arXiv:2601.01459v1 [cs.SD] 4 Jan 2026

OV-INSTRUCT 文本转语音：迈向开放词汇指令式文本转语音

Yong Ren[12] *Jiangyan Yi*[3†]、*Jianhua Tao*[34†]、*Haiyang Sun*[1],*Zhengqi Wen*[4†]、*Hao Gu*[12],*Le Xu*[1], *Ye Bai* [1]

[1]中国科学院多模态人工智能系统国家重点实验室、自动化研究所，中国[2]中国科学院大学人工智能学院，中国[3]清华大学自动化系，中国[4]清华大学北京信息科学与技术国家研究中心，中国

摘要

指令式文本转语音（InstructTTS）利用自然语言描述作为风格提示来引导语音合成。然而，现有的 InstructTTS 方法主要依赖音频相关标签的直接组合或其各种改写，因此难以处理灵活的高层指令。对于希望通过描述性指令引导生成的内容创作者等用户而言，这种僵化的控制方式并不够用。为了解决这些限制，我们提出了**OV-InstructTTS**，一种开放词汇指令式文本转语音的新范式。我们提出了一套完整的解决方案，包括新整理的数据集 OV-Speech 和一种新颖的推理驱动框架。OV-Speech 数据集将语音与开放词汇指令配对，并为每条指令补充推理过程，以连接高层指令与声学特征。该推理驱动框架会在合成语音之前，从开放词汇指令中推断情绪、声学和副语言信息。评估结果表明，这种推理驱动方法显著提升了指令遵循忠实度和语音表现力。我们相信，这项工作能够启发新一代更易用、泛化能力更强且更适用于现实场景的 InstructTTS 系统。数据集和演示已在我们的项目页面公开提供。[1]

索引词— InstructTTS、开放词汇、推理

1. 引言

近年来，文本转语音（TTS）技术取得了显著进展。先进的 TTS 系统已达到人类水平的语音质量，并展现出仅凭一小段音频样本即可克隆任意说话人音色的能力 [1–7]。这些系统的快速发展与普及加速了内容创作、个性化以及人机交互。随着 TTS 技术逐步走向实际部署，提升文本转语音可定制性已成为研究重点 [8]。

为了提升文本转语音可控性，近期研究开始探索将自然语言描述用作指令，这类方法被称为 InstructTTS [9–15]。PromptTTS [9] 首次引入了基于五种预定义属性的风格提示。随后，InstructTTS [10], 进行了扩展，采用了更长且描述性更强的句子。更新的系统，如 PromptTTS2 [11], 、Textrol-Speech [12], 和 SpeechCraft [13], ，利用大语言模型（大语言模型）生成详细且多样的风格提示。此外，VoxInstruct [14] 转向通过以自由形式的人类指令取代相互独立的内容提示和风格提示，从而更全面地捕捉用户意图。

然而，现有的 InstructTTS 系统仍受到其对直接声学属性（例如音高、语速和情绪）依赖的根本性限制。这些系统中的指令通常源自预定义声学标签的组合或多样化改写。尽管这些方法能够实现基本控制，但由于无法捕捉高层指令与低层声学特征之间的复杂映射，它们限制了创造力。正如 SpeakEasy [16], 所指出的，内容创作者仍然缺乏向 TTS 模型直观传达其表达意图的方式。最近，InstructTTSEval [17] 引入了角色扮演指令任务，但该任务仅限于一个小型评测集，并未提供解决方案。

本文提出了一种新范式**OV-InstructTTS**，旨在让模型能够根据开放词汇指令直接合成符合用户意图的语音，从而使 InstructTTS 更加易用、灵活。为此，我们基于 ContextSpeech 语料库[2]构建了**OV-Speech**这一多层数据集。OV-Speech 的一项关键特色是，其指令源自叙事语境，而非对声学或情感标签的改述。这种设计确保指令表达多样、不受拘束，同时与目标语音保持一致。为了弥合高层指令与底层声学特征之间的差距，我们为每个样本添加了由大语言模型生成的推理链。此外，我们还为转录文本增添了副语言标签。

