arXiv:2512.19090v1 [cs.SD] 22 Dec 2025

JoyVoice：面向拟人化多说话人对话合成的长上下文条件控制

SpeechTeam，京东

语音实验室，京东探索学院，京东集团

joyai audio llm@jd.com

摘要

大型语音生成模型正从单说话人、短句合成演进为多说话人、长对话生成。当前的长篇语音生成模型主要局限于双人、轮流式交互。为此，我们提出**JoyVoice**，一种新颖的拟人化基础模型，旨在灵活、无边界地合成多达八名说话人的语音。与传统级联系统不同，JoyVoice采用统一的**E2E-Transformer-DiT架构**，直接利用自回归隐表示作为扩散输入，从而实现整体的端到端优化。我们进一步提出一种以12.5赫兹低比特率运行的**MM-Tokenizer**，它融合多任务语义损失与MMSE损失，以有效建模语义信息与声学信息。此外，该模型通过大规模数据扰动实现稳健的文本前端处理。实验表明，JoyVoice在多语言生成（中文、英语、日语、韩语）和零样本语音克隆方面达到了最先进水平。JoyVoice在Seed-TTS-Eval 基准和多说话人长篇对话语音克隆任务上均取得顶尖结果，展现出卓越的音频质量与泛化能力。它在长篇语音的韵律连续性、多说话人对话的节奏丰富度、副语言自然度方面均取得显著提升，并具备更优的可懂度。我们鼓励读者前往https://jea-speech.github.io/JoyVoice收听演示。

1 引言

文本转语音（TTS）合成旨在根据输入文本生成清晰自然的合成语音。近年来，随着大语言模型（LLM）和扩散模型等先进生成模型的出现，以及训练数据规模和模型尺寸的不断扩大，文本转语音领域取得了显著进展，尤其是在零样本语音克隆方面。与传统的单元选择、拼接和统计参数方法相比，现代文本转语音模型如今已能生成与真人语音几乎难以区分的语音。

当前主流的文本转语音范式通常分为自回归方法和非自回归方法。自回归方法通常采用仅解码器Transformer，根据文本 [1, 2]预测离散语音标记。相比之下，扩散模型等非自回归模型则直接根据文本输入 [3, 4]预测声学特征。自回归生成在流式处理能力和降低延迟方面具有优势，而非自回归模型通常拥有更低的实时因子（RTF），因此各自适用于不同的应用

预印本。尚在进行中。

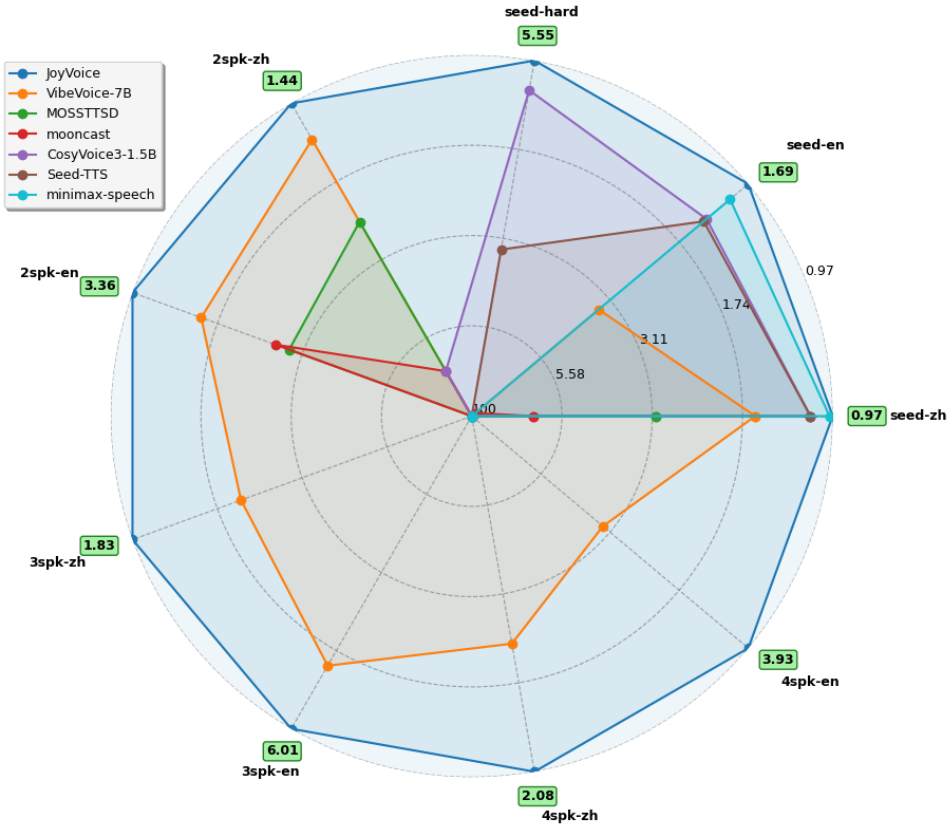


图1：在各种基准测试中，就内容一致性（CER/WER）而言，我们的 JoyVoice 与具有竞争力的语音生成模型之间的性能比较。错误率为 100 表示已发布的模型不支持相应测试任务。

需求。近期，已有研究提出了若干两阶段文本转语音模型 [5–12] 。这些模型结合了自回归大语言模型和非自回归模型：前者根据文本生成语音衍生语义标记，后者（例如扩散模型或其变体，如流匹配模型 [13] 和扩散变换器（DiTs） [14]）则根据这些预测出的标记合成声学特征。

这些离散语义标记充当文本与声学信息之间至关重要的中间表示。然而，固有的量化阻碍了从声学合成阶段（扩散模型）到标记预测阶段（大语言模型）的梯度反向传播，因此必须分别、独立地训练这些组件。此外，常见的语音词元化器 [6, 7] 保持 1:2 的语义标记与声学标记比率，导致大语言模型的输出序列过长，从而降低推理效率。更长的序列还会增加由大语言模型引发生成伪影的风险，这种现象通常称为幻觉。

与上述两阶段模型不同，近期出现的一种趋势是采用基于连续特征的创新型单阶段框架，通常利用变分自编码器（VAE）表示 [15, 16]。在这一范式中，大语言模型的隐藏状态对局部扩散变换器进行条件控制，以生成连续表示。该方法无缝整合了文本到标记和标记到声学阶段，从而能够对所有模型组件进行联合优化。通过避免量化，连续的 VAE 表示保留了原始信息的保真度，并可在输入大语言模型之前高效压缩为图块 [17]。这种压缩显著降低了大语言模型组件的有效帧率（最低可达 5–10 赫兹），提升了计算效率以及生成长语音的能力。

然而，这些单阶段模型也面临着自身的挑战。局部 DiT 组件的感受野有限，可能会损害音色相似度。此外，如果提示语音中含有噪声，这种污染可能会通过连续表示传播至大语言模型，

从而导致合成过程不稳定。因此，对于复杂文本和具有挑战性的场景，生成高质量、富有表现力的语音仍然受到限制。此外，最新的大规模文本转语音模型正开始从单说话人、短篇音频生成转向多说话人、长篇音频生成。然而，目前的工作主要聚焦于两人轮流对话合成 [18–20]。VibeVoice [15] 是少数能够支持生成包含两名以上说话人的长篇音频的大规模文本转语音模型之一。然而，在零样本文本转语音设置下，确保多说话人、长上下文、多轮对话音频合成的稳定性与自然度，仍是一个尚未解决且极具挑战性的问题。

为了解决这些局限性，并从根本上实现自然度高、说话人保真度高的多说话人长篇音频的可扩展、稳定合成，我们提出了 JoyVoice，这是一种高效且可扩展的零样本文本转语音合成器。JoyVoice 原生支持长上下文条件控制，以及多说话人、多轮对话合成。与其他单阶段文本转语音模型类似 [17], , JoyVoice 采用集成式架构，由大语言模型的隐藏状态对 DiT 进行条件控制，从而实现无缝的联合优化。不过，一个关键区别在于，我们采用了全局因果 DiT 模型，专为建模长上下文序列而设计，进而提升合成语音的说话人相似度。同时，我们有策略地重新引入了离散语音标记，类似于 CosyVoice 等先前两阶段模型中的标记，以增强系统稳定性。我们在 DiT 上采用动态分块注意力策略，在单一框架中统一流式与非流式合成，使系统能够灵活适应不同的延迟要求。受 CosyVoice 启发 3 [8], ，我们提出了一种通过多任务学习方法优化的监督式语音语义分词器（MM-tokenizer），将标记速率高效压缩至 12.5 赫兹。这种高压缩率对于高效的长上下文建模和长篇音频生成至关重要。此外，通过采用课程学习策略，JoyVoice 能够有效支持长上下文、多说话人和多轮对话生成。主观和客观评估结果均表明，与最先进的文本转语音合成器相比，所提出的 J oyVoice 在韵律自然度、内容一致性和说话人相似度方面表现更优。此外，JoyVoice 在后训练阶段引入强化学习，以减少发音错误和不自然的输出，进一步提升模型稳定性。

