# ITO-MASTER：推理时间优化用于音乐母带处理程序的音频效果建模

**鞠贤旭** [1] **马尔科·A·马丁内斯-拉米雷斯** [1] **廖伟翔** [1]
**乔尔乔·法布罗** [2] **米凯莱·曼库西** [2] **水藤由希** [1,3]
[1] Sony AI, Japan [2] Sony Europe B.V., Germany [3] Sony Group Corporation, Japan
{firstname.lastname}@sony.com



## ABSTRACT

音乐母带处理风格转换旨在建模并应用参考曲目的母带处理特征到目标曲目，模拟专业的母带处理过程。然而，现有的方法基于参考曲目进行固定处理，限制了用户调整结果以匹配其艺术意图的能力。本文介绍了 ITO-Master 框架，这是一个结合推理时优化（ITO）的基于参考的母带风格转换系统，使用户能够更精细地控制母带处理过程。通过在推理过程中优化参考嵌入 $z_{\text{ref}}$，我们的方法允许用户动态调整输出，进行微调以实现更精确的母带处理结果。我们探讨了建模母带处理器的黑盒和白盒方法，并证明 ITO 可以提高不同风格下的母带性能。通过客观评估、主观听觉测试以及使用基于文本条件的 CLAP 嵌入进行定性分析，我们验证了 ITO 增强了母带风格相似度并提供了更高的适应性。我们的框架为母带风格转换提供了一种有效且用户可控的解决方案，使用户能够超越初始风格转换来调整其结果。

## 1. 介绍

音乐母带处理是音频制作过程中的最后一步，确保专业音质并在各种音乐分发平台上保持一致的播放效果。这个过程涉及应用一系列音频效果，如均衡、压缩、立体声成像和限制器，这些共同塑造了声音特性并提升了整体音频质量 [1,2]。传统上，母带处理需要技术娴熟的工程师根据曲目的内容和预期的艺术效果仔细调整这些效果。然而，随着音乐制作量的增加以及对流媒体平台上一致性需求的增长，自动化母带处理解决方案的需求大幅上升。

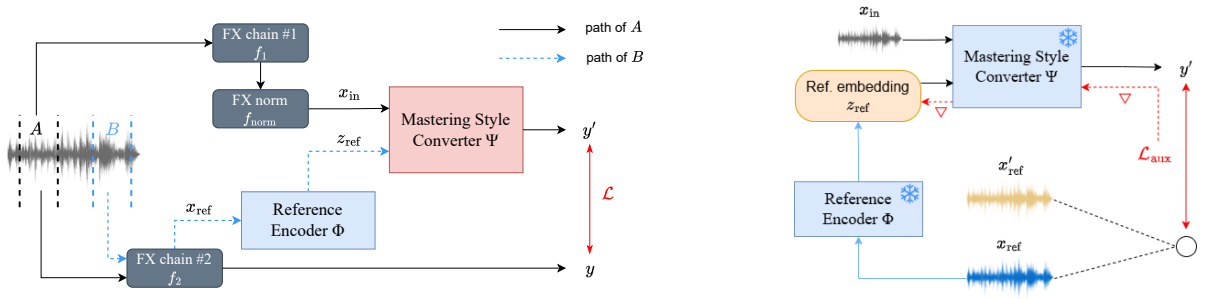


为满足这一需求，各种自动母带处理系统应运而生 [3–5]. 然而，这些系统以无条件的方式运行，在没有直接用户控制的情况下应用音频效果。为了引入适应性，已经探索了基于参考的方法，其中参考曲目的处理特性被应用于另一首曲目 [6,7]. 这些方法旨在匹配音频特征如动态范围、音调平衡和立体声宽度，提供了一种替代全自动母带处理的方案。然而，在实现高质量结果和可控性方面仍存在重大挑战。

现有的基于参考的方法可以广泛分为黑箱和白盒模型。黑盒模型通常基于端到端的神经网络 [6]，能够有效地捕捉高级音频模式，但缺乏透明度和可解释性，使得用户难以修改处理过程中的特定方面。相比之下，白盒模型利用特征匹配算法 [7] 或可微音频处理器 [5] 来提供对各个参数的更大控制权。虽然白盒方法提供了结构化和可解释的方法 [8]，但它们通常受到其可微处理器简单性的限制，这可能无法完全复制专业母带处理中使用的复杂工具。

在本文中，我们介绍了 ITO-Master 框架，这是一种基于参考的音乐母带处理音频效果建模方法，它集成了推理时间优化（ITO），以实现更精细的用户控制。与之前的应用固定处理的风格转换方法不同，我们的方法允许用户在初始结果未能完全符合其偏好时动态调整输出。通过在推理过程中优化参考嵌入 $z_{\text{ref}}$，ITO-Master 能够进行微调级别的调整，从而实现更加精确和有针对性的母带精修。

本工作的关键贡献包括：(1)**ITO-主框架**：一种使用 ITO 的基于参考的音频效果建模的新方法，适用于母带处理程序。(2) **黑盒模型与白盒模型的比较**：对两种评估其有效性和权衡方案的范式进行了系统性研究。(3) **真实的掌握处理器链**：实现了一个结构化的可微分母带处理流水线，以增强白盒处理的真实感。(4) **综合评价**：通过客观指标、听觉测试以及使用 CLAP 嵌入的文本条件下的定性分析进行性能验证。




