Tenlai-TTS 2.0：基于连续表征的大语言模型文本转语音系统与端到端训练

Changfeng Gao*、Yong Ren*、Jun Yuan、Zhengkun Tian、Ye Bai、Zhao You、Shidong Shan

腾讯元宝、腾讯天籁音频实验室

arXiv:2611.29226v1 [cs.SD] 22 Nov 2026

摘要

采用连续语音潜变量的自回归文本转语音（TTS）能够避免量化引起的信息损失，但通常依赖于单独学习、并在语言模型训练前冻结的潜空间。这样的潜空间必须满足两项相互竞争的要求：既要保留足够的声学细节以实现高保真重建，又要具备足够的结构性和正则化，以支持语言模型进行自回归预测。我们提出 Tenlai-TTS 2.0，通过端到端训练学习连续语音表示，在声学细节与自回归可预测性之间取得平衡。该模型由语音编码器（带有容量受控的高斯瓶颈）、带自回归扩散 Transformer（AR-DiT）头的大语言模型（LLM）、流匹配（FM）梅尔频谱图解码器，以及多任务识别模型（RM）组成。借助能量归一化和固定方差高斯噪声，该瓶颈的信息率具有闭式上界。通过联合优化重建、识别和下一块预测目标，我们学习适用于这三项任务的连续语音表示，并训练 LLM 和 AR-DiT 进行自回归生成。此外，我们还利用流匹配解码器和 RM 对预测潜变量进行可微重建，并通过识别反馈进一步优化 LLM 和 AR-DiT。在 Seed-TTS-eval 上，与仅进行联合训练相比，这种基于反馈的优化将 test-zh 上的字符错误率（CER）从 0.82% 降至 0.74%，将 test-hard-zh 上的字符错误率从 8.32% 降至 7.08%，并将 test-en 上的词错误率（WER）从 1.32% 降至 1.27%。后续的强化学习进一步将 test-hard-zh 上的字符错误率降至 6.29%，将 test-en 上的词错误率降至 1.08%；在 test-zh、test-en 和 test-hard-zh 上，说话人相似度得分分别为 0.811、0.766 和 0.784。Tenlai-TTS 2.0 的演示页面见https://tenlai-tts.github.io/Tenlai-TTS-2.0/。

关键词 文本转语音；连续表示；端到端训练

*同等贡献。

1 引言

如今，基于大语言模型的文本转语音系统通常通过自回归语言模型预测一系列语音表示，再将其合成为音频（Wang 等，2023a；Anastassiou 等，2024；Du 等人，2024b，2025）。这些系统大多使用离散语音标记。量化通过类别式的下一标记预测降低了大语言模型的任务难度，但离散标记的信息预算有限，制约了其所能承载的声学细节，从而限制了重建保真度和表现力。因此，越来越多的研究开始直接预测连续潜变量，并在大语言模型之上使用回归头或扩散头，以保留更丰富的声学信息（Meng 等，2024；Jia 等，2025年；Peng 等人，2025年；Zhou 等人，2025b）。

连续表示并未消除根本矛盾；更大的信息预算让语言结构与声学细节之间的平衡变得更加困难。一方面，潜变量必须包含丰富的声学信息，使解码器能够重建高质量、富有表现力的音频。另一方面，它必须呈现可预测的语言结构，以便自回归模型能够逐帧稳定地生成。面向重建的自编码器可以存储难以预测的精细声学细节，即使重建效果良好，也会使稳定生成变得困难。因此，大多数连续文本转语音系统都在寻找语义与声学信息之间经过人工调优的组合，例如通过对变分自编码器进行正则化、将语义特征蒸馏到潜变量中，或将语义规划与声学渲染分离，然后在训练大语言模型之前冻结所得空间（Lian 等人，2026年；Zhou 等人，2025b；Shen 等人，2026年；JoyFutureAcademy，2026年）。在这种分阶段方案中，大语言模型既无法塑造其必须建模的表示中的这种平衡，也不会获知解码后其预测误差会造成什么代价：回归或扩散损失并未明确区分会改变词语或说话人的误差，与几乎不会产生可听影响的误差。

我们之前的工作（Gao 等人，2026年）通过端到端训练分词器、大语言模型、流匹配（FM）解码器和基于识别的奖励模型（RM），并将来自大语言模型的直通式 Gumbel-Softmax 样本反馈到 FM 和 RM，从而解决了离散词元文本转语音中的这一分离问题。联合训练使文本转语音模型能够决定其表示应保留哪些信息，以用于识别、生成和重建，而不是预先固定语义与声学之间的平衡。本报告将这一思路扩展到连续表示；在这种表示中，可微潜变量无需通过量化和 Gumbel-Softmax 词元采样进行直通式梯度近似。不过，这种扩展仍需要明确的容量控制，因为连续潜变量原则上可以容纳无限量的信息；还需要一条可微反馈路径，使 FM 和 RM 的信息能够经由大语言模型的扩散头传回。

我们提出了 Tenlai-TTS 2.0，这是一种经过端到端训练的连续表征文本转语音模型，并作出以下贡献：

- **容量受控的连续瓶颈**。能量归一化加上固定方差高斯噪声，将潜变量转化为加性高斯信道，并得到一个闭式容量上界：在我们的主要设置下为 11.1 kbps，约为 Qwen3-TTS 的 RVQ标记器的 5 倍，以及 CosyVoice 3 的单码本标记器的 30–56 倍。

- **端到端表示学习**。 编码器与识别头、流匹配解码器和大语言模型共同训练，使表示同时受到重建和生成的塑造。

- **基于反馈的潜在表示预测优化**。 在预测的潜变量上评估的重建和识别目标，通过流匹配解码器和奖励模型为大语言模型提供可微分的反馈。

- **可控解码**。通过调整 CFG 强度，并将提示路由独立地应用于大语言模型、自回归 DiT 头部和流匹配解码器，可在推理时控制可懂度、说话人相似度和语速跟随。

- **结果**。与联合训练模型相比，基于反馈的优化阶段将 Seed-TTS-eval 字符错误率在中文测试集上从 0.82% 降至 0.74%，在 test-hard 上从 8.32% 降至 7.08%，且说话人相似度保持不变；随后进行的强化学习阶段进一步将困难集字符错误率降至 6.29%，英语词错误率降至 1.08%，同时提升了三个测试集上的说话人相似度。

