任何伴生：通过量化旋律瓶颈实现通用伴奏生成

Junan Zhang* Yunjia Zhang* Xueyao Zhang Zhizheng Wu

The Chinese University of Hong Kong, Shenzhen

ABSTRACT

歌唱伴奏生成（SAG）是为给定的干净人声输入生成器乐音乐的过程。然而，现有的 SAG 技术使用源分离的人声作为输入，并过度拟合于分离伪影。这导致了关键的训练和测试不匹配问题，使得在干净的真实世界人声输入上失败。我们引入了任意伴随 ，这是一个通过将伴奏生成与源依赖性伪影解耦来解决这一问题的框架。任何伴生 首先采用量化旋律瓶颈，使用音级图和 VQ-VAE 提取核心旋律的离散且不随音色变化的表现形式。随后的流匹配模型则基于这些稳健代码生成伴奏。实验显示任意伴随 在分离人声基准测试中表现出具有竞争力的性能，同时在干净工作室人声和显著地独奏乐器曲目的泛化测试集上大大优于基线模型。这展示了泛化的质变，实现了对乐器的稳健伴奏——这是现有模型完全失败的任务，并为更灵活的音乐共创工具铺平了道路。演示音频和代码：https://anyaccomp.github.io/

Index Terms— 音乐生成，伴唱生成，矢量量化，流匹配

1. 介绍

歌唱伴奏生成（SAG）是音乐创作中的一个关键任务，旨在生成与给定的声乐旋律在和声和节奏上相辅相成的器乐音乐 [1,2]。高保真的 SAG 模型拥有巨大的潜力，能够革新音乐创作过程，为艺术家提供强大的协作创作工具，使制作人能够快速原型化想法，并赋予业余爱好者将他们的音乐愿景变为现实的能力。

一种用于最先进的 SAG 模型 [1–3] 的常见训练范式依赖于通过音乐源分离（MSS）算法 [4, 5] 提取的语音数据。条件信号通常是直接的声学表示，例如 FastSAG [2] 中使用的梅尔频谱图，或是来自强大的自监督学习（SSL）模型的功能，这种方法由 SingSong [1] 以一个定制的 w2v-BERT [6] 引领。然而，无论是梅尔频谱图还是通用的 SSL 特征都旨在保留丰富的声学细节。因此，在使用分离不完美的数据进行训练时，它们不可避免地学习到一种人为关联，将有效的伴奏生成与分离伪影的存在联系起来。虽然以前的工作 [1,2] 试图通过注入白噪声来缓解这个问题，但这可能会引入新的、不现实的伪影。这导致了一个关键问题 **训练-测试不匹配**：在录音干净的真实世界场景中，这些模型会失败，因为它们缺少训练时所期待的那种“提示”。对不完美数据的依赖极大地限制了当前 SAG 系统的鲁棒性和实际应用性。

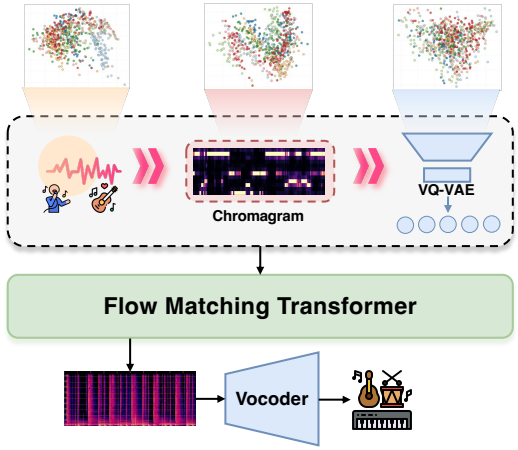


Fig. 1: 任何伴生 的概述。该过程包括两个主要阶段：（1）输入音频通过量化旋律瓶颈进行处理，在此过程中，VQ-VAE 将其色谱图编码为一系列离散标记。(2) 流匹配变换器根据这些离散标记生成一个梅尔频谱图，随后由声码器将其合成为最终的伴奏音频。

在本文中，我们将 SAG 任务视为一个条件生成问

*Equal Contribution.

