arXiv:2602.08233v1 [cs.SD] 9 Feb 2026

Tutti：通过结构级音色控制与人声质感建模实现表现力多歌手合成

Jiatao Chen [1 2] **Xing Tang** [1] **Xiaoyue Duan** [2] **Yutang Feng** [2] **Jinchao Zhang** [2] **Jie Zhou** [2]

摘要

尽管现有歌声合成系统能够实现高保真的独唱表演，但它们受限于全局音色控制，无法在同一首歌曲中处理动态多歌手编排和人声质感。为此，我们提出**Tutti**，这是一个专为结构化多歌手生成设计的统一框架。具体而言，我们引入结构感知歌手提示，使歌手调度能够灵活地随音乐结构变化；并提出基于条件引导VAE的互补质感学习，以捕捉与显式控制互补的隐式声学质感（例如，空间混响和频谱融合）。实验表明，Tutti 在精准多歌手调度方面表现出色，并显著提升了合唱生成的声学真实感，为复杂的多歌手编排提供了一种新范式。音频示例见

h
t
t
p
s://
annoauth123-ctrl.github.io/Tutii_演示/。

1. 引言

音乐是人类表达情感的通用语言。随着深度学习的迅速发展，生成式 AI 在音乐生成领域取得了显著突破（Dhariwal 等人，2020；Agostinelli 等人，2023；Liu 等人，2023），展现了创作完整音乐作品的潜力。

在音乐作品中，歌声是最具感染力和表现力的组成部分，承载着深刻的叙事力量。因此，作为一个将语言内容与音乐表现相结合的独特领域，歌声合成（SVS）也取得了同样显著的成就——著的成果。DiffSinger（Liu等人，2022）和 VISinger（Zhang等人，2022）等领先模型确立了高保真生成范式，使逼真的独唱表演得以合成。

然而，尽管现有歌声合成（SVS）模型在音色克隆方面表现出色，但它们仍在很大程度上受限于“独唱者范式”。绝大多数现有模型都假设单首歌曲仅有一位歌手，或只支持稳定的歌手身份，即将歌手身份定义为在整首作品中保持不变的全局条件。这种僵化的设定忽略了复杂音乐表现力的两个关键维度：（1）歌手编排——依据音乐结构对多位歌手进行时间调度。在真实的作品中，歌手的角色会随着歌曲的推进而变化：主歌部分可能由一位歌手独唱，而副歌则会加入其他歌手或整个合唱团。此外，不同独唱者之间也经常在不同段落间切换（例如，独唱者之间的切换）。现有模型根本缺乏在单首歌曲中对这类歌手切换进行建模，以及对合唱团所需的同步音色表征进行建模的机制。（2）人声质感——由人声互动和制作方式所产生的声学属性，包括空间混响、频谱融合和谐波叠加。多人合唱并非个体人声的简单线性混合，而是包含复杂互动，例如不同声部之间共振峰的融合。当前的歌声合成模型主要针对独唱合成进行设计，缺乏对这些互动进行建模的显式机制，因此生成的多人合唱片段听起来像是人为混合的独立音轨，而非和谐统一的合唱演出。

为弥补这一空白，我们将任务从传统的歌声合成（SVS）扩展至多歌手歌声合成，并提出 **Tutti**：一种面向复杂多歌手编排的结构感知框架。具体而言，我们设计了结构感知歌手提示，能够根据音乐结构（例如，主歌/副歌）构建歌手序列，从而灵活安排多位歌手的参与。为应对同时表征多位歌手的挑战，我们进一步引入自适应歌手提示融合器，以动态整合多歌手嵌入。

[1]中国武汉武汉理工大学计算机科学与人工智能学院 [2]中国 Tentenct 公司微信 AI。通讯作者：Xing Tang <tangx- ing@whut.edu.cn>, Jinchao Zhang <dayerzhang@tencent.com>。

预印本。*2026*年*2*月*10*日。

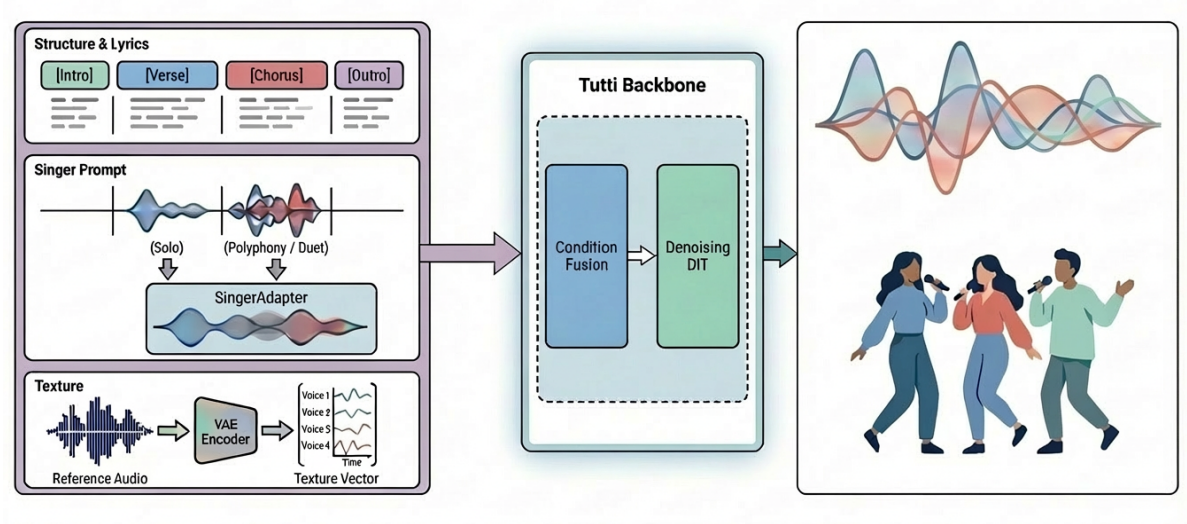


图*1*。 Tutti框架用于结构感知多歌手生成的概览。该流程首先构建结构感知歌手提示，并从参考音频中提取互补的质感特征。随后，将这些条件连同歌词和结构标签输入 DiT 主干网络，以生成目标人声潜变量。最后，Vocal VAE 解码器将潜变量重建为高保真的多歌手波形。

与此同时，为捕捉显式特征之外丰富的声学细节，我们提出基于条件引导变分自编码器的互补纹理学习。该模块学习提取隐式质感特征，例如混音效果和空间氛围；这些特征与歌词、歌手身份等显式条件互为补充，从而精准还原人声互动的声学品质。我们的贡献总结如下：

- 我们提出了 Tutti，这是首个专为在单首歌曲中合成多位歌手而设计的生成式框架。
- 我们设计了结构感知歌手提示，并结合自适应歌手提示融合器，以灵活处理独唱者与合唱团之间的复杂转换，从而精准控制每个乐段的歌手编排。

•我们提出了基于条件引导VAE的互补质感学习，通过引入条件引导的重建任务来捕捉与显式控制互补的隐式人声质感，从而显著增强生成结果的声学真实感。

2. 相关工作

2.1. 歌声合成

近年来，歌声合成（SVS）领域经历了从端到端生成向扩散模型的范式转变，在音质、控制粒度和输入灵活性方面取得了显著进展。VISinger（Zhang等人，2022）系列和 DiffSinger（Liu等人，2022确立了高保真生成范式

该范式以显式乐谱输入为基础，DiffSinger 成功将扩散模型引入歌声合成，大幅提升了合成自然度和稳定性。此后，研究人员致力于拓展歌声合成的任务边界与控制能力：MVoice（Huang等人，2023）提出了一个整合文本转语音、语音转换和歌声合成的统一框架，并通过说话人提示初步实现了零样本多歌手适配；TCSinger（Zhang等人，2024b）和 ExpressiveSinger（Dai等人，2024年）专注于跨语言风格迁移和细粒度演唱控制（例如颤音、响度），增强了生成音乐的表现力。在生成效率方面，InstructSing（Zeng等人，2024年）将可微分数字信号处理与对抗训练相结合，大幅降低了高采样率合成的计算成本。

