arXiv:2603.01565v1 [eess.AS] 2 Mar 2026

探索用于基于扩散变换器的文本转音频生成的组相对策略优化

Yi Gu[1*] **Yanqing Liu**[2†] **Chen Yang**[1] **Sheng Zhao**[21]清华大学[2]微软

摘要

近年来，文本转音频（T2A）生成取得了长足进展，但现有方法仍面临诸多挑战：难以准确呈现复杂的文本提示，尤其是涉及复杂音效的提示，也难以实现精准的文本-音频对齐。尽管以往方法探索了数据增强、显式时间条件和强化学习，整体合成质量仍受到限制。在本研究中，我们基于扩散变换器（DiT）架构，尝试通过强化学习进一步提升 T2A 生成质量。我们的方法首先利用大语言模型（LLM）生成高保真、内容丰富且细节详尽的音频描述，从而显著改善文本-音频语义对齐，尤其适用于含义模糊或信息不足的提示。随后，我们采用近期提出的强化学习算法——组相对策略优化（GRPO）——对 T2A 模型进行微调。通过针对多种奖励函数（包括 CLAP、KL散度、FAD 及其组合）开展系统实验，我们找出了音频合成中强化学习奏效的关键因素，并分析了奖励设计对最终音频质量的影响。实验结果表明，基于 GRPO 的微调显著提升了合成保真度和提示词遵循度。

1 引言

文本转音频（TTA）生成在内容创作和艺术应用领域前景广阔，既可用于播客和故事讲述平台，也能支持语音驱动的视频合成和沉浸式声音设计。然而，生成连贯且与提示对齐的音频——尤其是针对复杂或冗长提示时——历来需要大量数据标注和精准的文本-音频对齐。近期 TTA 模型的进展 [Zheng 等，2025，Jiang 等，2025a，Hung 等，2024，Jiang 等，2025b，Evans 等，2025，Guan 等，2024，Liu 等，2024b，Mo 等，2024] 带来了创新技术，例如：通过潜在一致性模型以最少的采样步骤实现快速推理 [Liu 等，2024b], ；采用流匹配范式加速生成并提升保真度 [Guan 等，2024，Hung 等， 2024], 免训练时间规划以实现长篇可控性 [Jiang 等，2025b], ；渐进式扩散以实现时间与语义控制的解耦 [Jiang 等，2025a], ；基于时间戳或自然语言的时间引导[Zheng 等， 2025], ；以及针对可变长度音效和音乐优化的开源架构 [Evans 等， 2025]。这些进展显著提升了多样性、表现力、推理效率、长篇合成能力和细粒度可控性，推动了同步视听内容和长篇叙事音频等应用的发展。尽管取得了这些进展，仍存在一个重大难题：如何生成保真度高且语义与时间对齐稳健的音频

*在微软语音团队实习期间完成的工作。
†通讯作者：yanqliu@microsoft.com

对于涉及多个连续事件、微妙细节或长篇叙事的复杂或冗长提示，实现这一目标仍然具有挑战性。现有方法经常难以捕捉细粒度的文本细节，导致事件遗漏、幻觉、顺序错乱或文本-音频对应关系减弱，尤其是在真实世界的长篇场景中。

我们认为，上述局限性主要源于数据质量和建模方法方面的不足。数据质量对文本转音频生成系统的可控性和保真度起着关键作用，而低质量字幕往往会导致音频事件错误、时间序列紊乱，或对输入提示产生语义误解。为解决这些问题，我们通过使用更高质量、更详尽的文本描述增强训练数据，从而提升提示词遵循能力。具体而言，我们首先设计了一组经过精心构造的提示，以从大语言模型（LLM）中获取丰富、准确且结构一致的音频描述。随后，我们将这些提示应用于基础音频数据集，利用 GPT-4 [OpenAI、 2023], DeepSeek-V2 [DeepSeek-AI、 2024], 以及 Qwen-2.5 [Team 等强大的大语言模型， 2025] 自动生成增强描述。最后，我们在该增强数据集上对文本转音频模型进行微调，从而显著提升对齐、可控性和整体合成质量。随后，我们采用强化学习，具体而言是组相对策略优化（GRPO），并结合多种奖励函数（包括 CLAP、KL散度、FAD 及其加权组合），进一步提升文本转音频（T2A）生成质量。通过系统性实验，我们分析了不同奖励设计如何影响 T2A 合成的关键方面，包括分布保真度、样本级对齐、语义提示遵循度以及整体音频质量。

