VIVIDVOICE：用于场景感知视觉驱动语音合成的统一框架

Chengyuan Ma[1]*, Jiawei Jin*[1]*, Ruijie Xiong*[2]*,Chunxiang Jin*[2]*,Canxiang Yan*[2]*,Wenming Yang*[1*]

[1] 中国清华大学深圳国际研究生院， [2]中国蚂蚁集团

arXiv:2602.02591v1 [cs.SD] 1 Feb 2026

摘要

我们提出并定义了一项新任务——场景感知视觉驱动语音合成，旨在解决现有语音生成模型在生成与真实物理世界相契合的沉浸式听觉体验方面的局限。针对数据稀缺和模态解耦这两大核心挑战，我们提出了统一生成框架 VividVoice。首先，我们构建了大规模、高质量的混合多模态数据集 Vivid-210K，并通过创新的程序化流水线，首次建立了视觉场景、说话人身份与音频之间的强关联。其次，我们设计了核心对齐模块 D-MSVA，该模块利用解耦记忆库架构和跨模态混合监督策略，实现从视觉场景到音色及环境声学特征的细粒度对齐。主观和客观实验结果均有力证明，VividVoice 在音频保真度、内容清晰度和多模态一致性方面显著优于现有基线模型。我们的演示可访问 https://chengyuann.github.io/VividVoice/ 。

索引词— 场景感知语音合成、多模态对齐、潜在扩散模型

1. 引言

场景感知沉浸式语音生成旨在实现两种关键声学组成部分的联合合成：(1) 内容可控的语音，以及 (2) 与特定场景相匹配的环境声学。这项任务超越了传统的语音合成范式——传统语音合成主要专注于生成纯净语音——通过拓展表达维度，同时传达语言信息和空间环境语境 [1,2, 3, 4]。实现语音与环境的同步生成，对于提升下一代应用（如虚拟现实和数字人）中的感知真实感和用户沉浸感至关重要 [5, 6]。

早期的环境感知语音生成方法主要通过将纯净语音与白噪声或固定混响相叠加，来模拟静态声学特征。为追求更真实的动态效果，后续研究转向声学特征迁移，例如通过提取房间脉冲响应（RIRs） [7], 扩展环境控制，或将参考音频中的噪声与混响特征迁移到目标语音，如 DiffRENT [2] 和 AST-LDM [8]。尽管这些方法提升了真实感，其可控性仍受限于参考信号的可用性，因而难以针对任意新场景灵活生成语音。为实现更灵活的生成控制，以 VoiceLDM [1] 为代表的研究采用了文本驱动方法，试图通过自然语言描述联合建模语音与环境声音。然而，文本模态在信息密度上存在固有局限：其抽象性与歧义性使其不足以捕捉视觉场景中蕴含的丰富声学细节。因此，转向信息更密集、细节更丰富的视觉模态，成为克服这一局限的自然选择。

然而，当前由视觉驱动的研究尚无法满足我们的任务需求。一类研究（例如 FaceTTS [9]）关注说话人，生成与视觉身份一致的音色，却缺乏感知或建模环境声学的能力。另一类研究（例如 SSV2A [10]）侧重于场景，生成逼真的环境音景，却无法合成指定的语言内容。无法联合生成语音和环境声学，源于数据和模型设计层面的两项根本挑战。首先，在数据层面，目前仍缺少一个高度对齐的数据集，能够同时包含视觉场景、说话人身份及相应的环境声学。例如，AudioSet [11] 提供了视听场景，但缺乏稳定的说话人身份；而 LRS3-TED [12] 等数据集虽包含对齐的语音和身份信息，却通常是在声学“干净”的条件下录制的，因此无法让模型学习场景与声学之间的相关性。其次，在模型设计层面，现有的对齐架构不足以学习从单一视觉输入到音色和环境声学这两个解耦成分的复杂多对一映射。因此，这些架构难以在实现对这两种属性独立控制的同时保留细粒度特征。