近期研究进一步打破了传统歌声合成（SVS）对乐谱的依赖。TechSinger（Guo 等人，2025年）提出了一种结合技巧预测器的流匹配框架，实现了对多种语言和演唱技巧的精细建模；Vevo（Zhang等人，2025c）和 Vevo2（Zhang等人，2025b）则通过自监督解耦和多模态输入机制，实现了零样本音色模仿和无乐谱生成，推动歌声合成迈向更灵活的生成范式。然而，尽管这些模型在单歌手音色克隆和风格控制方面表现出色，它们仍主要局限于“单曲单歌手”或“全局音色控制”框架，无法有效解决“单曲内动态多歌手编排”这一复杂问题，也缺乏对多人合唱中声部互动和人声质感的显式建模。

2.2. 多说话人对话生成

为克服传统歌声合成仅支持单人演唱的局限，并实现单曲内的动态多歌手编排，我们从多说话人对话生成领域汲取灵感。该领域致力于解决“说话人调度”和“轮流发言”问题，为动态歌手提示设计提供了重要的参考范式。早期探索如 dGSLM（Nguyen 等人，2023）采用双塔 Transformer 架构，通过交叉注意力机制建立离散单元流之间的联系，以隐式学习说话人互动节奏和副语言行为。随后，为更精确地控制重叠语音，CoVoMix（Zhang等人，2024年a）和 CoVoMix2（Zhang等人，2025年a）提出了完全非自回归的流匹配框架，直接将多流转录文本映射为混合梅尔频谱图，实现了精细的时间控制和说话人分离。近期研究进一步强化了对互动逻辑的显式建模。FD-SLMs（Wang 等人，2024年）提出了一种基于大语言模型（LLM）预测的神经有限状态机（Neural FSM），通过“说话/聆听”状态切换来决定轮流发言，实现了基于语义上下文的在线互动决策。DialoSpeech（Xie 等人，2025年）将大语言模型与双轨标记生成相结合，使用特殊的说话人切换标记（例如，‘[spkchange]’）显式控制对话流程。在此基础上，JoyVoice（Yu 等人，2025年）将显式说话人 ID 序列和嵌入作为条件，通过因果自回归 Transformer 实现长上下文对话的结构化调度，在说话人一致性方面取得了显著进展。然而，将上述对话生成范式直接迁移到多歌手合唱任务中，会面临根本性挑战。首先，对话模型通常将重叠语音视为时间上的抢占或简单的信号混合，缺乏对人声质感的声学建模——即频域中多个声部的谐波融合与共鸣。其次，对话调度逻辑主要由语义驱动，而演唱编排则受严格的音乐结构约束（例如主歌/副歌、和声进行）。现有对话模型无法感知这种高层次的音乐先验，导致生成的合唱片段往往缺乏声部之间的音高和谐与整体凝聚力。

3. 方法

3.1. 概述

我们提出 **Tutti**，一个基于潜在扩散 Transformer（扩散变换器）范式构建的结构感知多歌手生成框架。如图 2所示，我们的模型由两个核心组件组成：用于压缩的 Vocal VAE 将原始音频转换为潜在表示，以及用于条件生成的基于扩散变换器的主干网络。

模型架构。 为高效建模长序列音乐依赖关系，我们的骨干网络采用基于 LLaMA 的 Transformer 架构，遵循 DiffRhythm（Ning 等人，2025年）的设计。对于潜在空间，我们采用基于 Stable Audio 2.0（Evans 等人，2025年的专用 Vocal VAE，并使用人声数据对其重新训练，以便从低维潜在表示中忠实地重建高保真波形。

模型条件控制。 为了精确控制内容、结构和音色，我们在生成过程中注入多种条件。具体而言，歌词 L 通过预训练文本编码器进行编码，以提供内容指导；结构标签 S 表示音乐段落（例如主歌、副歌），用于引导结构推进。为处理复杂的歌手调度，我们采用结构感知型歌手提示 C_{singer}，将其作为多歌手身份的动态表示，并通过我们提出的自适应歌手提示融合器进行处理。此外，我们通过条件引导变分自编码器编码器从参考音频中提取人声质感 $Z_{texture}$，以捕捉气声感和空间效果等隐含的声学细节。最后，我们同时纳入扩散时间步 t 和音频起始时间，向模型提供去噪进度以及歌曲中的绝对时间位置。

基于这些条件，我们将生成任务表述为学习条件分布 $p_\theta(z_0|L,\ S,\ C_{singer},\ Z_{texture})$，其中 θ 表示可学习参数。

特征融合。 我们采用基于拼接的融合策略。连续条件（例如，歌手提示、纹理）和离散条件（例如，歌词、结构）通过广播或下采样对齐至潜变量帧率，然后沿特征维度与带噪潜变量 z_t 拼接。这样可确保 DiT 在每个步骤都能同时关注所有控制信号。

训练目标。 对于 Vocal VAE，我们采用由多分辨率短时傅里叶变换损失（Steinmetz 和 Reiss，2020）构成的复合目标，以实现谱收敛；采用 KL 散度损失（Yu 等人，2013）进行潜变量正则化；并采用对抗损失提升感知质量。随后，对于 DiT 主干网络，我们遵循条件流匹配范式（Lipman 等人，2023），使模型学习速度场 $v_\theta(z_t,\ t,\ c)$，将噪声分布 $p_0(z)$ 通过常微分方程输运至数据分布 $p_1(z)$：

$$\frac{dz_t}{dt} = v_\theta(z_t, t, c) \quad \text{with} \quad \begin{cases} z_0 \sim p_0(z) \\ z_1 \sim p_1(z) \end{cases} \tag{1}$$

训练目标是最小化期望平方误差

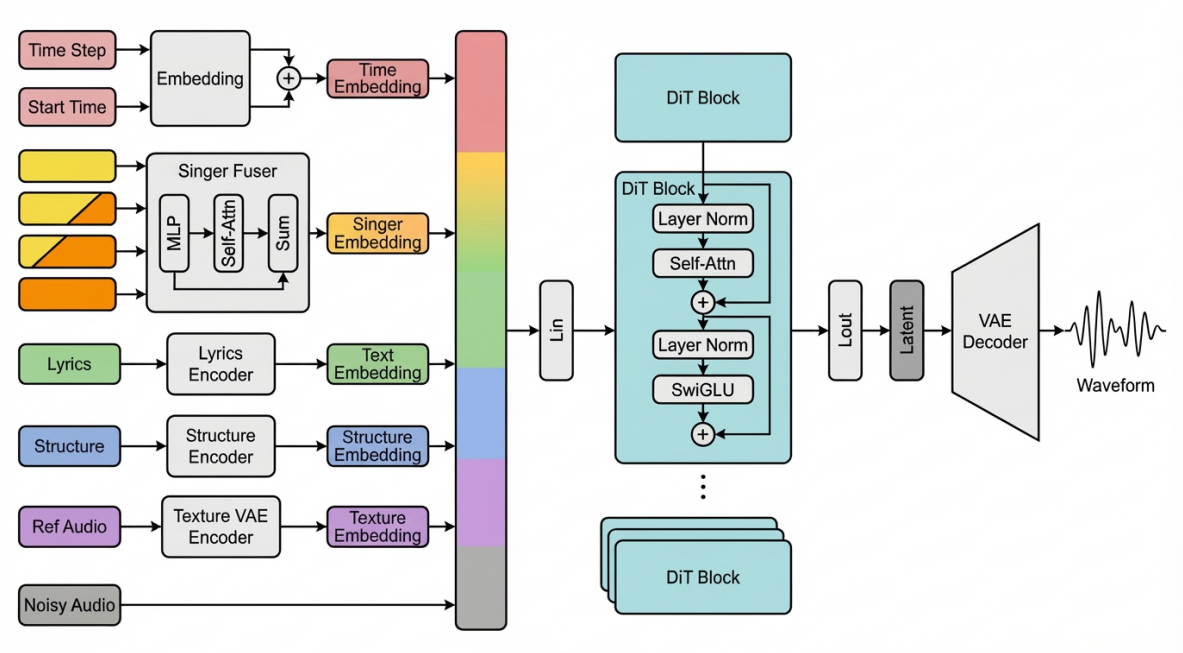


图 2. 基于 DiT 的骨干网络架构。左侧面板展示了条件处理模块：结构感知歌手提示（以橙色标示）使用自适应融合器整合多歌手信息，而纹理嵌入（以紫色标示）则从参考音频中提取，以提供互补的声学特征。这些嵌入与其他条件拼接后输入 DiT 模块。最后，预测潜变量由人声 VAE 解码器解码，生成波形。