我们的贡献如下：

• 集成式端到端架构：我们提出了一种统一框架，其中 AR-Transformer 的隐藏状态直接为全局因果 DiT 提供条件，从而确保无缝优化并实现卓越的多说话人合成。

• 高效的 MM-Tokenizer：我们推出了一种以 25 赫兹和 12.5 赫兹运行的分词器，通过多任务损失平衡语义理解与声学重建。

• 无需规范化也具备强鲁棒性：借助大规模数据扰动和合成生成，JoyVoice 降低了对文本规范化模块的依赖，同时最大限度地提升稳定性。

• 最先进的性能：JoyVoice 在 Seed-TTS-Eval 基准和复杂的对话式语音克隆任务中取得了最先进的成果，在韵律、一致性和相似度方面超越了现有基线。

2 JoyVoice

2.1 E2E-Transformer-DiT架构

以往的由粗到细 TTS 系统 [5–12] 通常采用两阶段流程：先使用自回归 Transformer 声学模型（AM）进行离散词元预测，再使用独立的流匹配（FM）模型实现词元到梅尔频谱的声学细节恢复。尽管这种解耦方法具有模块化优势，但也存在若干固有局限性：(1) 信息瓶颈：AM 生成的离散令牌可能会丢失对高保真合成至关重要的细粒度韵律和声学信息；(2) 优化不一致性：AM 的训练目标是最大化词元预测似然，而未考虑下游 FM 的重建质量，导致整体性能欠佳；(3) 误差传播：AM 的预测误差会直接影响 FM 的条件输入，且 FM 无法向 AM 提供反馈以提升其输出质量。

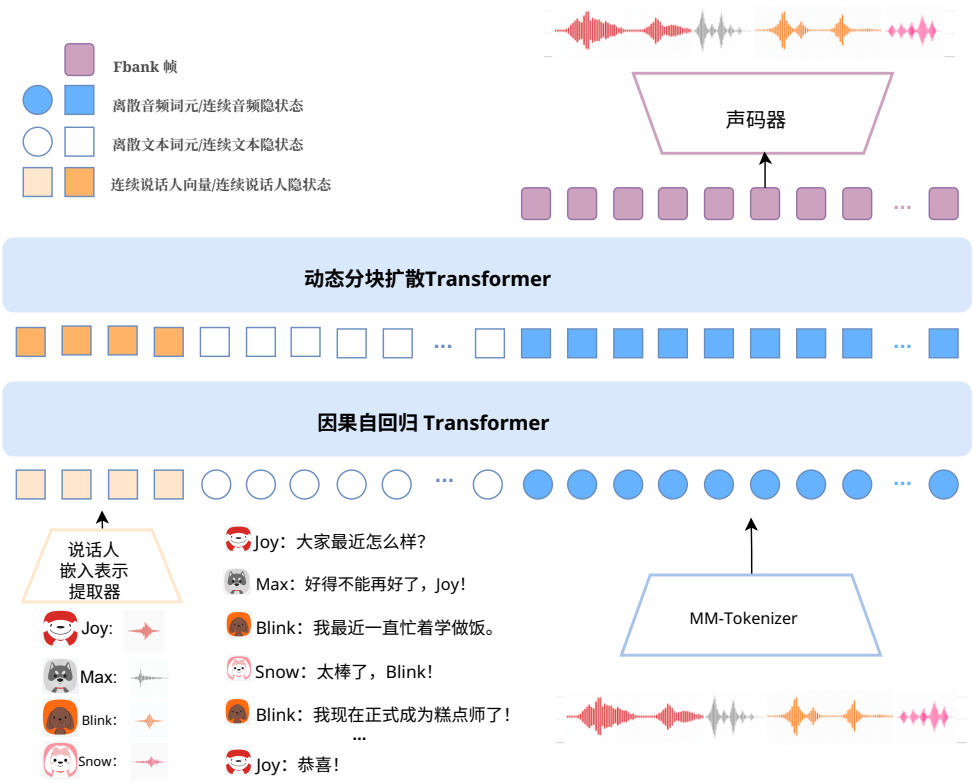


图2：JoyVoice 概览。它由几个关键组件构成：1）因果自回归 Transformer：该模块以系统提示和带有说话人标签的文本为输入，并预测由 MM-Tokenizer 生成的离散令牌。2）动态分块扩散 Transformer：该组件以因果自回归 Transformer 的隐藏表示为输入，并预测梅尔频谱图输出。3）MM-Tokenizer 和声码器：这两个模块分别负责将音频转换为离散令牌表示，以及将梅尔频谱图重建为波形音频。

我们提出了一种联合训练框架，将自回归声学建模阶段与基于流的声学细化阶段统一起来。通过将 FM 损失的梯度反向传递至 AM，我们的方法使声学模型能够学习到不仅适用于离散词元预测，而且对连续声学细节恢复具有最大信息量的表示。这种联合优化范式自然促使 AM 保留仅凭离散令牌无法捕获的声学信息，同时使 FM 能够引导 AM 生成更有利于重建的隐藏表示。

我们提出的架构如图2所示，由两个核心组件构成：（1）一个基于 Qwen 的 [21] 自回归声学模型，用于根据文本预测这些令牌；（2）一个流匹配模型，用于恢复细粒度的声学细节。其关键创新在于端到端联合训练策略：将 AM 的隐藏状态直接作为条件信号输入 FM，从而实现双向梯度传播。
联合训练框架具有几个关键优势：

• **指令感知声学细化用于流匹配**：不同于以往的级联系统仅由自回归声学模型通过语音词元形式的语音内容来为流匹配模型提供条件，我们的架构能够将完整的指令信息——包括风格指令、说话人特征和韵律控制——直接传递给流匹配模型。这使流

匹配模型能够利用高层语义指导来塑造超越单纯发音的声学细节，从而实现更具可控性且更能感知上下文的合成。

• **多说话人长篇音频建模能力**：这种端到端优化机制对于 JoyVoice 实现多说话人长音频生成至关重要。由于离散语义令牌包含的说话人相关信息有限，传统两阶段模型只能逐轮预测语义令牌，同时需要预先了解语义令牌中的说话人边界。相比之下，JoyVoice 直接将信息丰富的隐藏连续表示输入基础模型，并以端到端方式优化整个系统，使模型能够自动学习区分说话人特征和语义内容的信息。因此，无需明确了解说话人声学边界，JoyVoice 也能逼真地生成包含 1 至 8 位说话人的长篇音频。

• **通过端到端优化提升可懂度**：实验结果表明，与级联模型相比，我们联合训练的模型词错误率（WER）显著更低，这说明双向梯度流使自回归声学模型能够生成更有利于准确声学重建的表示。

• **对分词器压缩的鲁棒性**：当将分词器的帧率从 25 赫兹降至 12.5 赫兹时——即压缩 50%，并显著加剧信息瓶颈——级联基线的性能明显下降。相比之下，我们的端到端框架仍能保持与 25 赫兹设置相当的性能，这表明联合优化能够有效补偿离散令牌中的信息损失，使 AM 在其隐藏表示中编码互补的连续信息，供 FM 用于重建。

总体训练损失定义为：

$$\mathcal{L} = \mathcal{L}_{\text{AM}} + \lambda \cdot \mathcal{L}_{\text{FM}} \tag{1}$$

其中，λ 是一个平衡超参数，用于控制各组成部分的相对贡献。

自回归声学模型损失：AM 使用交叉熵损失，以标准的下一令牌预测目标进行训练：

$$\mathcal{L}_{\text{AM}} = -\sum_{t=1}^{T} \log p(s_t \mid \mathcal{I}_{<t}; \theta_{\text{AM}}) \tag{2}$$