(a) Mastering Style Converter 的训练流程 Ψ。在此阶段，参考编码器 Φ 使用通过随机 FX 操作生成的多种母带风格进行训练 $f$。目标信号 $y$ 是通过对片段 $A$ 应用与参考片段 $B$ 相同的操作合成的，这两个片段都来自同一首歌曲。

(b) 迭代优化是使用一个辅助（内容无关的）目标函数进行的，允许使用任何参考音乐。用户可以根据他们对参考音乐和目标函数的偏好来优化 $z_{\text{ref}}$。

**Figure 1**: ITO-Master 的整体流程。

## 2. 相关工作

### 2.1 音频效果风格转移

音频效果风格转换已经成为自动和增强音乐制作研究的重要领域。深度学习的最新进展带来了更为复杂的方法，其中神经网络被用于学习输入和输出音频信号之间的复杂映射。这些方法已被应用于单个音频效果（Fx）或多个 Fx 集合（Fx 链）[5,6,9–14] 的风格转换，有效地建模了时间依赖性，并基于参考轨道在波形级别上应用风格转换。虽然这些方法在受控环境中表现出色，但在将其应用扩展到多样化的现实世界场景中仍面临挑战，特别是在适应不同的母带制作风格和变化的输入信号方面。

### 2.2 推理时间优化

最近，[15,16] 在音乐生成任务中探索了推理时优化（ITO），其中初始潜在嵌入通过基于扩散的模型反向传播，并使用生成样本与参考曲目之间的损失进行优化。在音频效果风格迁移的背景下，ITO 已应用于诸如 ST-ITO [17,18] 等方法中，在这些方法中，可解释参数在一个白盒可微分 Fx 链中被优化。

我们的工作特别集中在掌握风格转换，其中处理高度压缩的音频——这在商业发布的音乐中很常见——需要特别注意限制器和动态范围管理。ITO-Master 框架引入了参考嵌入 $z_{\text{ref}}$ 上的 ITO，允许在黑盒模型和白盒模型中进行推理时优化。通过微调 $z_{\text{ref}}$，我们的方法能够适应参考曲目的母带风格而不需重新训练整个模型。ITO-Master 确保平稳地适应参考曲目的特征，同时保持观察和调整白盒模型下层参数的能力。这使得我们的框架特别适合母带制作任务，为专业人士和业余爱好者提供对音频母带处理过程的精确控制。

## 3. 方法论

本节描述了所提出的掌握风格转换框架的组成部分：训练流程、掌握风格转换器、可微掌握链和 ITO。训练流程通过应用随机 Fx 操作来模拟现实世界的掌握场景。掌握风格转换器 Ψ 将参考曲目的掌握风格转移到目标曲目上，可以使用黑盒方法和白盒方法进行实现。可微掌握链作为白盒处理器，以结构化的顺序建模各种 Fx。最后，ITO 过程在推理时优化 $z_{\text{ref}}$ 以提升风格转换性能。以下子节详细描述每个组件。

### 3.1 掌握风格转换的训练流程

如图 1(a) 所示，训练流程遵循风格迁移中的既定方法，利用带有随机 Fx 操作的自监督训练框架 [6,11–13]。基于单首歌曲在整首歌中保持一致母带处理风格的理解 [2]，一首歌首先被分割成两部分，$A$ 和 $B$。对于输入到 Ψ，应用随机操作 $f_1$ 来模拟随机风格，类似于在应用程序设置中的功能过程。然后，我们应用 Fx-规范化 [19] $f_{\text{norm}}$，该规范将某些 Fx 特性标准化为固定的目标水平，以促进风格转换的性能。$f_{\text{norm}}$ 仅应用于均衡器（EQ）、立体声成像器和响度级别，使模型能够捕捉到更广泛的非线性 FX 变换，例如压缩级别和失真。虽然从技术上讲可以规范化压缩，但这并不适合实时训练过程。对于失真，规范化将需要从给定的歌曲中移除所有失真或在整个轨道上应用一致的失真水平，这超出了本文的范围。总结来说，输入到 Ψ 定义为 $x_{\text{in}} = f_1(f_{\text{norm}}(A))$。

为了实现风格转换，参考轨道 $x_{\text{ref}}$ 的 Fx 信息被编码以创建条件为 Ψ 的参考嵌入。对片段 $B$ 进行第二次随机操作 $f_2$，然后由参考编码器 Φ 编码，产生参考嵌入 $z_{\text{ref}} = \Phi(f_2(B))$。训练过程最小化模型输出 $y' = \Psi(x_{\text{in}}, z_{\text{ref}})$ 和目标信号 $y = f_2(A)$ 之间的损失。由于 $A$ 和 $B$ 来源于同一首母带处理过的歌曲，我们假设