题，认为不匹配源自于条件表示对源特定特征的敏感性。因此，我们提出了一种设计来满足两个关键属性以实现鲁棒性的表示：**音色不变性**用于过滤出依赖于源的纹理，如音色或分离伪影，而**旋律聚类性**则以紧凑且结构化的方式保留了基本的旋律内容。这提供了一个统一的条件，解决了训练-测试不匹配问题，使得仅在人声数据上训练的模型能够稳健地泛化到干净的人声甚至乐器音轨。

为此，我们提出了一种新颖的两阶段框架任何伴生体 ，用于稳健的伴奏生成。任何伴生物 的第一阶段是一个瓶颈模块，它通过使用向量量化变分自编码器（VQ-VAE）[7] 将输入音频的时间不变色度图量化为一系列离散代码来提取泛化的旋律表示。这一过程有效地隔离了核心的旋律内容，形成了一个对来源依赖性伪影具有鲁棒性的表示。这种表示的有效性在图 2 中得到了实证证明，该图显示我们的代码实现了比传统梅尔频谱图显著更清晰的旋律聚类能力和时间不变性。

在第二阶段，一个 Flow-Matching Transformer [8] 根据这些稳健且通用的表示生成伴奏。通过将生成过程与保持人工制品的表示解耦，任意伴生 避免了限制现有方法的分离人工制品的相关性。我们的实验验证了这种方法：任何伴影 不仅在领域内分离人声基准上表现出色，还在具有干净录音室人声和重要的是独奏乐器曲目的泛化测试集上显著优于基线模型。这一成功表明我们的方法有效地解决了关键的训练-测试不匹配问题。这项工作的主要贡献如下：

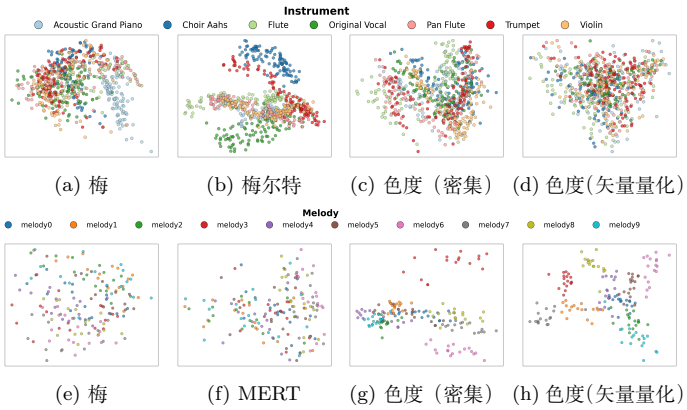
- 我们通过使用音色图和 VQ-VAE 瓶颈引入了一种量化的旋律表示，这种表示方法对音色不变且能抵御分离伪影。
- 我们提出了任意伴生 ，这是一个两阶段的框架，通过流匹配变换器将伴奏生成与易产生伪影的输入解耦，并展示了其在解决训练-测试不匹配方面的能力，优于基线模型，在干净的录音棚人声和独奏乐器轨道上表现更优。

2. 任何伴生物 框架

2.1. 量化旋律瓶颈

我们的方法核心是学习一种中间表示，这种表示能够捕捉到基本的旋律内容，同时丢弃不相关的音色信息和源分离伪影。为了获得这种泛化表示，我们采用了一种在 50 Hz 频谱图输入 $\boldsymbol{x}$ 上运行的向量量化变分自动编码器（VQ-VAE）。模型的编码器将频谱图映射到一个连续的潜在表示 $\boldsymbol{z}_e(\boldsymbol{x})$，然后通过 $k = \arg\min_j \|\boldsymbol{z}_e(\boldsymbol{x}) - \boldsymbol{e}_j\|_2$ 从学习到的码本 $\boldsymbol{E}$ 中将其量化为其最近邻 $\boldsymbol{e}_k$。接着解码器从得到的量化序列 $\boldsymbol{z}_q(\boldsymbol{x})$ 中重构频谱图 $\hat{\boldsymbol{x}}$。模型被训练以最小化一个联合损失函数：

$$\mathcal{L} = \|\boldsymbol{x} - \hat{\boldsymbol{x}}\|_2^2 + \beta\|\boldsymbol{z}_e(\boldsymbol{x}) - \boldsymbol{z}_q(\boldsymbol{x})\|_2^2 \qquad (1)$$