为解决上述局限性，我们提出了一种统一的解决方案，其核心是一项新任务——场景感知视觉驱动语音合成。我们的主要贡献如下:

(1) 我们首次提出了场景感知视觉驱动语音合成这一新任务，并设计了首个统一的生成框架 VividVoice，旨在生成与视觉场景在声学上紧密耦合的沉浸式语音。

(2) 为支持这项任务，我们构建了首个大规模数据集 Vivid-210K。该数据集通过创新的程序化流程构建，为训练提供了所需的强多模态关联信息。

(3) 我们提出了一种基于解耦的声学特征对齐方法。该方法的核心是我们设计的解耦多模态场景-语音对齐（D-MSVA）模块，该模块采用独特的混合监督策略进行训练，以实现对说话人音色和环境声学的独立控制。

* 通信作者 本工作部分得到了国家重点研发计划（2023YFB4302200）以及深圳市战略性新兴产业发展专项基金（KJZD20231023094700001）的支持。

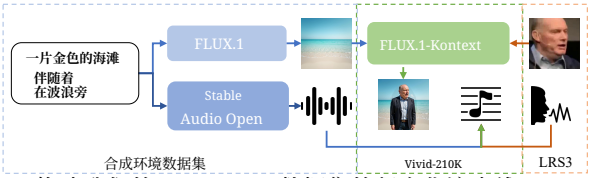


图1：构建我们的Vivid-210K数据集的程序化流水线。

2. 方法

2.1. Vivid-210K

为支持所提出的场景感知视觉驱动语音合成任务，我们构建了一个大规模多模态数据集 Vivid-210K，其中包含超过 21 万个样本和 800 多名说话人。该数据集由程序化合成的预训练集和旨在缩小领域差距的真实场景微调集组成。我们创新的程序化合成流程（如图 1 所示）遵循以下数据流：首先，从 LRS3 数据集中提取说话人身份信息（人脸图像和纯净语音）。为获得匹配的视听环境，我们考察了 VGGSound 等现有数据集，但发现它们通常在质量和对齐方面都存在局限。因此，我们采用“配对生成”策略——使用同一文本提示分别驱动文本到图像模型（FLUX.1）和文本到音频模型（Stable Audio Open [13]）以确保生成场景的语义一致性。随后，在数据融合阶段，我们采用指令引导图像编辑技术 FLUX.1-Kontext-dev [14]，将人脸图像无缝融入场景，并以随机信噪比将相应语音与环境声音混合。对于微调集，我们从真实视频中提取并分离独立的视觉背景、环境声音和语音音频元素，以提升模型的泛化能力。

为确保数据集质量与跨模态一致性，我们设计了一套基于视觉语言模型（视觉语言模型）与大语言模型（大语言模型）协作的自动化评估流程。具体而言，Qwen2.5VL [15] 生成合成图像的结构化描述，随后由一组大语言模型（Qwen3 [16], DeepSeek-V3 [17], 和GPT-4o[18]）联合评估，对图像内容与场景标签之间的语义对齐进行评分，并过滤低质量样本。通过这一流程，合成样本的自动场景匹配率达到98.6%。此外，领域专家开展的主观评估显示，视听一致性率高达95.4%。这些结果有力证明了Vivid-210K数据集的可靠性，为模型训练奠定了坚实基础。