基于 OV-Speech 数据集，我们进一步为 OV-InstructTTS 任务提出了一个推理驱动框架——OV-InstructTTS-Thinking-Emotion-Paralanguage（**OV-InstructTTS-TEP**）。其主要挑战在于弥合高层开放词汇指令与低层声学实现之间显著的语义鸿沟。Kimi-Audio [18], Qwen2.5-Omini [19],和 Step-Audio 2 [20], 等大型音频语言模型（LALMs）的最新进展，展现了它们在理解、推理和生成方面强大的统一能力。这些特点使 LALMs 尤其适用于 OV-InstructTTS，因为它们的推理能力使其能够在合成之前，根据开放式指令推断出恰当的声音风格。我们构建了一个推理驱动系统，将 OV-InstructTTS 重新构想为一种新范式：首先通过思维过程，从开放词汇指令中推导出明确的情绪和声学描述；随后生成音频与文本令牌交错排列的序列，其中的文本融入了情感标签和副语言标签。我们的主要贡献如下：

†通讯作者[1]

[1]https://y-ren16.github.io/OV-InstructTTS

[2]https://huggingface.co/datasets/Insects/ContextSpeech

• **范式：** 本文提出了 OV-InstructTTS 这一新颖范式，使指令式 TTS 摆脱对改述音频属性的依赖，推动可控语音合成迈向更灵活、更便于用户使用的现实应用。

• **数据集：** 我们构建了 OV-Speech，这是一个为这一范式奠定基础的大规模数据集。它包含源自叙事语境的开放词汇指令、将指令与声学特征联系起来的推理链，以及添加了副语言标签的转录文本。

• **方法：** 本文提出了 OV-InstructTTS-TEP，这是一种基于大语言音频模型的新型推理驱动型 OV-InstructTTS 框架。我们的方法旨在通过推理过程理解开放式指令，从而生成高度富有表现力且符合用户意图的语音。

• **实验：** 大量实验和消融研究证明了我们数据集的价值以及 OV-InstructTTS-TEP 框架的有效性。大语言模型即评判者和主观评估证实，合成语音与开放式指令保持一致。

2. OV-Speech数据集构建

开发 OV-InstructTTS 系统面临的首要挑战，是缺乏将语音与开放词汇指令配对的合适数据。为此，我们构建了 OV-Speech，这是一个基于 ContextSpeech语料库的大规模多层数据集。ContextSpeech语料库包含 476.8 小时的多说话人有声书，并配有对应的小说。我们以此为基础，利用其中的单句音频话语、文本转录，以及现有的情感标签和声学描述。如图 1 所示，OV-Speech 的构建遵循五阶段流程。

上下文信息提取。 开放词汇指令的基础，是围绕每段话语展开的丰富叙事内容。对于每个音频样本，我们首先将其与来源小说对齐，并提取目标话语前后各 1000 个词的上下文窗口。随后，我们使用具备强大理解与推理能力的 Qwen3-32B[3] [21]，将这些原始文本提炼为一组结构化上下文要素，包括环境描述、当前事件、说话人的性格、对话者的状态以及说话人的意图。这一过程有效捕捉了与语音相关的细微情境信息

性能。

开放词汇指令生成。 随后，我们基于提取出的上下文元素生成多样化的高层指令。对于每个样本，我们随机选择 2–5 个元素，并将其提供给 Qwen3-32B，提示其合成一条开放词汇指令，类似于导演可能会向配音演员提出的指令。这一包含随机选择和创意生成的两步过程，确保生成大量不受约束且具有描述性的指令，并使其扎根于叙事内容之中。

一致性筛选。 一项关键挑战是确保从小说叙述中提取的开放词汇指令确实与语音音频的声学风格相符。为此，我们采用大语言模型评审过滤策略：向 Deepseek-R1 [22]提供说话人身份、转录文本和上下文信息，并要求其预测情绪和声学属性。随后，大语言模型评审 Qwen3-32B 会将这些预测与 ContextSpeech 提供的真实标签进行比较，并舍弃一致性得分较低的样本（情绪 < 6 或声学 < 5方面）。