文本转音频（T2A）模型的预训练——无论采用自回归还是基于扩散的方法——通常都会导致文本-音频对齐不够精确。相比之下，后训练阶段的强化学习技术已被证明能有效缓解大语言模型（LLM）中的对齐失配。大语言模型取得的显著进展，很大程度上得益于对齐阶段：这些阶段通常采用基于人类反馈的强化学习（RLHF）或类似的基于奖励的优化方法，使模型输出与人类偏好紧密对齐。然而，这种对齐范式在 T2A 开发中仍未得到充分探索，主要原因是存在一个关键的工具缺口：可靠的大规模偏好对数据集稀缺。与拥有成熟奖励模型和丰富人类反馈资源的大语言模型领域不同，T2A 领域目前缺乏收集和利用此类偏好数据的标准化、可扩展方法，因而阻碍了有效的后训练对齐。在本研究中，我们通过探索用于 T2A 生成的基于强化学习的对齐方法，来解决这一局限。具体而言，我们研究了强化学习训练的数据集准备策略，并系统考察了不同奖励函数设计（包括 CLAP、KL散度、FAD 及其组合）对最终音频质量的影响。我们的目标是减少仅靠数据增强无法解决的对齐失配，并为文本到音频合成中的有效奖励建模提供见解。
在这项工作中，我们研究了文本转音频（T2A）生成中的强化学习，并开展系统实验，以探索未来基于强化学习的 T2A 建模的潜在方向。我们的方法涵盖三个关键方面：高质量数据构建、通过强化学习进行模型训练，以及奖励函数设计。在数据构建阶段，我们从大规模的〈文本，音频〉对集合着手，并利用大语言模型（LLM）生成更丰富、更详尽且语义更精准的字幕，以扩充数据集。在强化学习阶段，我们采用组相对策略优化（GRPO），系统地探索了多种奖励函数，包括 CLAP 分数、Fréchet 音频距离（FAD）、Kullback–Leibler（KL）散度及其加权组合；实验结果表明，采用不同奖励函数均能稳定提升生成质量。这些发现为如何通过有效的奖励工程改善 T2A 系统中的文本-音频对齐和整体合成保真度提供了有价值的见解。

总而言之，我们的贡献如下。

- 我们使用大语言模型（LLM）扩充文本描述，以生成更丰富、更详细的字幕。这一增强提升了渲染音效的鲁棒性，尤其是针对复杂或细微的声学现象；我们还将这些增益与通过强化学习取得的改进进行了比较。
- 我们进一步将组相对策略优化（GRPO）应用于基于扩散变换器（DiT）的 T2A 模型，以提升文本到音频（T2A）生成质量。我们系统分析了不同奖励函数之间的性能差异，并开展

了大量消融实验，以验证各组件对提升基准测试分数的贡献。

2 相关工作

2.1 生成式语音模型

基于自回归或扩散架构的生成式语音模型，显著提升了文本转语音（TTS）、语音识别、语音到语音转换及多模态建模等任务的端到端语音质量 [Hu et al., 2026, Xu et al., 2025, Wu et al., 2025, Wang et al., 2023, Chen et al., 2024, Zhang et al., 2023, Shen et al., 2023, Kanda et al., 2024, Liu et al., 2025d, Ju et al., 2024, Han et al., 2024, Eskimez et al., 2024, Li et al., 2024, Yang et al., 2025, Wang et al., 2025a, Li et al., 2018, 2019, 2020, Liu et al., 2021, Tan et al., 2024, Yuan et al., 2024, Eskimez et al., 2024, Meng et al., 2024, Liu et al., 2022]。连续语音特征——例如通过信号处理方法得到的梅尔频谱图，或通过基于残差向量量化（RVQ）的编解码器模型获得的离散语音标记——常被用作语音建模中的有效中间表示，并在文本转语音和语音到语音任务中展现出优异性能。这些语音合成领域的成功，启发了文本转音频（T2A）生成采用类似的架构、数据扩展和训练策略；这一领域已成为近期备受关注的研究趋势。

2.2 文本转音频

文本转音频（T2A）生成已成为一个重要领域，使模型能够可靠地合成与给定文本描述在语义上保持一致的音频。如今，该领域主要依赖扩散模型和流匹配模型：这些模型以文本嵌入和时间步为条件，在潜在空间中迭代地对纯噪声进行去噪。基于扩散的方法 [Yang et al., 2023, Liu et al., 2024a, Evans et al., 2025, Majumder et al., 2024, Novack et al., 2025, Liu et al., 2025b, Feng et al., 2025, Jeong et al., 2025, Ronchini et al., 2024, Lee et al., 2024, Niu et al., 2024, Jiang et al., 2025b, Zhao et al., 2025, Wang et al., 2025b, Yuan et al., 2025, Vora et al., 2025, Tian et al., 2025] 已成为主流范式，这得益于其出色的音频保真度和生成效率。近期研究日益关注长篇生成和精确的时间对齐音频合成。Stable Audio [Evans et al., 2025] 通过在大规模专有数据集上训练扩散模型并融入时间条件嵌入，提升了长篇生成能力；不过，这些嵌入主要用于调节整体时长，而非精细控制特定事件的时间。TANGO 和 TangoFlux [Majumder et al., 2024, Hung et al.,2024] 通过指令微调语言模型和基于偏好的微调，提升了对复杂提示的组合准确性和对齐效果。尽管取得了这些进展，在扩散框架中实现精确的事件级文本-音频对齐仍是一项关键挑战。近期有若干研究直接针对这一控制能力缺口展开探索。ControlAudio [Jiang et al., 2025a] 在大规模文本-音频配对数据上预训练扩散变换器（DiT），并通过统一语义表示整合时间和音素特征，从而实现可扩展的精细可控性。PicoAudio2 [Zheng et al., 2025] 采用定位模型为真实音频文本数据集标注事件时间戳，从而整理出一个时间对齐程度较高的语料库，以扩充现有合成数据，推进了以数据为中心的方法。最后，FreeAudio [Jiang et al., 2025b]提出了一种无需训练的 T2A 方法，能够同时根据复杂的自然语言提示实现精确的时间控制和连贯的长篇合成，标志着向结构一致且可控的音频生成迈出了重要一步。