其中，s_t 是位置 t处的真实离散语音令牌，$\mathcal{I}_{<t}$ 表示上下文，包括所有说话人信息、所有文本令牌以及此前生成的语音令牌 $\{s_i \mid i < t\}$。$\mathcal{I}$ 的详细公式将在下一节中介绍。

流匹配模型损失：FM 经过训练，可根据 AM 的隐藏表示恢复连续声学特征。按照标准的流匹配公式，我们优化：

$$\mathcal{L}_{\text{FM}} = \mathbb{E}_{t\sim\mathcal{U}(0,1),x_0,x_1}\left[|v_\theta(x_t, t, \mathbf{h}_{\text{AM}}) - (x_1 - x_0)|^2\right] \tag{3}$$

其中 $x_0 \sim \mathcal{N}(0, I)$ 从标准高斯先验中采样，x_1 表示真实声学特征（例如梅尔频谱图），$x_t = (1-t)x_0 + tx_1$ 是线性插值，v_θ是 FM 预测的速度场，而 $\mathbf{h}_{\text{AM}}$ 表示从 AM 中提取的隐藏状态，用作条件信号。对于梯度流机制，其关键创新在于以 $\mathbf{h}_{\text{AM}}$ 而非仅以离散令牌作为 FM 的条件。在反向传播过程中，来自 $\mathcal{L}_{\text{FM}}$ 的梯度会经由 $\mathbf{h}_{\text{AM}}$ 回传至 AM 参数 θ_{AM}：

$$\frac{\partial \mathcal{L}}{\partial \theta_{\text{AM}}} = \frac{\partial \mathcal{L}_{\text{AM}}}{\partial \theta_{\text{AM}}} + \lambda \cdot \frac{\partial \mathcal{L}_{\text{FM}}}{\partial \mathbf{h}_{\text{AM}}} \cdot \frac{\partial \mathbf{h}_{\text{AM}}}{\partial \theta_{\text{AM}}} \tag{4}$$

这种双向优化使自回归声学模型能够学习兼顾离散令牌预测准确率与连续声学重建质量的表征，有效解决级联架构固有的信息瓶颈。

2.2 基于长连续语音上下文的多说话人多轮建模

以往的多说话人、多轮对话合成流程采用模块化的话语级策略 [15, 18]。在这些框架中，对话被处理为一系列彼此独立的文本—语音对，每轮的文本输入 i都会单独转换为相应的语音输出。尽管这种处理方式明确标示了话语边界，却引入了人为分段，可能打乱自然的对话动态，并且需要在训练过程中进行精确的强制对齐。

相比之下，我们的方法引入了一种统一序列形式，在不依赖话语级分段的情况下，保持多轮对话的自然连续性。形式化地，对于涉及 N 位说话人和 M 轮对话的会话，我们构造输入序列 $\mathcal{I}$ 如下：

$$\mathcal{I} = \big[P;\ T;\ S\big] \tag{5}$$

其中

$$\begin{aligned} P &= \big[\text{spk}_0, \mathbf{e}_0, \text{spk}_1, \mathbf{e}_1, \ldots, \text{spk}_{N-1}, \mathbf{e}_{N-1}\big], \\ T &= \big[\text{spk}_{i_0}, \mathbf{t}_0, \text{spk}_{i_1}, \mathbf{t}_1, \ldots, \text{spk}_{i_{M-1}}, \mathbf{t}_{M-1}\big], \\ S &= \big[s_1, s_2, \ldots, s_T\big]. \end{aligned} \tag{6}$$

spk_k 表示第 k 位说话人的说话人标签，$\mathbf{e}_k$ 是相应的说话人嵌入，$\mathbf{t}^j$ 表示第 j 轮的文本词元序列（包含说话人索引 $i_j \in \{0, 1, \ldots, N-1\}$ 和可变序列长度），而 s_i 表示单个语音词元。需要注意的是，S 表示跨越整个多轮对话且未经分段的语音词元序列。

通过采用无需显式语音分段的隐式轮级对应关系学习，模型借助注意力机制，将说话人标签与其对应的嵌入、文本内容和语音词元跨度关联起来，从而实现灵活的说话人与内容绑定。将整个对话作为连续序列进行保留，既维持了对话上下文，也使模型能够捕捉跨轮依赖关系、轮流发言动态以及在分段方法中会被割裂的长程韵律模式。实验验证表明，现代 Transformer 注意力机制仅凭说话人标签，便能够准确建立未经分段的语音词元与其对应文本和说话人身份之间的对应关系。

2.3 MM-Tokenizer

如图3所示，JoyVoice 采用基于监督式多任务训练的语义词元方案的 MM-Tokenizer，与 CosyVoice 3 [8] 类似。我们采用 Whisper-large-v3 [22] ASR 模型作为预训练模型，并将有限标量量化（FSQ） [23] 模块集成到编码器第 12 个 Transformer 块的线性层中。该模型使用监督式多任务学习方法进行训练，涵盖多种音频理解任务，包括自动语音识别（ASR）、语音情感识别（SER）、音频事件检测（AED）、音频事件描述（AEC）、说话人验证（SV）、年龄检测（AD）和性别分类（GC）。我们在标准 Whisper-large-v3 词汇表中加入了专用离散令牌，以表示各种音频理解任务，例如 <|AED|> 表示音频事件检测（AED），等等。这使解码器能够同时针对多项任务执行序列到序列生成。我们还设计了一个外部音频解码器模块，用于根据编码器输出重建原始梅尔频谱图，从而确保学到的离散表示具有更好的泛化能力和鲁棒性，并能同时捕获语义信息和声学信息。
MM-Tokenizer 训练的最终目标函数是两个主要损失分量的加权组合：

$$\min_{\Theta} \mathcal{L}_{\text{total}} = \mathcal{L}_{\text{semantic}} + \beta \mathcal{L}_{\text{recon}}$$

，其中 β 是一个标量超参数，用于控制语义性能与声学重建质量之间的权衡。

值得注意的是，在实验中，我们探索了两种不同词元速率下的离散表示：25 赫兹和 12.5 赫兹。为实现这两种速率，我们分别通过添加额外的卷积神经网络层，对原始 Whisper-large-v3 编码器进行了修改，以进行 4× 和 8× 下采样。后续实验将比较这两种词元速率下的性能。一个值得注意的发现是，我们提出的

联合建模方法降低了对词元速率的依赖：在主观和客观指标上，12.5 赫兹的表示性能与 25 赫兹的表示相当。

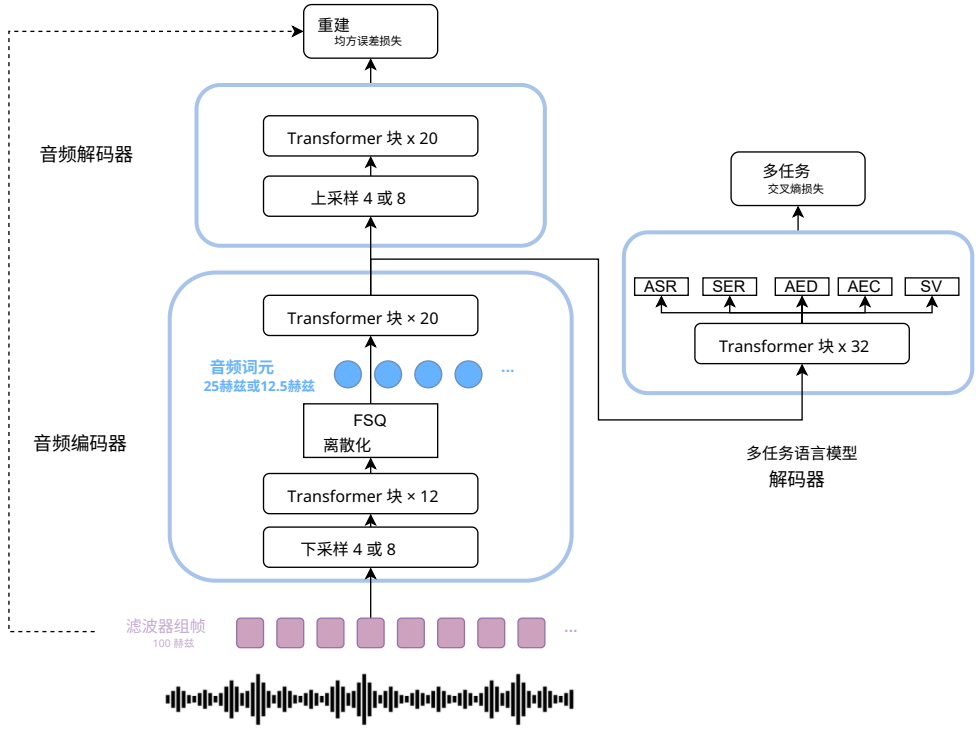


图3：JoyVoice MM-Tokenizer 框架示意图。

2.4 动态分块流匹配

为实现流式合成，我们采用因果分块流匹配模型，将大语言模型隐藏输出划分为较小的分块。受其他流式合成模型启发的分块因果掩码 [7, 18] 被应用于每个 Transformer 层中的自注意力模块，从而确保当前分块只能访问先前分块的历史上下文。为便于推理，我们在每个训练步骤中随机设定分块大小。由于训练阶段不再采用固定的分块大小，因此推理时可以灵活使用任意分块大小，从而支持广泛的应用场景和不同的延迟约束。