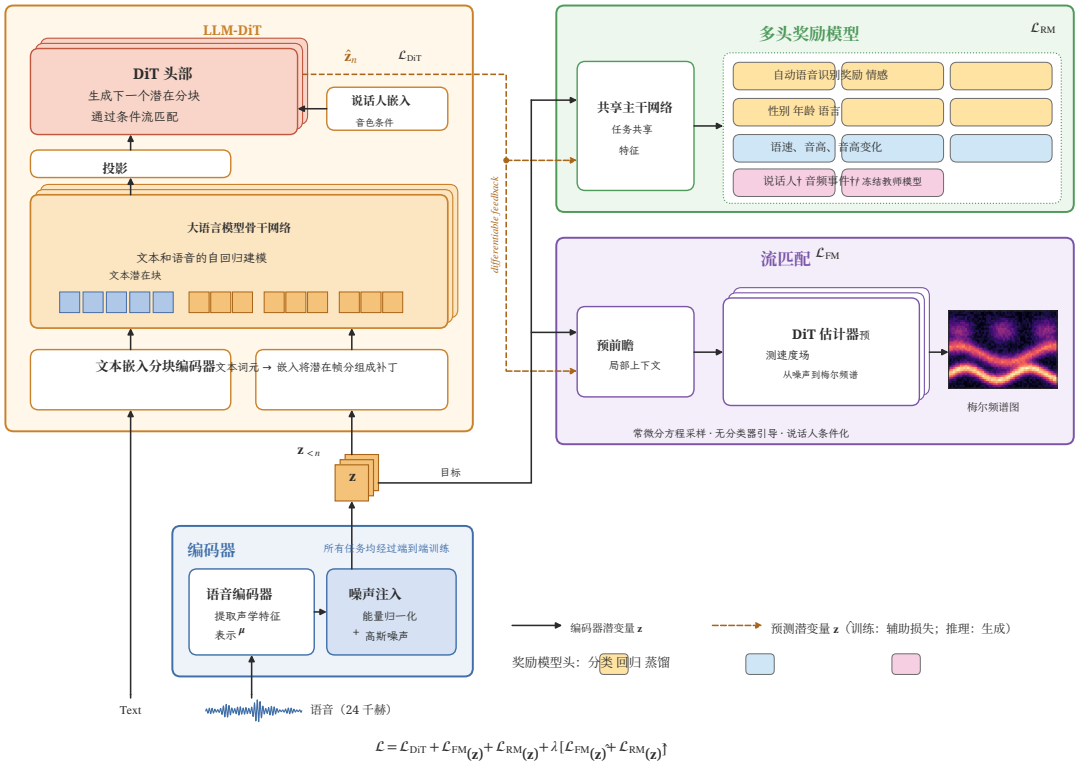
2 相关工作

基于大语言模型、采用离散和连续表示的文本转语音模型。近期的文本转语音模型主要使用离散令牌或连续表示来生成语音。Seed-TTS（Anastassiou 等，2024）、CosyVoice 2/3（Du 等人，2024b、2025）和 Qwen3-TTS（Hu 等人，2026）等离散令牌系统，将语音建模为以文本和参考音频为条件的量化令牌序列，然后将生成的令牌解码为语音。离散令牌支持分类式下一令牌预测，而连续表示则避免了量化引入的信息损失。在连续模型中，MELLE（Meng 等，2024）直接预测梅尔频谱图帧，而 DiTAR（Jia 等，2025）、VibeVoice（Peng 等人，2025）和 VoxCPM（Zhou 等人，2025b）则生成连续语音潜变量，并通过波形解码器重建音频。包括 Dots.TTS（Lian 等，2026）、FireRedTTS3（Shen 等，2026）和 JoyAI-Voice 2.0（Joy Future Academy，2026）在内的较新系统，进一步探索了兼具语义结构与声学细节的连续表示。

连续自回归文本转语音的稳定性与表现力。 连续潜变量能够保留精细的声学细节，但由于没有量化步骤将不完美的预测拉回有效值，因此细微误差可能会在长序列展开中不断累积（Lian 等人，2026年）。Dots.TTS（Lian 等人，2026年）在面向重建的变分自编码器基础上加入多任务识别头和 WavLM 对齐，向大语言模型输入每个补丁的语义摘要，以完整历史为条件驱动流匹配头，并通过无奖励自校正对该头进行后训练。FireRedTTS3（Shen 等人，2026年）取消 KL 正则化，以避免过度压缩声学细节；转而将其潜变量与冻结的音频理解编码器对齐，从表示层面解决误差累积问题。JoyAI-Voice 2.0（Joy Future Academy，2026年）将每个变分自编码器潜变量补丁拆分为两部分：一部分是来自冻结的自动语音识别骨干网络的语义表示，用于稳定规划；另一部分是联合训练的声学表示，用于保留细节，随后再进行有监督微调和强化学习。这些工作的主要区别在于如何平衡自回归生成的稳定性与音频重建的表现力。三者都在生成训练前固定潜空间，并在之后通过语义约束、专用模块或后训练来调和这两个目标。相比之下，Tenlai-TTS 2.0 以端到端方式训练表示和生成器：编码器同时接收来自重建、识别和生成的梯度，使这两个目标在瓶颈所设定的容量范围内自然达到平衡。

联合优化与基于反馈的优化。 若干工作利用下游目标来塑造上游模块：TadiCodec 和 DiffCodec 使用扩散重建来训练分词器（Wang 等，2026a，b）；JoyVoice 联合优化大语言模型和流匹配解码器（Yu 等，2025年）；DiffRO 则通过可微编解码器后端反向传播基于识别的奖励（Gao 等人，2025年）。我们此前的工作（Gao 等人，2026年）对离散标记系统中的分词器、大语言模型、流匹配和奖励模型进行端到端训练，并将大语言模型产生的直通式 Gumbel-Softmax 样本输入流匹配和奖励模型。Tenlai-TTS 2.0 将这一方法扩展到连续潜变量；在此过程中，它还需要显式的容量控制，并通过扩散头的一步潜变量预测传递来自重建和识别目标的可微反馈，而非通过采样标记传递。

3 方法



$$\mathcal{L} = \mathcal{L}_{\text{DiT}} + \mathcal{L}_{\text{FM}(\mathbf{z})} + \mathcal{L}_{\text{RM}(\mathbf{z})} + \lambda[\mathcal{L}_{\text{FM}(\hat{\mathbf{z}})} + \mathcal{L}_{\text{RM}(\hat{\mathbf{z}})}]$$

图1. Tenlai-TTS 2.0 概览。实线箭头表示编码器潜变量$\mathbf{z}$；虚线箭头表示由大语言模型预测的潜变量 $\hat{\mathbf{z}}$，该潜变量会传递给流匹配解码器和奖励模型，以便在训练期间进行基于反馈的优化（第 3.5节）。

Tenlai-TTS 2.0 的核心特点是端到端表示学习。不同于先训练变分自编码器以获得连续语音表示、再训练大语言模型来预测该固定表示的分阶段方案，Tenlai-TTS 2.0 在文本转语音系统内部直接学习表示。语音编码器与大语言模型及其自回归扩散头、流匹配（FM）解码器联合优化，而不是作为独立的表示模型进行训练，并在生成训练前冻结。因此，潜变量同时受到自回归预测和音频重建的塑造：它既必须便于大语言模型预测，又要保留流匹配解码器所需的声学信息。识别模型（RM）进一步约束语音内容和说话人身份的保留。

这些目标在明确的信息预算内取得平衡，该预算通过能量归一化和固定方差高斯噪声注入（第3.1节）加以约束。在此预算范围内，联合训练决定保留哪些信息用于预测和重建，而不是要求大语言模型拟合一种仅为重建而选定的表示。图1 展示了编码器、识别模型（RM）（第3.2节）、流匹配解码器和声码器（第3.3节），以及带有自回归扩散头的大语言模型（第3.4节）。第3.5节介绍基于反馈的潜在表示预测优化；在两个训练阶段中，编码器始终保持可训练状态（第4节）。

3.1 容量受控连续瓶颈

不受信息约束的连续潜变量可以保留有助于重建、但难以通过自回归方式预测的精细声学细节。相比之下，过度压缩可能会丢弃高保真重建所需的信息。我们通过噪声注入来限制编码器可传输的信息量，使联合训练能够在可控的容量上界内学习适用于这两项任务的表征。

实现。 语音编码器将输入音频映射为一系列D维向量 $\tilde{\boldsymbol{\mu}}_t$，采样频率为 25Hz（$D = 384$）。在将潜变量传递给任何下游模块之前，会执行两项操作：

$$\begin{aligned} \boldsymbol{\mu}_t &= \tilde{\boldsymbol{\mu}}_t \,/\, \mathrm{RMS}(\tilde{\boldsymbol{\mu}}), \\ \mathbf{z}_t &= \boldsymbol{\mu}_t + \sigma \boldsymbol{\epsilon}_t, \quad \boldsymbol{\epsilon}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I}). \end{aligned} \tag{1}$$

第一项操作使用单一全局缩放因子，将潜变量归一化为每个维度的单位平均功率；第二项操作则注入标准差为 σ 的固定 高斯噪声，以限制下游预测和重建模块可获取的信息量。

分析。对于表示z，每个维度都可以视为一个输入功率为单位值的加性高斯噪声信道。根据香农容量公式（香农，1948年），每帧的信息量上限为

$$C = \frac{D}{2} \log_2 \left(1 + \frac{1}{\sigma^2} \right) \text{ bits} \tag{2}$$