[3]https://huggingface.co/Qwen/Qwen3-32B

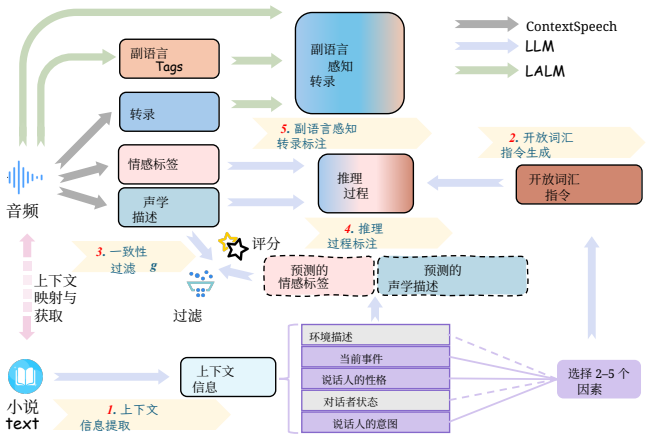


图 1。构建 OV-Speech 数据集的流程。

表 1. 带副语言标签的转录生成方法评估。**红色**表示最佳性能。

	NVSpeech170k-Testset				OV-Speech
方法	**C-F1↑**	**P-F1↑**	**APS↓**	**S-CER(%) ↓**	**S-CER(%) ↓**
PASR	0.87	0.60	7.72	18.91	12.95
PRI	**0.88**	**0.83**	**6.56**	0.12	4.04
PC-PTI	**0.88**	0.82	6.67	**0.03**	**0.35**

推理过程标注。 这一核心标注步骤使我们的推理驱动型 InstructTTS 模型成为可能。对于每个通过一致性过滤器的样本，我们要求 Qwen3-32B 生成一条循序渐进的“推理链”，将高层指令与低层声学属性联系起来。这条推理链分为两个阶段：首先是指令解构，模型在此阶段识别指令所暗示的关键上下文要素；其次是属性推断，模型根据这些要素推断情感标签和声学描述。

副语言感知转录标注。 为了向文本转语音模型提供信息更丰富的预测目标，我们使用 NVSpeech [23] 中的 18 个副语言标签丰富原始转录文本（例如，[笑声]，[咳嗽]）。为此，我们在 NVSpeech170k [5] 数据集上对 Qwen2-Audio-7B [24][4] 模型进行了微调。针对这一任务，我们探索了三种策略：副语言感知语音识别（**PASR**）：直接从音频预测带有副语言标签的转录文本。副语言识别与插入（**PRI**）：同时输入音频和原始转录文本，预测最终的带标签转录文本。副语言分类后标签插入（**PC-PTI**）：一种两阶段方法，首先从音频中预测副语言事件是否出现，然后将相应标签插入原始转录文本中的正确位置。如表 1 所示，我们使用类别*F1*分数（*C-F1*）评估标签分类性能，使用位置*F1*分数（*P-F1*）和平均绝对位置误差（*APS*）评估标签插入位置，并使用稳定性字符错误率（*S-CER*）评估标签插入后的转录完整性。尽管 PRI 方法在标签定位准确性（P-F1 和 APS ）方面略胜一筹，但 PC-PTI 方法展现出压倒性的稳定性优势，几乎完好地保留了原始转录文本（在我们的 OV-Speech 上，S-CER 为 0.35%）。鉴于保留真实文本的完整性至关重要，我们采用 PC-PTI 方法处理

[4]https://huggingface.co/Qwen/Qwen2-Audio-7B [5] https://huggingface.co/datasets/Hannie0813/NVSpeech170k