即预测速度场与目标速度场之间的期望平方误差：

$$\mathcal{L} = \mathbb{E}_{t\sim\pi_{\ln},z_t\sim p_t(z_t)}\left[\|v_\theta(z_t,t,c)-(z_1-z_0)\|_2^2\right], \quad (2)$$

其中，c 表示条件集合，时间步采样分布 $\pi_{\ln}(t;\ m,\ s)$ 遵循 logit-正态密度：

$$\pi_{\ln}(t;m,s)=\frac{1}{s\sqrt{2\pi}}\frac{1}{t(1-t)}\exp\left(-\frac{(\mathrm{logit}(t)-m)^2}{2s^2}\right), \quad (3)$$

其中 $\mathrm{logit}(t)\ =\log\frac{t}{1-t}$。正如 Stable Diffusion 3（Esser 等人，2024年）中所述，logit-正态采样可提供自适应加权，其中尺度参数 s 控制中间时间步附近的集中程度，而位置参数 m 则使采样偏向数据域或噪声域。实际操作中，我们采样 $u\sim\mathcal{N}(m,\ s)$ 并通过逻辑函数 $t=\sigma(u)\ =1/(1+e^{-u})$。

3.2. 结构感知歌手提示

为了在单首歌曲中灵活编排多位歌手，我们设计了一种结构感知的提示构建策略，根据音乐结构自动调度歌手资源，随后通过自适应融合器整合多歌手信息。

提示构建。 我们提出一种结构引导流程来构建歌手提示。首先，我们使用 SongPrep(Tanet al.,2025年)将输入音频时间线 T 划分为 M 互不重叠的结构单元 $\{\tau_1,\tau_2,\ldots,\tau_M\}$。1）对于每个片段 τ_m，我们使用 CAM++(Wang et al., 2023)提取一个具有代表性的歌手嵌入$\mathbf{v}_m$。2）鉴于主歌段落通常由单人声演唱者演唱这一假设，我们仅计算主歌片段嵌入之间的余弦相似度 $s_{i,j}$。满足 $s_{i,j}\ <\ \delta_{id}$ 的配对被视为来自不同歌手，由此形成该歌曲的全局歌手集合$U_{global}\ =\ \{\mathbf{u}_1,\ldots,\mathbf{u}_K\}$。3）我们采用基于结构的分配：对于主歌片段，我们严格地将已识别聚类中的统一歌手嵌入分配给它们，以增强零样本音色稳定性；对于副歌或桥段，如果全局歌手数量 $K>1$，我们就计算当前片段的嵌入$\mathbf{v}_m$ 与$\mathbf{u}_k$ 中每位歌手的相似度，其计算方式为 $s_{m,k}$。当匹配到的歌手数量满足 $\sum_{k=1}^{K}\mathbb{I}(s_{m,k}>\delta_{multi})\ \geq 2$时，其中 $\mathbb{I}(\cdot)$为指示函数，该片段将被识别为多歌手段落，并且我们会同步更新其结构标签以反映这一状态。

自适应歌手提示融合器。 为了有效地将每个片段中识别出的歌手嵌入融合为统一表示，我们设计了一种基于自注意力机制的自适应融合器。给定第m个片段的歌手集合 U_m，融合器会将其映射为单个条件嵌入 C_{singer}。具体而言，生成的条件

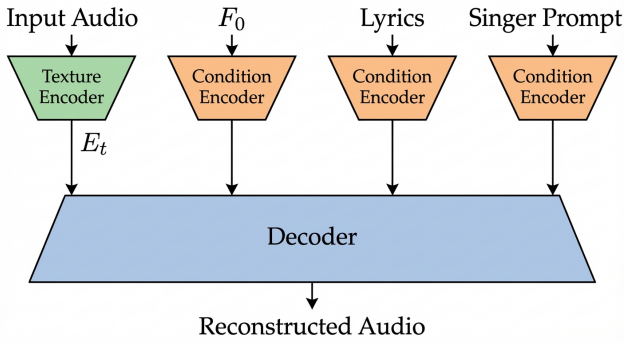


图3。 用于互补纹理学习的条件引导变分自编码器架构。

嵌入计算如下：

$$C_{singer} = \sum_{i=1}^{k_m} \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)_i \mathbf{u}_i \tag{4}$$

这里，查询（Q）、键（K）和值（V）均由输入歌手嵌入 U_m经过线性投影得到。这种参数化加权方案使模型能够自适应地处理主唱与和声之间的主次关系。随后，将得到的嵌入广播至片段内的所有时间步，为音色提供稳定且具备结构感知能力的控制信号。

3.3. 基于条件引导VAE的互补质感学习

在复杂的多歌手场景中，仅依赖显式条件（例如歌词、结构标签和歌手提示）通常不足以生成具有真实声学细节的高保真音频。这些显式控制能够定义内容和身份，却无法捕捉“人声质感”——即空间混响、混音效果和高频谐波等决定演出声学真实感的声学属性。为此，我们引入条件引导变分自编码器，以学习一种互补的纹理表征，弥合显式控制与真实音频之间的差距。

首先，我们训练一个专用的 Vocal VAE。由于标准音乐变分自编码器（Evans 等人，2025年）针对通用音乐进行了优化，可能无法捕捉细粒度的人声特征，因此我们针对自己的人声数据集，重新训练第 3.1节所述的 VAE 架构。为使学习到的潜在空间能够捕捉与显式条件互补的纹理信息，我们提出一种条件引导的重建任务。如图3所示，我们将 VAE 训练表述为一个潜在扰动重建问题。给定一个干净的人声波形 y，编码器 $\mathcal{E}$ 首先将其映射为潜在表示 z。为防止编码器保留本应由其他条件提供的显式信息（如音高、内容和音色），我们对 z 施加高斯噪声扰动，得到受扰动的潜在表示 $\tilde{z}$。关键在于，为了将纹理与内容及音色解耦，我们向解码器 $\mathcal{D}$ 明确提供一组显式条件 $C_{ex} = \{c_{lyrics}, c_{singer}, c_{F0}\}$，其中$c_{lyrics}$、 c_{singer}和 c_{F0} 分别对应歌词、结构级歌手提示和基频的嵌入。随后，解码器需要结合受扰动的潜在表示和显式条件，重建原始干净音频 y：

$$\hat{y} = \mathcal{D}(\tilde{z}, C_{ex}) = \mathcal{D}(\mathcal{E}(y) + \epsilon, c_{lyrics}, c_{singer}, c_{F0}) \tag{5}$$

其中， ϵ从高斯分布 $\mathcal{N}(0,\ \sigma^2 I)$中采样。通过将 C_{ex} 直接提供给解码器并扰动潜在空间，我们促使编码器丢弃这些条件中已经包含且由 C_{ex} 稳健提供的信息，转而专注于提取恢复无损目标 y所需的剩余声学细节——人声质感。

在推理过程中，这种设计支持灵活的人声质感控制。我们可以选择向编码器提供参考音频，以提取人声质感潜变量 $z_{texture}$。由于编码器经过训练，能够将人声质感与音高、内容和音色解耦，该潜变量主要捕获声学属性，例如混音效果、空间氛围和音色稳定性。这确保参考音频能够增强人声质感的声学真实感与丰富度，而音乐结构则主要由扩散变换器主干网络和显式条件控制。