2.3 组相对策略优化（GRPO）

强化学习（RL）已成为跨模态对齐大语言模型（LLM）的关键范式，涵盖文本、图像，以及最近的语音。轻量级的策略方法组相对策略优化（GRPO） [Shao 等人， 2024] 在大语言模型后训练中表现尤为出色，能够提升安全性和效用。在此基础上，近期研究探索了强化学习在语音处理任务中的潜力，包括自动语音识别（ASR）和文本转语音（TTS）。针对 ASR， [Gao 等人，2025b] 应用组相对策略优化框架，尝试了多种基于规则的奖励函数，并考察了强化学习数据构建策略的影响。在 TTS 领域， [Gao 等人，2025a] 提出了可微奖励

优化（DiffRO），该方法直接基于神经编解码器令牌计算奖励，显著提高了 TTS 系统的发音准确度。在 ASR 研究的基础上， [Liu 等，2025a]提出了一种信息量更丰富的复合奖励函数，将字符错误率（CER）与 ASR 模型的负对数似然（NLL）相结合，从而提供更准确、稳定的奖励信号。

3 方法

我们采用预训练的 TangoFlux 模型 Hung 等人 [2024] 作为实验基线。TangoFlux 构建于扩散 Transformer（DiT）和多模态扩散 Transformer（MMDiT）架构之上，可生成时长最长达 30 秒、采样率为 44.1 千赫的高保真音频片段。该模型在连续音频潜空间中学习整流流轨迹；音频由 Stable Audio Open 的变分自编码器（VAE）进行编码，同时以 FLAN-T5 的文本表示作为条件。为构建强化学习实验，我们首先构建一个包含更详细且语义更精确的文本描述的增强数据集，随后将其用作后续强化学习的训练集。

3.1 描述增强

获取多样且高质量的训练数据，仍然是开发高性能文本转音频（T2A）生成模型的一项主要挑战。现有的带标注音频数据集通常受限于字幕过短、多样性不足以及偶尔出现的不准确问题。尽管人工标注员能够生成详细且可靠的描述，但人工标注成本高昂，难以扩展。为构建适用于强化学习实验的有效训练集，我们采用数据增强流程，利用大语言模型（LLM）根据稀疏的原始标签和简短字幕生成丰富、详细的音频描述。我们的方法包括以下步骤：我们设计多组提示规则，旨在促使 LLM 生成更长、更具描述性且语义更忠实的输出。我们使用一个经过人工验证的小型评估子集，在多个大语言模型上系统地评估这些规则集。人工标注员会根据质量和保真度对生成的描述进行评分，我们则依据平均分选择表现最佳的配置。确定最优提示模板和大语言模型后，我们将其规模化应用于全部原始标签和标注，从而生成一个内容显著丰富、包含详细文本描述的语料库。在训练过程中，我们按照可控的混合比例，从两类来源中随机采样，将新生成的描述与原始字幕结合起来。我们会持续在留出的验证集上监测模型性能，以确定最有效的比例，并避免对合成描述过拟合。该流程实现了可扩展的高质量数据增强，能够在后续强化学习阶段显著提升提示词遵循度和整体合成保真度。

3.2 结合组相对策略优化的流匹配

3.2.1 流匹配

基于扩散的生成模型对噪声调度器的选择高度敏感；次优的调度方案会大幅降低性能。相比之下，流匹配（FM）框架 [Lip- man 等， 2022] 在超参数变化下表现出卓越的鲁棒性，近期更是在多种音频生成任务中取得了最先进的成果，包括文本转音频（T2A）、文本转语音（TTS）以及语音到语音转换。流匹配学习一个随时间变化的速度场，可将样本直接从简单的基分布（通常为高斯噪声）传输到目标数据分布。该方法通过简单的速度回归目标实现直接训练，并支持通过常微分方程（ODE）进行精确的似然计算。这一范式也推动了文本转图像生成领域近期的突破，生成了保真度前所未有的照片级真实感图像。此外，强化学习（RL）此前已在自回归语言建模中带来变革性进步，近期也成功融入流匹配目标，进一步显著提升了可控且富有表现力的文本转语音合成效果 [Sun 等，2025，Liu 等，2025c]。