2.2. 系统概述

所提出的 VividVoice 框架旨在实现端到端的视觉驱动场景到语音合成。如图 2 所示，整体架构由三个关键组件构成：内容生成路径、场景感知路径和潜在扩散主干网络。在内容生成路径中，我们遵循 VITS [19]确立的成功范式。输入文本首先经过文本编码器和时长预测器处理，再通过单调对齐搜索（MAS）与声学特征帧对齐，生成与声学表示在时间上对齐的文本潜变量。该路径确保合成语音在语言内容方面准确度高，同时具有自然的韵律和节奏。在场景感知路径中，模型接收多模态场景输入，包括语音（Voice）和图像（Image）。为提取语义表示，我们采用预训练的 CLAP 模型作为音频编码器，并采用预训练的 MetaCLIP [20] 模型作为视觉编码器。随后，这些表示被送入我们的核心模块 D-MSVA，该模块对跨模态场景信息进行细粒度对齐与融合，最终生成统一的召回场景嵌入。最后，内容生成路径中的文本潜变量与场景感知路径中的召回场景嵌入共同对潜在扩散模型进行条件控制，由该模型通过迭代去噪生成声学潜变量 z_0。然后，这些潜变量经预训练的 VAE解码器和声码器解码，生成与视觉场景在声学上高度匹配的高保真语音波形。

2.3. 基于解耦记忆的场景-语音对齐

受MFVA [21], 启发，我们提出一种新颖的解耦多模态场景–语音对齐（D-MSVA）模块，以实现复杂视觉场景信息与混合音频信号之间的细粒度对齐。与传统对齐模块不同，D-MSVA的核心思想是通过多个专用记忆库，在架构层面将高度纠缠的多模态特征解耦为两个独立且具有语义意义的成分——说话人音色与场景音效。

该模块由四个可学习的记忆库（$M_{\{\cdot\}} \in \mathbb{R}^{N\times D}$）组成，共同搭建从视觉空间到听觉空间的桥梁：用于编码视觉概念的角色键记忆库（M_{pk}）和环境键记忆库（M_{ek}），以及用于存储听觉基元的音色值记忆库（M_{tv}）和声音值记忆库（M_{sv}）。其中，N 表示每个解耦概念记忆库（即角色/音色和环境/声音）中的槽位数量，D 表示嵌入维度。在训练期间，D-MSVA通过并行双路径机制进行优化。听觉路径旨在确保听觉记忆库的表征能力。给定同时包含音色和音效的混合听觉嵌入$a \in \mathbb{R}^D$，该路径查询值记忆库（M_{tv} 和 M_{sv}）以解耦其中的成分，并计算两组参考注意力权重 w'_t 和 w'_s。例如，音色权重$w'_{t,\,i}$ 的计算方式为：

$$w'_{t,i} = \mathrm{softmax}\left(\frac{a^T m_{tv,i}}{||a||_2||m_{tv,i}||_2}\right) \tag{1}$$

其中 $m_{tv,i}$ 表示声音值记忆库中第 i 个可学习槽向量。声音权重 w'_s 的计算方式与此类似。利用这两组权重，可从相应的值记忆库中重建音色成分和音效成分，并将二者组合为最终重建的听觉嵌入$\hat{a}$。该过程通过最小化听觉重建损失 L_{rec}进行监督。同时，与公式（1）类似，给定同时包含角色和环境信息的混合视觉嵌入$v \in \mathbb{R}^D$，视觉路径查询关键记忆库，以计算相应的角色权重 w_p 和环境权重 w_e。获得两组解耦的视觉注意力权重后，该路径的核心任务是跨模态检索与合成：使用 w_p 查询音色值记忆库（M_{tv}），使用 w_e 查询声音值记忆库（M_{sv}）。随后将所得结果组合，形成“召回”的场景嵌入：$\hat{a}_v = M_{tv}^T w_p + M_{sv}^T w_e$。为促使两条路径协同工作，我们引入额外的注意力对齐损失 L_{align}，通过 KL散度约束二者之间的一致性