（附录A）。因此，连续潜变量的容量并非无限：它由单个超参数决定，且可在训练前计算。

表 1 将该下界与离散语音分词器的标称比特率进行了比较，并考虑了它们不同的帧率。在 $\sigma = 0.5$ 时，我们的下界为 11.1 kbps，约为 Qwen3-TTS 多码本 RVQ 标记器（2.2 kbps）的五倍，也是单码本标记器（0.20–0.38 kbps）的 30–56 倍（Du 等人，2025年；Hu 等人，2026年）。正如我们的实验（第 5节）所示，降低 σ 可提升编码器潜变量上的 RM 识别和 FM 重建能力，却会削弱大语言模型的潜变量预测能力，并最终降低合成语音的质量。因此，更有利于识别和重建的表示，仍可能更难进行自回归预测。因此，选择 σ 时，必须权衡大语言模型的预测能力与 FM 解码器的重建能力，而不能只追求更大的容量或更高的重建质量。

表1。 基于大语言模型的文本转语音系统所使用的语音表示的信息容量。单位为 kbps 的数值表示离散令牌和连续表示的理论信息速率上限。

系统	瓶颈	帧率 (Hz)	码率上限 (kbps)
CosyVoice 3（Du 等人，2025年）	FSQ，3^8	25	0.32
Qwen-Audio-3.0-TTS（Xiang 等人，2026年）	FSQ，3^{10}	12.5	0.20
Qwen3-TTS-25Hz（Hu 等人，2026年）	矢量量化，32,768	25	0.38
Qwen3-TTS-12Hz（Hu 等人，2026年）	残差矢量量化，$16 \times 2,048$	12.5	2.2
Ours, $\sigma = 0.5$	高斯分布，384 维	25	11.1
Ours, $\sigma = 0.2$	高斯分布，384 维	25	22.6

3.2 识别模型

识别模型（RM）用于检查潜变量是否保留了输入中的内容、说话人以及副语言特征。它由一个作用于 $\mathbf{z}$ 之上的 Conformer骨干网络和一组任务头组成；RM 损失 $\mathcal{L}_{\text{RM}}(\cdot)$ 是给定潜变量序列中所有启用任务头的损失之和。基础模型仅使用两个任务头：基于CTC的语音识别（Graves等人，2006）以及针对冻结的 CAM++ 编码器的说话人嵌入预测（Wang 等，2023b）。这两个任务头约束了零样本克隆最重要的要素：语音内容和说话人身份；整个预训练语料库都提供了相应标签。在有监督微调中，RM 会扩展任务头，以涵盖指令可控的属性：情感、语言和方言、性别、年龄、语音质量、语速、音高及音频事件（如果缺少标签，则从冻结的 PANNs 标注器（Kong 等人，2020）中蒸馏得到音频事件标签）。借助这些任务头，第 3.5节中的反馈目标的识别部分不仅检查说了什么以及由谁来说，还检查如何表达。

3.3 FM解码器和声码器

遵循 CosyVoice 3（Du 等人，2025年）的流匹配（FM）解码器根据说话人嵌入和提示词作为条件，从潜变量生成梅尔频谱；预训练的 HiFT 声码器（Du 等人，2024a）则将其转换为波形。对于给定潜变量序列作为条件时的流匹配损失，我们记作 $\mathcal{L}_{\text{FM}}(\cdot)$ 。该 FM–声码器路径用于重建和文本到语音合成。

3.4 带有自回归 DiT 头的大语言模型

大语言模型初始化自 Qwen3-1.7B-Base（Yang 等人，2025年），并读取可选的指令、文本和语音潜变量。每三个潜在帧合并为一个输入嵌入，因此大语言模型以约 8.3 Hz 的频率运行。在每个语音步长中，一个小型扩散Transformer头（AR-DiT）生成由三个帧组成的下一个补丁$\mathbf{z}_n$ ，同时辅助语音结束头通过损失 $\mathcal{L}_{\text{eos}}$预测终止标记。训练目标 $\mathbf{z}_n$ 是$\mathbf{z}$ 在两个阶段中的三帧补丁。该头部以大语言模型隐藏状态 $\mathbf{y}_{\le n}$、前置潜在补丁$\mathbf{z}_{<n}$ 和说话人嵌入$\mathbf{s}$为条件，并采用条件流匹配（Lipman 等人，2023年）进行训练：其中$\mathbf{x}_0 \sim \mathcal{N}(\mathbf{0},\ \mathbf{I})$ 和$\mathbf{x}_t = t\,\mathbf{z}_n + (1-t)\,\mathbf{x}_0$，

$$\mathcal{L}_{\text{DiT}} = \mathbb{E}\left\|\hat{\mathbf{v}}_n(\mathbf{x}_t, t \mid \mathbf{y}_{\le n}, \mathbf{z}_{<n}, \mathbf{s}) - (\mathbf{z}_n - \mathbf{x}_0)\right\|_2^2. \tag{3}$$

训练期间会随机丢弃条件，以启用无分类器引导（Ho 和 Salimans，2022年），并在推理时使用少量欧拉步。大语言模型的输入和历史信息取自$\mathbf{z}$，即模型在推理时反馈所使用的相同噪声空间。解码时，CFG 缩放系数以及三个独立开关分别控制是否向大语言模型、AR-DiT 头部和流匹配解码器提供参考音频。

3.5 基于反馈的潜在预测优化

流匹配解码器通常仅使用编码器潜变量进行训练，而在推理时，它接收的是由大语言模型生成的潜变量。反过来，大语言模型仅使用式（3）进行训练；该式将所有潜在维度一视同仁，且未说明预测误差会如何影响解码语音。我们利用来自重建和识别目标的可微反馈优化大语言模型的潜变量预测，这些目标通过流匹配解码器和冻结的识别模型进行评估。

对于插值块 $\mathbf{x}_t$ 和预测速度 $\hat{\mathbf{v}}_n$，流匹配恒等式给出目标块$\mathbf{z}_n$ 的一步估计，其形式为 $\mathbf{z}$，

$$\hat{\mathbf{z}}_n = \mathbf{x}_t + (1-t)\,\hat{\mathbf{v}}_n, \tag{4}$$

该估计关于大语言模型和 AR-DiT 头部均可微。将所有块拼接起来，即可得到预测潜变量序列 $\hat{\mathbf{z}}$，并据此评估反馈目标

$$\mathcal{L}_{\text{fb}} = \lambda_{\text{FM}}\,\mathcal{L}_{\text{FM}}(\hat{\mathbf{z}}) + \lambda_{\text{RM}}\,\mathcal{L}_{\text{RM}}(\hat{\mathbf{z}}). \tag{5}$$

$\mathcal{L}_{\text{FM}}(\hat{\mathbf{z}})$ 要求预测的潜变量能够被解码为目标梅尔频谱图，而$\mathcal{L}_{\text{RM}}(\hat{\mathbf{z}})$ 要求冻结的识别模型仍能从中正确读取内容和说话人信息。梯度通过公式（4）传递到大语言模型，因此生成器能够学习解码后哪些潜变量误差更为重要，而流匹配解码器则适应预测潜变量的分布。由于预测的潜变量是连续的，因此不需要直通估计器。