$y$ 和 $x_{\text{ref}}$ 具有相同的母带处理风格。

### 3.2 掌握风格转换器

$\Psi$ 可以使用两种不同的建模方法来实现：黑盒建模和白盒建模。在黑盒方法中，$\Psi$ 直接对波形信号 $y'$ 进行建模。相反，白盒方法估计可微母带处理 Fx 链的不同参数 $\Theta$。在白盒模型中的可微母带处理 Fx 链的公式与用于处理随机母带音频的 Fx 操纵器 $f$ 中使用的相同。

训练目标对于 $\Psi$ 是多尺度频谱损失 $\mathcal{L}_{\text{MSS}}$，应用于左右和中侧通道（其中中间=左+右，侧=左-右），如在 [12] 中所使用的。

### 3.3 可微掌握链建模

掌握链被设计为完全可微分的白盒处理器，实现双重功能：既可以作为掌握风格转换器，也可以作为训练转换器的随机掌握操作模块。通过建模广泛的掌握风格变化性，该链使系统能够稳健地处理和复制现实世界音乐掌握中的复杂情况。为了模拟一个真实的音乐掌握过程 [3]，该链包括六个不同的 Fx 模块：1. 六段参量均衡器，2. 失真，3. 三段压缩器，4. 增益补足，5. 立体成像器和 6。限制器。这些模块的顺序是固定的，在训练过程中对每个 Fx 模块进行随机操作的概率分别设定为 90%，30%，80%，85%，60% 和 100%。这些概率被采用以引入更大的变异性，同时防止合成过于不现实的演奏风格，从而增强 $\Psi$ 的建模能力。该链总共包含 46 个可控参数。

为确保可微分性，Fx 模块使用支持基于梯度优化技术的开源库 [1,2] 来实现。对于 3 频带多频带压缩建模，首先应用四阶 Linkwitz-Riley 交叉滤波器 [20] 将信号分割成三个频段，随后是可微分全极点滤波器。链中的一个关键组件是使用这些由 [21] 建模的可微分全极点滤波器，这使得在多频带压缩器和限制器中都能计算压缩和扩展效果。这种能力对于实际应用至关重要，使系统能够管理限幅器和定界器，在大多数商业发布的音乐都经过大量压缩 [2] 的现实世界场景下尤为重要，需要在这种条件下进行有效的风格转换。

### 3.4 推理时间参考嵌入优化

本工作的主要贡献是引入了在 $z_{\text{ref}}$ 上的 ITO。不需要对整个模型 $\Psi$ 进行微调，而是专注于仅优化 $z_{\text{ref}}$ 而保持预训练的 $\Psi$ 固定，如图 1(b) 所示。尽管在黑盒模型的推理时间内优化 $z_{\text{ref}}$ 并不能从 Fx 处理器的角度提供可解释性，但白盒模型保留了这种可解释性。事实上，在 ITO 过程前后可以看到参数 $\Theta$ 的变化情况，这提供了如何调整 $\Theta$ 的洞察。此外，用户可以使用另一种参考信号 $x'_{\text{ref}}$ 来优化系统，提供了一个不同于传统母带处理风格转移的方法，因为该方法是基于新的参考信号和优化目标函数的结合来融合母带处理风格的。ITO 在 $z_{\text{ref}}$ 上的优势在于显著减少了与从头开始优化整个可微链的 $\Theta$ 相比的优化步骤数量，如我们在第 5 节中比较所见。

对于 ITO，使用了 [13] 提出的音频特征（AF）损失作为辅助目标函数 $\mathcal{L}_{\text{aux}}$。AF 损失是一种内容无关的损失，结合了各种音频特征变换，捕捉到了音频的动力学、空间化和频谱特性。AF 损失中的每个变换都有其预定义的权重因子，并将这些加权变换相加以计算总体损失。在我们的实验中，遵循参考论文中原有的权重。我们使用梯度下降迭代优化 $z_{\text{ref}}$：$z_{\text{ref}}^{(t+1)} = z_{\text{ref}}^{(t)} - \eta \nabla_z \mathcal{L}_{\text{aux}}(\Psi(x_{\text{in}}, z_{\text{ref}}^{(t)}), x_{\text{ref}})$，其中 $\eta$ 是学习率。对于客观和主观评估，我们仅专注于将 AF 损失作为 ITO 的优化目标。由于 ITO 可以使用任何损失函数进行优化，我们在第 5.3 节中还定性地探讨了使用带有 CLAP 嵌入 [22] 的文本提示进行优化。

## 4. 实验

### 4.1 数据集

我们使用了 MoisesDB 数据集 [23] 进行训练，并用 MUSDB18 验证子集 [24] 进行了验证。来自这些数据集的混合样本被采用，因为它们尚未完全掌握，允许通过 $f$ 随机操控 Fx 来创建合成已掌握样本。对于 Fx-归一化，在 MoisesDB 数据集上预先计算了均值统计，并应用归一化以匹配 EQ、立体声成像器和响度级别。为了评估，从 MTG-Jamendo 数据集 [25] 中随机选择了 200 首歌曲。其中，100 首歌曲用作 $x_{\text{in}}$，剩余的 100 首作为 $x_{\text{ref}}$ 输入到 $\Psi$。在训练和验证过程中，两个片段 $A$ 和 $B$ 都是 11.8 秒长。对于评估，使用 30 秒样本，因为 $\Psi$ 的全卷积架构可以处理变长输入。