2.5 无 TTS 前端优化

传统文本转语音（TTS）系统通常包含一个前端文本预处理模块，负责规范化输入文本，处理标点符号和文本规范化/逆文本规范化（TN/ITN）等内容。这对于中文文本转语音系统尤为重要，因为它们还必须处理生僻字和多音词。在 JoyVoice 中，我们旨在模拟基于文本的基础模型的鲁棒性，使系统无需依赖专门的 TTS 前端即可处理多样化的输入文本。因此，本文所展示的所有实验和演示样例均未使用任何预处理模块。为实现这一目标，除了使用 Qwen 文本基础模型初始化模型外，我们还针对文本规范化/逆文本规范化（TN/ITN）、多音字和生僻字进行了有针对性的数据增强。

2.6 课程学习

我们采用两阶段课程学习策略，提升 JoyVoice 对多说话人长篇音频建模时的训练稳定性。第一阶段使用大规模单说话人短音频片段对模型进行预训练，每个片段的时长最多为一分钟。该预训练阶段将产出一个高质量基础模型，能够有效合成单说话人语音。第二阶段，我们将完整数据集（包括长篇单说话人音频、多说话人长篇音频和原始的单说话人短音频数据）结合起来，继续对模型进行微调。预训练后的 JoyVoice 模型作为这一阶段的初始化检查点。微调过程中，音频时长延长至最多五分钟，说话人数增加至最多八人。通过这种两阶段课程学习方法，JoyVoice 不仅在零样本单说话人语音合成方面表现出色，也能高质量地零样本生成多说话人长篇音频。

2.7 强化学习

2.7.1 偏好对齐

语言模型与人类偏好的对齐通常被表述为一个强化学习（RL）问题。给定输入提示 x 和模型的回答 y，目标是学习一个策略 π_θ，在保持与参考策略 π_{ref} 接近的同时，使预期奖励最大化。其形式化定义如下。

$$\max_{\pi_\theta} \mathbb{E}_{y\sim\pi_\theta(y|x)}[r(x,y)] - \beta D_{\text{KL}}(\pi_\theta(y|x)\|\pi_{\text{ref}}(y|x)) \tag{7}$$

其中，$r(x,\ y)$ 表示奖励函数，而 KL散度项由超参数 β 控制，用于约束策略偏离程度。由于真实奖励函数通常未知，因此需要从人类偏好数据中进行估计。人类偏好数据由若干元组 $(x,\ y_w,\ y_l)$ 组成，其中 y_w 表示更受偏好的回答，而 y_l 表示不受偏好的回答。奖励函数可以通过最大似然估计近似得到：

$$\hat{r} \in \arg\min_r \mathbb{E}_{(x,y_w,y_l)}[-\log\sigma(r(x,y_w) - r(x,y_l))] \tag{8}$$

其中 σ 是 Sigmoid函数。

2.7.2 直接偏好优化

直接偏好优化（DPO） [24]无需显式奖励建模，即可通过闭式形式求解这一优化问题。该方法利用KL约束目标的最优解对奖励函数进行重新参数化：

$$r(x,y) = \beta\log\left(\frac{\pi_\theta(y|x)}{\pi_{\text{ref}}(y|x)}\right) + \beta\log Z(x) \tag{9}$$

在Bradley-Terry模型下，偏好概率表示为：

$$P(y_w \succ y_l|x) = \sigma\left(\beta\log\frac{\pi_\theta(y_w|x)\pi_{\text{ref}}(y_l|x)}{\pi_\theta(y_l|x)\pi_{\text{ref}}(y_w|x)}\right) \tag{10}$$

DPO目标函数直接最大化这一似然：

$$\mathcal{L}_{\text{DPO}} = \mathbb{E}_{(y_w,y_l,x)}\left[-\log\sigma\left(\beta\log\frac{\pi_\theta(y_w|x)\pi_{\text{ref}}(y_l|x)}{\pi_\theta(y_l|x)\pi_{\text{ref}}(y_w|x)}\right)\right] \tag{11}$$

2.7.3 APO：面向文本转语音的声学偏好优化

我们提出声学偏好优化（APO），这是一种专为文本转语音（TTS）系统设计的新方法，可将偏好优化扩展到声学令牌级别。在现代基于神经编解码器的文本转语音系统中，神经音频编解码器会将语音波形转换为离散声学令牌序列。尽管基于 DPO 的方法无法直接用于优化最终波形输出，APO 却能对生成的语音序列中的单个声学令牌预测进行细粒度优化。对于文本转语音而言，这种方法尤其有价值，因为令牌级别的质量差异会显著影响整体语音自然度和可懂度。

声学令牌级偏好构建 对于给定的文本输入 x，我们生成 N 候选声学令牌序列 $\{s_1, s_2, \ldots, s_N\}$，其中每个序列 $s_i = \{t_{i,1}, t_{i,2}, \ldots, t_{i,n}\}$ 对应一种可能的语音输出。我们使用字符错误率（CER）指标评估每个序列。偏好集基于这些评估结果构建：

$$\mathcal{S}_{\text{chosen}} = \{s_i \in \{s_1, \ldots, s_N\} | \text{CER}(s_i) = 0\} \tag{12}$$

$$\mathcal{S}_{\text{rejected}} = \{s_j \in \{s_1, \ldots, s_N\} | \text{CER}(s_j) > 0\} \tag{13}$$

其中，$\mathcal{S}_{\text{chosen}}$ 包含所有字符准确率完美（CER = 0）的序列，而 $\mathcal{S}_{\text{rejected}}$ 包含存在转录错误的序列。在训练过程中，我们通过对来自 $\mathcal{S}_{\text{chosen}}$ 和 $\mathcal{S}_{\text{rejected}}$的序列进行所有可能的组合，构建偏好对。形式上，所有偏好对的集合记为 $\mathcal{P}$，其构造如下：

$$\mathcal{P} = \mathcal{S}_{\text{chosen}} \times \mathcal{S}_{\text{rejected}} = \{(s_{\text{chosen}}, s_{\text{rejected}}) \mid s_{\text{chosen}} \in \mathcal{S}_{\text{chosen}}, s_{\text{rejected}} \in \mathcal{S}_{\text{rejected}}\} \tag{14}$$

这种方法确保模型学会区分完全准确的令牌序列与存在错误的序列，从而直接优化生成语音的令牌级正确性。

3 多说话人语音克隆评测基准

随着语音生成模型的快速发展，现有评测基准已无法满足模型评估需求，尤其是在多说话人长篇音频建模方面。首先，当前的多说话人对话评测基准（如 zipvoice-dialog [20]）所涉及的说话人数有限——通常只有两位参与者，限制了多样性。在有声书、播客和直播等实际应用领域，涉及两到四位甚至更多参与者的多说话人对话十分常见。因此，现有基准主要针对双说话人场景设计，无法充分评估模型在这些多样且复杂场景中的表现。其次，现有基准中的对话脚本大多由时长不足一分钟的短片段构成，因此缺乏评估持续对话互动所需的上下文连贯性。最后，传统基准侧重于字符准确率和单句说话人相似度，不足以评估多说话人长对话。它未能考虑到长篇音频评估中的关键需求：综合评估轮次间的音色稳定性、字符准确率和整体连贯性，而这些问题恰恰会因频繁的说话人切换而加剧。