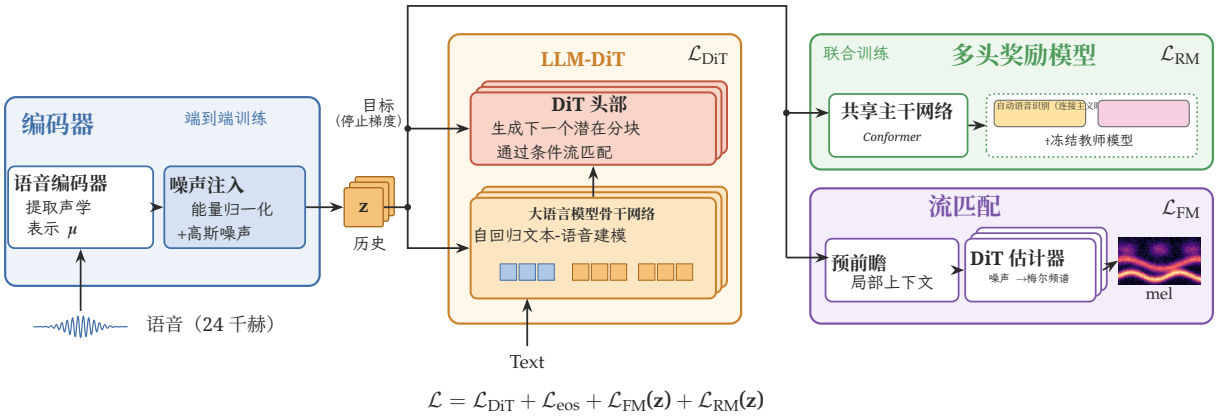
4 训练

4.1 端到端训练

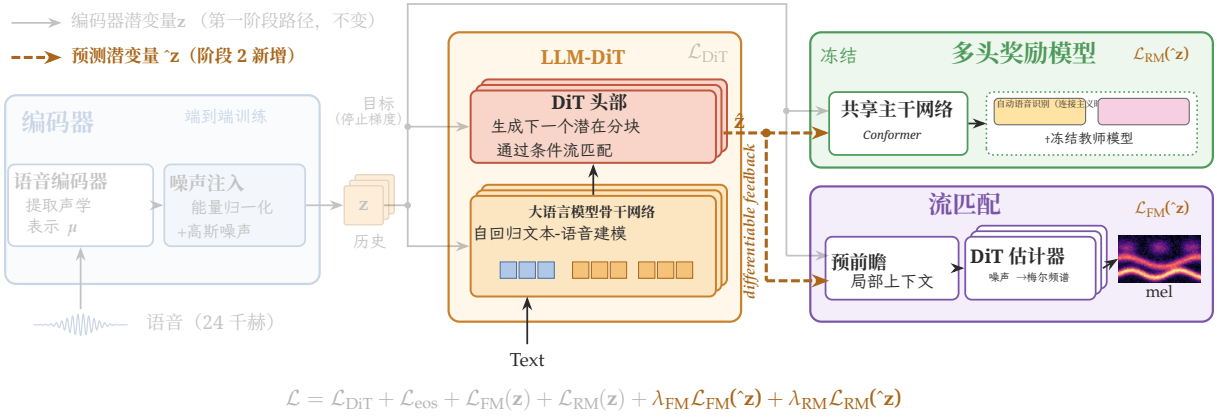
我们的端到端训练学习一种连续表示，使其既能被大语言模型预测，也能由流匹配解码器重建，同时满足噪声注入所施加的信息预算。我们采用两个连续阶段（表 2）：首先通过生成、重建和识别目标进行联合训练，以塑造编码器潜变量；随后通过流匹配解码器和冻结的奖励模型，为大语言模型自身的潜变量预测增加监督，进行基于反馈的优化。

表2。 端到端（E2E）训练和目标说话人监督微调（SFT）的训练计划。

组件	端到端训练		目标说话人监督微调	
	阶段1	阶段 2	阶段 A	阶段 B
编码器	训练	训练	冻结	冻结
RM	训练	冻结	训练	冻结
流匹配解码器	训练	训练	训练	冻结
大语言模型 + AR-DiT	训练	训练	冻结	训练
反馈损失 $\mathcal{L}_{\text{fb}}$	–	✓	–	✓



(a) 阶段1：联合训练。



(b) 阶段 2：基于反馈的优化。

图 2。 训练流程对比。实线箭头表示编码器潜变量；粗虚线箭头突出显示阶段 2 中加入的预测潜变量，该潜变量会传递给流匹配解码器和冻结的识别模型，以提供重建和识别反馈（棕色项）

阶段1：联合训练。 编码器、识别模型、流匹配解码器、大语言模型和 AR-DiT 头部通过重建损失和识别损失进行联合训练，$\mathcal{L}_{\text{FM}}(\mathbf{z}) + \mathcal{L}_{\text{RM}}(\mathbf{z})$，以及 $\mathcal{L}_{\text{DiT}}$ 和辅助终止令牌损失 $\mathcal{L}_{\text{eos}}$。编码器同时接收来自重建、识别和生成的梯度，因此表征由所有任务共同决定，而不是预先固定。AR-DiT 目标与编码器分离，再加上能量归一化，可防止潜变量坍缩。在我们的实验中，各项任务并未相互冲突，且似乎相辅相成。

阶段 2：基于反馈的优化。 编码器继续与流匹配解码器、大语言模型和 AR-DiT 头部联合训练，而识别模型则保持冻结，作为识别参照。加入公式（5）中的反馈目标，使预测潜变量也通过重建和识别获得监督。冻结识别模型至关重要：否则，$\mathcal{L}_{\text{RM}}(\hat{\mathbf{z}})$ 可能会通过调整识别器来适应生成器的错误而降低，而不是纠正这些错误。

4.2 目标说话人微调

能够独立训练或冻结各个模块，从而支持分两个阶段进行适配（表 2），以微调目标说话人；整个过程中语音编码器始终保持冻结，以保留已学习的潜在表示。

在阶段 A 中，我们使用重建损失和识别损失，基于编码器潜变量对流匹配解码器和奖励模型进行适配，同时保持大语言模型和 AR-DiT 不变。在阶段 B 中，我们冻结已适配的流匹配解码器和奖励模型，并使用下一块预测损失、终止损失以及反馈目标 $\mathcal{L}_{\text{fb}}$，对大语言模型、潜变量适配器和 AR-DiT 进行微调。冻结的声学模块和识别模块为预测的潜变量提供可微监督，引导生成结果贴合目标说话人的声学特征，同时保留语言内容。

4.3 强化学习

我们使用强化学习来优化内容准确度、说话人保真度和生成稳定性。我们将扩散负向感知微调（DiffusionNFT）（Zheng 等人，2025年）应用于自回归语音生成，在保持语音编码器、声学解码器和奖励模型固定的同时，更新大语言模型、潜变量输入适配器和 AR-DiT。

对于每个文本与说话人条件 c，行为策略会生成 K 个潜变量序列 $\{Z_i\}_{i=1}^{K}$，并对其解码后分配话语级奖励。在每个组内对奖励进行标准化，并将其映射为权重 $u_i \in [0,1]$，奖励较高的样本获得更大的权重。对于每个生成的块 $\mathbf{z}_{i,n}$，我们对噪声 $\boldsymbol{\epsilon}$ 和流时间 t 进行采样，以构造 $\mathbf{x}_t = t\mathbf{z}_{i,n} + (1-t)\boldsymbol{\epsilon}$。以 c 和生成的前缀为条件，当前速度预测器和行为速度预测器分别定义正向和负向重建分支：

$$\mathbf{v}_\theta^{+} = \mathbf{v}_\theta, \qquad \mathbf{v}_\theta^{-} = 2\mathbf{v}_b - \mathbf{v}_\theta, \qquad \hat{\mathbf{z}}_{i,n}^{\pm} = \mathbf{x}_t + (1-t)\mathbf{v}_\theta^{\pm}, \tag{6}$$

其中$\mathbf{v}_b$ i 在每次更新期间保持固定。奖励加权目标（略去常数缩放因子后） is

$$\mathcal{L}_{\text{NFT}} = \mathbb{E}\left[\frac{1}{K}\sum_{i=1}^{K}\frac{1}{N_i}\sum_{n=1}^{N_i}\left\{u_i\, d(\hat{\mathbf{z}}_{i,n}^{+}, \mathbf{z}_{i,n}) + (1-u_i)\, d(\hat{\mathbf{z}}_{i,n}^{-}, \mathbf{z}_{i,n})\right\}\right], \tag{7}$$

其中N_i 表示补丁数量，d 表示经停止梯度的平均绝对误差归一化后的平方重建误差。这两个分支分别鼓励生成奖励较高的内容，并抑制奖励较低的内容；对补丁取平均则可确保每个话语的权重相同，不受长度影响。这种前向过程更新使用生成的历史记录，无需显式似然，也无需通过展开过程和波形奖励反向传播梯度。我们维护一个移动平均行为策略，并将学习器正则化至其初始化状态，以稳定优化过程。