3.2.2 组相对策略优化

强化学习（RL）推动了大语言模型（LLM）与人类偏好对齐方面的显著进展。直接偏好优化（DPO）和组相对策略优化（GRPO）等基于偏好的优化方法已被证明非常有效，显著提升了生成文本的安全性、有用性和效用。这些技术近期已扩展至语音和音频生成领域。将强化学习应用于音频合成的一个关键因素，是审慎选择能够准确反映感知音频质量和客观音频质量的奖励函数。在文本转语音（TTS）系统中，基于强化学习的对齐通常采用说话人相似度（SIM）或词错误率（WER）作为奖励。这些指标利用语音编码器（例如 WavLM）捕捉韵律、语调、音高和停顿等声学属性，同时借助自动语音识别模型确保内容一致且清晰易懂。对于文本转音频（T2A）生成，合适的奖励信号包括 CLAP 相似度（用于文本-音频语义对齐）、Kullback–Leibler（KL）散度（用于衡量样本层面的分布相似性），以及弗雷歇音频距离（FAD）（用于衡量整体分布保真度）。在我们基于 GRPO 的增强流程中，我们采用这三项指标作为主要奖励。GRPO 通过更新预训练生成策略，使其倾向于生成能够最大化指定奖励的样本。更新所依据的优势估计，是根据每个训练批次（或组）生成样本的相对奖励计算得出的：该优势反映了某个输出相较于批次平均水平的表现优劣。按照标准做法，我们使用相同的预训练权重（例如 TangoFlux 的权重）初始化策略模型和冻结参考模型。在整个 GRPO 训练过程中，参考模型保持不变，为优势计算提供稳定的基准。我们主要采用 CLAP 相似度、KL散度和 FAD 作为奖励组成部分，因为它们共同涵盖了 T2A 质量最重要的维度：语义提示遵循度、样本层面的分布对齐，以及全局分布真实性。

奖励函数我们的奖励函数整合了三个互补组成部分——CLAP（用于文本-音频语义对齐）、FAD（弗雷歇音频距离，用于全局分布保真度）和KL（库尔贝克-莱布勒散度，用于样本级分布相似性）——以全面捕捉生成音频质量的不同方面。我们展示了实验结果，说明在单独使用各奖励配置及其加权组合时的训练进展和最终性能。

$$\mathcal{R} = \mathcal{R}_{\text{CLAP}} + \mathcal{R}_{\text{KL}} + \mathcal{R}_{\text{FAD}} \tag{1}$$

CLAP CLAP（对比语言-音频预训练）分数通过计算生成音频样本与其文本提示在共享表示空间中嵌入之间的余弦相似度，量化二者之间的语义对齐程度。该指标能够有效捕捉音频与所描述事件、内容及整体意图的匹配程度——这类似于CLIP（对比语言-图像预训练）在图像生成任务中对文本-图像对齐程度的评估。CLAP通常基于大规模音频-文本对进行预训练，使其能够学习一个联合嵌入空间，将原始波形（或其特征）与相应的自然语言描述进行对齐。得益于这一对比学习目标，CLAP已在包括零样本音频分类、自动音频字幕生成以及文本到音频（T2A）生成模型评估在内的下游应用中展现出良好效果。

FAD 弗雷歇音频距离（FAD）将弗雷歇初始距离（FID）——一种最初为图像生成开发、广泛使用的指标——应用于音频领域。与 FID 一样，FAD 已成为评估文本转音频（T2A）和通用音频生成模型质量的标准客观指标，尤其侧重于衡量感知保真度以及与真实音频的分布相似性。FAD 的计算方法如下：使用预训练音频分类器（例如 VGGish、PANNs 等）分别提取生成音频样本和参考（真实）音频样本的嵌入。然后将这些嵌入的分布建模为多元高斯分布，并计算两个分布之间的 Fréchet 距离（也称为高斯分布之间的 2-瓦瑟斯坦距离）。FAD 值越低，表示生成音频在保真度和多样性方面都越接近参考集。

KL散度 Kullback–Leibler（KL）散度在文本转音频（T2A）评估中用作语义一致性指标。对于每个生成音频与参考音频对，提取嵌入并传入

模型	FD_{Openl3} ↓	KL_{PaSST} ↓	CLAP 分数 ↑
TangoFlux	181.49 ± 5.66	2.1420 ± 0.0202	0.3259 ± 0.0254
+数据增强	161.93 ± 5.52	1.0279 ± 0.0110	0.5456 ± 0.0342
+加权奖励	158.76 ± 5.74	1.0117 ± 0.0192	0.5673 ± 0.0299

表 1：比较基线 TangoFlux 模型、采用数据增强的 TangoFlux 模型、采用加权 GRPO（加权奖励）的 TangoFlux 模型以及采用数据增强的模型的结果。

通过AudioSet预训练分类器，以获得各语义类别上的softmax概率向量。随后计算从参考分布到生成分布的KL散度，用以量化用生成音频的语义近似参考语义时所产生的信息损失。由于KL散度在逐样本层面运作，并直接衡量预测语义标签分布之间的不匹配，它能够有效刻画生成音频是否忠实保留了文本提示所描述（或任何条件信号中存在）的细粒度内容、事件和概念，从而补充FAD等全局分布指标。