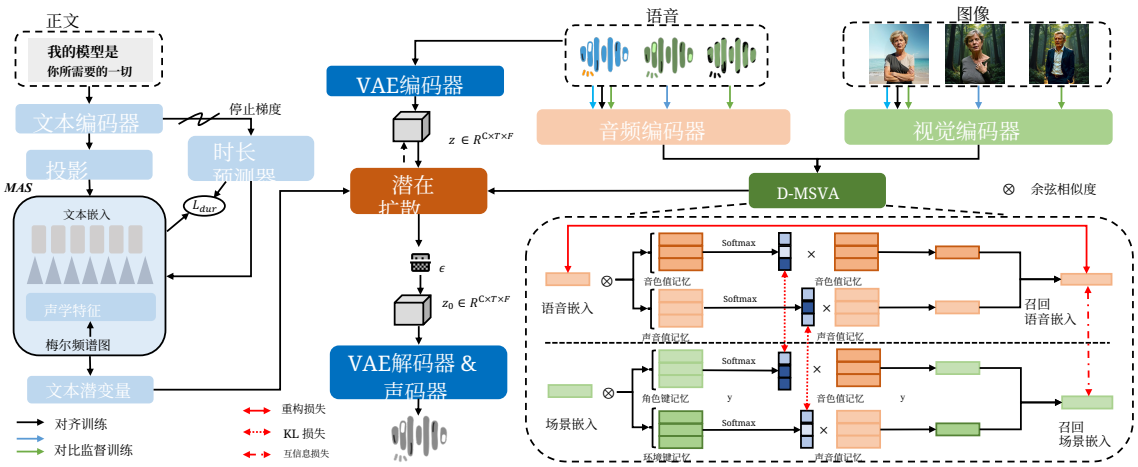


图 2：VividVoice框架的整体架构。

即通过 KL散度约束听觉和视觉路径的注意力权重分布：

$$L_{align} = D_{KL}(w'_t||w_p) + D_{KL}(w'_s||w_e) \quad (2)$$

通过重建与对齐的联合监督，D-MSVA模块以端到端方式进行训练，从而学习从视觉场景到其对应声学特征的细粒度解耦映射。

2.4. 跨模态混合解耦监督策略

为了充分发挥 D-MSVA 模块的解耦潜力，我们设计了一种混合监督策略。该策略分为对齐监督和对比解耦监督。当训练样本由标准配对数据构成时，我们采用对齐监督，旨在学习模态之间的精确映射。该损失由三部分组成：听觉重建损失（L_{rec}）、注意力对齐损失（L_{align}）和模仿损失（L_{imi}）。模仿损失确保视觉路径生成的解耦成分能够精确模仿听觉路径解耦出的对应成分：

$$L_{imi} = ||a'_{timbre} - \hat{a}_{timbre}||_2^2 + ||a'_{sound} - \hat{a}_{sound}||_2^2 \quad (3)$$

当我们从数据集中采样具有特定属性差异的数据对时，会引入对比解耦监督，明确引导模型学习解耦。具体而言，给定来自同一角色但处于不同环境的数据对时，我们施加音色一致性损失（L_{timbre_c}），以确保同一角色的视觉输入生成的音色特征保持一致：

$$L_{timbre_c} = ||\hat{a}_{timbre_A} - \hat{a}_{timbre_B}||_2^2 \quad (4)$$

另一方面，当给定来自不同角色但处于相同环境的数据对时，我们施加环境一致性损失（L_{env_c}），以确保由同一环境的视觉输入生成的环境声音特征尽可能相似：

$$L_{env_c} = ||\hat{a}_{sound_A} - \hat{a}_{sound_B}||_2^2 \quad (5)$$

总之，训练期间 D-MSVA模块的总损失函数是所有损失分量的加权和，如下所示：

$$L = L_{rec} + \lambda_1 L_{align} + \lambda_2 L_{imi} + \lambda_3 L_{timbre_c} + \lambda_4 L_{env_c} \quad (6)$$