对于目标说话人适配，奖励综合考量内容准确度、与目标说话人的相似度，以及与参考录音的一致语速和时长，同时对截断生成施加惩罚。对于零样本 TTS，奖励综合考量内容准确度和话语完成度、与说话人提示的相似度、语速一致性，并对异常时长、静音和削波施加惩罚。

5 实验

5.1 训练数据

数据分布。通用语音语料库包含 160 万小时的中文和英文语音，涵盖粤语、潮州话、上海话、四川话、东北话和河南话等汉语方言及地域变体。来源领域包括教育语音、有声读物、播客、影视和动画以及会议语音；也包含耳语。为进一步提升文本转语音模型的性能上限，我们还额外收集了

50,000 小时的挑战性样本，涵盖极端语速、长篇语音、多样化领域、方言和说话风格，以及儿童和老年人语音（表3）。

表3。 用于提升文本转语音性能的额外5万小时挑战性样本采集的主要组成部分。

组成部分	覆盖范围与筛选
质量筛选语音	通过客观的整句质量估计和 局部窗口质量估计筛选出的语音。
极端语速	快语速和慢语速，并有针对性地强化 语速覆盖范围。
长篇语音	通过拼接相邻片段构成的单说话人示例 来自源录音的语音活动检测片段。
广覆盖的补充 ment	来自其他领域、方言和说话 风格的语音。
儿童和老年人语音	涵盖自动语音识别、会议、故事 讲述和视听语音的专用数据源。

数据筛选。 为支持使用这批多样化数据进行可靠的文本转语音训练，我们将基于质量的筛选与基于波形和对齐的筛选相结合。对于目标补充数据中经过质量筛选的部分，我们依据整体语句和局部窗口级别的客观质量估计，预先筛选语音。训练期间，我们会丢弃明确标记为多说话人的样本，以及包含非有限值或接近满量程样本的波形；缺少说话人数标签或标签未知的样本则予以保留。在存在有效对齐信息的情况下，我们会裁去边界静音，同时保留语音前最多 150 毫秒和语音后 200 毫秒的音频，并保留裁剪后时长为 0.05–199.8 秒的片段。为限制突发的响度变化，我们会拒绝电平范围或首尾漂移超过 15 dB 的样本；在有声内容充足时，测量范围限定在根据能量选出的有声帧上。最后，保留的音频取自第一声道，并重采样至 24 千赫；对于以该采样率运行的组件，还会相应生成 16 千赫版本。

5.2 评估协议

我们在 Seed-TTS-eval 的中文、英语和困难集上评估零样本语音克隆（Anastassiou 等，2024年）。中文字符错误率使用 Paraformer（Gao 等人，2023）计算，并报告词错误率；英语词错误率使用 Whisper-large-v3（Radford 等，2023），并对两段文本都应用 Whisper 英文文本规范化器。相似度为整段语音与参考提示之间余弦相似度的均值。
为验证模型在富有表现力的音频上的效果，我们还在部分实验中使用了 CV3-Eval 的 198 条语音作为测试集（Du 等人，2025年）。

5.3 表示重建与识别

我们在重建和识别任务上评估编码器潜变量。FM 重建使用 $\mathbf{z}$和源说话人嵌入，在 Seed-TTS 中文/英语和 CV3-Subject 划分上进行。直接 RM 识别使用完整的 CommonVoice 中文/英语测试集，以便与离散 FSQ 表示进行比较。

表 4. RM 识别和 FM 重建结果。词错误率以百分比表示。粗体数字表示各列中的最佳结果。

	RM 识别		FM 重建					
	CMV-zh	CMV-en	Seed-TTS 中文		Seed-TTS 英语		CV3-Subject	
表示	WER↓	WER↓	词错误率↓	相似度↑	词错误率↓	相似度↑	词错误率↓	相似度↑
FSQ 3^8	7.16	14.77	3.05	0.825	3.26	0.704	11.67	0.781
FSQ 3^{10}	**6.50**	14.53	2.93	0.826	**3.08**	0.702	11.56	0.780
$\sigma = 0.5$	6.57	14.65	1.78	0.939	4.07	0.897	7.70	0.909
$\sigma = 0.2$	6.55	**14.22**	**1.76**	**0.952**	4.03	**0.923**	**7.38**	**0.920**

表格4 表明，RM 识别和 FM 重建从更大的信息预算中获益的方式不同。减小 σ 可以放宽容量约束，但带来的识别收益有限。这表明，更紧的瓶颈已经保留了 RM 所需的大部分语言信息。在连续瓶颈内，减小 σ 能够提升 FM 重建的说话人相似度，并在全部三个测试集上略微降低重建 WER。然而，与 FSQ 相比，连续表示始终能提升说话人相似度，并降低中文和 CV3 上的重建 WER；但英语重建 WER 更高。因此，额外容量似乎更多地保留了声学细节和说话人特有细节，而非改善语言内容的恢复。RM 性能大致相当，但重建相似度高于 FSQ，这与这一解释一致。

5.4 Seed-TTS-eval 上的零样本文本转语音

表5 显示，提高表示容量并不一定能改善零样本文本转语音。在阶段1中，将 σ 从 0.5 降至 0.2，会使容量上界从 11.1 kbps 提高到 22.6 kbps（表1），但中文字符错误率从 0.82% 升至 0.92%，英语词错误率从 1.32% 升至 1.63%。中文说话人相似度也从 0.787 降至 0.765，英语说话人相似度则从 0.745 降至 0.697。可懂度和说话人相似度均有所下降，这与第3.1节讨论的容量—可预测性权衡一致：更丰富的编码器表示有助于识别和重建，但也会增加大语言模型预测的难度。因此，信息预算应同时根据自回归可预测性和流匹配重建来确定，而不应只考虑重建容量；我们将 $\sigma = 0.5$ 作为默认值。

在 $\sigma = 0.5$，阶段 2 将中文字符错误率/英语词错误率从 0.82%/1.32% 降至 0.74%/1.27%，同时相似度小幅下降，从 0.787/0.745 降至 0.782/0.744。我们的最终模型在第2阶段检查点上进行了 1,000 次强化学习更新，达到 0.77% 的中文字符错误率和 1.08% 的英语词错误率，说话人相似度分别为 0.811 和 0.766。在表 5所列的系统中，该模型实现了最低的英语词错误率和最高的中文说话人相似度；各训练阶段的效果将在第 5.6 节中进行考察。

5.5 零样本声学细节跟随

随着文本转语音质量提升，WER 和 SIM 单独使用越来越难以区分语音克隆的差异。主观听音测试可以揭示这些差异，但评估工作量很大，且容易受到评分者和测试条件的影响，从而限制了可复现性。因此，我们

表5。 零样本 Seed-TTS 评测结果。基线方法数据由作者报告。我们的最终模型是通过对第2阶段检查点应用强化学习得到的。粗体数字表示各列中的最佳结果（包括并列结果）。