(a) 梅 (b) 梅尔特 (c) 色度（密集） (d) 色度(矢量量化)

(e) 梅 (f) MERT (g) 色度（密集） (h) 色度(矢量量化)

Fig. 2: 不同表示的可视化。顶部一行评估**音色不变性**，其中更好的表示显示出乐器颜色更加充分地混合。底部一行评估**旋律聚类性**，其中更好的表示为每种旋律颜色形成更紧密、更分明的聚类。

此量化过程迫使模型捕捉关键的旋律特征同时丢弃伪影，在生成阶段形成一个强大的瓶颈。

为了评估我们表示方法的有效性，我们在 M4Singer 数据集的旋律上进行了一次可视化实验 [9]。对于每个旋律，我们使用原始人声发音，并通过使用 pretty_midi[1] 从其相应的 MIDI 数据合成几个乐器版本来增强它。这创建了一个受控集合，在该集合中旋律保持不变而音色有所变化。从此综合收集的数据

[1] https://github.com/craffel/pretty-midi

中，我们提取并比较了四种表示：标准的 mel 谱图、来自预训练的 MERT 模型的特征 [10]、密集的音阶图和我们的 VQ 量化音阶图。图 2 可视化了它们 PCA 降维后的嵌入来评估两个关键属性。

音色不变性。一种鲁棒的表示必须对来源音色不敏感，这是泛化的重要属性。如图 2 的顶部所示，保留丰富声学特征的表示无法通过此测试。传统的梅尔谱图（图 2a）表现出强烈的乐器类型聚类现象。这种效果在 MERT 表示（图 2b）中更加显著，因为其预训练目标设计为捕捉细粒度声学特征。密集的音色图（图 2c）通过打破这些明显的集群显著减轻了这一偏差。然而，我们的最终向量量化音色图表示（图 2d）达到了理想的结果：所有乐器的点在特征空间中彻底混合，展示了旋律内容与来源音色的成功分离。

旋律聚类能力。有效的表示必须将相同的旋律映射到特征空间中的紧凑区域，无论使用何种乐器。图 2 的底部行说明了一个明显的进展。梅尔频谱图（图 2e）和 MERT 表示（图 2f）表现不佳，未能形成连贯的旋律聚类。密集音色图（图 2g）显示出显著改进，不同的旋律聚类变得可见。我们的最终向量量化音色图表示（图 2h）通过生成特别紧密且区分明显的聚类进一步优化了这一点，证明量化不仅保留而且增强了旋律结构，为生成模型创造了高度稳健和离散的条件。

2.2. 流匹配伴奏生成

在第二阶段，一个流匹配（FM）Transformer [8] 根据离散的旋律代码生成伴奏。至关重要的是，该策略将生成过程与输入源伪影解耦，确保输出完全基于音乐旋律和节奏。

我们将和声生成公式化为一个条件流匹配问题，目标是学习一条连续时间的概率路径 $p_t(\boldsymbol{x})$，将一个简单的先验分布转换为目标数据分布。令 $\boldsymbol{x}_1$ 表示目标和声梅尔频谱图。该路径从 $\boldsymbol{x}_0 \sim \mathcal{N}(0, \boldsymbol{I})$ 开始，这是一个标准高斯分布的样本。我们定义这两点之间的轨迹为：

$$\boldsymbol{x}_t = (1 - (1 - \sigma)t)\boldsymbol{x}_0 + t\boldsymbol{x}_1 \tag{2}$$

其中 $t \in [0, 1]$ 是时间步长，σ 是一个小常数（例如，10^{-5}）。在训练过程中，时间步长 t 是通过首先抽取 $t' \sim U[0, 1]$，然后应用 $t = 1 - \cos(t' \cdot \frac{\pi}{2})$ 来从余弦调度中采样，以使采样密度集中在轨迹的开始部分。