整个 OV-Speech 数据集。

通过这一五阶段流程，OV-Speech 数据集在 ContextSpeech 的基础上扩展了开放词汇指令、推理链以及带有副语言标签的细粒度转录，为 OV-InstructTTS 奠定了全面而坚实的基础。

3. OV-INSTRUCTTTS-TEP 框架

我们提出了 OV-InstructTTS-TEP，这是一种推理驱动框架，利用大型音频语言模型（LALMs）统一的理解、推理和生成能力来实现 OV-InstructTTS。该框架将 OV-InstructTTS 从直接的高层指令到语音的映射，重新构建为结构化的两步过程。模型不再隐式地学习映射，而是通过明确引导，先生成推理链进行“思考”，再以此为条件合成语音。这一设计利用了 LALMs 强大的上下文学习能力，弥合高层用户指令与低层声学实现之间的语义鸿沟。我们采用 Step-Audio-2-mini-Base[6] [20]作为预训练音频语言模型（ALM）骨干模型，因其具备强大的音频理解能力且可公开获取，并在 OV-Speech 数据集上进行监督微调（SFT）。如图2(a)所示，我们的模型以开放词汇指令和文本转录为输入，并分两步进行：

• **思考令牌生成：**模型首先生成文本推理过程，将开放式指令映射为情感标签、声学描述和副语言标签。它会解析高层指令，识别其中包含哪些上下文要素（环境描述、当前事件、说话人的性格、对话者状态和说话人的意图），然后推断在给定上下文中适用于语音表现的属性。

• **交错文本音频令牌生成：**模型以此前的推理历史为条件，生成文本令牌与音频令牌交错排列的序列。文本令牌表示经过增强的原始转录文本，并以如下格式加入推断出的属性： [情感标签]带有 <副语言标签>的转录文本。随后，使用与 Step-Audio-2-mini-Base 相同的流匹配模型和 HiFiGAN 声码器，将离散音频令牌合成为波形。

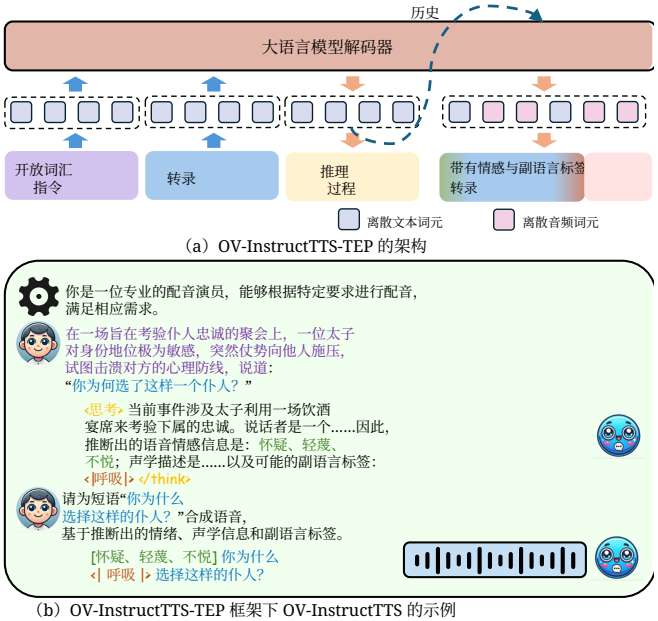
图2(b)给出了一个具体示例。给定一个复杂的坐姿——在 uational 指令下，我们的模型首先生成一个 <think> 块，分析上下文以推断情感标签（[怀疑、蔑视、不悦]）、声学描述和副语言标签（<呼吸>）。随后，它预测增强转录文本的词元（[怀疑、蔑视、不悦] 你为什么要 <呼吸> 选择这样一个仆人？）以及相应的音频，最终合成语音。这种由推理驱动的方法是生成富有表现力语音的关键，使其与开放式自然语言描述相契合，并确保生成的声音能够恰当地体现超越字面文本内容的细腻情绪和副语言线索。