为更好地评估 JoyVoice，我们建立了一个多说话人、多轮次长篇中文和英语音频基准，名为 JoyVoice-MSMT-eval。具体如下：

• **多说话人：**：该基准涵盖由 2 至 4 位说话人参与的对话。中文和英语子集均包含 50 段双人对话、50 段三人对话和 25 段四人对话。源音频取样自 SEED-TTS-eval。为确保长对话中的所有源音频（提示音频）片段均来自不同说话人，我们进行了说话人相似度验证和人工验证。

• **长篇音频**：该基准包含与时长为 1 至 5 分钟的音频片段相对应的脚本，旨在评估多说话人对话语音生成模型的长篇合成能力。所有长篇脚本均由大语言模型生成，并由人工标注员仔细审核和批准，以确保内容质量高且连贯。

• **评估指标**：由于仅凭字符错误率（CER）不足以捕捉说话人频繁轮替时的音色变化，我们引入拼接最小排列词错误率（cpWER）[1] [25, 26] 作为客观评估指标，该指标同时受语音识别和说话人分离结果的影响。

[1]cpWER 的计算方法如下：(i) 按时间顺序将每位说话人的所有参考转写拼接起来。(ii) 按时间顺序将每位检测到的说话人的所有假设转写拼接起来。(iii) 针对所有可能的说话人排列，计算拼接后的参考转写与假设转写之间的 WER，并选取其中最低的 WER。

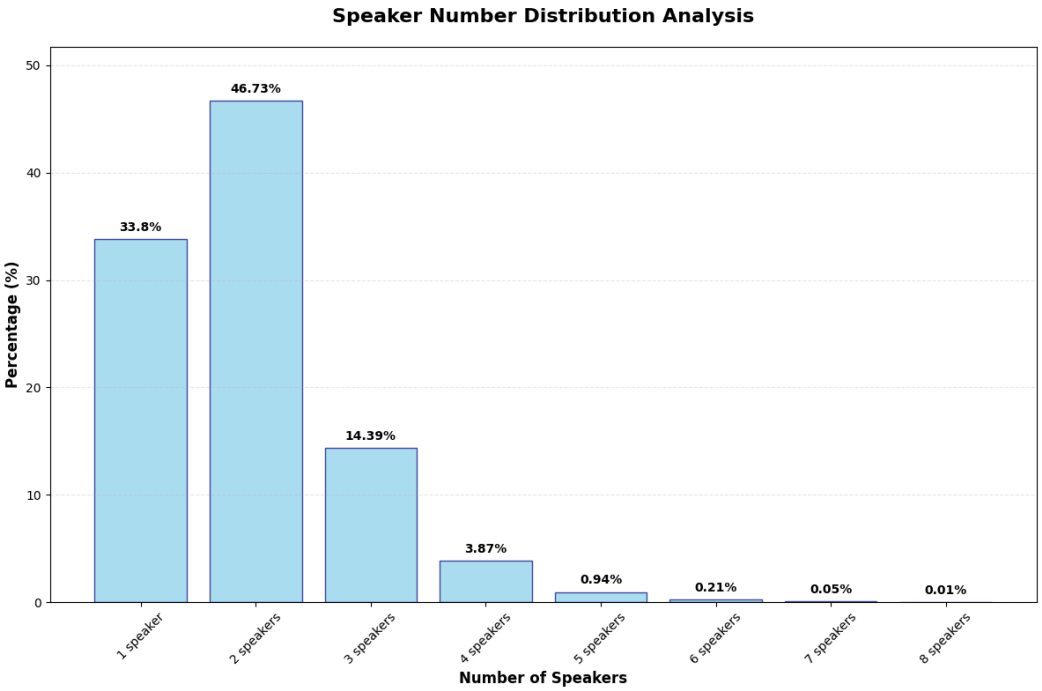


图4：JoyVoice 长篇音频训练数据中的说话人分布。

4 实验结果

4.1 训练数据

我们开发了一套改进的多语言数据生产流程，用于构建 JoyVoice训练数据集。语言覆盖范围包括中文（其中少量数据为粤语和其他汉语方言）、英语、日语和韩语，中文和英语合计占整个数据集的 90% 以上。我们的自动化流程依次执行以下处理阶段：音频归一化、质量评估、增强与分离、语音活动检测（VAD）时间戳标注、说话人分离、使用多个语音识别系统进行文本转录，以及标点恢复。我们综合利用前序阶段生成的元数据，筛选并挑选高质量样本。

我们构建了两种不同的训练数据格式。**短语句数据（BaseData）**：时长通常不足 30 秒的音频片段，严格来说仅包含单说话人语句。**长篇音频数据（LongData）**：时长为 30 秒至 5 分钟的音频片段，包含 1 至 8 位说话人。LongData 子集中说话人的分布如图 4 所示。

4.2 模型配置

除了基于 E2E-Transformer-DiT架构的 JoyVoice 外，我们还开发了一个两阶段级联系统，作为对比基线。为便于在实验部分中表述，我们将端到端系统称为*JoyVoice*，将两阶段变体称为*JoyVoice-Cascade*。

两种架构均使用 Qwen2.5-0.5B大语言模型 [27]初始化自回归（AR）Transformer。对于流匹配组件，我们采用 F5-TTS [3], 中的扩散Transformer（DiT）架构，该架构约包含 3 亿个参数。值得注意的是，JoyVoice-Cascade 仅使用单说话人数据进行训练，因此不支持多说话人语音合成。

模型采用 AdamW优化器进行训练，并使用线性预热和余弦退火调度。具体而言，预热步数设为 10,000，峰值学习率为 1×10^{-4}。

4.3 在 MM-Tokenizer 上的评估

基准测试	Whisper-large-v3	多模态分词器（25赫兹）	多模态分词器（12.5赫兹）
自动语音识别任务（词错误率/字符错误率）↓			
AISHELL-1	5.14	**2.53**	2.60
AISHELL-2	4.96	**3.56**	3.84
LibriSpeech-clean	**1.80**	2.35	2.67
LibriSpeech-other	**3.60**	5.24	6.62
Fleurs-zh	7.70	**4.25**	4.72
Fleurs-en	**4.25**	8.00	9.52
音频理解*(准确率)* ↑			
ESD（语音情感识别）	-	**0.886**	0.850
FSDKaggle（音频事件检测）	-	**0.548**	0.530

表 1：**音频理解任务对比**。 我们将提出的多模态分词器（25 赫兹和 12.5 赫兹）与 Whisper-large-v3 基线进行比较。 ↓表示越低越好（词错误率/字符错误率），↑表示越高越好（准确率）。最佳结果以**粗体**标出。

模型	架构。	中文测试集		英文测试集		困难测试集	
		字符错误率↓	说话人相似度↑	词错误率↓	说话人相似度↑	字符错误率↓	说话人相似度↑
帧率：*25*赫兹							
S^3-分词器 [6]	级联	1.28	0.821	2.33	0.768	6.10	0.802
MM-Tokenizer	级联	1.16	0.832	1.74	0.782	6.14	**0.817**
MM-Tokenizer	E2E	**0.97**	**0.836**	1.70	**0.794**	6.06	0.799
帧率：*12.5*赫兹							
MM-Tokenizer	级联	1.57	0.794	2.19	0.718	6.98	0.778
MM-Tokenizer	E2E	1.01	0.828	**1.63**	0.785	**6.07**	0.798

表2：**SEED-TTS-Eval 基准上的零样本文本转语音性能**。我们比较了所提出的 MM-Tokenizer 与 S^3-Tokenizer 基线 [6] 在不同帧率下的表现。**粗体** 表示最佳结果。“架构”指采用 JoyVoice 主干网络的生成架构（级联与端到端）。大语言模型使用**说话人嵌入**进行解码。

如第 2.3 节所述，MM-Tokenizer 通过在 Whisper-large-v3 主干网络中集成有限标量量化（FSQ）层和基于卷积神经网络的下采样层构建而成。模型使用结合有监督判别任务损失与无监督重建损失的混合目标进行优化。表 1 全面评估了 MM-Tokenizer 在各类音频理解基准测试中的表现，包括自动语音识别（ASR）、语音情感识别（SER）和音频事件检测（AED）。结果表明，与 Whisper-large-v3 基线相比，MM-Tokenizer 在 25 赫兹和 12.5 赫兹帧率下均保持了有竞争力的语义保留能力。