模型被训练以预测生成此路径的矢量场 $\boldsymbol{v}_t$。真实的速度是 $\boldsymbol{x}_t$ 的时间导数：

$$\boldsymbol{v}_t = \frac{d\boldsymbol{x}_t}{dt} = \boldsymbol{x}_1 - (1 - \sigma)\boldsymbol{x}_0 \tag{3}$$

我们的模型 f_θ，在离散旋律代码 $\boldsymbol{c}$ 条件下，学习近似这个速度场。训练目标是一个简单的均方误差损失：

$$\mathcal{L}_{FM}(\theta) = \mathbb{E}_{t, \boldsymbol{x}_1, \boldsymbol{x}_0, \boldsymbol{c}} \left\| f_\theta(\boldsymbol{x}_t, t, \boldsymbol{c}) - \boldsymbol{v}_t \right\|^2 \tag{4}$$

我们增加了一个表示对齐（REPA）损失 [11]，它将一个中间 FM Transformer 层与预训练音乐模型的表示 [10] 进行对齐。最终目标是两个损失的加权和：$\mathcal{L} = \mathcal{L}_{FM} + \lambda \mathcal{L}_{REPA}$。

在推理过程中，我们从一个随机噪声样本 $\boldsymbol{x}_0$ 开始生成伴奏梅尔频谱图 $\boldsymbol{x}_1$，并使用前向欧拉方法将预测的速度 $\hat{\boldsymbol{v}}_t = f_\theta(\boldsymbol{x}_t, t, \boldsymbol{c})$ 从 $t = 0$ 积分到 $t = 1$：$\boldsymbol{x}_{t+h} = \boldsymbol{x}_t + h\hat{\boldsymbol{v}}_t$，其中 h 是作为专用于积分的总时间步长的逆来计算的步长大小。为了利用无分类器引导（CFG）[12]，在训练过程中以 0.1 的概率随机丢弃旋律条件 $\boldsymbol{c}$。

3. 实验

3.1. 实验设置

训练数据。我们准备了 8k 小时的配对歌声-伴奏数据来训练任何伴生物 ，遵循 SingNet 流程 [13]。数据来源于互联网上的野生歌曲，分离成声乐和伴奏，然后切分成 3 秒到 30 秒的片段。所有音频以 24 kHz 的采样率进行处理。

实现细节。我们的 VQ-VAE（44M 参数），改编自 Amphion [14, 15]，将 24 个频段的色谱图量化为一个具有 512 项代码簿的 50 Hz 序列，并使用大小为 200 秒的小批量训练了 0.5M 步。Flow-Matching（FM）Transformer 基于 Vevo [16, 17]，由 10 个隐藏维度为 1024 的 LLaMA 解码器层组成，总计 220M 参数。我们在音乐数据上微调来自 Vevo 的 vocoder 以在我们的系统中使用。FM 模型每 GPU 小批量大小为 100 秒，训练了 1M 步，并在 $\lambda = 0.5$ 重量的第 4 层中采用 REPA 损失，将 MERT-330M [10] 作为对齐目标。所有模型都使用 AdamW [18]（学习率 1e-4，32k 预热步长）进

行优化，并在单个 GPU 上训练。在推理过程中，我们使用 50 个采样步骤和 CFG 比例为 3。

比较方法。为了评估我们的旋律瓶颈，我们将任意伴音 与使用声乐梅尔频谱图的最先进非自回归模型 FastSAG [2] 进行比较，并且还对比了两个共享任何伴生物 的 FM Transformer 架构的受控变体：FM-Mel，基于添加了白噪声（信噪比为 15-20 分贝）的梅尔频谱图进行条件处理，以及 FM-Chroma，基于原始的 24 位音调图进行条件处理。任意伴生 则使用量化色度图代码。

Table 1: 客观评估结果显示，任何伴生体 在域内分离人声（YuE）方面具有竞争力，同时在泛化到干净的人声（MUSDB18）和乐器（MoisesDB）方面显著优于基线。