4. 实验

4.1. 实验设置

数据集。我们将提出的 OV-Speech 数据集划分为训练集和测试集。训练集包含从 83部小说中抽取的 316,807 条话语

[6]https://github.com/stepfun-ai/Step-Audio2



（a）OV-InstructTTS-TEP 的架构

（b）OV-InstructTTS-TEP 框架下 OV-InstructTTS 的示例

图2。我们提出的基于推理的 OV-InstructTTS 模型架构及示例。

，每条话语都配有三条不同的开放词汇指令。测试集包含来自 3 部留出小说的 1,500 条话语，以确保训练期间未见过这些说话人和叙事上下文。

实现细节。我们的系统基于 Step-Audio-2-mini-Base 模型构建。我们使用 OV-Speech 对模型进行微调，学习率为 1×10^{-5}，全局批量大小为 32。所有实验均在 8 块 NVIDIA A100（80GB）GPU 上进行。

基线模型。我们将模型与多个高性能文本转语音或大语言音频模型系统进行比较。GPT-4o 通过其官方 API 进行评估。Higgs-Audio-V2[7]，、CosyVoice2[8]，和 Step-Audio-2-mini[9] 使用其公开的代码库和检查点进行评估。我们选择这些基线模型是因为它们性能出色：CosyVoice2 是广受认可的文本转语音模型，以合成语音的高自然度著称；其余三个模型则展现出显著的指令遵循能力，能够根据开放词汇描述合成语音。

4.2. 评估指标

我们从多个角度评估性能：**指令遵循**。为了评估合成语音遵循开放词汇指令的程度，我们采用大语言模型评审方法。Gemini [10] 担任 AI 评估器，输入一组语音样本后，Gemini 会根据语音与开放词汇指令的一致程度进行评分，并给出*Gemini*评分（0–100）和该组样本中的*Gemini*排名 。**可懂度**。我们使用 ASR 模型 Paraformerzh [25] 计算字符错误率（*CER*），以评估语音可懂度。**音色相似度**。我们计算 WavLM-large [26] 模型提取的说话人嵌入之间的余弦相似度（*SIM*），以衡量音

[7]https://github.com/boson-ai/higgs-audio [8] https://github.com/FunAudioLLM/CosyVoice [9] https://github.com/stepfun-ai/Step-Audio2 10gemini-2.5-pro-preview-05-06

表2。与基线系统的整体性能比较。**红色**表示性能最佳，**蓝色**表示次佳。◇表示使用 CosyVoice2 的分词器和流匹配将说话人音色转换为目标说话人。

	Gemini评分↑	Gemini排名↓	字符错误率(%)↓	SIM↑	MOS↑	ICMOS↑
真实值	75.43	2.94/6	3.10	-	4.10 (±0.14)	4.33 (±0.15)
Cosyvoice2（无指令）	66.99	3.59/6	**3.09**	0.659	**3.84 (±0.19)**	2.94 (±0.23)
GPT-4o◇	**68.31**	**3.48/6**	3.89	0.701	3.23 (±0.24)	2.42 (±0.23)
Higgs Audio V2◇	65.10	3.73/6	8.42	**0.707**	3.81 (±0.20)	**3.00 (±0.20)**
Step-Audio-2-mini	67.59	3.56/6	5.49	0.701	3.53 (±0.24)	2.40 (±0.21)
OV-InstructTTS-TEP	**70.42**	**3.39/6**	**3.61**	**0.722**	**4.28 (±0.14)**	**3.91 (±0.17)**

表3。关键组件影响的消融研究：（T）思考过程，以及带有（E）情感标签和（P）副语言标签的增强转录文本中的预测词元。**红色**表示最佳性能，**蓝色**表示次佳性能。