4. 实验

4.1. 实验设置

数据准备。 为了训练我们的框架，我们分别为 Vocal VAE 和 DiT 主干网络构建了两个独立的数据集。对于 Vocal VAE，我们收集了约 14 万首歌曲。为获取干净的训练目标，我们统一使用 BS-Roformer（Lu 等人，2024年）从原始音频中提取人声轨道。具体而言，为确保重建过程稳定，我们对所有提取的人声片段进行响度归一化，以统一动态范围。对于 DiT 主干网络，我们将数据规模扩展至约 30 万首歌曲，主要来源于 AudioBox（Vyas 等人，2023）和 SongEval（Yao 等人，2025年）。我们采用相同的人声提取流程，但引入了更复杂的标注流程。具体而言，我们使用 WhisperX（Bain 等人，2023）进行歌词转录和强制对齐，根据置信度分数筛选样本，并仅保留中文和英语歌曲。为构建多歌手序列，我们使用 CAM++ （Wang 等人，2023）提取每个片段的声纹嵌入，并将余弦相似度阈值严格设为 0.4，以区分不同歌手，确保数据集包含显著的音色变化和歌手

表 1. 人声波形重建性能的比较评估。我们报告了无损到无损和有损到无损重建场景下的 STOI、宽带感知语音质量评估、MCD 和弗雷歇音频距离。

模型	无损 → 无损				有损 →无损			
	STOI↑	宽带感知语音质量评估↑	MCD↓	弗雷歇音频距离↓	STOI↑	宽带感知语音质量评估↑	MCD↓	弗雷歇音频距离↓
DiffRhythm VAE	0.7557	2.7461	2.9580	0.8252	0.7551	2.6934	3.0601	1.4162
LeVo VAE	0.7246	2.5426	3.3435	1.2382	0.7211	2.4792	3.2666	1.8931
SongBloom VAE	0.7232	2.5384	3.6137	1.1872	0.7195	2.4757	3.6721	1.8239
我们的（Vocal VAE）	**0.8175**	**3.1602**	**2.4696**	**0.5676**	**0.8163**	**3.0805**	**2.4458**	**1.1697**

过渡。

基线。 为验证所提出框架的有效性，我们设计了包含外部基线和内部变体的对比实验。在声学重建方面，我们选取 DiffRhythm VAE（Ning 等人，2025年）、LeVo VAE（Lei 等人，2025年）和 SongBloom VAE（Yang 等人，2025年）作为基线，它们代表了音乐和歌声领域最先进的音频压缩模型。在歌声生成方面，鉴于目前缺乏现成的多歌手生成模型，且我们的任务不需要特定的音高输入，我们选取 Vevo2（Zhang等人，2025b）作为强基线，它代表了多模态生成和零样本音色模仿领域的最新水平。内部基线方面，我们纳入了 Tutti 的不同变体进行比较：（1）Tutti-w/o-Texture：移除互补纹理学习模块，仅依赖基础 VAE重建；（2）Tutti-w/o-Fuser：移除自适应歌手提示融合器，并使用简单的线性求和来融合歌手特征。此设置旨在同时评估模型相较于现有 SOTA 的优势，以及其核心组件的贡献。

评估数据。 为评估模型在不同任务上的性能，我们构建了三个测试子集。对于人声重建，我们随机选取了30首采样率为44.1kHz的歌曲。通过音乐源分离提取人声轨道，并将其作为重建目标，以评估不同的VAE架构。对于基础生成指标的评估，我们从数据集中随机抽取了50首包含人声的歌曲。对于多歌手相关指标的评估，我们首先在数据标注流程中通过声纹相似度阈值筛选出一批疑似包含多位歌手的候选样本，随后进行人工核验，以排除因单人演唱中存在显著音色变化而被误判的样本。最终，我们得到了一组包含20首歌曲的多歌手测试集。该子集涵盖了多种典型的多歌手场景，包括乐队合唱和男女二重唱，能够有效考察模型的能力包括歌手调度和合唱建模。

评估指标。 为全面评估生成质量，我们采用结合客观指标与主观指标的评估体系。对于声学重建，我们使用 STOI（Taal 等，2010）来衡量可懂度保持，使用宽带感知语音质量评估（WB-PESQ）（Rix 等，2001）进行感知质量评估，使用 MCD（Kubichek，1993）评估频谱失真，并使用弗雷歇音频距离（FAD）（Kilgour 等，2019）评估分布层面的保真度。对于歌声生成，我们使用词错误率来评估歌词可懂度（Radford 等，2023）。需要注意的是，由于我们的任务被定义为无乐谱生成，旋律由模型自由决定，因此标准的音高准确度指标（例如F0均方根误差）并不适用。我们还计算生成音频与参考音频之间的说话人相似度（SIM）。考虑到多歌手场景，我们根据歌手时长计算加权平均值，以衡量身份一致性。对于主观评估，我们采用 MOS-Q 和 MOS-N 分别评估音频质量和自然度。此外，针对本文的多歌手任务，我们引入了两个采用 1—5 分制评定的附加指标：（1）多歌手MOS（MS-MOS）：评估多歌手音色的融合效果、每位歌手的辨识度及其稳定性；（2）旋律MOS（Mel-MOS）：专门评估演唱表现的旋律和音高准确度。

4.2. 主要结果

声学重建。 我们首先在纯人声场景中评估波形重建性能，将重新训练的 Vocal VAE 与三个代表性基线进行比较。评估方案涵盖两种实验设置：无损到无损重建，以无损音频作为输入；以及有损到无损重建，以经 MP3 压缩的有损音频作为输入，同时以无损音频作为参考目标。如表 1所示，我们的 Vocal VAE 在两种条件下的全部指标上均取得最优表现。在可懂度保持方面，STOI 较次优基线提升约 8%。在感知质量与

表2. 客观与主观评估结果。无纹理表示移除了互补纹理学习模块，无融合器表示以线性求和替代自适应歌手提示融合器。

模型	客观指标		主观指标（95% 置信区间）			
	WER↓	SIM↑	MOS-Q↑	MOS-N↑	多歌手平均意见分↑	Mel-MOS↑
真实值（真实值）	12.45%	-	4.50±0.05	4.65±0.05	4.30±0.05	4.16±0.05
Vevo2	16.80%	0.657	3.85±0.12	4.01±0.12	-	-
Tutti（我们的方法）	**13.50%**	0.691	**4.12±0.06**	**4.12±0.06**	**4.02±0.10**	**3.89±0.05**
无纹理	13.85%	**0.705**	3.99±0.08	3.97±0.07	3.87±0.06	3.57±0.12
无融合器	17.25%	0.649	4.02±0.07	4.10±0.07	3.61±0.12	3.87±0.06

频谱重建方面，宽带感知语音质量评估与 MCD 均显著提升，表明重建音频在感知上更接近原始信号。在分布级保真度方面，弗雷歇音频距离大幅下降，说明重建音频的整体分布更接近真实音频。更值得注意的是，在处理有损 MP3 输入时，基线模型因缺乏压缩伪影恢复能力而出现显著性能下降，而我们的 Vocal VAE 仅出现适度的性能下降，展现出对有损输入的稳健性。高质量的声学重建为后续歌声生成任务确立了坚实的性能上界。

多歌手生成性能。表2展示了所有模型在客观和主观指标上的比较结果。需要指出的是，Tutti 是首个支持在单首歌曲中进行动态多歌手编排的生成框架，而 Vevo2 等现有基线模型只能处理单歌手场景，因此无法直接比较与多歌手相关的指标。在可比较的基础指标上，Tutti 在歌词可懂度方面表现出显著优势，其词错误率（WER）明显低于 Vevo2，这得益于我们采用的词级歌词对齐策略。在音色一致性方面，由于我们采用时长加权平均策略来计算多歌手场景中的相似度，Tutti 的相似度（SIM）高于 Vevo2，表明结构感知歌手提示能够在复杂的歌手切换场景中保持稳定的身份表征。在主观评估中，Tutti 的音频质量和自然度均优于 Vevo2。值得注意的是，Vevo2 在较长音频中偶尔会出现中断问题，而 Tutti 凭借基于扩散变换器（DiT）的长序列建模能力，展现出更稳定流畅的表现。针对多歌手演唱任务，我们引入了两项专用指标：多歌手平均意见分（MS-MOS）和 Mel-MOS。就多歌手辨识度而言，Tutti 获得了较高的 MS-MOS 分数，表明在多歌手场景中，各歌手的音色能够清晰区分，同时声部之间保持了良好的融合与平衡。在旋律表现方面，尽管 Mel-MOS 与真实值仍有差距，但整体旋律性和音高准确度已达到较高水平。