模型	APA↑	**脂肪酸脱氢酶**↓	CE↑	CU↑	PQ↑	**个人电脑**
余音（领域分离人声）						
Ground Truth	-	-	7.270	7.784	7.734	5.752
FastSAG	0.444	0.598	6.351	6.821	6.814	6.321
FM-Mel	0.806	0.416	6.964	7.725	7.758	5.614
FM-Chroma	0.633	0.418	7.151	7.801	7.909	5.436
任何伴生	0.713	0.414	7.283	7.903	7.989	5.742
MUSDB18（**干净人声**）						
Ground Truth	-	-	7.164	7.616	7.485	5.957
FastSAG	0.000	1.115	4.853	5.789	6.315	5.778
FM-Mel	0.167	0.999	5.202	6.616	6.841	4.090
FM-Chroma	0.704	0.798	7.017	7.598	7.744	5.104
任意伴生	0.710	0.788	7.277	7.804	7.891	5.498
莫伊塞斯 DB（**仪器**）						
Ground Truth	-	-	7.236	7.791	7.778	5.694
FastSAG	0.000	0.904	5.966	6.507	6.696	5.952
FM-Mel	0.000	0.936	5.424	6.923	7.151	3.804
FM-Chroma	0.157	0.849	6.308	7.377	7.508	4.104
任何伴生	0.203	0.890	6.660	7.581	7.581	4.782

评估数据集。我们在三个不同的 10 秒片段数据集上评估我们的模型以测试不同能力。对于领域内性能，我们使用来自 YuE 数据的 3,000 个分离的伴唱对 [19]。为了评估对无瑕疵音频的泛化能力，我们使用了 MUSDB18 测试集中的 2,777 个干净人声轨道 [20]。最后，为了推动泛化的边界，我们在 MoisesDB 的 2,500 个独奏乐器曲目上进行评估 [21]。

评估指标。我们使用客观和主观指标来评估伴奏。客观衡量标准包括 Fréchet Audio Distance (FAD) [22] 用于分布相似性，Accompaniment Prompt Adherence (APA) [23] 用于条件对齐，以及 audiobox-aesthetics [24] 在内容享受 (CE)、内容实用性 (CU)、制作复杂度 (PC) 和生产质量 (PQ) 上的得分。对于主观评估，我们在每个测试集上运行了 20 个样本的 MOS 测试，其中听众对整体质量和一致性进行了 1 至 5 分的评分。

3.2. 结果

表 1 中的目标结果证实了我们的瓶颈表示在鲁棒泛化中的关键作用。虽然任意伴生 在域内 YuE 数据集上具有竞争力，但当训练-测试不匹配变得至关重要时，它的真正优势才显现出来。在干净的 vocals（MUSDB18）上，由 mel-频谱图条件化的模型（FastSAG、FM-Mel）遭遇了灾难性的崩溃，APA 分数下降至 0。这个分数表明条件生成完全失败，因为生成的伴奏与输入旋律没有任何关联——这是严重过拟合到分离伪影的直接后果。这种失败在 MoisesDB 中更离域的乐器轨道上更为明显。

相比之下，任何伴生物 通过其量化的旋律瓶颈，保持了在所有泛化集合中的高度一致性和质量。这证明了它卓越的泛化能力（由图 2 中展示的音色不变性和旋律聚类能力验证），使其能够克服训练-测试不匹配，并稳健地为其他模型完全失败的新音色生成伴奏。我们还注意到，生产复杂度（PC）衡量的是纹理复杂度，而不是感知质量，因此不是一个“越高越好”的指标。

如图 3 所示的案例研究，FM-Mel 遭受严重的频谱泄漏，直接复制了输入人声中的伪影。相比之下，任何伴生物 生成了一个连贯的伴奏，突出了其对抗过拟合的强大性能。

主观听音结果（表 2）证实了我们的客观发现。听众在所有数据集中都显著地将任何伴生 的品质和连贯性评为高于 FastSAG。至关重要的是，任何伴生物在具有挑战性的 MUSDB18 和 MoisesDB 泛化集上保