表 2 展示了采用 JoyVoice 主干网络在 SEED-TTS-Eval 基准上的零样本文本转语音性能结果。在标准的 25 赫兹帧率下，所提出的多模态分词器 在大多数指标上始终优于基线 S^3-Tokenizer。具体而言，多模态分词器的端到端（E2E）变体在中文数据集（中文测试集）上实现了更高的可懂度和相似度，字符错误率（CER）最低，为 0.97；说话人相似度（SS）最高，为 0.836。与此同时，25 赫兹下的级联架构表现也颇具竞争力，在 困难测试集 子集上取得了最高相似度分数（0.817）。当帧率降至 12.5 赫兹时，E2E 模型展现出卓越的鲁棒性。尽管时间分辨率较低，12.5 赫兹的 E2E 模型仍在英文测试集上取得了最先进的词错误率（WER）1.63

（英文测试集）上取得了最低的 CER（6.07），并在具有挑战性的 困难测试集 上验证了所提出分词器设计的效率。

4.4 单说话人零样本语音克隆

模型	中文测试集		英文测试集		困难测试集	
	字符错误率（%）↓	SS ↑	词错误率（%）↓	SS ↑	字符错误率（%）↓	SS ↑
人类	1.26	0.755 (0.775)	2.14	0.734 (0.742)	-	-
声码器重合成	1.27	0.720	2.17	0.700	-	-
单说话人模型						
Seed-TTS [5]	1.12	0.796	2.25	**0.762**	7.59	0.776
MiniMax-Speech [12]	0.99	**0.799**	1.90	0.738	-	-
MaskGCT [28]	2.27	0.774 (0.752)	2.62	0.714 (0.730)	10.27	0.748 (0.720)
F5-TTS（32 次神经函数求值） [3]	1.56	0.741 (0.794)	1.83	0.647 (0.742)	8.67	0.713 (0.762)
Spark TTS [9]	1.20	0.672	1.98	0.584	-	-
Qwen2.5-Omni-7B$_{RL}$ [29]	1.42	0.754	2.33	0.641	6.54	0.752
CosyVoice 2 [7]	1.45	0.748 (0.806)	2.57	0.652 (0.736)	6.83	0.724 (0.776)
CosyVoice 3-0.5B [8]	1.16	0.780 (0.825)	2.02	0.718 (0.789)	6.08	0.758 (0.815)
CosyVoice 3-0.5B$_{RL}$ [8]	0.75	0.774 (0.836)	1.76	0.695 (0.783)	**5.09**	0.750 (0.809)
多说话人模型						
Mooncast* [18]	6.72	0.550 (0.559)	9.77	0.437 (0.463)	-	-
MOSS-TTSD* [19]	3.05	0.678 (0.669)	17.64	0.594 (0.628)	23.21	0.644 (0.633)
VibeVoice-1.5B* [15]	2.86	0.689 (0.656)	6.61	0.587 (0.595)	25.59	0.612 (0.587)
VibeVoice-7B* [15]	1.60	0.707 (0.676)	4.42	0.604 (0.594)	21.65	0.640 (0.618)
JoyVoice-Cascade	1.13	0.780 (0.825)	1.75	0.710 (0.778)	6.07	0.764 (0.805)
JoyVoice-Cascade$_{RL}$	**0.73**	0.786 (0.831)	**1.45**	0.718 (0.786)	5.56	0.774 (0.813)
JoyVoice	0.97	0.786 (0.827)	1.69	0.736 (0.790)	5.55	0.746 (0.773)
JoyVoice-Streaming	0.97	0.790 (0.831)	1.70	0.738 (0.793)	5.62	0.777 (0.809)
+ 大语言模型 说话人嵌入。	0.97	0.797 (**0.838**)	1.69	0.737（**0.795**）	5.97	**0.784** (**0.820**)
JoyVoice-12.5Hz	0.95	0.778 (0.824)	1.64	0.717 (0.783)	5.90	0.745 (0.780)

表3：**在 SEED-TTS-Eval 基准测试上的零样本文本转语音性能评估**。我们将 JoyVoice 与最先进的基线模型进行比较，评估内容一致性（字符错误率/词错误率 ↓）和说话人相似度（SS ↑）。SS 分数按以下格式报告：*WavLM* (*ERes2Net*)。**粗体** 和下划线分别表示最佳和次佳结果。**"-"**表示数据不可用。**"*"**表示通过我们自行使用的开源模型解码得到的结果。

表3展示了 JoyVoice 与近期多个模型在 SEED 测试集上的文本转语音性能。评估重点关注内容一致性（词错误率/字符错误率）和说话人相似度（SS）。单说话人模型仅能生成单说话人音频，而多说话人模型则具备生成涉及多个说话人的长篇交互式音频的能力。

尽管 JoyVoice 的设计目标是成为能够进行多说话人长篇音频合成的通用模型，但它在单说话人语音克隆基准上也达到了最先进的性能。与 JoyVoice-Cascade 相比，JoyVoice 实现了显著提升，相对增幅在中文测试集 上为 14.2%（1.13% → 0.97%），在英文测试集上为 3.4%（1.75% → 1.69%）。在困难测试集 上，JoyVoice 将字符错误率从 6.07% 降至 5.55%（相对提升 8.6%）。与其他无强化学习基线模型相比，JoyVoice 在所有指标上均持续表现出色。值得注意的是，将我们的**强化学习策略** 应用于 JoyVoice-Cascade 模型后，在中文测试集、英文测试集和困难测试集上的字符错误率分别相对降低了 48.4%、17.2% 和 8.4%。

关于**流式处理**，我们以 JoyVoice 为初始化模型，冻结大语言模型组件，并仅对动态分块流匹配模块进行训练。如表3所示，JoyVoice-Streaming 不仅在内容一致性方面保持了与 JoyVoice 相当的性能，还在具有挑战性的困难测试集上显著提升了说话人相似度。通过采用动态分块训练策略，JoyVoice-Streaming 支持在推理过程中使用可变的分块大小，从而能够精确控制计算延迟，以满足不同应用的需求。相关性能指标汇总于表4。

此外，表 4 中的消融研究比较了大语言模型组件在使用和不使用说话人嵌入时的性能。结果一致表明，加入说话人嵌

入会降低字符错误率/词错误率的表现，但仍能显著提升说话人相似度。此外，在块大小为 24/48 时，采用说话人嵌入的 JoyVoice-Streaming 在内容一致性和说话人相似度两方面均优于 CosyVoice 3-0.5B，后者是当前表现最佳的流式模型之一。

模型	块	中文测试集		英文测试集		困难测试集	
		字符错误率↓	说话人相似度↑	词错误率↓	说话人相似度↑	字符错误率↓	说话人相似度↑
CosyVoice 3-0.5B	full	1.16	0.825	2.02	0.789	6.08	0.815
JoyVoice	full	0.97	0.827	**1.69**	0.790	**5.55**	0.773
+ 大语言模型 说话人嵌入。	full	0.97	0.836	1.70	0.794	6.06	0.799
JoyVoice-流式处理	48	0.97	0.831	1.70	0.793	5.62	0.809
+ 大语言模型 说话人嵌入。	48	0.97	**0.838**	**1.69**	**0.795**	5.97	**0.820**
JoyVoice-流式处理	24	0.98	0.831	1.70	0.793	5.69	0.809
+ 大语言模型 说话人嵌入。	24	**0.96**	**0.838**	1.73	0.794	5.97	**0.820**

表4：在 SEED-TTS-Eval 数据集上对零样本文本转语音性能进行评估，比较采用不同分块大小的大语言模型组件的表现。粗体数字表示各列的最佳结果。

4.5 多说话人零样本语音克隆

表5展示了 JoyVoice 与近期若干模型在我们的 JoyVoice-MSMT-eval 测试集上的多说话人长篇音频合成表现，这些测试集包括中文*test-zh-msmt*和英语*test-en-msmt*。评估不仅关注 CER/WER，也关注 cpCER/cpWER。在 2 至 4 位说话人的 MSMT 测试集上取得最低 WER 和 cpWER 的结果，凸显了 JoyVoice 忠实保留内容和说话人音色的卓越能力。观察发现，与 2 位和 3 位说话人的结果相比，4 位说话人结果的 cpWER 明显下降。这主要是因为我们的训练数据分布中，说话人数超过四人的对话占比不足，详情见图4。为解决这一局限，未来工作将优先通过增加多说话人样本在数据集中的占比，提升模型在此类场景下的表现。

对于说话人识别，我们采用开源模型 pyannote [30], , 解读相应指标时应将此上下文考虑在内，因为这些指标反映了该外部系统的能力与局限性。我们的主观听测网站提供了互补视角，有助于进行全面评估。