4 实验

4.1 训练细节

在训练中，我们使用AdamW优化器，并结合余弦退火学习率调度器，以确保稳定且高效的收敛。在数据增强预训练阶段及后续GRPO微调阶段，初始学习率均设为1e-4，并采用余弦衰减将其逐步降至最小值5e-6。所有实验均在NVIDIA A100 GPU上进行。我们使用内部音频数据集作为数据增强和强化学习实验的数据来源。

4.2 评估

4.2.1 指标分析

为评估模型性能，我们采用若干客观指标，从不同方面衡量文本到音频（T2A）生成的生成质量。对于每次实验，我们报告三项关键分数：弗雷歇距离（FD）、库尔贝克-莱布勒散度（KL散度）和CLAP 分数。

- 弗雷歇距离（FD）衡量生成音频样本与真实音频样本之间的分布相似性。该指标基于从预训练音频分类器（例如PANNs、VGGish或其他分类器）中提取的特征计算；不同的骨干分类器会产生不同的FD值，反映出特征敏感度的差异。
- 库尔贝克-莱布勒散度（KL）用于衡量生成音频与参考音频的概率分布之间的差异，通常在单样本层面进行计算。KL 值越低，表示生成音频在分布特征上越接近真实数据。
- CLAP 分数通过计算文本提示嵌入与音频嵌入之间的余弦相似度得出（这些嵌入由 CLAP 等对比式语言-音频预训练模型生成）。分数越高，表示生成音频与输入文本提示之间的语义对齐越强。

这些指标综合评估了分布保真度（FD）、样本级相似度（KL）和文本-音频相关性（CLAP）。

表 1 展示了三种模型变体的对比评估结果：基线 TangoFlux、采用数据增强进行 GRPO 实验的 TangoFlux，以及采用加权 GRPO（wGRPO）的 TangoFlux，后者结合了 CLAP、FAD 和 KL 奖励。与基线 TangoFlux 相比，数据增强变体在 FAD、KL 和 CLAP 指标上始终表现更优。加权 GRPO（wGRPO）变体相较于数据增强基线，进一步取得了幅度不大但稳定的提升。表 2 报告了在 GRPO 中使用单一奖励函数时的性能。正如预期，

模型	FD_{Openl3} ↓	KL_{PaSST} ↓	CLAP 分数 ↑
TangoFlux	161.93 ± 5.52	1.0279 ± 0.0110	0.5456 ± 0.0342
+CLAP	171.78 ± 5.45	1.0600 ± 0.0144	0.5652 ± 0.0329
+FAD	157.72 ± 5.31	1.0900 ± 0.0176	0.5531 ± 0.0271
+KL	167.45 ± 6.42	1.0210 ± 0.0183	0.5588 ± 0.0297
+加权奖励	158.76 ± 5.74	1.0117 ± 0.0192	0.5673 ± 0.0299

表 2：比较基线 TangoFlux 与其强化学习版本在 CLAP、FAD、KL散度和加权奖励（加权奖励）作为奖励时的结果。

由于优化直接针对这一目标，CLAP-only 奖励在 CLAP 指标上取得了最出色的结果，FAD 和 KL 奖励也分别对应各自的指标。这证实了单目标组相对策略优化能够可靠地提升相应的评估指标。加权组相对策略优化配置在所有指标（FAD、KL、CLAP）上的表现均与其他方法相当或更优，展现了多奖励加权在提升整体文本转音频生成质量方面的有效性。

5 结论

在本研究中，我们探讨将组相对策略优化（GRPO）作为一种强化学习方法，用于文本转音频（T2A）生成。我们系统地研究了一系列奖励函数——包括 CLAP、KL散度、弗雷歇距离（FD）及其加权组合——并分析它们对关键评估指标（FAD、KL、CLAP）的影响。结果表明，GRPO 尤其是在采用多奖励加权时，能够有效平衡分布保真度（FD）、样本级语义一致性（KL）和提示词遵循度（CLAP），从而显著提升整体生成质量。未来的工作将优先扩大训练数据集规模，以增强模型鲁棒性和性能。我们计划开展受控实验，考察增加数据量如何影响 GRPO 的训练动态、收敛稳定性、样本效率，以及文本转音频合成的最终音频质量。

参考文献

Sanyuan Chen、Shujie Liu、Long Zhou、Yanqing Liu、Xu Tan、Jinyu Li、Sheng Zhao、Yao Qian 和 Furu Wei。VALL-E 2：神经编解码器语言模型是达到人类水平的零样本文本转语音合成器。*arXiv* 预印本 *arXiv:2406.05370*，2024。

DeepSeek-AI。Deepseek-v2：一种强大、经济且高效的混合专家语言模型。*arXiv*预印本 *arXiv:2405.04434*，2024。

Sefik Emre Eskimez、Xiaofei Wang、Manthan Thakker、Canrun Li、Chung-Hsien Tsai、Zhen Xiao、Hemin Yang、Zirun Zhu、Min Tang、Xu Tan 等。E2 TTS：极其简单的全非自回归零样本文本转语音。*arXiv*预印本 *arXiv:2406.18009*，2024。

Zach Evans、Julian D Parker、CJ Carr、Zack Zukowski、Josiah Taylor 和 Jordi Pons。Stable Audio Open。发表于*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。