其中，λ_1、λ_2、λ_3和 λ_4 为常数权重，用于控制各个分量的重要性。

3. 实验

3.1. 实现细节

我们的所有实验均基于提出的Vivid-210K数据集。我们使用16万个样本进行训练，其余样本留作验证，并将真实场景微调集中的12名未见说话人作为测试集，以评估模型在真实场景中的性能。所有音频样本均下采样至16kHz；当音频与环境音效融合时，应用4至20 dB之间的随机信噪比（SNR）。我们采用与 AudioLDM [22] 扩散模型相同的U-Net结构，并从头开始训练。主要训练阶段使用8块NVIDIA A100 GPU进行，共训练800,000步，每块GPU的批量大小为8。采用AdamW优化器，固定学习率为1e-5。对于D-MSVA模块，内部记忆库中的槽位数量设为 $N = 128$，各个损失分量的权重超参数设置如下：$\lambda_1 = 10$、$\lambda_2 = 2$、$\lambda_3 = 0.5$ 和 $\lambda_4 = 0.5$。随后进行微调阶段，共持续100,000步，使用指数移动平均（EMA）[23] 和自动混合精度（AMP） [24] 处理，有效批量大小为16，学习率调整为5e-5。

3.2. 指标

客观指标： 为定量评估 VividVoice 模型的性能，我们采用一组客观指标。词错误率（WER）用于衡量内容准确度和清晰度（由预训练的 Whisper-Large-v3 模型计算）。整体音频保真度通过 Fr´echet 音频距离（FAD）和 KL散度（KL）进行评估，二者分别比较生成音频与真实音频的特征分布和概率分布的相似度。为评估核心的视听一致性，我们使用 CLAP$_{cap}$ 分数 [25]，通过计算生成音频与原始图像描述之间的匹配分数，衡量视觉到声学转换的质量。

主观指标： 在主观评估中，我们开展了一项众包听力测试，由25名评估者从四个维度按照1至5分的量表给出平均意见分数（MOS）。四个维度分别为：整体自然度（MOS-NA），用于评估音频的整体感知质量；场景一致性（MOS-SC），用于评估音频的声学环境与视觉场景之间的对齐程度；音色一致性（MOS-TI），用于评估生成语音的音色特征与图像所定义的特征之间的匹配程度；以及内容保真度（MOS-CO），用于

表1：不同方法的基准评估实验结果。这里，↑表示数值越高越好，而↓表示数值越低越好。

模型	词错误率(%)↓	FAD↓	KL↓	CLAP$_{cap}$ ↑	MOS-CO↑	MOS-TI ↑	MOS-SC ↑	MOS-NA ↑
GT	10.62	-	-	0.39	4.36	4.03	4.11	4.25
VoiceLDM	9.23	4.74	1.79	**0.27**	3.23	1.75	2.56	3.41
VividVoice（我们的方法）	**7.15**	**3.98**	**1.53**	0.25	**3.95**	**3.08**	**4.30**	**3.88**

衡量语音的可懂度和内容准确度。

3.3. 评估结果

由于现有视觉驱动模型通常只处理不含语言内容的单属性映射，因此因任务不匹配而未纳入定量比较。因此，我们选择 VoiceLDM 作为主要基线进行比较，因为其最终目标——生成具备环境场景感知能力的沉浸式语音——与我们的工作最为接近。如表1所示，在真实世界测试集上，我们的模型在大多数客观和主观评估指标上均显著优于基线。在语音清晰度和音频保真度方面，VividVoice 的词错误率（WER）为7.15%，不仅低于 VoiceLDM 的9.23%，也优于真实录音（真实值）。此外，该模型在 FAD 和 KL散度上的得分也更优，分别为3.98和1.53，展现出卓越的内容保持能力和更高的音频生成质量。主观评估显示，我们的模型在音色一致性（MOS-TI）和场景一致性（MOS-SC）上的得分分别为3.08和4.30，远高于 VoiceLDM 的1.75和2.56。尽管由于描述文本中的无声视觉细节造成干扰，CLAP$_{cap}$ score 略低，但这一指标偏差并不能反映音频保真度，未来可通过相关性过滤予以缓解。这些结果有力地证明了 VividVoice 框架在整体性能上的优势：它既能实现高保真音频生成，又能与视觉场景保持高度一致，从而带来更加沉浸的听觉体验。