模型	*test-zh-msmt*（字符错误率 / cpCER）↓			*test-en-msmt*（词错误率 / cpWER）↓		
	2spk	3spk	4spk	2spk	3spk	4spk
Mooncast*	7.57 / 12.62	-	-	5.32 / 17.38	-	-
MOSS-TTSD*	3.00 / 6.02	-	-	5.56 / 10.04	-	-
VibeVoice-1.5B*	5.54 / 14.88	6.30 / 16.77	5.29 / 17.20	4.26 / 7.73	8.56 / 12.22	8.89 / 14.11
VibeVoice-7B*	1.80 / 7.57	3.15 / 9.37	3.66 / 16.60	4.19 / 6.58	6.66 / 10.60	6.42 / 13.81
JoyVoice	**1.44 / 1.88**	**1.83 / 4.94**	**2.08 /13.34**	**3.36 /3.61**	**6.01/ 9.57**	**3.93 / 12.18**

表5：**在 JoyVoice-MSMT-eval 上评估的多说话人文本转语音模型性能基准**。 我们将 JoyVoice 与最先进的基线模型进行比较。结果按以下格式报告：中文数据集为**字符错误率/cpCER** ，英语数据集为**词错误率/cpWER** （越低越好）。**粗体** 表示最佳结果。“*”表示使用我们自行采用的开源模型进行解码所得的结果。

4.6 说话人微调模型

为研究预训练基础模型对多个目标说话人的适配，我们采用了 CosyVoice2 提出的多说话人微调（mSFT）范式，旨在从单个模型中生成各具特色且高质量的语音特征。具体而言，本研究选取了五位说话人（分别标记为说话人 1 至说话人 5）。训练期间，我们通过在系统提示中加入说话人提示标签（例如“你是说话人 1”），明确指定每位说话人的身份。用于微调的语音数据总时长约为 12 小时，每位说话人贡献了约 2 至 3 小时的音频。

在多说话人微调过程中，我们始终采用 1×10^{-5} 的学习率。我们的研究重点之一是比较不同微调策略的效果与效率。为此，我们分别尝试了全参数监督微调和参数高效的低秩适配（低秩适配监督微调）方法。研究结果表明，策略的选择应取决于数据情况。根据我们的实证观察，当目标说话人的数据有限时（例如少于 1 小时），我们建议使用低秩适配监督微调，以有效降低过拟合风险。图5展示了这些方法的结果对比。

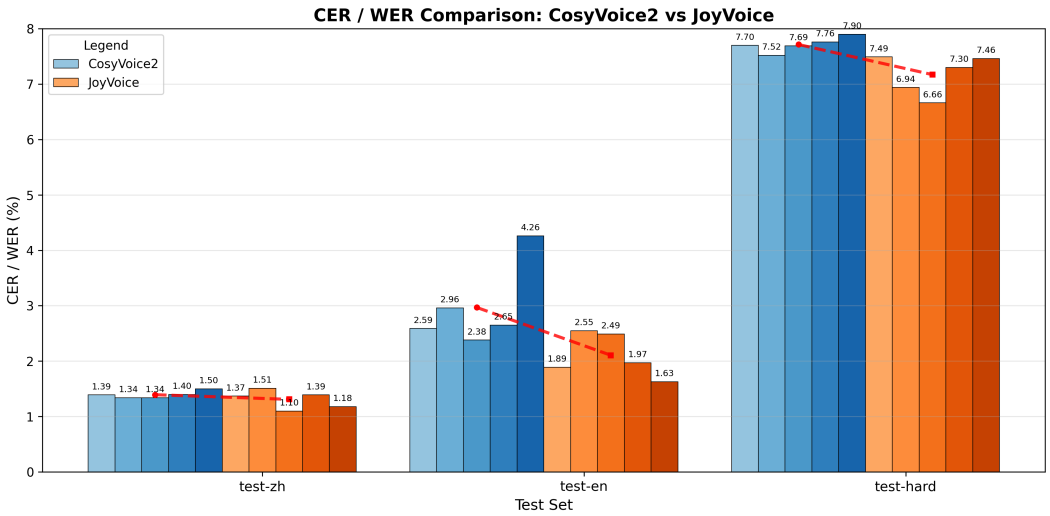


图5：五位说话人的全参数监督微调结果对比。

传统的文本转语音监督微调（SFT）需要使用单说话人数据训练各个模型。JoyVoice 基础模型支持多说话人生成，因此我们创新性地使用约一小时的双说话人对话播客数据对其进行微调。由此得到的双说话人文本转语音模型能够精准捕捉源语音的音色、说话风格和韵律自然度。主观听音测试结果请参阅演示页面。

5 结论

在本研究中，我们推出了 JoyVoice，这是一款新颖且强大的对话语音合成基础模型。JoyVoice 旨在克服以往模型的局限性，是一个高度拟人化、支持多说话人和长上下文的系统，能够生成逼真的对话。模型展现了先进的能力：在单次生成中，它可以合成最多包含八位不同说话人的自然对话，时长最长可达五分钟。JoyVoice 的卓越性能源于多项关键技术进展。我们提出了一种创新的端到端 Transformer-DiT 架构，专为处理复杂的对话动态和长期上下文而优化。作为对这一架构的补充，我们还推出了新颖的 MM-Tokenizer 音频离散化方案。此外，通过策略性地运用大规模数据并进行广泛的数据增强，我们成功开发出一种稳健的音频生成模型，大幅降低或消除了对传统文本预处理流程的依赖，而这类流程往往

脆弱不可靠。全面评估证实，JoyVoice 的表现显著优于同类语音基础模型。具体而言，JoyVoice 在 Seed-TTS-Eval 基准和多说话人长篇对话语音克隆任务上均取得了顶尖成绩，展现出卓越的音频质量和泛化能力。

6 局限性

尽管 JoyVoice 在高度拟人化的多说话人对话语音合成方面达到了最先进的性能，但仍存在一些固有局限，这也促使我们持续开展研究。

首先，在生成涉及四位以上说话人的对话时，合成音频的稳定性和质量目前会有所下降。我们认为，主要原因是当前训练语料库对高度动态、自然的多说话人互动数据覆盖不足。未来的工作将优先获取规模更大、内容更复杂的对话数据集，以确保模型在日益复杂的对话场景中具备稳健性和高保真度。
其次，JoyVoice 的大规模强化学习（RL）训练阶段仍在进行中。除了优化合成输出中的文本一致性和说话人音色相似度等传统目标外，我们还计划进一步探索强化学习的效用。具体而言，我们旨在研究如何有效利用强化学习，提升多说话人合成的长期稳定性，并显著增强模型在较长对话上下文中表达拟人化情感和细腻情感的能力。

最后，JoyVoice 目前仅限于纯语音生成，尚不支持合成通用音频事件或音乐。要实现统一的大规模音频生成模型，涵盖语音、音频和音乐，需要在两个主要技术领域开展大量深入探索：通用音频分词器的设计，以及能够对多种音频模态进行建模的高度泛化模型架构的开发。

按字母顺序排列的作者

核心贡献者： Fan Yu、Tao Wang、You Wu、Lin Zhu、Wei Deng、Weisheng Han、Wenchao Wang、Lin Hu、Xiangyu Liang、Xiaodong He、Yankun Huang、Yu Gu、Yuan Liu、Yuxuan Wang、ZhangyuXiao、Ziteng Wang
贡献者： Boya Dong、Feng Dang、Jinming Chen、Jingdong Li、Jun Wang、Yechen Jin、Yuan Zhang、ZhengyanSheng、Xin Wang