为了直观验证模型在多歌手调度和合唱建模方面的能力，我们对生成音频进行了声学可视化分析。如图4所示，我们展示了同时包含独唱和合唱片段的音高显著性图和梅尔频谱图。(a) 在音高显著性图中，白色虚线清楚地标示出独唱A（0:00–0:20）、独唱B（0:20–0:40）和副歌（0:40起）之间的结构边界，表明模型精准遵循了输入的结构控制。在独唱A和独唱B区域，音高图呈现出清晰的单条基频轨迹，主要集中在C3–C4音域，这正是单个歌手的典型特征。相比之下，副歌区域则呈现出明显不同的模式：音高能量分布显著扩展至C3–C5，同时可见多条交织的旋律轮廓。这种独特的谐波分层现象——即不同频率层级上出现平行的音高轨迹——直接体现了合唱的多声部和声特征。(b) 梅尔频谱图从频谱角度提供了补充证据。独唱区域的能量分布相对稀疏，共振峰结构清晰；而副歌区域在整个频率范围内都呈现出显著更高的能量密度、更丰富的高频谐波成分，以及更连续的时间能量模式。这与多人演唱中声压叠加的声学特征一致，进一步表明Tutti成功建模了合唱中复杂的声学纹理与交互。

自适应歌手提示融合器的作用。 移除此组件会导致多歌手辨识度指标显著下降，同时分数方差明显增大。

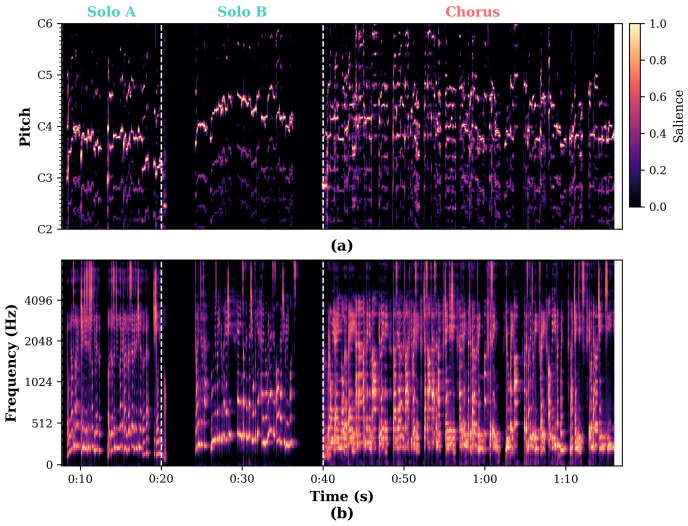


图*4*。生成的多歌手音频的声学可视化分析。(a) 音高显著性图展示了鲜明的旋律模式：独唱部分呈现单一旋律线，而副歌部分则呈现交织的旋律线。(b) 梅尔频谱图展示了时频域中的能量分布，突显了频谱连贯性和丰富的共鸣结构。

这表明，简单的线性加权求和无法处理合唱场景中主唱与和声之间复杂的主次关系。缺少自注意力机制所提供的非线性交互建模后，各声部之间的音色边界会变得模糊，导致生成的合唱片段中的歌手辨识度不稳定。此外，该变体的歌词可懂度也有所下降，部分音节变得模糊不清，这可能是因为缺少有效的歌手调度机制，使音色表征与内容生成之间的耦合不足。此外，定性分析表明，在需要极快歌手切换的场景中（例如，< 0.2s），缺少该模块会导致严重的音色拖尾，即前一位歌手的音色不自然地延续到后续片段中。

互补纹理学习的效果。移除此模块会导致旋律表现指标显著下降，且分数方差大幅增加。这一现象验证了纹理建模对于捕捉合唱中声学细节的重要性。缺少质感特征的引导时，生成音频会缺乏空间混响和频谱融合等声学特性，导致旋律丰富度和真实感下降。值得注意的是，该变体的音频质量指标也有所降低，这表明纹理建模不仅有助于营造整体合唱氛围，也能提升帧级声学质量。有趣的是，该变体的音色一致性略有提高。我们推测，这是因为移除纹理建模会迫使模型在生成时更加依赖明确的歌手条件，从而在音色维度上表现出更强的条件遵循度，但代价是牺牲声学质感丰富度。

5. 局限性与未来工作

我们承认所提出的框架存在局限性。首先，我们的方法依赖于一个强假设，即主歌段落仅包含一位歌手。然而，在现实世界的音乐作品中，主歌可能由多位歌手演唱，也可能包含二重唱。这一简化假设限制了模型处理主歌段落中此类复杂编曲场景的能力。其次，在歌声表征方面，我们发现所采用的特征未能全面涵盖主流歌手的人声特征，因此限制了我们在演唱表现中运用某些特定音色的能力。第三，我们观察到，生成歌声的旋律性仍有很大的提升空间，目前尚不足以呈现真正动人且富有情感共鸣的演唱表现。因此，未来工作将聚焦于探索更高效的歌声特征压缩表征，并致力于还原更具感染力和表现力的演唱表现。

6. 结论

本文介绍了 Tutti，这是首个专为单首歌曲中的动态多歌手编排而设计的生成框架，旨在突破传统歌声合成中独唱者范式的局限。通过引入结构感知歌手提示，我们实现了随音乐结构演变的动态角色调度，使模型能够灵活处理独唱与合唱之间的复杂转换。此外，我们提出了基于条件引导 VAE 的互补纹理学习，借助条件引导重建任务，成功捕捉多声部互动中的复杂声学纹理。实验结果表明，Tutti 不仅在多歌手编排的精准度方面表现出色，还显著提升了合唱生成的声学真实感与艺术凝聚力，为复杂的多歌手编排开辟了新途径。

参考文献

Agostinelli, A., Denk, T. I., Borsos, Z., Engel, J., Verzetti, M., Caillon, A., Huang, Q., Jansen, A., Roberts, A., Tagliasacchi, M, 等。Musiclm：从文本生成音乐。*ArXiv* 预印本，abs/2301.11325，2023 年。网址：https://arxiv.org/abs/2301.11325。

Bain、M.、Huh、J.、Han、T. 和 Zisserman、A. WhisperX：长篇音频的精准计时语音转录。载于 Harte、N.、Carson-Berndsen、J. 和 Jones、G.（编），第*24*届国际语音通信协会年会，*Interspeech 2023*，爱尔兰都柏林，*2023*年*8*月*20*日至*24*日，第4489–4493页。ISCA，2023年。doi：10.21437/INTERSPEECH.2023-78。网址：https://doi.org/10.21437/Interspeech.2023-78。

Dai、S.、Liu、M.-Y.、Valle、R. 和 Gururani、S. Expressivesinger：具有表现力演绎控制的多语言、多风格基于乐谱的歌声合成。载于第*32*届 *ACM* 国际多媒体会议论文集，第3229–3238页，2024年。

Dhariwal, P., Jun, H., Payne, C., Kim, J. W., Radford, A. 和 Sutskever, I. Jukebox：一种音乐生成模型。*ArXiv* 预印本，abs/2005.00341，2020年。网址 https://arxiv.org/abs/2005.00341。

Esser,P., Kulal, S., Blattmann, A., Entezari, R., M¨uller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., Podell, D., Dockhorn, T., English, Z. 和 Rombach, R. 扩展用于高分辨率图像合成的整流流变换器。载于第四十一届国际机器学习会议（*ICML 2024*），奥地利维也纳，*2024*年*7*月*21*日至*27*日。OpenReview.net，2024年。网址 https://openreview.net/forum?id=FPnUhsQJ5B。