### 4.2 实验设置

实验设置包括两种主要的训练配置用于 $\Psi$：黑盒方法和白盒方法。这两种配置都使用了时间卷积网络（TCN）[26] 架构来处理 $x_{\text{in}}$，拥有 1050 万可训练参数。预训练权重的 *FX 编码器* [12] 被采用为 $\Phi$ 并在两种条件下进行测试：有和没有与 $\Phi$ 一起训练 $\Psi$。所有模型均以批次大小为 4 的情况下进行了 72,000 次迭代的训练。

[1] https://github.com/csteinmetz1/dasp-pytorch

[2] https://github.com/DiffAPF/torchcomp

除了训练 $\Phi$ 和 $\Psi$ 模型外，还对每个测试数据执行 ITO 以优化维度为 2048 的参考嵌入 $z_{\text{ref}}$。ITO 过程最多运行 100 步，或者如果损失值增加则提前停止，这表明收敛。为了进行比较，还应用了一种替代的 ITO 方法，在该方法中仅对可微掌握链 $f$ 的参数 $\Theta$ 进行优化。在这种情况下，$\Theta$ 最多优化 2K 步以评估其相对于专注于 $z_{\text{ref}}$ 的提议 ITO 方法的有效性。所有方法均使用学习率为 $2 \cdot 10^{-4}$ 的 RAdam 优化器 [27] 进行优化。更多详情请参见我们的开源仓库 [3]。

### 4.3 评估指标

为了客观评估风格迁移的掌握情况，采用了内容独立的目标，因为正在比较不同的内容。使用了以下指标：

- **音频特征（AF）损失**：如第 3.4 节所述，AF 损失衡量输出 $y'$ 与所需音频特征的匹配程度。
- **动态范围变异性（DRV）**：DRV 是评估音频压缩程度的关键指标，特别是在音乐母带处理中，限制器发挥着重要作用。DRV 指标通过首先使用高频内容起始检测函数识别音频信号中的峰值来计算。DRV 是在过滤掉最低的 25%值后这些峰值的标准差，定义为：

$$\text{DRV} = \frac{1}{C}\sum_{c=1}^{C} \text{std}\left(\{p_i^c : p_i^c > \text{percentile}(p^c, 75)\}\right) \quad (1)$$

  其中 $p^c$ 表示通道 $c$ 中的峰值集 $\{p_1^c, p_2^c, ...\}$，而 $C$ 是总通道数。该指标反映了动态范围的变化性，数值越高表示压缩一致性越低。
- **Fx 嵌入相似性（余弦相似度）**：余弦相似性度量参考嵌入 $\Phi(x_{\text{ref}})$ 和输出嵌入 $\Phi(y')$ 之间的相似性。我们采用预训练的 FXencoder 作为 $\Phi$。该指标评估输出在学习到的 Fx 特征方面与参考的匹配程度。
- **弗雷歇音频距离（FAD）**：FAD [28] 通过比较模型输出的统计分布与参考分布来评估生成音频的感知质量。FAD 使用三种深度音频嵌入进行计算：CLAP [29]，DAC [30] 和 EnCodec [31]。我们选择了这些嵌入，因为它们在声学质量方面与人类偏好表现出很强的相关性，尤其是像 DAC 和 EnCodec 这样的基于编解码器的模型对声学效果特别敏感，如 [32] 所示。该指标计算从模型输出中提取的特征分布与从 Jamendo 数据集子集中提取的特征之间的距离，衡量风格转换后的音频听起来有多自然，相对于真实录音而言。

[3] https://github.com/SonyResearch/ITO-Master

### 4.4 基线方法

以下基线方法代表现有的掌握风格转换系统，用于进行比较：

- *声学特征匹配方法*：**特征匹配**方法旨在通过直接对齐特定音频特征，将输入轨道的 Fx 调整到与参考轨道相匹配。
  - **Fx-标准化** [19]: 不是对给定的音频进行标准化以匹配目标数据分布的平均统计量，而是直接将 Fx 级别与参考歌曲的级别相匹配。官方实现 [4] 用于依次匹配音频效果的顺序，分别是均衡器、压缩、立体声成像和响度。
  - **匹配处理** [7]：一个开源库，将给定歌曲的均方根值、频率响应、峰值幅度和立体声宽度与参考曲目相匹配。官方实现 [5] 用于推断处理过的歌曲。
- **端到端重构** [6]：这种端到端的重新制作方法是一个黑盒模型，直接在波形级别预测信号 $y'$。该模型使用大量发布的流行歌曲数据集以自监督方式训练。它利用预训练的编码器和投影判别器来鼓励生成准确反映参考曲目母带风格的真实音频。