	Gemini评分↑	Gemini排名↓	字符错误率(%)↓	SIM↑	MOS↑	ICMOS↑
真实值	76.02	3.41/8	3.10	-	4.10 (±0.14)	4.33 (±0.15)
(a) Step-Audio-2-mini（无指令）	61.49	4.85/8	8.06	0.684	3.70 (±0.22)	2.57 (±0.22)
(b) Step-Audio-2-mini	63.18	4.75/8	5.49	0.701	3.53 (±0.24)	2.40 (±0.21)
(c) 文本转语音（无指令）	66.34	4.48/8	3.78	0.715	4.15 (±0.21)	3.61 (±0.20)
(d) OV-InstructTTS	67.70	4.40/8	**3.56**	**0.720**	4.23 (±0.16)	3.74 (±0.18)
(e) OV-InstructTTS-EP	66.98	4.45/8	3.65	**0.722**	**4.27 (±0.16)**	3.81 (±0.16)
(f) OV-InstructTTS-T	**68.71**	**4.26/8**	**3.45**	**0.722**	**4.27 (±0.18)**	**3.90 (±0.16)**
(g) **OV-InstructTTS-TEP**	**71.57**	**3.89/8**	3.61	**0.722**	**4.28 (±0.14)**	**3.91 (±0.17)**

色相似度。**主观评估**。我们还邀请 8 位普通话母语者进行了人工听测。听者从两个维度对音频样本进行评分：整体平均意见分（*MOS*），以 5 分制反映感知到的自然度和音频质量；以及指令一致性平均意见分（*ICMOS*），同样采用 5 分制，衡量生成语音与给定指令的符合程度。

4.3. 主要结果

如表2所示，我们提出的 OV-InstructTTS-TEP 在客观和主观指标上均取得了最佳综合性能。在基于 Gemini 的评估中，除真实值外，我们提出的模型获得了最高的 Gemini评分（70.42）和最佳 Gemini 排名（3.39/6）。这表明，与现有文本转语音方法相比，该模型在开放词汇指令遵循方面有了显著提升。在主观评估中，该模型在自然度（平均意见分 4.28）和指令一致性（ICMOS 3.91）两项指标上均优于所有基线，其中平均意见分甚至超过了真实值录音。此外，该模型在保持字符错误率（3.61%）具有竞争力的同时，实现了最佳说话人相似度（相似度 0.722），证实其能够生成清晰且说话人一致的语音。所有这些指标上的优异表现验证了我们方法的有效性。

4.4. 消融研究

为剖析 OV-Speech 数据集及 OV-InstructTTS-TEP 框架中各组件的贡献，我们进行了全面的消融研究，结果汇总于表3。

OV-Speech 数据集。将未经指令微调的 Step-Audio-2-mini（a、b）与在我们的 OV-Speech 数据集上微调的模型（c、d）进行比较，可以看到显著提升：在无指令设置下，Gemini评分从 61.49 提升至 66.34；在指令设置下，则从 63.18 提升至 67.70。平均意见分和 ICMOS 也有大幅提升。此外，在微调后的模型中，基于指令的变体（d）始终优于其无指令对应版本（c），这证实了 OV-Speech 数据集的有效性。

推理的作用。引入明确的推理步骤能够带来稳定的改进。比较（d）和（f）可见，Gemini评分从 67.70 提升至 68.71，同时平均意见分和 ICMOS 也有所提升（分别为 4.23→4.27 和 3.74→3.90）。这支持了我们的假设：推理对于衔接高层指令与低层声学属性至关重要，使模型能够超越记忆关联进行泛化。

推理与细粒度转录文本的协同作用。有趣的是，仅预测带有情感标签和副语言标签的转录文本令牌（e）并不总是优于基线（d），Gemini评分反而略有下降（67.70 → 66.98）。这可能是由于缺乏推理，影响了情绪和标签预测的准确性。然而，当与推理结合时（g），模型取得了最佳结果：Gemini评分上升至 71.57，Gemini排名提升至 3.89/8，平均意见分和 ICMOS 也均有所提高。这表明，推理使模型能够从开放词汇指令中推断细粒度属性，从而使预测出的增强转录文本更具表现力，并与指令保持一致。