Evans, Z., Parker, J. D., Carr, C., Zukowski, Z., Taylor, J. 和 Pons, J. Stable Audio Open。在 *ICASSP 2025-2025 IEEE*声学、语音与信号处理国际会议 *(ICASSP)*，第 1–5 页。IEEE，2025 年。

Guo, W., Zhang, Y., Pan, C., Huang, R., Tang, L., Li, R., Hong, Z., Wang, Y. 和 Zhao, Z. TechSinger：通过流匹配实现技术可控多语言歌声合成。在 *AAAI*人工智能会议论文集，第 39 卷，第 23978–23986 页，2025年。

Huang, R., Zhang, C., Wang, Y., Yang, D., Liu, L., Ye, Z., Jiang, Z., Weng, C., Zhao, Z. 和 Yu, D. Make-a-voice：基于离散表示的统一语音合成。*ArXiv* 预印本，abs/2305.19269，2023年。URLhttps://arxiv.org/abs/2305.19269。

Kilgour、K.、Zuluaga、M.、Roblek、D. 和 Sharifi、M.《Fr》´*echet* 音频距离：一种用于评估音乐增强算法，pp 2350 . – 2354。ISCA，2019。doi：10.21437/interspeech. 2019-2219。Kubichek，R. 用于客观语音质量评估的梅尔倒谱距离度量。载于*IEEE*太平洋地区通信、计算机与信号处理会议论文集，第125–128页。IEEE，1993。doi：10.1109/pacrim.1993.407206。

Lei，S.、Xu，Y.、Lin，Z.、Zhang，H.、Tan，W.、Chen，H.、Yu，J.、Zhang，Y.、Yang，C.、Zhu，H.、Wang，S.、Wu，Z. 和 Yu，D. Levo：通过多偏好对齐实现高质量歌曲生成。*ArXiv* 预印本 *arXiv:2506.07520*，2025年。

Lipman，Y.、Chen，R. T. Q.、Ben-Hamu，H.、Nickel，M. 和 Le，M. 用于生成建模的流匹配。载于第十一届国际学习表征会议，*ICLR 2023*，卢旺达基加利，*2023*年*5*月*1*日至*5*日。OpenReview.net，2023年。网址：https://openreview.net/pdf?id=PqvMRDCJT9t。

Liu, H., Chen, Z., Yuan, Y., Mei, X., Liu, X., Mandic, D. P., Wang, W., and Plumbley, M. D. Audioldm：基于潜扩散模型的文本到音频生成。载于 Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J.（编），国际机器学习会议，*ICML 2023*，*2023*年*7*月*23*日至*29*日，檀香山，夏威夷，美国，第 202 卷，机器学习研究论文集，第 21450–21474 页。PMLR，2023。网址：https://proceedings.mlr.press/v202/liu23f.html。

Liu, J., Li, C., Ren, Y., Chen, F., and Zhao, Z. Diff-singer：通过浅层扩散机制实现的歌声合成。载于第三十六届 *AAAI* 人工智能会议，*AAAI 2022*，第三十四届人工智能创新应用会议，*IAAI 2022*，第十二届人工智能教育进展研讨会，*EAAI 2022* 虚拟会议，*2022*年*2*月*22*日至*3*月*1*日，第 11020–11028 页。AAAI出版社，2022。网址 https://ojs.aaai.org/index.php/AAAI/article/view/21350。

Loshchilov, I. 和 Hutter, F. 解耦权重衰减正则化。载于 第七届国际学习表征会议，*ICLR 2019*，美国路易斯安那州新奥尔良，*2019*年*5*月*6*日至*9*日。OpenReview.net，2019年。网址 https://openreview.net/forum?id=Bkg6RiCqY7。

Lu, W. T., Wang, J., Kong, Q., 和 Hung, Y. 基于频带分割 RoPE Transformer 的音乐源分离。载于*IEEE* 声学、语音与信号处理国际会议（*ICASSP 2024*），*2024*年*4*月*14*日至*19*日，大韩民国首尔，第 481–485 页。IEEE，2024年。doi：10.1109/ICASSP48485.2024。

10446843.
URLhttps://doi.org/10.1109/ICASSP48485.2024.10446843。

Nguyen, T. A., Kharitonov, E., Copet, J., Adi, Y., Hsu, W.-N., Elkahky, A., Tomasello, P., Algayres, R., Sagot, B., Mohamed, A., and Dupoux, E. 生成式口语对话语言建模。计算语言学协会学报，11:250–266，2023年。doi：10.1162/tacl_a_00545。URL https://aclanthology.org/2023.tacl-1.15。

Ning, Z., Chen, H., Jiang, Y., Hao, C., Ma, G., Wang, S., Yao, J., and Xie, L. DiffRhythm：借助潜在扩散实现快速且极其简洁的端到端全长歌曲生成。*ArXiv* 预印本，abs/2503.01183，2025年。URL https://arxiv.org/abs/2503.01183。

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., and Sutskever, I. 通过大规模弱监督实现鲁棒语音识别。载于 国际机器学习会议，第28492–28518页。PMLR，2023年。Rix, A., Beerends, J., Hollier, M., and Hekstra, A. 语音质量的感知评估（PESQ）——一种用于电话网络和编解码器语音质量评估的新方法。载于 *2001 IEEE* 国际声学、语音与信号处理会议论文集（目录号：*01CH37221*），第2卷，第749–752页，第2卷，2001年。doi：10.1109/ICASSP.2001.941023。Steinmetz, C. J. and Reiss, J. D. auraloss：PyTorch 中的音频专用损失函数。载于 数字音乐研究网络一日研讨会（*DMRN+ 15*），第124页，2020年。Taal, C. H., Hendriks, R. C., Heusdens, R., and Jensen, J. 短时客观可懂度测量：用于时频加权噪声语音。载于 *2010 IEEE* 国际声学、语音与信号处理会议，第4214–4217页，2010年。doi：10.1109/ICASSP.2010.5495701。Tan, W., Lei, S., Zhang, H., Li, G., Zhang, Y., Chen, H., Yu, J., Gu, R., and Yu, D. SongPrep：用于整首歌曲结构解析和歌词转录的预处理框架及端到端模型，2025年。URL https://arxiv.org/abs/2509.17404。Vyas, A., Shi, B., Le, M., Tjandra, A., Wu, Y.-C., Guo, B., Zhang, J., Zhang, X., Adkins, R., Ngan, W., et al. Audiobox：通过自然语言提示词实现统一音频生成。*ArXiv* 预印本，abs/2312.15821，2023年。URLhttps://arxiv.org/abs/2312.15821。

Wang, H., Zheng, S., Chen, Y., Cheng, L., and Chen, Q. CAM++：一种使用上下文感知掩码进行说话人验证的快速高效网络。载于 Harte, N., Carson-Berndsen, J.与Jones, G.（编），国际语音通信协会第*24*届年会，*Interspeech 2023*，爱尔兰都柏林，*2023*年*8*月*20*日至*24*日，第5301–5305页。ISCA，2023年。doi: 10.21437/ INTERSPEECH.2023-1513。URL https://doi.org/10.21437/Interspeech.2023-1513。