## 5. 结果

### 5.1 目标评估

所提方法及基线方法的性能总结如表 1 所示。特征匹配方法，特别是 Fx-标准化和 Matchering，在 AF 和 FAD 指标上表现出色。这是意料之中的，因为这些方法直接应用了与 Fx 相关的变换以匹配参考轨道的特点。然而，这些方法在 DRV 指标上的表现不佳，因为他们缺乏对动态范围调整所需控制的把控，这在涉及限定的真实世界母带处理任务中是至关重要的。端到端重制 [6] 在使用 CLAP 和 EnCodec 嵌入时，在 FAD 上表现出色，可能是由于其在训练过程中采用了对抗目标，有助于生成接近参考分布的逼真音频。然而，该系统在 AF 和 DRV 上的表现不足，表明在捕捉精确音频特征变换和管理动态范围方面存在挑战。

在所提出的方法中，黑盒方法在 AF 方面优于白盒方法，表明其通过直接建模 $y'$ 与 $\mathcal{L}_{\text{MSS}}$ 能够更有效地捕捉音频特征转换。然而，白盒方法在所有 FAD 指标上表现出更好的结果，这表明它生成的音频更加符合现实世界的分布。这可能意味着虽然黑盒模型捕获了更多细节的转换，但白盒方法产生的输出在感知上与现实世界中的制作风格更为一致。

[4] https://github.com/sony/FxNorm-automix

[5] https://github.com/sergree/matchering

| | 方法 | AF (↓) | 驱动程序 (↓) | 余弦相似度 (↑) | 脂肪酸脱氢酶 CLAP(↓) | $FAD_{DAC}$(↓) | $FAD_{EnCodec}$(↓) |
|---|---|---|---|---|---|---|---|
| 特征匹配 | Fx-Normalization [19] | 0.157 | 0.801 | 0.941 | 161.4 | 177.4 | 84.53 |
| | Matchering [7] | 0.160 | 0.823 | 0.942 | 110.8 | 126.1 | 59.34 |
| 基线 | E2E Remastering [6] | 0.288 | 0.858 | 0.942 | 104.3 | 176.7 | 37.19 |
| 提议 | Black-box | 0.346 | 0.685 | 0.944 | 160.8 | 378.4 | 51.12 |
| | + train $\Phi$ | 0.125 | 0.577 | 0.945 | 159.8 | 177.4 | 46.94 |
| | + ITO on $z_{\text{ref}}$ | **0.099** | 0.567 | **0.946** | 182.2 | 180.5 | 42.82 |
| | White-box | 0.253 | 0.598 | **0.946** | 93.7 | 144.8 | **36.22** |
| | + train $\Phi$ | 0.186 | 0.521 | 0.945 | **93.2** | **101.4** | 38.90 |
| | + ITO on $z_{\text{ref}}$ | 0.139 | **0.474** | **0.946** | 105.2 | 109.1 | 42.99 |
| | ITO on $\Theta$ | 0.250 | 0.609 | 0.927 | 216.8 | 294.8 | 101.60 |

**Table 1**: 掌握 Jamendo 数据集上的风格转换（现实世界场景）。

当训练 $\Psi$ 而保持预训练的 FXencoder$\Phi$ 固定时，性能通常较差。这可能是因为 FXencoder 是在一组不同的 Fx 链上训练的，并且可能无法完全捕捉到在这种情况下应用于参考歌曲的操作。然而，当同时训练 $\Phi$ 和 $\Psi$ 时，性能有了显著提高，因为这种联合训练使得编码器能够更好地适应特定的 Fx 操作，从而实现更准确的母带风格转换。

将 ITO 应用于 $z_{\text{ref}}$ 可以增强 AF 性能，但会引入 FAD 得分的权衡，这表明虽然 ITO 能够优化风格转换，但仍需仔细校准优化步骤的数量以平衡竞争目标。相反，在 $\Theta$ 上直接应用 ITO 在所有指标上的结果都很差，即使增加优化步骤的数量也是如此。有趣的是，在逆向工程任务中——即输入和输出内容相同的情况下——优化 $\Theta$ 效果很好，尽管 Fx 链的复杂性为 [33–35]。然而，在风格转换过程中，使用了与内容无关的损失函数仅复制参考轨道的风格。这种区别解释了为什么在该情况下 ITO 对 $\Theta$ 的效果较差。由于 $\Psi$ 是通过依赖内容的目标进行训练的，它利用内容信息来增强风格转换效果。相比之下，在 ITO 中使用的 $\mathcal{L}_{\text{aux}}$ 无法捕捉任务的复杂性，使其不适合在这种情境下优化整个可微分链。

