地址：https: // arxiv.org/abs/2409.09213[27]D. Niizumi 等，《M2D-CLAP：掩码建模双人组与 CLAP 结合，学习通用音频-语言表征》，载于 Interspeech 2024 论文集，2024年。 [在线]。获取地址：https://arxiv.org/abs/2406. 02032[28] A. Hanif 等，《Palm：面向音频语言模型的少样本提示学习》，载于 2024年自然语言处理实证方法会议论文集，2024年。 [在线]。获取地址：https: // aclanthology.org/2024.emnlp-main.1030/[29] Y. Chu、J. Xu、X. Zhou、Q. Yang、S. Zhang、Z. Yan、C. Zhou 和 J.Zhou，《通义千问音频：通过统一的大规模音频语言模型推进通用音频理解》，arXiv 预印本 arXiv:2311.07919，2023年。 [在线]。获取地址：https://arxiv.org/abs/2311. 07919[30]Z. Tang、Z. Tian、T. Shi、Z. Ouyang、C. Wu、Z. Zhu、J. Wang、Y. Shen、Y. Wang 和 N. Jaitly，《SALMONN：迈向大语言模型的通用听觉能力》，arXiv 预印本 arXiv:2310.13289，2023年。 [在线]。获取地址：https://arxiv.org/abs/2310. 13289[31] J. Bruna 和 S. Mallat，《基于散射矩的音频纹理合成》，arXiv 预印本 arXiv:1311.0407，2013年。 [在线]。获取地址：https://arxiv.org/abs/1311. 0407[32] J. H. McDermott 和 E. P. Simoncelli，《通过听觉外周统计特征感知声音纹理：来自声音合成的证据》，《神经元》，第 71 卷，第 5 期，第 926–940 页，2011年。[33] A. Natsiou 和 S. O’Leary，《深度学习在声音合成中的音频表征：综述》，arXiv 预印本 arXiv:2201.02490，2022年。 [在线]。获取地址：https://arxiv.org/abs/2201.02490[34]B. Weck 等，《Wikimute：音乐音频语义描述的网络来源数据集》，arXiv 预印本 arXiv:2312.09207，2023年。 [在线]。获取地址：https://arxiv.org/abs/2312.09207[35] A. Srivatsan 等，《通过多模态前缀进行检索引导的音乐描述生成》，载于第三十三届国际人工智能联合会议论文集，2024年。 [在线]。获取地址：https://www. ijcai.org/proceedings/2024/859[36]M. Gonzalez 等，《一种用于文本生成音乐的检索增强方法》，载于第三届

音乐与音频自然语言处理研讨会（*NLP4MusA*），2024年。 [在线]。获取地址：https://aclanthology.org/2024. nlp4musa-1.6/[37]S. Doh 等，《大语言模型能否根据自然语言预测音频效果参数？》*arXiv* 预印本 *arXiv:2505.20770*，2025年。 [在线]。获取地址：https://arxiv.org/abs/2505.20770[38] Y. Liu 等，《HM-RAG：分层多智能体多模态检索增强生成》，*arXiv* 预印本 *arXiv:2504.12330*，2025年。 [在线]。获取地址：https://arxiv.org/abs/2504.12330