Wang, P.、Lu, S.、Tang, Y.、Yan, S.、Xia, W.与Xiong, Y. 基于大语言模型的全双工语音对话方案。载于第*38*届神经信息处理系统国际会议论文集，NIPS '24，美国纽约州雷德胡克，2024年。Curran Associates Inc. ISBN 9798331314385。Xie, H.、Guo, D.、Wang, C.、Li, Y.、Tian, W.、Zhu, X.、Wang, X.、Li, X.、Miao, G.、Liu, B.等。DialoSpeech：基于大语言模型与流匹配的双说话人对话生成。*ArXiv*预印本，abs/2510.08373，2025年。URLhttps://arxiv.org/abs/2510.08373。Yang, C.、Wang, S.、Chen, H.、Tan, W.、Yu, J.与Li, H. Songbloom：通过交错自回归草拟与扩散精炼实现连贯歌曲生成。*arXiv*预印本 *arXiv:2506.07634*，2025年。Yao, J.、Ma, G.、Xue, H.、Chen, H.、Hao, C.、Jiang, Y.、Liu, H.、Yuan, R.、Xu, J.、Xue, W.等。SongEval：用于歌曲美学评估的基准数据集。*ArXiv*预印本，abs/2505.10793，2025年。URLhttps://arxiv.org/abs/2505.10793。Yu, D.、Yao, K.、Su, H.、Li, G.与Seide, F. 面向改进大词汇量语音识别的KL散度正则化深度神经网络适应。载于*2013*年*IEEE*声学、语音与信号处理国际会议，第7893–7897页。IEEE，2013年。Yu, F.、Wang, T.、Wu, Y.、Zhu, L.、Deng, W.、Han, W.、Wang, W.、Hu, L.、Liang, X.、He, X.等。JoyVoice：面向拟人化多说话人对话合成的长上下文条件控制。*ArXiv*预印本，abs/2512.19090，2025年。URLhttps://arxiv.org/abs/2512.19090。Zeng, C.、Wang, C.、Miao, X.、Zhao, J.、Jiang, Z.与Chen, Y. InstructSing：通过自我指令实现高保真歌声生成。载于*2024*年*IEEE*口语语言技术研讨会（*SLT*），第675–681页。IEEE，2024年。Zhang, L.、Qian, Y.、Zhou, L.、Liu, S.、Wang, D.、Wang, X.、Yousefi, M.、Qian, Y.、Li, J.、He, L.、Zhao, S.与Zeng, M. CoVoMix：推进面向类人多说话人对话的零样本语音生成。载于Globersons, A.、Mackey, L.、Belgrave, D.、Fan, A.、Paquet, U.、Tomczak, J. M.与Zhang, C.（编），神经信息处理系统进展第*38*卷：*2024*年神经信息处理系统年度会议，*NeurIPS*

*2024*年，加拿大不列颠哥伦比亚省温哥华，*2024*年*12*月*10*日至*15*日，2024a。

Zhang, L., Qian, Y., Wang, X., Thakker, M., Wang, D., Yu, J., Wu, H., Hu, Y., Li, J., Qian, Y., et al. Cov- omix2：以完全非自回归流匹配推进零样本对话生成。*ArXiv* 预印本，abs/2506.00885，2025a。网址https://arxiv.org/abs/2506.00885。

Zhang, X., Zhang, J., Wang, Y., Wang, C., Chen, Y., Jia, D., Chen, Z., and Wu, Z. Vevo2：通过统一韵律学习衔接可控语音与歌声生成。*arXiv* 电子预印本，第 arXiv–2508 页，2025b。

Zhang, X., Zhang, X., Peng, K., Tang, Z., Manohar, V., Liu, Y., Hwang, J., Li, D., Wang, Y., Chan, J., Huang, Y., Wu, Z., and Ma, M. Vevo：通过自监督解耦实现可控的零样本语音模仿。收录于*ICLR*。OpenReview.net，2025c。

Zhang, Y., Cong, J., Xue, H., Xie, L., Zhu, P.和Bi, M. Visinger：面向端到端歌声合成的变分推断与对抗学习。见*IEEE*国际声学、语音与信号处理会议，*ICASSP 2022*，线上及新加坡，*2022*年*5*月*23–27*日，第7237–7241页。IEEE，2022。doi: 10.1109/ ICASSP43922.2022.9747664。URLhttps://doi.org/10.1109/ICASSP43922.2022.9747664。

Zhang, Y.、Jiang, Z.、Li, R.、Pan, C.、He, J.、Huang, R.、Wang, C. 与 Zhao, Z. TCSinger：基于风格迁移与多层级风格控制的零样本歌声合成。载于*EMNLP*，2024b。

A. 实现与训练细节

A.1. 模型配置

对于 Vocal VAE，我们采用 Stable Audio 2.0 的架构（Evans 等人，2025年），以 44.1kHz 的采样率运行，潜在维度为 64，与原始配置保持一致。DiT 主干网络实现为 16 层 Transformer，隐藏维度为 2048，具有 32 个注意力头和 4 层前馈网络。关于条件嵌入，各维度设置如下：歌词嵌入为 512，结构标签嵌入为 20，歌手提示嵌入为 192，纹理嵌入为 64，时间嵌入为 512。梅尔频谱的潜在维度与 VAE 的潜在大小 64 相匹配。

A.2. 训练设置

所有模型均使用 16 块 NVIDIA A100 GPU 进行训练。我们始终采用 AdamW 优化器（Loshchilov 和 Hutter，2019）贯穿所有训练阶段。

VAE 训练。对于 Vocal VAE，优化器超参数设置为 $\beta_1 = 0.8$ 和 $\beta_2 = 0.99$。模型采用 32 的批量大小和 1.5×10^{-4} 的初始学习率进行训练，共计 970K 步。在条件引导变分自编码器训练阶段（纹理学习）中，我们将批量大小调整为 8，并将学习率降至 5×10^{-5}，额外训练 200K 步。为促进稳健的纹理特征提取，在此阶段对潜在空间施加信噪比（SNR）约为 10dB 的高斯噪声扰动。

DiT 主干网络训练。DiT 模型采用 4 的批量大小和 1×10^{-5} 的学习率进行训练。我们使用结合线性预热和线性衰减的学习率调度器。训练过程共持续 950K 步。

A.3. 推理配置

推理时，我们采用 Euler 常微分方程求解器生成样本。采样过程共进行 50 个推理步骤。为提高生成质量和条件遵循度，我们使用引导尺度为 4.0 的无分类器引导（CFG）。

B. 主观评估详情

为全面评估生成的歌声，我们组织了 30 名参与者进行主观评估，分为专业组（10 名专家）和普通组（20 名志愿者）。我们遵循标准的平均意见分（MOS）流程，请参与者使用 5 点李克特量表（1：差，2：较差，3：一般，4：良好，5：优秀）为样本评分，并以 0.5 分为增量。

专业组成员来自音乐学院和专业艺术团体。鉴于声乐表演需要长期的生理训练，我们对成员的领域专业知识制定了严格的筛选标准。如表3所示，专业组涵盖三种主要演唱风格：美声唱法、民族唱法和流行/爵士，并邀请合唱指挥专家专门评估多人和声。这些专家的平均训练年限为13.7年。普通组由20名背景各异、听力正常的志愿者组成，代表普通听众的视角。他们受邀重点评估歌曲的整体自然度和旋律。

表3. 10 名专业评估者的人口统计资料。

ID	专业/职业	学术身份	训练年限	专业领域
P01	美声唱法（女高音）	博士候选人	18	歌剧表演、共鸣控制
P02	美声唱法（男高音）	专业人士	15	声乐技巧、舞台表演
P03	中国民族声乐	硕士生	12	民族风格、吐字、地域特色
P04	流行声乐表演	指导教师	20	流行风格、麦克风技巧、表现力
P05	爵士声乐	高年级本科生	10	即兴创作、节奏、爵士和声
P06	合唱指挥	专业人士	16	合唱平衡、声部融合、排练
P07	合唱指挥	硕士生	13	音准、多声部协调
P08	声乐创作	博士候选人	15	声乐写作、对位法、和声
P09	声乐教学法	硕士生	11	嗓音生理学、训练方法
P10	音乐制作（人声）	自由职业	9	人声混音、音色质感处理

C. 纹理学习中无泄漏的验证

为验证 Conditio n 引导的 VAE 学习到的纹理表示独立于旋律内容、歌词信息和说话人身份实体，我们进行了一项系统性的纹理交换实验。

C.1. 实验设置

我们随机选取了 50 个生成样本进行评估。对于每个样本，我们将随机种子固定为 42，并通过生成不含任何纹理参考的音频建立基线（记为“无纹理”）。随后，我们引入 10 个不同的参考音频片段作为纹理输入，这些片段经过随机挑选，覆盖了多种声学场景，包括独唱表演、二重唱和多歌手合唱，并包含风格各异的旋律和歌词。此设置共生成 500 个变体。我们将这些带纹理的生成结果与基线进行比较，以评估旋律一致性，并计算标准客观指标，以衡量内容和身份稳定性。