音频样本可在我们的演示页面 [6] 上获取。

### 5.2 主观评价

为了进一步主观验证我们提出的方法，我们进行了一个包含 10 名参与者的 MUSHRA 类型听音测试。

[6] https://tinyurl.com/ITO-Master

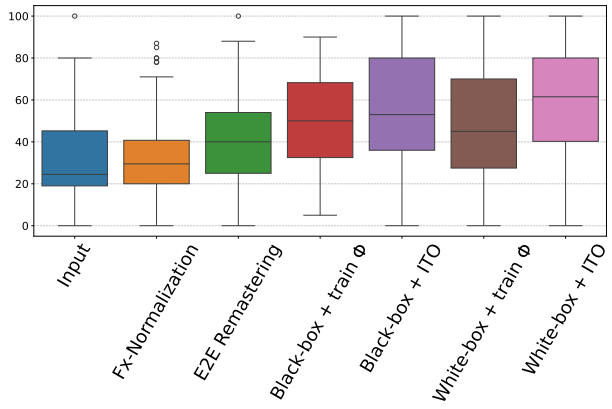
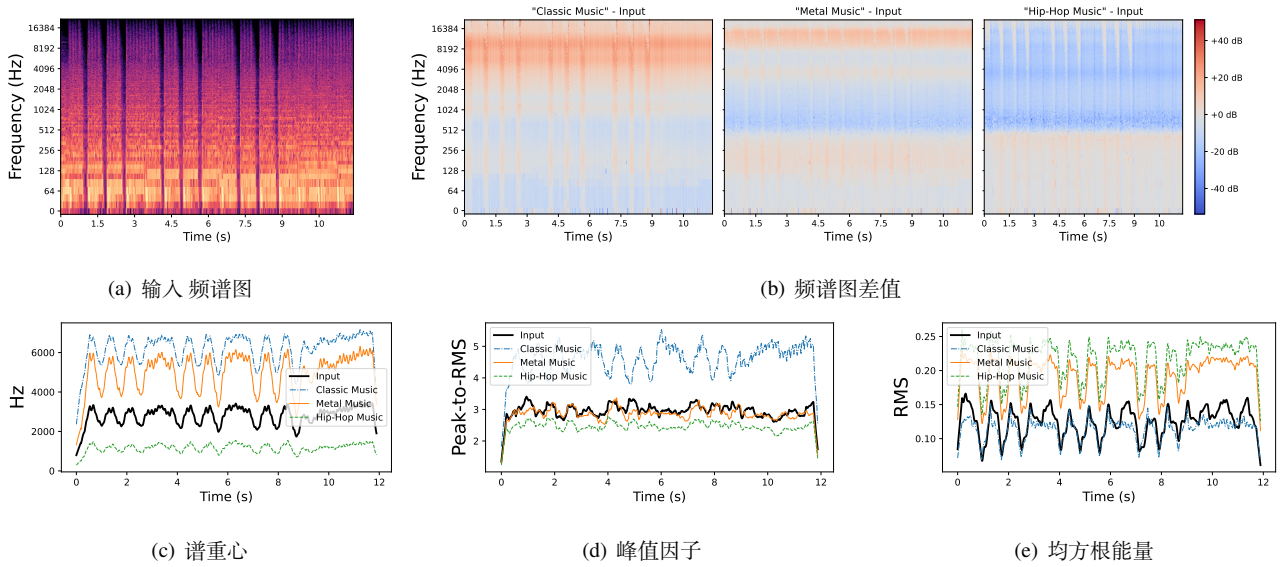



**Figure 2**: 主观评估结果。

所有参与者都熟悉音乐后期制作和数字效果，并且在录音、混音或母带处理方面具有 2 到 5 年的经验。参与者根据各种经过处理的曲目与参考曲目的母带音频效果相似性进行评分。评估包括 8 个问题，所有刺激均使用 30 秒长的音乐录音。参考音频包含与刺激不同的内容，但我们确保参考和刺激在流派或乐器方面不太相似。作为低锚点，最初输入到风格转换系统之前的音乐曲目被呈现出来。没有高锚点，因为评估设置旨在模拟使用 Jamendo 数据集中的音乐曲目进行真实世界的效果风格转移。

如图 2 所示，主观测试结果与我们客观评估中观察到的趋势一致。所有提出的方均超越了基线方法，并且通过 ITO 进一步增强了效果，显示出更接近参考音频特征的特性。所提系统的相似度分数范围从 0 到 100，表明即使对于具有领域知识的专家而言，听力测




(a) 输入 频谱图　(b) 频谱图差值

(c) 谱重心　(d) 峰值因子　(e) 均方根能量

**Figure 3**: 输入音乐与使用文本提示古典音乐、金属音乐和嘻哈音乐处理的 ITO 轨道之间的不同音频特征比较。

试也非常有挑战性。然而，所提系统始终优于基线，显示出显著改进（配对 t 检验，$p < 0.05$）。

### 5.3 ITO 的定性分析带文本提示

为了评估在不同 $\mathcal{L}_{\text{aux}}$ 下 ITO 的有效性，我们使用带有 CLAP 嵌入的文本提示进行定性分析 [22]，类似于在 [36] 中展示的应用。给定一个输入音乐片段，我们利用基于文本的条件调整来优化所提出的白盒模型的 $z_{\text{ref}}$，以 CLAP 嵌入余弦相似性作为优化目标。具体来说，我们计算引导输出的音频嵌入 $\text{CLAP}_{\text{aud}}$ 和给定参考文本提示的文本嵌入 $\text{CLAP}_{\text{txt}}$，然后最大化它们的余弦相似性以指导转换。