Linfeng Feng、Lei Zhao、Boyu Zhu、Xiao-Lei Zhang 和 Xuelong Li。AudioSpa：通过文本实现声音事件的空间化。*arXiv*预印本 *arXiv:2502.11219*，2025年。

Changfeng Gao、Zhihao Du 和 Shiliang Zhang。面向基于大语言模型的语音合成系统的可微奖励优化。*arXiv* 预印本 *arXiv:2507.05911*，2025a。

Changfeng Gao、Yabin Li、Keyu An、Zhifu Gao、Zhihao Du、Han Zhao 和 Xiangang Li。探索基于大语言模型的自动语音识别和语音合成系统中的强化学习。*arXiv* 预印本 *arXiv:2509.18569*，2025b。

Wenhao Guan、Kaidi Wang、Wangjin Zhou、Yang Wang、Feng Deng、Hui Wang、Lin Li、Qingyang Hong 和 Yong Qin。Lafma：一种用于文本到音频生成的潜在流匹配模型。*arXiv* 预印本 *arXiv:2406.08203*，2024。

Bing Han、Long Zhou、Shujie Liu、Sanyuan Chen、Lingwei Meng、Yanming Qian、Yanqing Liu、Sheng Zhao、Jinyu Li 和 Furu Wei。Vall-e r：通过单调对齐实现稳健且高效的零样本文本到语音合成。*arXiv* 预印本 *arXiv:2406.07855*，2024。

Hangrui Hu、Xinfa Zhu、Ting He、Dake Guo、Bin Zhang、Xiong Wang、Zhifang Guo、Ziyue Jiang、Hongkun Hao、Zishan Guo 等。Qwen3-tts 技术报告。*arXiv* 预印本 *arXiv:2601.15621*，2026年。

Chia-Yu Hung、Navonil Majumder、Zhifeng Kong、Ambuj Mehrish、Amir Ali Bagherzadeh、Chuan Li、Rafael Valle、Bryan Catanzaro 和 Soujanya Poria。TangoFlux：通过流匹配和 CLAP 排序偏好优化，实现超快速且忠实的文本到音频生成。*arXiv* 预印本 *arXiv:2412.21037*，2024。

Yujin Jeong、Yunji Kim、Sanghyuk Chun 和 Jiyoung Lee。阅读、观看并尖叫！从文本和视频生成声音。载于*AAAI* 人工智能会议论文集，第39卷，第17590–17598页，2025年。

Yuxuan Jiang、Zehua Chen、Zeqian Ju、Yusheng Dai、Weibei Dou和Jun Zhu。ControlAudio：通过渐进式扩散建模应对文本引导、时间指示且可懂的音频生成。*arXiv*预印本 *arXiv:2510.08878*，2025a。

Yuxuan Jiang、Zehua Chen、Zeqian Ju、Chang Li、Weibei Dou和Jun Zhu。FreeAudio：面向可控长篇文本到音频生成的免训练时间规划。载于第*33*届*ACM*国际多媒体会议论文集，第9871–9880页，2025b。

Zeqian Ju、Yuancheng Wang、Kai Shen、Xu Tan、Detai Xin、Dongchao Yang、Yanqing Liu、Yichong Leng、Kaitao Song、Siliang Tang等。Naturalspeech 3：基于因子化编解码器与扩散模型的零样本语音合成。*arXiv*预印本 *arXiv:2403.03100*，2024。

Naoyuki Kanda、Xiaofei Wang、Sefik Emre Eskimez、Manthan Thakker、Hemin Yang、Zirun Zhu、Min Tang、Canrun Li、Chung-Hsien Tsai、Zhen Xiao 等。让基于流匹配的零样本文本转语音随心而笑。*arXiv* 预印本 *arXiv:2402.07383*，2024。

Junwon Lee、Modan Tailleur、Laurie M Heller、Keunwoo Choi、Mathieu Lagrange、Brian McFee、Keisuke Imoto 和 Yuki Okamoto。声音场景合成挑战：评估文本到音频生成。*arXiv* 预印本 *arXiv:2410.17589*，2024。

Jiaqi Li、Dongmei Wang、Xiaofei Wang、Yao Qian、Long Zhou、Shujie Liu、Midia Yousefi、Canrun Li、Chung-Hsien Tsai、Zhen Xiao 等。探究用于基于语音语言模型的语音生成的神经音频编解码器。*arXiv* 预印本 *arXiv:2409.04016*，2024。

Naihan Li、Shujie Liu、Yanqing Liu、Sheng Zhao、Ming Liu 和 Ming Zhou。接近人类品质的 TTS：基于 Transformer 的方法。*arXiv* 预印本 *arXiv:1809.08895*，第 2 卷，2018 年。

Naihan Li、Shujie Liu、Yanqing Liu、Sheng Zhao 和 Ming Liu。基于 Transformer 网络的神经语音合成。载于 *AAAI* 人工智能会议论文集，第 33 卷，第 6706–6713 页，2019 年。