3.4. 消融研究

为验证所提出的 D-MSVA 模块在多模态对齐和特征解耦方面的有效性，我们进行了消融研究。如表2所示，我们将完整的 VividVoice 模型与两种主流融合方法进行了比较：（1）w/ Concat-Fusion，一种直接拼接多模态嵌入的简单基线；（2）w/ Attn-Fusion，一种采用标准交叉注意力机制 [26] 进行融合的强基线。实验结果表明，我们提出的 D-MSVA 在所有客观指标上均取得了最佳性能。具体而言，与更强的 Attn-Fusion 基线相比，我们的方法在 FAD 和 KL 散度上分别取得了 16.0% 和 9.5% 的相对提升。这一结果有力地证明，D-MSVA 凭借其独特的解耦记忆库设计和混合监督策略，能够更有效地融合多模态信息，并将其转化为高质量音频生成。相比之下，简单的拼接或注意力机制难以达到同等水平的细粒度对齐与控制。我们进一步研究了 D-MSVA模块中记忆库槽位数量（N）的影响，该参数共同决定了两个解耦概念“角色/音色”和“环境/声音”的表征能力。如图3所示，增大 N 虽然能提高对齐精度（CLAP$_{cap}$），但当 N 超过 128 时，由于过拟合，音频生成质量（FAD）开始下降。因此，我们建议将 N 设为 128。

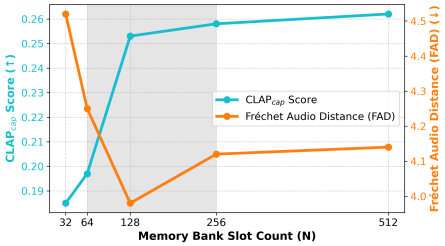


图3：记忆库槽位数量的消融研究。**表2**：D-MSVA模块的消融实验结果。w/ Attn-Fusion 使用标准的基于注意力的融合模块，而 w/ Concat-Fusion 使用简单的拼接融合模块。

模型	词错误率(%)↓	FAD↓	KL↓	CLAP$_{cap}$ ↑
VividVoice w/ Concat-Fusion	7.85	4.87	1.85	0.18
VividVoice 搭配 Attn-Fusion	7.31	4.74	1.69	0.24
VividVoice(D-MSVA)	**7.15**	**3.98**	**1.53**	**0.25**

3.5. 解耦能力评估

为明确评估 VividVoice 的解耦能力，我们将模型与基线模型（w/ Attn-Fusion）进行了 A/B偏好测试。测试包含两种情形：固定角色、变化环境（FC-VE），以及固定环境、变化角色（FE-VC）。评估者需要选择在保持固定属性不变的同时，能够更好地独立控制变化属性的模型。如图4所示，在“固定角色”和“固定环境”测试中，VividVoice 的偏好率分别达到 64% 和 53%。这些结果有力地证明，我们的模型在独立控制音色和环境方面具有显著优势。

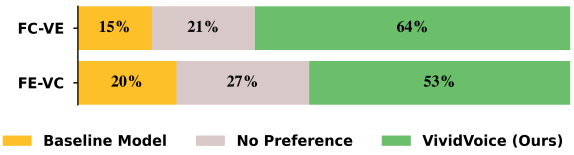


图4：解耦能力的A/B偏好测试结果。

4. 结论

本文提出并定义了一项新任务——场景感知的视觉驱动语音合成，并将 VividVoice 框架作为该任务的统一解决方案。为支持这项任务，我们系统地构建了一个名为 Vivid-210K 的大规模多模态数据集。核心模块 D-MSVA 凭借其独特的解耦式记忆库设计和跨模态混合监督策略，成功实现了从视觉场景到音色及环境声学特征的细粒度对齐。主观和客观实验结果均有力证明，VividVoice 在音频保真度、内容清晰度和多模态一致性方面显著优于现有模型。消融实验进一步验证了，与传统融合方法相比，D-MSVA 模块在提升音频质量方面更具优势。