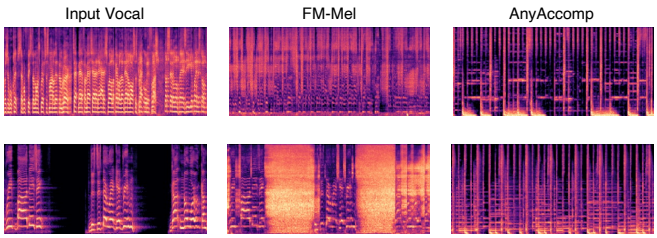


Fig. 3: 一项关于干净 MUSDB18 人声的案例研究。FM-Mel 的输出表现出严重的**谱泄露**输入，这是过度拟合到源分离伪影的一个迹象。相比之下，任何伴生生成了一个连贯的乐器伴奏。

Table 2: 三个测试集的主观评估结果。质量指的是整体伴奏质量，而连贯性衡量的是伴奏与输入的匹配程度。

模型	**质量↑**	**一致性↑**
宇月（领域分离人声）		
Ground Truth	3.92	3.88
FastSAG	1.98	1.82
任意伴生	3.12	3.05
MUSDB18（**干净人声**）		
Ground Truth	3.65	3.48
FastSAG	1.73	1.48
任何伴生	3.23	2.75
莫伊塞斯 DB（**仪器**）		
Ground Truth	4.05	4.08
FastSAG	1.62	1.52
任何伴奏	3.00	2.70

持高分，验证了我们的客观改进转化为更优质且更稳健的听音体验。

4. 结论

在这项工作中，我们提出了任何伴生 ，一个通过使用量化色谱瓶颈来分离生成过程与源分离伪影和音色变化的框架，从而解决 SAG 模型中的关键训练-测试不匹配问题。我们的模型在领域内表现出竞争性能，同时在推广到干净的人声和未见过的独奏乐器时远超基线表现。未来的工作将集中在优化瓶颈参数（例如帧率和词汇量大小），并探索替代表示方法如常 Q 变换（CQT）。

5. REFERENCES

[1] Chris Donahue, Antoine Caillon, Adam Roberts, Ethan Manilow, Philippe Esling, Andrea Agostinelli, Mauro Verzetti, Ian Simon, Olivier Pietquin, Neil Zeghidour, et al., “Singsong: Generating musical accompaniments from singing,” arXiv preprint arXiv:2301.12662, 2023.

[2] Jianyi Chen, Wei Xue, Xu Tan, Zhen Ye, Qifeng Liu, and Yike Guo, “FastSAG: towards fast non-autoregressive singing accompaniment generation,” in Proc. IJCAI, 2024.

[3] Junmin Gong, Sean Zhao, Sen Wang, Shengyuan Xu, and Joe Guo, “Ace-step: A step towards music generation foundation model,” arXiv preprint arXiv:2506.00045, 2025.

[4] Yi Luo and Jianwei Yu, “Music source separation with band-split RNN,” IEEE/ACM TASLP, vol. 31, pp. 1893–1901, 2023.

[5] Wei-Tsung Lu, Ju-Chiang Wang, Qiuqiang Kong, and Yun-Ning Hung, “Music source separation with band-split rope transformer,” in Proc. ICASSP, 2024.

[6] Yu-An Chung, Yu Zhang, Wei Han, Chung-Cheng Chiu, James Qin, Ruoming Pang, and Yonghui Wu, “W2v-bert: Combining contrastive learning and masked language modeling for self-supervised speech pre-training,” in IEEE ASRU, 2021.

[7] Aaron Van Den Oord, Oriol Vinyals, et al., “Neural discrete representation learning,” NeurIPS, vol. 30, 2017.

[8] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le, “Flow matching for generative modeling,” arXiv preprint arXiv:2210.02747, 2022.

[9] Lichao Zhang, Ruiqi Li, Shoutong Wang, Liqun Deng, Jinglin Liu, Yi Ren, Jinzheng He, Rongjie Huang, Jieming Zhu, Xiao Chen, et al., “M4singer: A multi-style, multi-singer and musical score provided mandarin singing corpus,” NeurIPS, 2022.