模型	中文测试集		英文测试集	
	字符错误率 (%)↓	相似度↑	词错误率 (%)↓	相似度↑
Seed-TTS（Anastassiou 等，2025年）	1.12	0.796	2.25	0.762
MaskGCT（Wang 等，2025年）	2.27	0.774	2.62	0.714
F5-TTS（Chen 等，2025年）	1.53	0.760	2.00	0.670
DiTAR（Jia 等，2025年）	1.02	0.753	1.69	0.735
MegaTTS3（Jiang 等，2025年）	1.52	0.790	2.79	0.771
MiniMax-Speech（Zhang 等，2025年）	0.83	0.783	1.65	0.692
CosyVoice3-1.5B（Du 等人，2025年）	1.12	0.781	2.21	0.720
IndexTTS2（Zhou 等人，2025a）	1.03	0.765	2.23	0.706
OpenAudio-s1-mini（Liao 等人，2026年）	1.18	0.685	1.94	0.550
VibeVoice（Peng 等人，2025年）	1.16	0.744	3.04	0.689
Qwen3-TTS-1.7B-Base（Hu 等人，2026年）	1.22	0.770	1.23	0.717
MOSS-TTS（Gong 等人，2026年）	1.46	0.762	1.92	0.693
VoxCPM2（Zhou 等人，2026年）	0.97	0.795	1.84	0.753
Dots.TTS（Lian 等人，2026年）	0.94	0.810	1.30	0.771
Qwen-Audio-3.0-TTS（Xiang 等人，2026年）	0.84	0.792	1.54	0.762
FireRedTTS3-Base（Shen 等人，2026年）	1.01	0.809	1.64	**0.772**
JoyAI-Voice2.0-Base（Joy Future Academy，2026年）	**0.74**	0.771	1.27	0.739
FishAudio-Drama3（Han 等人，2026年）	0.83	–	1.54	–
Tenlai-TTS 2.0				
阶段1，$\sigma = 0.2$	0.92	0.765	1.63	0.697
阶段1，$\sigma = 0.5$	0.82	0.787	1.32	0.745
阶段 2，$\sigma = 0.5$	**0.74**	0.782	1.27	0.744
RL, $\sigma = 0.5$	0.77	**0.811**	**1.08**	0.766

通过比较合成语音与其参考提示在语速、音高和音量方面的相关性，评估更细粒度的声学细节克隆。语速描述语音的快慢；我们将移除边界静音后的每秒文本单位数作为语速的度量。音高是由 pYIN 估计的有声帧 F_0 的中位数；音量则是活跃帧均方根电平的中位数，单位为 dBFS，且不进行增益归一化。我们分别使用语速、对数 F_0 和 dBFS，计算有效参考语音–输出语音配对的皮尔逊相关系数 r；相关性越高，表示对参考语音的跟随程度越强。

在模型方面，我们比较了单码本离散模型 CosyVoice3-0.5B（Du 等人，2025年）、多码本离散模型 FishAudio-Drama3 （Han 等人，2026 年）以及连续表征模型 Dots.TTS （Lian 等人，2026 年）、FireRedTTS3（Shen 等人，2026年）和 Tenlai-TTS 2.0。测试集包括 Seed-TTS zh、Seed-TTS en 和 CV3-Subject。

表 6. 零样本语速、音高和音量跟随（皮尔逊 r；数值越高越好）。粗体数字表示各测试集和指标的最佳结果。

测试集	模型	皮尔逊 r↑		
		语速	音高（对数F_0）	音量（dBFS）
Seed-TTS zh	FishAudio-Drama3	0.443	0.928	0.818
	CosyVoice3-0.5B	0.599	0.966	0.908
	Dots.TTS	0.633	0.961	0.918
	FireRedTTS3	0.669	0.968	**0.930**
	Tenlai-TTS 2.0	**0.760**	**0.973**	0.919
Seed-TTS 英语	FishAudio-Drama3	0.371	0.910	0.813
	CosyVoice3-0.5B	0.332	0.933	0.859
	Dots.TTS	**0.425**	0.924	0.890
	FireRedTTS3	0.410	**0.944**	0.925
	Tenlai-TTS 2.0	0.401	0.942	**0.927**
CV3-sub	FishAudio-Drama3	0.730	0.766	0.836
	CosyVoice3-0.5B	0.790	0.913	0.884
	Dots.TTS	**0.868**	0.927	0.901
	FireRedTTS3	0.837	0.918	0.934
	Tenlai-TTS 2.0	0.850	**0.961**	**0.948**

Tenlai-TTS 2.0 在所有三个维度上都具有竞争力：它的中文语速相关性最高（0.760），英语音量相关性最高（0.927），CV3 的音高/音量相关性也最高（0.961/0.948）。在全部三个测试集上，连续模型的语速和音量相关性均高于两种离散基线方法。与语速和音量相比，音高差异较小。这一模式与以下观点一致：有限的离散信息预算更容易保留音高相关线索，而时序和音量细节则较难保留；相比之下，连续表示能够保留范围更广的声学信息。这表明，使用连续表示进行细粒度语音克隆的性能上限更高，超越了词错误率或相似度所能体现的差异。

5.6 基于反馈的优化和强化学习的效果

表 7 报告了联合训练（阶段1）、基于反馈的优化（阶段 2）以及在高质量训练数据的精选子集上进行强化学习后的 Seed-TTS-eval 结果。

表 7. Seed-TTS-eval 在连续训练阶段中的结果。强化学习从第2阶段检查点开始。粗体数字表示每列中的最佳结果。

训练条件	中文测试集		英文测试集		test-hard	
	字符错误率 (%)↓	相似度↑	词错误率 (%)↓	相似度↑	字符错误率 (%)↓	相似度↑
阶段1（联合）	0.82	0.787	1.32	0.745	8.32	0.757
阶段 2	**0.74**	0.782	1.27	0.744	7.08	0.752
RL	0.77	**0.811**	**1.08**	**0.766**	**6.29**	**0.784**

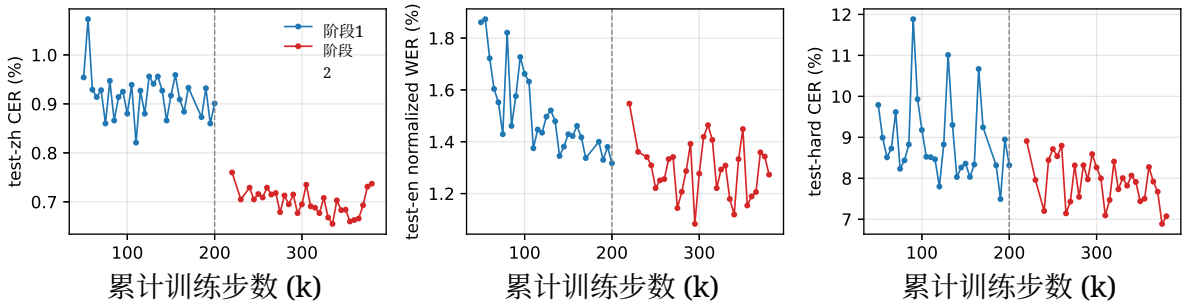


图 3。 Seed-TTS-eval 训练曲线：阶段1（蓝色，起始于 50k）和阶段 2（红色，偏移 200k）。虚线表示阶段转换。

阶段 2 降低了三个测试集上的识别错误，但说话人相似度略有下降。图 3 中的训练曲线显示，在阶段1进入平台期后，性能仍持续提升，这与重建和识别反馈对潜在预测的益处一致。

强化学习进一步将英语词错误率从 1.27% 降至 1.08%，将困难集字符错误率从 7.08% 降至 6.29%，相应的相对降幅分别为 15.0% 和 11.2%。在全部三个数据集上，说话人相似度均有所提升：在中文/英语/困难集上分别从 0.782/0.744/0.752 提升至 0.811/0.766/0.784，而中文字符错误率则从 0.74% 小幅上升至 0.77%。这些结果表明，基于反馈的优化与话语级奖励优化能够带来互补的收益。

5.7 目标说话人微调

我们在第4.2 节使用 Seed-TTS-eval 文本，在三个匿名化目标说话人上评估两阶段微调流程。图 4 将其内容一致性与表5中报告的阶段 2 零样本模型进行了比较。适配后的模型中文字符错误率为 0.71–0.80%，接近 0.74% 的参考值；英语词错误率则从 1.27% 降至 0.80–0.95%，困难集字符错误率从 7.08% 降至 4.70–5.06%。这些结果表明，目标说话人适配能够保持跨语言的可懂度，并且相较于零样本参考模型，在英语和具有挑战性的中文文本上识别错误更少。