Naihan Li、Yanqing Liu、Yu Wu、Shujie Liu、Sheng Zhao 和 Ming Liu。RobuTrans：一种基于 Transformer 的鲁棒文本转语音模型。载于 *AAAI* 人工智能会议论文集，第 34 卷，第 8228–8235 页，2020 年。

Yaron Lipman、Ricky TQ Chen、Heli Ben-Hamu、Maximilian Nickel 和 Matt Le。Flow matching 生成式建模。*arXiv* 预印本 *arXiv:2210.02747*，2022。

Chang Liu、Ya-Jun Hu、Ying-Ying Gao、Shi-Lei Zhang 和 Zhen-Hua Ling。基于大语言模型的文本转语音组相对策略优化。*arXiv* 预印本 *arXiv:2509.18798*，2025a。

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、语音和语言处理汇刊，32：2871–2883，2024a。

Huadai Liu、Rongjie Huang、Yang Liu、Hengyuan Cao、Jialei Wang、Xize Cheng、Siqi Zheng 和 Zhou Zhao。Audiolcm：利用潜在一致性模型进行文本到音频生成。*arXiv* 预印本 *arXiv:2406.00356*，2024b。

Huadai Liu、Jialei Wang、Rongjie Huang、Yang Liu、Heng Lu、Zhou Zhao 与 Wei Xue。Flashaudio：面向快速高保真文本到音频生成的整流流。载于第*63*届计算语言学协会年会论文集（第*1*卷：长文），第13694–13710页，2025b。

Jie Liu、Gongye Liu、Jiajun Liang、Yangguang Li、Jiaheng Liu、Xintao Wang、Pengfei Wan、Di Zhang 与 Wanli Ouyang。Flow-grpo：通过在线强化学习训练流匹配模型。*arXiv*预印本 *arXiv:2505.05470*，2025c。

Y Liu、Z Xu、G Wang、K Chen、B Li、X Tan 等。DelightfulTTS：面向暴雪挑战赛的微软语音合成系统。*DelightfulTTS*：面向暴雪挑战赛的微软语音合成系统，2021。

Yanqing Liu、Ruiqing Xue、Lei He、Xu Tan 和 Sheng Zhao。Delightfultts 2：采用对抗式向量量化自编码器的端到端语音合成。*arXiv* 预印本 *arXiv:2207.04646*，2022。

Yanqing Liu、Ruiqing Xue、Chong Zhang、Yufei Liu、Gang Wang、Bohan Li、Yao Qian、Lei He、Shujie Liu 和 Sheng Zhao。用于语音合成的下一词元去噪。*arXiv* 预印本 *arXiv:2507.22746*，2025d。

Navonil Majumder、Chia-Yu Hung、Deepanway Ghosal、Wei-Ning Hsu、Rada Mihalcea 和 Soujanya Poria。Tango 2：通过直接偏好优化对齐基于扩散的文本到音频生成。载于第 *32* 届 *ACM* 多媒体国际会议论文集，第 564–572 页，2024。

Lingwei Meng、Long Zhou、Shujie Liu、Sanyuan Chen、Bing Han、Shujie Hu、Yanqing Liu、Jinyu Li、Sheng Zhao、Xixin Wu 等。无需向量量化的自回归语音合成。*arXiv* 预印本 *arXiv:2407.08551*，2024。

Shentong Mo、Jing Shi 和 Yapeng Tian。与视频同步的文本到音频生成。*arXiv* 预印本 *arXiv:2403.07938*，2024。

Xinlei Niu、Jing Zhang、Christian Walder 和 Charles Patrick Martin。Soundlocd：一种用于文本到声音生成的高效条件离散对比潜在扩散模型。载于*ICASSP 2024-2024 IEEE*声学、语音与信号处理国际会议（*ICASSP*），第 261–265 页。IEEE，2024。

Zachary Novack、Zach Evans、Zack Zukowski、Josiah Taylor、CJ Carr、Julian Parker、Adnan Al-Sinan、Gian Marco Iodice、Julian McAuley、Taylor Berg-Kirkpatrick 等。通过对抗式后训练实现快速文本到音频生成。载于*2025 IEEE*音频与声学信号处理应用研讨会（*WASPAA*），第 1–5 页。IEEE，2025年。

OpenAI。GPT-4技术报告。*arXiv*预印本 *arXiv:2303.08774*，2023年。

Francesca Ronchini、Luca Comanducci与Fabio Antonacci。使用文本到音频模型进行环境声音分类的合成训练集生成。*arXiv*预印本 *arXiv:2403.17864*，2024。

Zhihong Shao、Peiyi Wang、Qihao Zhu、Runxin Xu、Junxiao Song、Xiao Bi、Haowei Zhang、Mingchuan Zhang、YK Li、Yang Wu等。DeepSeekMath：推动开放语言模型中数学推理的极限。*arXiv*预印本 *arXiv:2402.03300*，2024。

Kai Shen、Zeqian Ju、Xu Tan、Yanqing Liu、Yichong Leng、Lei He、Tao Qin、Sheng Zhao与Jiang Bian。Naturalspeech 2：潜在扩散模型是自然且零样本的语音与歌声合成器。*arXiv*预印本 *arXiv:2304.09116*，2023年。