5. 结论

本文提出了 OV-InstructTTS，这是一种全新范式，旨在解决现有 InstructTTS 系统的局限性：其指令通常局限于低层声学属性。为支持这一范式，我们构建了 OV-Speech 数据集，其中包含开放词汇指令、推理过程标注以及副语言感知转录文本。在此资源的基础上，我们进一步提出了 OV-InstructTTS-TEP，这是一种由推理驱动的框架：它首先对给定指令进行“思考”，随后交错生成丰富文本和音频令牌，最后合成语音。大量实验表明，我们的方法在指令遵循、自然度和说话人相似度方面均显著优于基线方法，其中推理和丰富文本令牌预测都发挥了关键作用。这项工作为更加直观、灵活且以用户为中心的可控文本转语音铺平了道路。

6. 参考文献

[1] Chengyi Wang、Sanyuan Chen、Yu Wu、Ziqiang Zhang、Long Zhou、Shujie Liu、Zhuo Chen、Yanqing Liu、Huaming Wang 等，“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。[2] Yuancheng Wang、Haoyue Zhan、Liwei Liu、Ruihong Zeng、Haotian Guo、Jiachen Zheng、Qiang Zhang、Xueyao Zhang、Shunsi Zhang 和 Zhizheng Wu，“Maskgct：基于掩码生成式编解码器Transformer的零样本文本转语音”，发表于*ICLR*，2025年。[3] Shijia Liao、Yuxuan Wang、Tianyu Li、Yifan Cheng、Ruoyi Zhang、Rongzhi Zhou 和 Yijin Xing，“Fish-speech：利用大语言模型实现先进的多语言文本转语音合成”，*arXiv* 预印本 *arXiv:2411.01156*，2024年。[4] Philip Anastassiou、Jiawei Chen、Jitong Chen、Yuanzhe Chen、Zhuo Chen、Ziyi Chen、Jian Cong、Lelai Deng、Chuang Ding 等，“Seed-tts：高质量多功能语音生成模型系列”，*arXiv* 预印本 *arXiv:2406.02430*，2024年。[5] Zhihao Du、Yuxuan Wang、Qian Chen、Xian Shi、Xiang Lv、Tianyu Zhao、Zhifu Gao、Yexin Yang、Changfeng Gao、Hui Wang 等，“Cosyvoice 2：基于大语言模型的可扩展流式语音合成”，*arXiv* 预印本 *arXiv:2412.10117*，2024年。[6] Yushen Chen、Zhikang Niu、Ziyang Ma、Keqi Deng、Chunhui Wang、Jian Zhao、Kai Yu 和 Xie Chen，“F5-tts：借助流匹配伪造流畅且忠实语音的童话讲述者”，*arXiv* 预印本 *arXiv:2410.06885*，2024年。[7] Siyi Zhou、Yiquan Zhou、Yi He、Xun Zhou、Jinchao Wang、Wei Deng 和 Jingchen Shu，“Indextts2：情感表达和时长控制自回归零样本文本转语音领域的突破”，*arXiv* 预印本 *arXiv:2506 21619*，2025。。[8] Tianxin Xie、Yan Rong、Pengfei Zhang 和 Li Liu，“大语言模型时代的可控语音合成：综述”，*arXiv* 电子论文，第 arXiv–2412 页，2024年。[9] Zhifang Guo、Yichong Leng、Yihan Wu、Sheng Zhao 和 Xu Tan，“Prompttts：通过文本描述实现可控文本转语音”，发表于*ICASSP 2023-2023*。IEEE，2023年，第 1–5 页。[10] Dongchao Yang、Songxiang Liu、Rongjie Huang、Chao Weng 和 Helen Meng，“Instructtts：利用自然语言风格提示在离散潜空间中建模表达性文本转语音”，*IEEE/ACM*音频、语音与语言处理汇刊，第 32 卷，第 2913–2925 页，2024年。[11] Yichong Leng、Zhifang Guo、Kai Shen、Zeqian Ju、Xu Tan、Eric Liu、Yufei Liu、Dongchao Yang、leying zhang、Kaitao Song、Lei He、Xiangyang Li、sheng zhao、Tao Qin 和 Jiang Bian，“PromptTTS 2：使用文本提示描述并生成声音”，发表于*ICLR*，2024年。