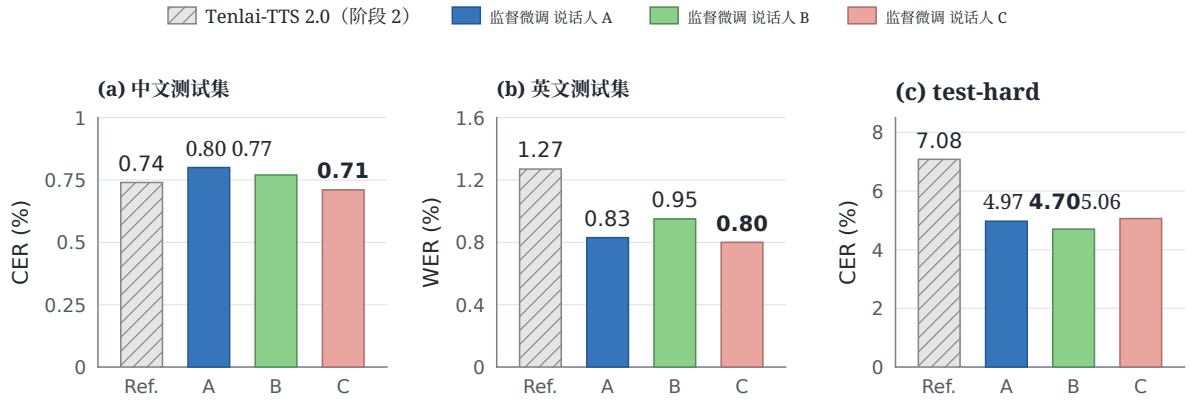


图 4. 在 Seed-TTS-eval 上，三个匿名目标说话人模型的内容一致性，与 Tenlai-TTS 2.0 阶段 2 零样本参考模型进行比较。CER/WER 越低越好；各面板使用不同的纵轴刻度。

5.8 推理时控制：分类器自由引导与提示路由

我们评估无分类器引导（CFG）和参考音频路由对可懂度及声学参考的影响。路由由三个二元指标指定，分别表示是否对大语言模型、AR-DiT 头和流匹配解码器进行参考音频条件控制；1 表示启用条件控制，0 表示禁用。例如，路由 111 会将参考音频提供给全部三个模块。我们分别评估这两个因素：固定路由为 111 时改变 CFG，在 CFG 为 1.0 时改变路由。表8中，r 表示参考语音与输出语音的对数语速相关性，报告于表6的语速列中。

表 8. 推理时的分类器自由引导与提示路由控制。使用路由 111 对分类器自由引导进行扫描；在分类器自由引导为 1.0 时对路由进行扫描。粗体数字表示每个面板中的最佳结果。词错误率以百分比表示。

（a）分类器自由引导扫描（路由 111）

分类器自由引导	词错误率↓	相似度↑	r ↑
0.0	11.55	0.7526	0.864
0.1	9.23	0.7544	0.858
0.5	7.07	**0.7644**	**0.866**
0.7	6.70	0.7621	0.836
1.0	6.01	0.7581	0.832
1.5	5.61	0.7615	0.789
2.0	**5.46**	0.7522	0.715

(b) 提示路由（分类器自由引导 1.0）

路由	词错误率↓	相似度↑	r ↑
111	6.01	0.7581	**0.832**
011	3.59	**0.7661**	0.829
001	**2.99**	0.7555	0.827

AR-DiT 的 CFG。使用路由 111 时，将 CFG 从 0 增至 2.0，可使词错误率从 11.55% 降至 5.46%（相对降幅为 52.7%）；说话人相似度和语速相关性则在 CFG 为 0.5 时达到峰值（分别为 0.7644 和 0.866；表8）。语速相关性从 CFG 为 0.5 时的 0.866 降至 2.0 时的 0.715，表明更强的引导有利于提升可懂度，但会削弱对声学参考的遵循。

提示路由。移除大语言模型提示（011）后，词错误率从 6.01% 降至 3.59%，相似度从 0.7581 升至 0.7661；仅 FM 提示（001）的词错误率达到 2.99%，但相似度降至 0.7555（表8）。语速相关性仍相近（0.832/0.829/0.827）。这些结果支持将自回归规划与声学参考的使用分离；001 仍通过 FM 提示和说话人嵌入进行条件控制。

CFG 与提示路由相辅相成，能够在无需重新训练的情况下，通过推理时控制使语音合成适应不同需求。CFG 可调节发音准确度与表现力之间的平衡，这体现在 WER 与参考语音语速相关性之间的权衡上。选择性参考音频注入可在保留明确的说话人身份条件控制的同时，调节音色、韵律和风格的迁移。这种将说话人身份与声学参考的使用分离的方式，有助于在不同风格、语言乃至方言间灵活进行语音克隆，而不必将参考音频中的所有特征一并复现。

6 结论

我们介绍了 Tenlai-TTS 2.0，这是一种基于大语言模型的文本转语音模型，能够生成连续语音潜变量，并采用端到端训练。具有可计算信息容量

上界的高斯瓶颈支持用于重建、识别和生成的联合表征学习。容量扫描表明，增大信息预算可以改善重建，但不一定能提升零样本合成效果，这凸显了平衡声学保真度与自回归可预测性的必要性。实验表明，基于反馈的优化和强化学习具有互补优势：前者可提高可懂度，后者则进一步提升英语语音和困难案例的可懂度，同时改善说话人相似度。CFG 强度和提示路由提供了互补的推理时控制，无需重新训练即可调节可懂度、说话人相似度和声学细节跟随。总体而言，这些结果支持采用容量受控的连续表示，并结合互补的基于反馈和基于奖励的优化，以实现高质量、可控的文本转语音。

参考文献

Philip Anastassiou、Jiawei Chen、Jitong Chen、Yuanzhe Chen、Zhuo Chen、Ziyi Chen、Jian Cong、Lelai Deng、Chuang Ding、Lu Gao 等。Seed-tts：一系列高品质、多用途语音生成模型。arXiv 预印本 arXiv:2406.02430，2024年。Yushen Chen、Zhikang Niu、Ziyang Ma、Keqi Deng、Chunhui Wang、Jian Zhao、Kai Yu 和 Xie Chen。F5-TTS：一款借助流匹配生成流畅、忠实语音的讲故事者。CoRR，abs/2410.06885，2024年。Zhihao Du、Qian Chen、Shiliang Zhang、Kai Hu、Heng Lu、Yexin Yang、Hangrui Hu、Siqi Zheng、Yue Gu、Ziyang Ma、Zhifu Gao 和 Zhijie Yan。Cosyvoice：基于监督语义标记的可扩展多语言零样本文本转语音合成器。CoRR，abs/2407.05407，2024a。Zhihao Du、Yuxuan Wang、Qian Chen、Xian Shi、Xiang Lv、Tianyu Zhao、Zhifu Gao、Yexin Yang、Changfeng Gao、Hui Wang 等。CosyVoice 2：利用大语言模型实现可扩展的流式语音合成。arXiv 预印本 arXiv:2412.10117，2024b。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。Changfeng Gao、Zhihao Du 和 Shiliang Zhang。基于大语言模型的文本转语音系统的可微奖励优化。arXiv 预印本 arXiv:2507.05911，2025年。Changfeng Gao、Yong Ren、Jun Yuan、Ye Bai、Zhao You 和 ShiDong Shang。离散标记大语言模型文本转语音的端到端训练。审稿中手稿，2026年。Zhifu Gao、Shiliang Zhang、Ian McLoughlin 和 Zhijie Yan。Paraformer：用于非自回归端到端语音识别的快速、准确并行Transformer，2023年。网址：https://arxiv.org/abs/22 06.08317。Yitian Gong、Botian Jiang、Yiwei Zhao、Yucheng Yuan、Kuangwei Chen、Yaozhou Jiang、Cheng Chang、Dong Hong、Mingshu Chen、Ruixiao Li 等。Moss-tts 技术报告。arXiv 预印本 arXiv:2603.18090，2026年。Alex Graves、Santiago Fernández、Faustino Gomez 和 Jürgen Schmidhuber。连接时序分类：使用循环神经网络标注未分段的序列数据。载于 ICML 会议录，2006 年。Dawei Han 、 Shidong Li 、 Shijia Liao 、 Jiahua Liu 、 Songting Liu 、 Qingzheng Wang、Yuxuan Wang、Xin Chen、Yifan Cheng、Tianyu Li 等。Fish Audio Drama 3 技术报告。arXiv 预印本 arXiv:2610.18010，2026年。Jonathan Ho 和 Tim Salimans。无分类器扩散引导。arXiv 预印本 arXiv:2207.12598，2022年。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。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。Ziyue Jiang 等。用于零样本语音合成的稀疏对齐增强潜变量扩散Transformer。arXiv 预印本 arXiv:2502.18924，2025年。Joy Future Academy。JoyAI-Voice 2.0：一种采用语义声学联合表示的全连续自回归语音生成模型。arXiv 预印本 arXiv:2610.08834，2026年。Qiuqiang Kong、Yin Cao、Turab Iqbal、Yuxuan Wang、Wenwu Wang 和 Mark D. Plumbley。PANNs：用于音频模式识别的大规模预训练音频神经网络。《IEEE/ACM 音频、语音与语言处理汇刊》，28：2880–2894，2020年。Shi Lian、Changtao Li、Bohan Li、Hankun Wang、Da Zheng、Junfeng Tian、Yufeng Ma、Colin Zhang 和 Kai Yu。dots.tts 技术报告。arXiv 预印本 arXiv:2606.07080，2026年。Shijia Liao、Yuxuan Wang、Songting Liu、Yifan Cheng、Ruoyi Zhang、Tianyu Li、Shidong Li、Yisheng Zheng、Xingwei Liu、Qingzheng Wang 等。Fish Audio S2 技术报告。arXiv 预印本 arXiv:2603.08823，2026年。Yaron Lipman、Ricky T. Q. Chen、Heli Ben-Hamu、Maximilian Nickel 和 Matthew Le。用于生成建模的流匹配。载于 ICLR。OpenReview.net，2023年。