用于此分析的输入音乐作品是一段长 11.8 秒的器乐摇滚曲目。

由于不同引导结果中的输入内容保持不变，我们可以直接评估每个文本提示对各种音乐特征的影响。

我们使用三个不同的提示来探索 ITO：古典音乐、金属音乐和嘻哈音乐进行分析。

此实验设置可以通过我们的交互式演示 [7] 进行探索。

如图 3 所示，优化结果表现出与每种类型的一般预期相符的显著特征。在 3(b) 中的频谱差值图（引导输出-输入）突出了受 ITO 影响最明显的频率范围。具体而言，古典音乐提示在中高频段发生了显着变化，这与古典录音通常关联的明亮和清晰特性相一致。金属音乐提示显示了低频和高频的变化，反映了该类型对强劲的贝斯和尖锐高音的重点强调，以配合激进的乐器编配。相比之下，嘻哈音乐提示主要影响低频范围，强化了这种类型在深度贝斯和次低音元素上的特征重点，这对推动节奏感强烈的节拍至关重要。

这些观察结果进一步得到了频谱质心、峰值因子和均方根能量分析的支持。频谱质心的结果遵循预期趋势，其中嘻哈由于其低音丰富的特性具有最低的质心，其次是金属，而古典音乐具有最高的质心，反映了其对谐波丰富度和高频清晰度的强调。峰值因子代表峰-均方根比，嘻哈的峰值因子最低，表明其动态结构更为压缩且低音丰富，而古典音乐具有最高的峰值因子，与其通常不压缩、更宽广的动态范围相一致。RMS 能量从经典到嘻哈呈上升趋势，金属位于其间，这与这些流派各自的响度和动态特性一致。这些发现表明，由 $\text{CLAP}_{\text{txt}}$ 指导的 ITO 成功地引导母带处理效果链与给定文本提示预期的声音特征保持一致，展示了其作为音乐后期制作创意工具的潜力。

## 6. 结论

在本文中，我们介绍了利用 $z_{\text{ref}}$ 上的 ITO 来掌握风格转换的 ITO-Master 框架。我们的实验表明，同时训练参考编码器 $\Phi$ 和 $\Psi$ 可以提高性能。使用 ITO 优化 $z_{\text{ref}}$ 在很少的步骤内带来了有意义的进步，比直接优化 $\Theta$ 更有效率。主观评估确认了我们的方法可以产生感知上一致的 mastering 效果，并且定性结果突出了基于文本条件的 ITO 在创意应用中的潜力。作为未来工作的一部分，我们计划在评估中加入生产质量和可用性的考量，并结合参考一致性。由于 mastering 是一项策展任务，糟糕的参考选择会导致即使高度对齐也产生次优结果，这强调了感知偏好度量的重要性。

[7] https://huggingface.co/spaces/jhtonyKoo/ITO-Master

## 7. REFERENCES


[1] U. Zölzer, X. Amatriain, D. Arfib, J. Bonada, G. De Poli, P. Dutilleux, G. Evangelista, F. Keiler, A. Loscos, D. Rocchesso *et al.*, *DAFX-Digital audio effects*. John Wiley & Sons, 2002.

[2] M. Shelvock, *Audio mastering as musical practice*. The University of Western Ontario (Canada), 2012.

[3] M. Piotrowska, S. Piotrowski, and B. Kostek, "A study on audio signal processed by" instant mastering" services," in *Audio Engineering Society Convention 142*. Audio Engineering Society, 2017.

[4] J. Sterne and E. Razlogova, "Machine learning in context, or learning from landr: Artificial intelligence and the platformization of music mastering," *Social Media+ Society*, vol. 5, no. 2, p. 2056305119847525, 2019.

[5] M. A. M. Ramírez, O. Wang, P. Smaragdis, and N. J. Bryan, "Differentiable signal processing with black-box audio effects," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 66–70.

[6] J. Koo, S. Paik, and K. Lee, "End-to-end music remastering system using self-supervised and adversarial training," in *Proc. ICASSP*, 2022, pp. 4608–4612.

[7] S. Grishakov, C.-Y. Yu, and Zicklag, "Matchering: Audio matching and mastering python library," https://github.com/sergree/matchering.

[8] J. Engel, C. Gu, A. Roberts *et al.*, "DDSP: Differentiable digital signal processing," in *International Conference on Learning Representations*, 2020.

[9] J. Koo, S. Paik, and K. Lee, "Reverb conversion of mixed vocal tracks using an end-to-end convolutional deep neural network," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 81–85.

[10] S. Lee, J. Park, S. Paik, and K. Lee, "Blind estimation of audio processing graph," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[11] C. J. Steinmetz, N. J. Bryan, and J. D. Reiss, "Style transfer of audio effects with differentiable signal processing," *J. Audio Eng. Soc*, vol. 70, no. 9, pp. 708–721, 2022.

[12] J. Koo, M. A. Martínez-Ramírez, W.-H. Liao, S. Uhlich, K. Lee, and Y. Mitsufuji, "Music mixing style transfer: A contrastive learning approach to disentangle audio effects," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[13] S. S. Vanka, C. Steinmetz, J.-B. Rolland, J. Reiss, and G. Fazekas, "Diff-MST: Differentiable mixing style transfer," in *Proc. ISMIR*, 2024.