参考文献

[1] Sanyuan Chen、Chengyi Wang、Yu Wu、Ziqiang Zhang、Long Zhou、Shujie Liu、Zhuo Chen、Yanqing Liu、Huaming Wang、Jinyu Li等。神经编解码器语言模型是零样本文本转语音合成器。*IEEE 音频、语音与语言处理汇刊*，2025年。[2] Zhen Ye、Xinfa Zhu、Chi-Min Chan、Xinsheng Wang、Xu Tan、Jiahe Lei、Yi Peng、Haohe Liu、Yizhu Jin、Zheqi DAI等。Llasa：扩展基于 Llama 的语音合成的训练时和推理时计算量。*arXiv 电子预印本*，页码 arXiv–2502，2025年。[3] Yushen Chen、Zhikang Niu、Ziyang Ma、Keqi Deng、Chunhui Wang、JianZhao JianZhao、Kai Yu和Xie Chen。F5-TTS：一种借助流匹配生成流畅、忠实语音的神奇造梦者。载于第*63*届计算语言学协会年会论文集（第*1*卷：长篇论文），第6255–6271页，2025年。[4] Han Zhu、Wei Kang、Zengwei Yao、Liyong Guo、Fangjun Kuang、Zhaoqing Li、Weiji Zhuang、Long Lin和Daniel Povey。Zipvoice：借助流匹配实现快速、高质量的零样本文本转语音。*arXiv*预印本 *arXiv:2506.13053*，2025年。[5] Philip Anastassiou、Jiawei Chen、Jitong Chen、Yuanzhe Chen、Zhuo Chen、Ziyi Chen、Jian Cong、Lelai Deng、Chuang Ding、Lu Gao等。Seed-TTS：一系列高质量、多功能的语音生成模型。*arXiv*预印本 *arXiv:2406.02430*，2024年。

[6] Zhihao Du、Qian Chen、Shiliang Zhang、Kai Hu、Heng Lu、Yexin Yang、Hangrui Hu、Siqi Zheng、Yue Gu、Ziyang Ma等。Cosyvoice：一种基于监督语义令牌、可扩展的多语言零样本文本转语音合成器。*arXiv*预印本 *arXiv:2407.05407*，2024年。[7] Zhihao Du、Yuxuan Wang、Qian Chen、Xian Shi、Xiang Lv、Tianyu Zhao、Zhifu Gao、Yexin Yang、Changfeng Gao、Hui Wang等。CosyVoice 2：借助大语言模型实现可扩展的流式语音合成。*arXiv*预印本 *arXiv:2412.10117*，2024年。[8] Zhihao Du、Changfeng Gao、Yuxuan Wang、Fan Yu、Tianyu Zhao、Hao Wang、Xiang Lv、Hui Wang、Chongjia Ni、Xian Shi等。CosyVoice 3：通过扩展规模和后训练迈向真实场景语音生成。*arXiv*预印本 *arXiv:2505.17589*，2025年。[9] Xinsheng Wang、Mingqi Jiang、Ziyang Ma、Ziyu Zhang、Songxiang Liu、Linqin Li、Zheng Liang、Qixi Zheng、Rui Wang、Xiaoqin Feng等。Spark-tts：一种基于大语言模型、采用单流解耦语音令牌的高效文本转语音模型。*arXiv*预印本 *arXiv:2503.01710*，2025年。[10] Hao-Han Guo、Kun Liu、Fei-Yu Shen、Yi-Chen Wu、Feng-Long Xie、Kun Xie和Kai-Tuo Xu。Fireredtts：面向工业级生成式语音应用的基础文本转语音框架。*arXiv* 电子预印本，页码 arXiv–2409，2024年。[11] Siyi Zhou、Yiquan Zhou、Yi He、Xun Zhou、Jinchao Wang、Wei Deng和Jingchen Shu。Indextts2：在富有情感表现力和时长控制的自回归零样本文本转语音方面实现突破。*arXiv*预印本 *arXiv:2506.21619*，2025年。[12] Bowen Zhang、Congchao Guo、Geng Yang、Hang Yu、Haozhe Zhang、Heidi Lei、Jialong Mai、Junjie Yan、Kaiyue Yang、Mingqi Yang等。Minimax-speech：采用可学习说话人编码器实现内在零样本文本转语音。*arXiv*预印本 *arXiv:2505.07916*，2025年。[13] Yaron Lipman、Ricky TQ Chen、Heli Ben-Hamu、Maximilian Nickel和Matt Le。生成建模的流匹配。*arXiv*预印本 *arXiv:2210.02747*，2022年。[14] William Peebles和Saining Xie。基于Transformer的可扩展扩散模型。载于*IEEE/CVF*国际计算机视觉会议论文集，第4195–4205页，2023年。[15] Zhiliang Peng、Jianwei Yu、Wenhui Wang、Yaoyao Chang、Yutao Sun、Li Dong、Yi Zhu、Weijiang Xu、Hangbo Bao、Zehua Wang等。Vibevoice技术报告。*arXiv*预印本 *arXiv:2508.19205*，2025年。[16] Zhijun Liu、Shuai Wang、Sho Inoue、Qibing Bai和Haizhou Li。用于文本转语音合成的自回归扩散Transformer。*arXiv*预印本 *arXiv:2406.05551*，2024年。[17] Dongya Jia、Zhuo Chen、Jiawei Chen、Chenpeng Du、Jian Wu、Jian Cong、Xiaobin Zhuang、Chumin Li、Zhen Wei、Yuping Wang等。Ditar：用于语音生成的扩散Transformer自回归建模。*arXiv*预印本 *arXiv:2502.03930*，2025年。[18] Zeqian Ju、Dongchao Yang、Jianwei Yu、Kai Shen、Yichong Leng、Zhengtao Wang、Xu Tan、Xinyu Zhou、Tao Qin和Xiangyang Li。MoonCast：高质量零样本播客生成。*arXiv*预印本 *arXiv:2503.14345*，2025年。[19] OpenMOSS。文本转口语对话生成。2025年。[20] Han Zhu、Wei Kang、Liyong Guo、Zengwei Yao、Fangjun Kuang、Weiji Zhuang、Zhaoqing Li、Zhifeng Han、Dong Zhang、Xin Zhang、Xingchen Song、Long Lin和Daniel Povey。Zipvoice-dialog：借助流匹配实现非自回归口语对话生成。*arXiv*预印本 *arXiv:2507.09318*，2025年。[21] Jinze Bai、Shuai Bai、Yunfei Chu、Zeyu Cui、Kai Dang、Xiaodong Deng、Yang Fan、Wenbin Ge、Yu Han、Fei Huang等。Qwen技术报告。*arXiv*预印本 *arXiv:2309.16609*，2023年。[22] Alec Radford、Jong Wook Kim、Tao Xu、Greg Brockman、Christine McLeavey和Ilya Sutskever。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。[23] Fabian Mentzer、David Minnen、Eirikur Agustsson和Michael Tschannen。有限标量量化：让VQ-VAE化繁为简。*arXiv*预印本 *arXiv:2309.15505*，2023年。[24] Rafael Rafailov、Archit Sharma、Eric Mitchell、Christopher D Manning、Stefano Ermon和Chelsea Finn。直接偏好优化：你的语言模型其实暗中就是一个奖励模型。神经信息处理系统进展，36：53728–53741，2023年。

[25] Fan Yu、Zhihao Du、ShiLiang Zhang、Yuxiao Lin 和 Lei Xie。《多方会议中说话人归属自动语音识别的比较研究》。载于 *INTERSPEECH* 会议论文集，第 560–564 页。ISCA，2022年。
[26] Naoyuki Kanda、Xiong Xiao、Jian Wu、Tianyan Zhou、Yashesh Gaur、Xiaofei Wang、Zhong Meng、Zhuo Chen 和 Takuya Yoshioka。《面向单声道长篇音频的说话人归属自动语音识别：模块化方法与联合方法的比较研究》。载于 *ASRU* 会议论文集，第 296–303 页。IEEE，2021年。
[27] An Yang、Baosong Yang、Beichen Zhang、Binyuan Hui、Bo Zheng、Bowen Yu、Chengyuan Li、Dayiheng Liu、Fei Huang、Haoran Wei、Huan Lin、Jian Yang、Jianhong Tu、Jianwei Zhang、Jianxin Yang、Jiaxi Yang、Jingren Zhou、Junyang Lin、Kai Dang、Keming Lu、Keqin Bao、Kexin Yang、Le Yu、Mei Li、Mingfeng Xue、Pei Zhang、Qin Zhu、Rui Men、Runji Lin、Tianhao Li、Tingyu Xia、Xingzhang Ren、Xuancheng Ren、Yang Fan、Yang Su、Yichang Zhang、Yu Wan、Yuqiong Liu、Zeyu Cui、Zhenru Zhang 和 Zihan Qiu。Qwen2.5 技术报告。*arXiv* 预印本 *arXiv:2412.15115*，2024年。[28] 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变换器的零样本文本转语音。*arXiv* 预印本 *arXiv:2409.00750*，2024年。[29] 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。[30] Herv´e Bredin。pyannote.audio 2.1 说话人分离流程：原理、基准测试与实践方案。载于 *Interspeech 2023*，第 1983–1987 页，2023年。