5. 参考文献

[1] Yeonghyeon Lee、Inmo Yeon、Juhan Nam 和 Joon Son Chung，“Voiceldm：结合环境上下文的文本转语音”，载于 *ICASSP 2024-2024 IEEE*声学、语音与信号处理国际会议 *(ICASSP)*。IEEE，2024年，第 12566–12571 页。[2] Jaekwon Im 和 Juhan Nam，“Diffrent：用于语音录音环境迁移的扩散模型”，载于*ICASSP 2024- 2024 IEEE*声学、语音与信号处理国际会议 *(ICASSP)*。IEEE，2024年，第 7425–7429 页。[3] Jaemin Jung、Junseok Ahn、Chaeyoung Jung、Tan Dat Nguyen、Youngjoon Jang 和 Joon Son Chung，“Voicedit：用于环境感知语音合成的双条件扩散 Transformer”，载于 *ICASSP 2025-2025 IEEE*声学、语音与信号处理国际会议 *(ICASSP)*。IEEE，2025年，第 1–5 页。[4] Apoorv Vyas、Bowen Shi、Matthew Le、Andros Tjandra、Yi-Chiao Wu、Baishan Guo、Jiemin Zhang、Xinyue Zhang、Robert Adkins、William Ngan 等，“Audiobox：使用自然语言提示的统一音频生成”，*arXiv*预印本 *arXiv:2312.15821*，2023年。[5] Mahmoud Ahmed Mahmoud Ahmed、Karim Ahmed Elghamrawy 和 Zeinab Abd El Haliem Taha，“(voick)：通过声音克隆技术提升有声书的无障碍体验”，载于*2024*年第六届计算与信息学国际会议 *(ICCI)*。IEEE，2024年，第 46–52 页。[6] Evelyn Kurniawati、Luca Celetto、Nicola Capovilla 和 Sapna George，“多模态用户界面中的个性化语音命令系统”，载于 *2012*年*IEEE*新兴信号处理应用国际会议。IEEE，2012年，第 45–47 页。[7] Daxin Tan、Guangyan Zhang 和 Tan Lee，“环境感知文本转语音合成”，载于 *Interspeech 2022*，2022年，第 481–485 页。[8] Miseul Kim、Soo-Whan Chung、Youna Ji、Hong-Goo Kang 和 Min-Seok Choi，“置身场景之中：面向沉浸式语音生成的基于扩散模型的声学场景迁移”，载于*Interspeech2024*，2024年，第 4883–4887 页。[9] Shunsuke Goto、Kotaro Onishi、Yuki Saito、Kentaro Tachibana 和 Koichiro Mori，“Face2speech：利用从人脸图像预测的嵌入向量实现多说话人文本转语音合成”，载于*Interspeech*，2020年，第 1321–1325 页。[10] Wei Guo、Heng Wang、Weidong Cai 和 Jianbo Ma，“声音全收：具备声源感知能力的视觉到音频生成”，*arXiv*预印本，第 arXiv–2411 页，2024年。[11] Jort F. Gemmeke、Daniel P. W. Ellis、Dylan Freedman、Aren Jansen、Wade Lawrence、R. Channing Moore、Manoj Plakal 和 Marvin Ritter，“Audio set：面向音频事件的本体与人工标注数据集”，载于*2017*年*IEEE*声学、语音与信号处理国际会议 *(ICASSP)*，2017年，第 776–780 页。[12] Triantafyllos Afouras、Joon Son Chung 和 Andrew Zisserman，“Lrs3-ted：用于视觉语音识别的大规模数据集”，*arXiv*预印本 *arXiv:1809.00496*，2018年。[13] Zach Evans、Julian D Parker、CJ Carr、Zack Zukowski、Josiah Taylor 和 Jordi Pons，“Stable audio open”，载于 *ICASSP 2025- 2025 IEEE*声学、语音与信号处理国际会议 *(ICASSP)*。IEEE，2025年，第 1–5 页。