[14] Y.-H. Chen, Y.-T. Yeh, Y.-C. Cheng, J.-T. Wu, Y.-H. Ho, J.-S. R. Jang, and Y.-H. Yang, "Towards zero-shot amplifier modeling: One-to-many amplifier modeling via tone embedding control," in *Proc. ISMIR*, 2024.

[15] Z. Novack, J. McAuley, T. Berg-Kirkpatrick, and N. J. Bryan, "DITTO: Diffusion inference-time t-optimization for music generation," in *Proc. ICML*, 2024.

[16] Z. Novack, J. McAuley, T. Berg-Kirkpatrick, and N. Bryan, "DITTO-2: Distilled diffusion inference-time t-optimization for music generation," in *Proc. ISMIR*, 2024.

[17] C. Steinmetz, S. Singh, I. Ibnyahya, S. Yuan, E. Benetos, J. Reiss *et al.*, "ST-ITO: Controlling audio effects for style transfer with inference-time optimization," in *Proc. ISMIR*, 2024.

[18] C.-Y. Yu, M. A. Martínez-Ramírez, J. Koo, W.-H. Liao, Y. Mitsufuji, and G. Fazekas, "Improving inference-time optimisation for vocal effects style transfer with a gaussian prior," *arXiv preprint arXiv:2505.11315*, 2025.

[19] M. A. Martínez-Ramírez, W.-H. Liao, G. Fabbro, S. Uhlich, C. Nagashima, and Y. Mitsufuji, "Automatic music mixing with deep learning and out-of-domain data," in *Proc. ISMIR*, 2022.

[20] S. H. Linkwitz, "Active crossover networks for non-coincident drivers," *Journal of the Audio Engineering Society*, vol. 24, no. 1, pp. 2–8, 1976.

[21] C.-y. Yu, C. Mitcheltree, A. Carson, S. Bilbao, J. Reiss, and G. Fazekas, "Differentiable all-pole filters for time-varying audio systems," in *27th International Conference on Digital Audio Effects (DAFx)*, 2024.

[22] Y. Wu, K. Chen, T. Zhang, Y. Hui, T. Berg-Kirkpatrick, and S. Dubnov, "Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[23] I. Pereira, F. Araújo, F. Korzeniowski, and R. Vogl, "MoisesDB: A dataset for source separation beyond 4-stems," in *Proc. ISMIR*, 2023.

[24] Z. Rafii, A. Liutkus, F.-R. Stöter, S. I. Mimilakis, and R. Bittner, "MUSDB18-HQ - an uncompressed version of MUSDB18," Aug. 2019. [Online]. Available: https://doi.org/10.5281/zenodo.3338373

[25] D. Bogdanov, M. Won, P. Tovstogan, A. Porter, and X. Serra, "The mtg-jamendo dataset for automatic music tagging," in *Proc. ICML*, 2019.

[26] C. J. Steinmetz and J. D. Reiss, "Efficient neural networks for real-time modeling of analog dynamic range compression," *arXiv preprint arXiv:2102.06200*, 2021.

[27] L. Liu, H. Jiang, P. He, W. Chen, X. Liu, J. Gao, and J. Han, "On the variance of the adaptive learning rate and beyond," in *International Conference on Learning Representations*, 2020.

[28] K. Kilgour, M. Zuluaga, D. Roblek, and M. Sharifi, "Fréchet audio distance: A metric for evaluating music enhancement algorithms," *arXiv preprint arXiv:1812.08466*, 2018.

[29] B. Elizalde, S. Deshmukh, M. Al Ismail, and H. Wang, "CLAP: Learning audio concepts from natural language supervision," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[30] R. Kumar, P. Seetharaman, A. Luebs, I. Kumar, and K. Kumar, "High-fidelity audio compression with improved rvqgan," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[31] A. Défossez, J. Copet, G. Synnaeve, and Y. Adi, "High fidelity neural audio compression," *Transactions on Machine Learning Research*, 2023.

[32] A. Gui, H. Gamper, S. Braun, and D. Emmanouilidou, "Adapting frechet audio distance for generative music evaluation," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 1331–1335.

[33] S. Lee, M. A. Martínez-Ramírez, W.-H. Liao, S. Uhlich, G. Fabbro, K. Lee, and Y. Mitsufuji, "Searching for music mixing graphs: A pruning approach," in *27th International Conference on Digital Audio Effects (DAFx)*, 2024.

[34] ——, "Reverse engineering of music mixing graphs with differentiable processors and iterative pruning," *Journal of the Audio Engineering Society*, vol. 73, pp. 344–365, June 2025.

[35] C.-Y. Yu, M. A. Martínez-Ramírez, J. Koo, B. Hayes, W.-H. Liao, G. Fazekas, and Y. Mitsufuji, "Diffvox: A differentiable model for capturing and analysing professional effects distributions," in *28th International Conference on Digital Audio Effects (DAFx)*, 2025.

[36] A. Chu, P. O' Reilly, J. Barnett, and B. Pardo, "Text2FX: Harnessing clap embeddings for text-guided audio effects," in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025, pp. 1–5.