Xiaohui Sun、Ruitong Xiao、Jianye Mo、Bowen Wu、Qun Yu 和 Baoxun Wang。F5r-tts：通过组相对策略优化改进基于流匹配的文本转语音。*arXiv* 预印本 *arXiv:2504.02407*，2025年。

Xu Tan、Jiawei Chen、Haohe Liu、Jian Cong、Chen Zhang、Yanqing Liu、Xi Wang、Yichong Leng、Yuanhao Yi、Lei He 等。Naturalspeech：具有人类级质量的端到端文本转语音合成。*IEEE*模式分析与机器智能汇刊，2024。

Qwen团队、An Yang等人。Qwen2.5技术报告。*arXiv*预印本 *arXiv:2412.15115*，2025年。

Zeyue Tian、Yizhu Jin、Zhaoyang Liu、Ruibin Yuan、Xu Tan、Qifeng Chen、Wei Xue和Yike Guo。Audiox：用于任意内容到音频生成的扩散 Transformer。*arXiv*预印本 *arXiv:2503.10522*，2025年。

Jayneel Vora、Aditya Krishnan、Nader Bouacida、Prabhu RV Shankar和Prasant Mohapatra。Ptq4adm：面向高效文本条件音频扩散模型的训练后量化。载于*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。

Chengyi Wang、Sanyuan Chen、Yu Wu、Ziqiang Zhang、Long Zhou、Shujie Liu、Zhuo Chen、Yanqing Liu、Huaming Wang、Jinyu Li等人。神经编解码器语言模型是零样本文本转语音合成器。*arXiv*预印本 *arXiv:2301.02111*，2023年。

Hui Wang, Shujie Liu, Lingwei Meng, Jinyu Li, Yifan Yang, Shiwan Zhao, Haiyang Sun, Yanqing Liu, Haoqin Sun, Jiaming Zhou 等。《Felle：采用逐词元由粗到细流匹配的自回归语音合成》。载于第*33*届*ACM*国际多媒体会议论文集，第 10229–10238 页，2025a。

Zehan Wang, Ke Lei, Chen Zhu, Jiawei Huang, Sashuai Zhou, Luping Liu, Xize Cheng, Shengpeng Ji, Zhenhui Ye, Tao Jin 等。《T2a-feedback：通过细粒度人工智能反馈提升文本到音频生成的基础能力》。载于第*63*届计算语言学协会年会（第*1*卷：长篇论文）论文集，第 23535–23547 页，2025b。

Boyong Wu, Chao Yan, Chen Hu, Cheng Yi, Chengli Feng, Fei Tian, Feiyu Shen, Gang Yu, Haoyang Zhang, Jingbei Li 等。《Step-audio 2 技术报告》。*arXiv* 预印本 *arXiv:2507.16632*，2025年。

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu 等。Qwen3-omni 技术报告。*arXiv* 预印本 *arXiv:2509.17765*，2025年。

Dongchao Yang, Jianwei Yu, Helin Wang, Wen Wang, Chao Weng, Yuexian Zou 和 Dong Yu。Diffsound：用于文本到声音生成的离散扩散模型。*IEEE/ACM*音频、语音和语言处理汇刊，31：1720–1733，2023年。

Yifan Yang, Shujie Liu, Jinyu Li, Yuxuan Hu, Haibin Wu, Hui Wang, Jianwei Yu, Lingwei Meng, Haiyang Sun, Yanqing Liu 等。用于高效零样本文本到语音合成的伪自回归神经编解码器语言模型。发表于第*33*届*ACM*国际多媒体会议论文集，第9316–9325页，2025年。

Yi Yuan、Xubo Liu、Haohe Liu、Xiyuan Kang、Zhuo Chen、Yuxuan Wang、Mark D Plumbley 和 Wenwu Wang。Dreamaudio：基于扩散模型的定制化文本到音频生成。*arXiv* 预印本 *arXiv:2509.06027*，2025年。

Ze Yuan、Yanqing Liu、Shujie Liu 和 Sheng Zhao。连续语音词元使大语言模型成为稳健的多模态学习器。*arXiv* 预印本 *arXiv:2412.04917*，2024。

Ziqiang Zhang、Long Zhou、Chengyi Wang、Sanyuan Chen、Yu Wu、Shujie Liu、Zhuo Chen、Yanqing Liu、Huaming Wang、Jinyu Li 等。用你自己的声音说外语：跨语言神经编解码器语言建模。*arXiv* 预印本 *arXiv:2303.03926*，2023年。

Junqi Zhao、Jinzheng Zhao、Haohe Liu、Yun Chen、Lu Han、Xubo Liu、Mark Plumbley 和 Wenwu Wang。Audioturbo：基于整流扩散的快速文本到音频生成。*arXiv* 预印本 *arXiv:2505.22106*，2025年。Zihao Zheng、Zeyu Xie、Xuenan Xu、Wen Wu、Chao Zhang 和 Mengyue Wu。PicoAudio2：基于自然语言描述的时间可控的文本到音频生成。*arXiv* 预印本 *arXiv:2509.00683*，2025年。