Lingwei Meng、Long Zhou、Shujie Liu、Sanyuan Chen、Bing Han、Shujie Hu、Yanqing Liu、Jinyu Li、Sheng Zhao、Xixin Wu、Helen Meng 和 Furu Wei。无需矢量量化的自回归语音合成。CoRR，abs/2407.08551，2024年。Zhiliang Peng 等。VibeVoice 技术报告。arXiv 预印本 arXiv:2508.19205，2025年。Alec Radford、Jong Wook Kim、Tao Xu、Greg Brockman、Christine McLeavey 和 Ilya Sutskever。通过大规模弱监督实现稳健的语音识别。载于《国际机器学习大会》，第28492–28518页。PMLR，2023年。克劳德·香农。通信的数学理论。《贝尔系统技术期刊》，27(3)：379–423，1948年。Feiyu Shen、Kun Xie、Yichen Wu、Ziqi Dai、Yichen Han、Junjie Li、Xuelong Geng、Fenglong Xie、Lei Xie、Xu Tang 和 Yao Hu。FireRedTTS3：利用语义丰富的语音表示统一语音生成与编辑。arXiv 预印本 arXiv:2608.17492，2026 年。Chengyi Wang 、 Sanyuan Chen 、 Yu Wu 、 Ziqiang Zhang 、 Long Zhou、Shujie Liu、Zhuo Chen、Yanqing Liu、Huaming Wang、Jinyu Li、Lei He、Sheng Zhao 和 Furu Wei。神经编解码器语言模型是零样本文本转语音合成器。CoRR，abs/2301.02111，2023a。doi：10.48550/ARXIV.2301.02111。网址：https://doi.org/10.48550/arXiv.2301.02111。Hui Wang、Siqi Zheng、Yafeng Chen、Luyao Cheng 和 Qian Chen。CAM++：一种利用上下文感知掩码进行说话人验证的快速高效网络。载于 INTERSPEECH，第5301–5305页。ISCA，2023b。Yuancheng Wang、Haoyue Zhan、Liwei Liu、Ruihong Zeng、Haotian Guo、Jiachen Zheng、Qiang Zhang、Xueyao Zhang、Shunsi Zhang 和 Zhizheng Wu。MaskGCT：使用掩码生成式编解码器Transformer实现零样本文本转语音。载于《国际学习表征会议》，第2025卷，第47127–47150页，2025年。Yuancheng Wang、Dekun Chen、Xueyao Zhang、Junan Zhang、Jiaqi Li 和 Zhizheng Wu。TadiCodec：用于语音语言建模的文本感知扩散语音分词器。《神经信息处理系统进展》，38：147494–147523，2026a。Yuancheng Wang、Zhenyu Tang、Yun Wang、Arthur Hinsvark、Yingru Liu、Yinghao Li、Kainan Peng、Junyi Ao、Mingbo Ma、Mike Seltzer 等。使用扩散自编码器扩展语音分词器。arXiv 预印本 arXiv:2602.06602，2026b。Bajian Xiang、Cheng Wen、Han Zhao、Hao Wang、Haoxu Wang、Jiawei Jin、Jiayan Cui、Jie Chen、Mengxi Nie、Tianyu Zhao 等。Qwen-Audio-3.0-TTS：采用多阶段训练范式，实现自由可控且高度稳健的语音合成。arXiv 预印本 arXiv:2607.23938，2026年。An Yang 等。Qwen3 技术报告。arXiv 预印本 arXiv:2505.09388，2025年。Fan Yu、Tao Wang、You Wu、Lin Zhu、Wei Deng、Weisheng Han、Wenchao Wang、Lin Hu、Xiangyu Liang、Xiaodong He 等。JoyVoice：面向拟人化多说话人对话合成的长上下文条件控制。arXiv 预印本 arXiv:2512.19090，2025年。Bowen Zhang 等。MiniMax-Speech：借助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。Kaiwen Zheng、Huayu Chen、Haotian Ye、Haoxiang Wang、Qinsheng Zhang、Kai Jiang、Hang Su、Stefano Ermon、Jun Zhu 和 Ming-Yu Liu。DiffusionNFT：基于前向过程的在线扩散强化。arXiv 预印本 arXiv:2509.16117，2025年。Siyi Zhou 等。IndexTTS2：情感表达力强且时长可控的自回归零样本文本转语音的突破。arXiv 预印本 arXiv:2506.21619，2025a。Yixuan Zhou、Guoyang Zeng、Xin Liu、Xiang Li、Renjie Yu、Jiancheng Gui、Jiaheng Wu、Ziyang Wang、Xudong Shen、Runchuan Ye、Zhisheng Zhang、Jiuyang Zhou、Bingsong Bai、Weiyue Sun、Mengyuan Deng、Qundong Shi、Zhiyong Wu 和 Zhiyuan Liu。VoxCPM2 技术报告。arXiv 预印本 arXiv:2606.06928，2026年。Yixuan Zhou 等。VoxCPM：用于上下文感知语音生成和逼真语音克隆的无分词器文本转语音。arXiv 预印本 arXiv:2509.24650，2025b。

A 瓶颈的容量

设 $\mu_{t,\ d}$ 为归一化潜变量的第 d 维，其 $\mathbb{E}[\mu_{t,d}^2] = P_d$ 和 $^1{}_D \sum_d P_d = 1$由公式（1）给出。观测值 $z_{t,d} = \mu_{t,d} + \sigma\epsilon_{t,d}$ 是一个加性高斯信道。对于固定功率 P_d，互信息 $I(\mu_{t,d}; z_{t,d})$ 在采用高斯输入时达到最大值，并等于 1 2 $\log_2$（$1 + P_d/\sigma^2$）。对各维求和并应用詹森不等式，

$$\sum_{d=1}^{D} \tfrac{1}{2} \log_2\left(1 + \frac{P_d}{\sigma^2}\right) \le \frac{D}{2} \log_2\left(1 + \frac{1}{\sigma^2}\right),$$

当所有维度具有相等功率时等号成立。对于$D = 384$，在 25 Hz 下，$\sigma = 0.5$ 可实现每帧 446 比特（11.1 kbps），而 $\sigma = 0.2$ 可实现每帧 902 比特（22.6 kbps）。对于具有K 个大小为V 的码本、帧率为f的离散分词器，相应的上界为$f K \log_2 V$ 比特/秒，在码本使用均匀时达到。