[10] Yizhi Li, Ruibin Yuan, Ge Zhang, Yinghao Ma, Xingran Chen, Hanzhi Yin, Chenghao Xiao, Chenghua Lin, Anton Ragni, Emmanouil Benetos, et al., “MERT: Acoustic music understanding model with large-scale self-supervised training,” in ICLR, 2024.

[11] Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie, “Representation alignment for generation: Training diffusion transformers is easier than you think,” in ICLR, 2025.

[12] Jonathan Ho and Tim Salimans, “Classifier-free diffusion guidance,” arXiv preprint arXiv:2207.12598, 2022.

[13] Yicheng Gu, Chaoren Wang, Junan Zhang, Xueyao Zhang, Zihao Fang, Haorui He, and Zhizheng Wu, “Singnet: Towards a large-scale, diverse, and in-the-wild singing voice dataset,” arXiv preprint arXiv:2505.09325, 2025.

[14] Xueyao Zhang, Liumeng Xue, Yicheng Gu, Yuancheng Wang, Jiaqi Li, Haorui He, Chaoren Wang, Ting Song, Xi Chen, Zihao Fang, Haopeng Chen, Junan Zhang,

Tze Ying Tang, Lexiao Zou, Mingxuan Wang, Jun Han, Kai Chen, Haizhou Li, and Zhizheng Wu, "Amphion: An open-source audio, music and speech generation toolkit," in Proc. IEEE SLT Workshop, 2024.

[15] Jiaqi Li, Xueyao Zhang, Yuancheng Wang, Haorui He, Chaoren Wang, Li Wang, Huan Liao, Junyi Ao, Zeyu Xie, Yiqiao Huang, Junan Zhang, and Zhizheng Wu, "Overview of the Amphion Toolkit (v0.2)," arXiv preprint arXiv:2501.15442, 2025.

[16] Xueyao Zhang, Xiaohui Zhang, Kainan Peng, Zhenyu Tang, Vimal Manohar, Yingru Liu, Jeff Hwang, Dangna Li, Yuhao Wang, Julian Chan, et al., "Vevo: Controllable zero-shot voice imitation with self-supervised disentanglement," in ICLR, 2025.

[17] Xueyao Zhang, Junan Zhang, Yuancheng Wang, Chaoren Wang, Yuanzhe Chen, Dongya Jia, Zhuo Chen, and Zhizheng Wu, "Vevo2: Bridging controllable speech and singing voice generation via unified prosody learning," arXiv preprint arXiv:2508.16332, 2025.

[18] Ilya Loshchilov and Frank Hutter, "Decoupled weight decay regularization," in ICLR, 2019.

[19] Ruibin Yuan, Hanfeng Lin, Shuyue Guo, Ge Zhang, Jiahao Pan, Yongyi Zang, Haohe Liu, Yiming Liang, Wenye Ma, Xingjian Du, et al., "Yue: Scaling open foundation models for long-form music generation," arXiv preprint arXiv:2503.08638, 2025.

[20] Zafar Rafii, Antoine Liutkus, and Fabian-Robert Stöter, "The MUSDB18 corpus for music separation," 2017.

[21] Igor Pereira, Felipe Araújo, Filip Korzeniowski, and Richard Vogl, "MoisesDB: A dataset for source separation beyond 4-stems," arXiv preprint arXiv:2307.15913, 2023.

[22] Kevin Kilgour, Mauricio Zuluaga, Dominik Roblek, and Matthew Sharifi, "Fréchet audio distance: A metric for evaluating music enhancement algorithms," arXiv preprint arXiv:1812.08466, 2018.

[23] Maarten Grachten and Javier Nistal, "Accompaniment prompt adherence: A measure for evaluating music accompaniment systems," in Proc. ICASSP, 2025.

[24] Andros Tjandra, Yi-Chiao Wu, Baishan Guo, John Hoffman, Brian Ellis, Apoorv Vyas, Bowen Shi, Sanyuan Chen, Matt Le, Nick Zacharov, et al., "Meta Audiobox Aesthetics: Unified automatic quality assessment for speech, music, and sound," arXiv preprint arXiv:2502.05139, 2025.