[14] Stephen Batifol、Andreas Blattmann、Frederic Boesel、Sak- sham Consul、Cyril Diagne、Tim Dockhorn、Jack English、Zion English、Patrick Esser、Sumith Kulal 等，“Flux. 1 kontext：用于潜在空间中上下文图像生成与编辑的流匹配”，*arXiv* 预印本，第 arXiv–2506 页，2025年。[15] Shuai Bai、Keqin Chen、Xuejing Liu、Jialin Wang、Wenbin Ge、Sibo Song、Kai Dang、Peng Wang、Shijie Wang、Jun Tang 等，“Qwen2. 5-vl 技术报告”，*arXiv* 预印本 *arXiv:2502.13923*，2025年。[16] An Yang、Anfeng Li、Baosong Yang、Beichen Zhang、Binyuan Hui、Bo Zheng、Bowen Yu、Chang Gao、Chengen Huang、Chenxu Lv 等，“Qwen3 技术报告”，*arXiv* 预印本 *arXiv:2505.09388*，2025年。[17] Aixin Liu、Bei Feng、Bing Xue、Bingxuan Wang、Bochao Wu、Chengda Lu、Chenggang Zhao、Chengqi Deng、Chenyu Zhang、Chong Ruan 等，“ Deepseek-v3 技术报告”，*arXiv* 预印本 *arXiv:2412.19437*，2024年。[18] Aaron Hurst、Adam Lerer、Adam P Goucher、Adam Perelman、Aditya Ramesh、Aidan Clark、AJ Ostrow、Akila Welihinda、Alan Hayes、Alec Radford 等，“Gpt-4o 系统卡”，*arXiv* 预印本 *arXiv:2410.21276*，2024年。[19] Jaehyeon Kim、Jungil Kong 和 Juhee Son，“结合对抗学习的条件变分自编码器，用于端到端文本转语音”，载于国际机器学习会议。PMLR，2021年，第5530–5540页。[20] Hu Xu、Saining Xie、Xiaoqing Ellen Tan、Po-Yao Huang、Rus- sell Howes、Vasu Sharma、Shang-Wen Li、Gargi Ghosh、Luke Zettlemoyer 和 Christoph Feichtenhofer，“揭开 CLIP 数据的神秘面纱”，*arXiv* 预印本 *arXiv:2309.16671*，2023年。[21] Zheng-Yan Sheng、Yang Ai、Yan-Nian Chen 和 Zhen-Hua Ling，“基于人脸驱动并结合记忆的人脸—语音对齐零样本语音转换”，载于第*31*届 *ACM*国际多媒体会议论文集，2023年，第8443–8452页。

[22] Haohe Liu、Zehua Chen、Yi Yuan、Xinhao Mei、Xubo Liu、Danilo Mandic、Wenwu Wang 和 Mark D Plumbley，《Audi- oldm：基于潜在扩散模型的文本生成音频》，*arXiv* 预印本 *arXiv:2301.12503*，2023年。

[23] Frank Klinker，《指数移动平均与移动指数平均》，*Mathematische Semesterberichte*，第58卷，第1期，第97–107页，2011年。[24] Paulius Micikevicius、Sharan Narang、Jonah Alben、Gregory Diamos、Erich Elsen、David Garcia、Boris Ginsburg、Michael Houston、Oleksii Kuchaiev、Ganesh Venkatesh 等，《混合精度训练》，*arXiv* 预印本 *arXiv:1710.03740*，2017年。[25] Yusong Wu、Ke Chen、Tianyu Zhang、Yuchen Hui、Taylor Berg-Kirkpatrick 和 Shlomo Dubnov，《通过特征融合与关键词到描述增强进行大规模对比式语言-音频预训练》，载于*ICASSP 2023-2023 IEEE* 国际声学、语音与信号处理会议 *(ICASSP)*。IEEE，2023年，第1–5页。[26] Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszko- reit、Llion Jones、Aidan N Gomez、Lukasz Kaiser 和 Illia Polosukhin，《注意力就是你所需要的一切》，神经信息处理系统进展，第30卷，2017年。