[12]Shengpeng Ji、Jialong Zuo、Minghui Fang、Ziyue Jiang、Feiyang Chen、Xinyu Duan、Baoxing Huai 和 Zhou Zhao，“Textrolspeech：结合编解码器语言文本转语音模型的文本风格控制语音语料库”，发表于*ICASSP2024-2024*。IEEE，2024年，第 10301–10305 页。[13] Zeyu Jin、Jia Jia、Qixin Wang、Kehan Li、Shuoyi Zhou、Song- tao Zhou、Xiaoyu Qin 和 Zhiyong Wu，“Speechcraft：具有自然语言描述的细粒度表达性语音数据集”，发表于*ACMMM*，2024年，第 1255–1264 页。[14] Yixuan Zhou, Xiaoyu Qin, Zeyu Jin, Shuoyi Zhou, Shun Lei, Songtao Zhou 等，“Voxinstruct：通过统一的多语言编解码器语言建模，实现富有表现力的人类指令转语音生成”，载于 *ACM MM*，2024年，第 554–563 页。[15] Guanrou Yang, Chen Yang, Qian Chen, Ziyang Ma, Wenxi Chen, Wen Wang, Tianrui Wang, Yifan Yang, Zhikang Niu, Wenrui Liu 等，“Emovoice：基于大语言模型的情感文本转语音模型，支持自由形式文本提示”，*arXiv* 预印本 *arXiv:2504.12867*，2025年。[16] Stephen Brade, Sam Anderson, Rithesh Kumar, Zeyu Jin 和 Anh Truong，“Speakeasy：增强文本转语音交互，助力表现力内容创作”，载于 *2025*年 *ACM CHI* 人类因素计算系统会议论文集，2025年，第 1–19 页。[17] Kexin Huang, Qian Tu, Liwei Fan, Chenchen Yang, Dong Zhang, Shimin Li, Zhaoye Fei, Qinyuan Cheng 和 Xipeng Qiu，“Instructttseval：文本转语音系统中复杂自然语言指令遵循能力的基准测试”，*arXiv* 预印本 *arXiv:2506.16381*，2025年。[18] Ding Ding, Zeqian Ju, Yichong Leng, Songxiang Liu, Tong Liu, Zeyu Shang, Kai Shen, Wei Song, Xu Tan, Heyi Tang 等，“Kimi-audio 技术报告”，*arXiv* 预印本 *arXiv:2504.18425*，2025年。[19] Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang 等，“Qwen2. 5-omni 技术报告”，*arXiv* 预印本 *arXiv:2503.20215*，2025年。[20] 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。[21] 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。[22] Daya Guo, Dejian Ya ng, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi 等，“Deepseek-r1：通过强化学习激励大语言模型的推理能力”，*arXiv* 预印本 *arXiv:2501.12948*，2025年。[23] Huan Liao, Qinke Ni, Yuancheng Wang, Yiheng Lu, Haoyue Zhan, Pengyuan Xie, Qiang Zhang 和 Zhizheng Wu，“Nvspeech：用于类人语音建模（含副语言发声）的集成式可扩展流水线”，*arXiv* 预印本 *arXiv:2508.04195*，2025年。[24] Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin 等，“Qwen2-audio 技术报告”，*arXiv* 预印本 *arXiv:2407.10759*，2024年。[25] Zhifu Gao, Zerui Li, Jiaming Wang, Haoneng Luo, Xian Shi, Mengzhe Chen, Yabin Li, Lingyun Zuo, Zhihao Du 和 Shiliang Zhang，“Funasr：端到端语音识别基础工具包”，载于 *INTERSPEECH*，2023年。[26] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao 等，“Wavlm：面向全栈语音处理的大规模自监督预训练”，*IEEE*信号处理精选主题期刊，第 16 卷，第 6 期，第 1505–1518 页，2022年。