C.2. 结果与分析

表4。 纹理学习中无泄漏的验证。我们报告了使用 10 种不同纹理参考生成的全部 500 个样本的平均指标。基频指标衡量其相对于“无纹理”基线的偏差，而词错误率（WER）和相似度（SIM）则评估绝对表现。

指标	值
旋律一致性（相较于无纹理）	
F0均方根误差（音分）↓	7.85
基频平均绝对误差（音分）↓	6.21
相关性 ↑	0.9997
内容与身份（绝对值）	
词错误率 ↓	13.53%
SIM ↑	0.690

如表 4所示，引入多样化的纹理参考对生成音频的显式属性影响微乎其微。带纹理生成结果的基频轮廓与无纹理基线高度一致，均方根误差小于 8 音分，相关性接近完美，达到 0.9997，这表明纹理编码器不会泄漏旋律信息。此外，词错误率（13.53%）和说话人相似度（0.690）与表2 中报告的主要结果（分别为 13.50% 和 0.691）相当，证实纹理表征能够捕捉互补的声学属性，同时不会干扰显式条件所定义的语言内容或歌手身份。

这证实纹理表征能够捕捉互补的声学属性，同时不会干扰显式条件所定义的语言内容或歌手身份。

D. 自适应歌手提示融合器机制分析

为探究自适应歌手提示融合器如何处理多歌手场景，并验证其能否有效汇集多歌手信息而不丢失各自特征，我们对融合器模块中的注意力权重进行了可视化。

图 5 展示了注意力机制的可视化结果。图5（a）展示了三个候选歌手（歌手A、B、C）在四个结构片段中获得的注意力权重。在独唱部分（主歌1、主歌2和主歌3）中，融合器呈现出高度选择性的注意力模式，为特定独唱歌手分配主导权重（4.00），同时抑制其他歌手（0.00）。这证实了该模型精准进行歌手调度的能力。尤为重要的是，在多歌手段落（副歌1）中，注意力分布于三位歌手之间（1.09、1.18、1.73）。这表明，最终的歌手嵌入并非坍缩为单一身份，而是一种加权复合表示，保留了所有参与者的声学特征。

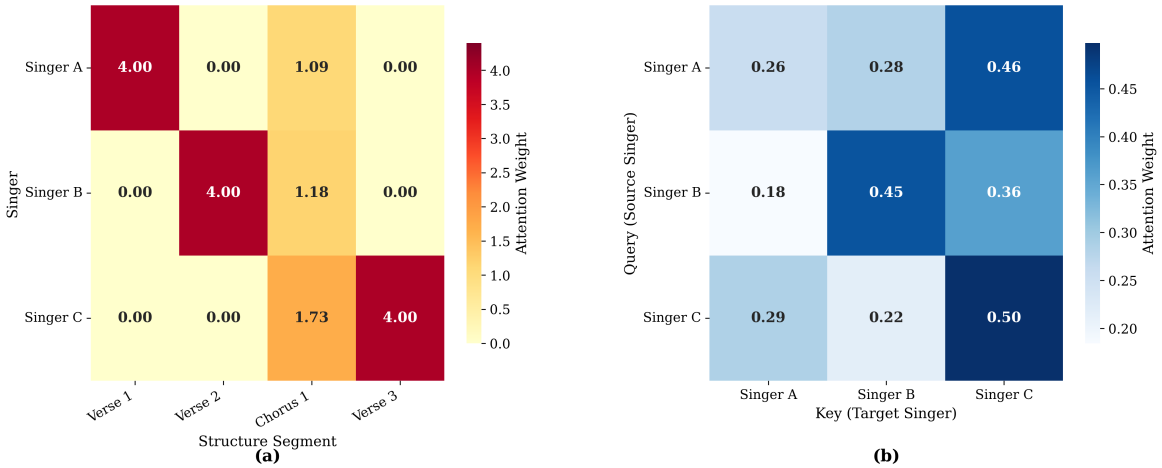


图*5.* 自适应歌手提示融合器中注意力权重的可视化。(a) 歌手注意力矩阵展示了候选歌手在不同结构片段中的权重分布。(b) 歌手间交叉注意力矩阵详细展示了多歌手副歌1片段内部的交互权重。

此外，图 5（b）详细展示了副歌1片段内部的自注意力机制，揭示了歌手之间的交互逻辑。该矩阵显示出不均匀的注意力权重，表明这种融合并非简单平均。例如，歌手A（作为查询）会显著关注歌手C（0.46）和歌手B（0.28），而非仅关注自身。这说明融合器对歌手之间的和声关系与音色融合进行建模，有效地将“合唱质感”编码到单个嵌入中。该机制确保即使压缩为单一向量，该表示仍能封装多种声音之间复杂的交互，使后续生成模型能够准确解码并重建多歌手演唱。

E. 主歌段落中歌手一致性的数据分析

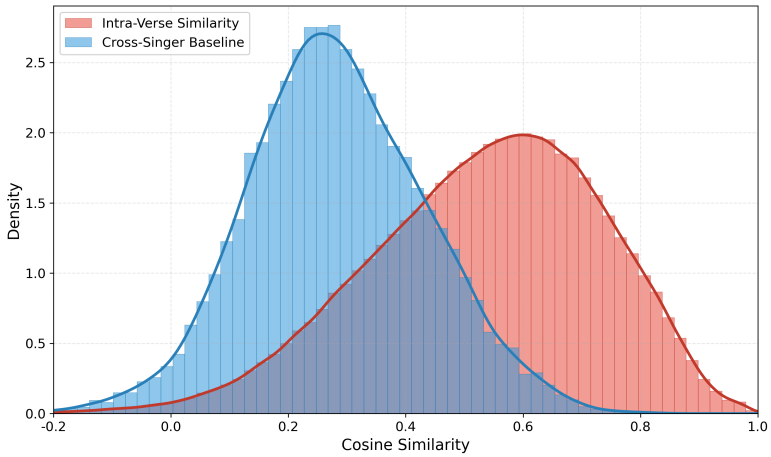


图*6.* 主歌内片段（红色）与跨歌手基线（蓝色）的余弦相似度分布。两种分布之间的显著差异支持了以下假设：主歌段落主要由单人声演唱者演唱。

为通过实证验证主歌段落通常由单人声演唱者演唱这一假设（第 3.2 节），我们针对主歌片段中的歌手一致性开展了大规模统计分析。

E.1. 实验设置

我们从数据集中随机抽取了约 10,000 首歌曲，确保每个样本都包含有效的结构标签，并通过语音活动检测。我们设计了两个对比组，以分析歌手相似度的分布：

诗节内相似度： 对于每首歌曲，我们从同一诗节部分中提取了多个非静音片段（每个片段约 3 秒，重叠 1.5 秒）。随后，我们计算了这些片段的说话人嵌入之间的余弦相似度，共得到 323,996 对。该组体现了单个主歌内的歌手一致性。

跨歌手基线： 我们从不同歌曲中随机配对片段，以模拟不同歌手之间的相似度分布。该组由 100,000 对配对组成，并作为非同一身份匹配的基线。

E.2. 统计证据与讨论

图 6 展示了两组余弦相似度的概率密度函数。定量结果显示出两种分布之间存在明显差异。具体而言，诗节内配对的相似度显著更高，均值为 0.5355，中位数为 0.5526（标准差=0.1973）。相比之下，跨歌手基线的分布明显更低，均值为 0.2835，中位数为 0.2778（标准差=0.1515）。

尽管歌唱领域中同一歌手的绝对相似度分数（约为 0.54）低于语音任务中通常观察到的分数（通常为 > 0.7）——这一现象归因于歌唱的声学变异性（例如音高变化、发声变化）以及语音预训练 CAM++ 模型的领域差距——但两者之间的相对差距仍然显著。诗节内配对与跨歌手配对之间显著的分布差距，从经验上支持了我们的假设：诗节部分主要由单人声演唱者演唱。此外，这种分布也证明了我们在数据处理流程中选择阈值（例如 0.4）的合理性；该阈值能够有效地区分大多数同歌手片段与不同歌手配对。