FlexiVoice：通过自然语言指令实现零样本文本转语音的灵活风格控制

Dekun Chen[1], **Xueyao Zhang**[1], **Yuancheng Wang**[1],**Kenan Dai**[2],**Li Ma**[2],**Zhizheng Wu**[1]1香港中文大学（深圳） [2]华为技术有限公司

arXiv:2601.04656v1 [cs.SD] 8 Jan 2026

摘要

本研究提出了 ***FlexiVoice***，这是一种文本转语音（TTS）合成系统，可通过零样本语音克隆实现灵活的风格控制。说话风格由自然语言指令控制，语音音色则以零样本方式通过语音参考提供。FlexiVoice 以大语言模型为核心，以文本为输入，同时还可接收可选的自然语言指令和语音参考，分别控制风格和音色。FlexiVoice 采用了一种新颖的 **渐进式后训练（PPT）**方案，逐步实现精准而灵活的可控性。具体而言，它首先采用直接偏好优化（DPO），使 FlexiVoice 能够同时准确遵循自然语言指令和语音参考；随后使用多目标组相对策略优化（GRPO），将风格指令、参考音色和文本内容解耦；最后，采用指令 GRPO 以实现更高级的指令遵循。实验结果表明，FlexiVoice 超越了其他基线方法，并展现出强大的控制因素解耦能力。人工评估进一步证实了其自然度、可控性和鲁棒性。音频样本可在 https://flexi-voice.github.io/ 获取。

1 引言

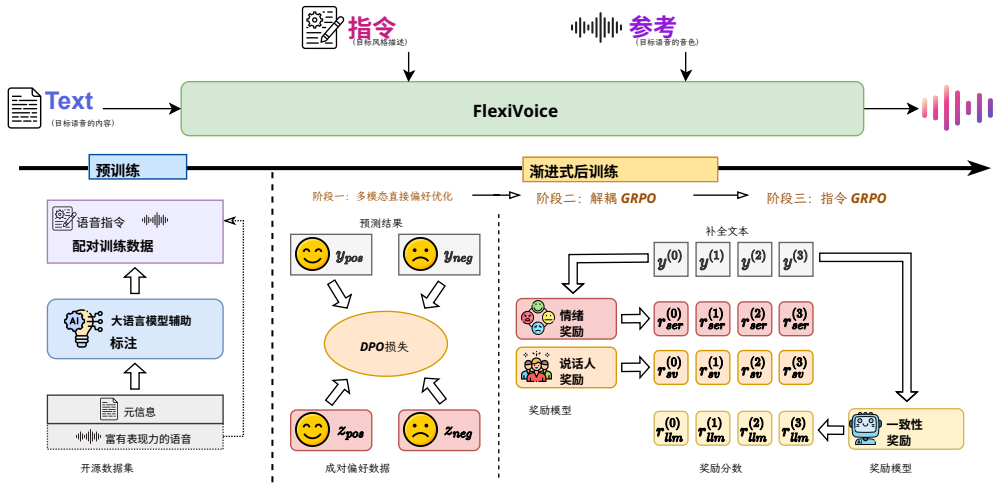

图 1：FlexiVoice 概览，支持使用任意语音音色生成多样化风格。它接受可选的自然语言风格**指令**以及可选的音色**参考** 语音。它由预训练和渐进式后训练（PPT）阶段组成。PPT 流程包括三个阶段：**阶段一：** 多模态直接偏好优化、**阶段二：** 解耦 GRPO、**阶段三：** 指令 GRPO。

近年来，文本转语音（TTS）的进展主要得益于大语言模型（LLM）的兴起以及不断完善的后训练技术。一项重要突破是零样本文本转语音（Du 等，2024；Zhou 等，2025），它仅凭一段简短的参考语音即可实现语音克隆，有效捕捉并复现说话人的音色。除了音色之外，如何控制说话风格也已成为一项重要挑战。一个方向以 Vevo（Zhang et al., 2025年c）和 IndexTTS2（Zhou 等，2025）为代表，使用两段独立的语音参考，分别控制音色和风格。另一条研究路线是基于指令的文本转语音（Vyas 等，2023；Zhou 等，2024b；Ji 等，2024b），通过自然语言指令指定目标风格。然而，现有的指令驱动模型往往难以准确遵循指令，或难以维持稳定的音色一致性。

在零样本文本转语音中实现灵活的风格控制面临一项独特挑战：**“风格-音色-内容冲突”**。在标准监督训练中，模型往往过度依赖参考语音中的强声学先验（音色泄漏），或根据文本推断韵律（内容泄漏），常常忽略明确的风格指令。仅仅采用指令条件化不足以解决这些纠缠的模态。因此，我们需要一个稳健的框架，不仅要对模型进行条件化，还要主动解耦这些因素，并在声学线索相互冲突时确保模型遵循指令。

在本研究中，我们提出了**FlexiVoice**，一种能够利用零样本声音灵活控制说话风格的文本转语音系统。具体而言，它可以接收自然语言指令和语音参考，也可以只接收其中之一，从而实现灵活的可控性。自然语言指令用于控制说话风格（例如情绪、语速），语音参考则用于控制说话人身份对应的音色。FlexiVoice 基于预训练大语言模型（LLM）构建。其核心 LLM 赋予 FlexiVoice 稳健、全面的指令遵循能力。为实现灵活的可控性，我们首先构建了一个大规模、多样化的语音数据集，其中包含自然语言指令，并将其命名为**FlexiVoice-Instruct**。FlexiVoice-Instruct 数据集借助 LLM 进行标注。随后，我们使用 Emilia（He 等，2024）和 FlexiVoice-Instruct 对 FlexiVoice 进行预训练。在后训练阶段，我们提出了一种新颖的**渐进式后训练（PPT）**框架。与通用的大语言模型对齐不同，PPT 专为解决文本转语音中的模态冲突而设计，并通过系统化的课程学习实现这一目标。它包含三个阶段：（1）**多模态直接偏好优化**为指令遵循建立初始对齐；（2）**解耦 GRPO**引入包含冲突数据场景的多目标优化，从数学上确保风格与音色、内容相分离；（3）**指令 GRPO**利用音频语言模型（ALM）奖励，将这一能力推广至复杂、开放式指令。这种渐进式方案将指令遵循任务从简单的条件化转变为严格的解耦过程。

我们使用情感数据集和 InstructTTSEval（Huang 等，2025）基准测试，从灵活可控性和指令遵循能力两个方面评估 FlexiVoice。实验结果表明，FlexiVoice 能够将说话风格（例如情绪）与说话人身份解耦。与基线方法相比，FlexiVoice 在多模态控制评估中的指令遵循度和鲁棒性均有显著提升，并且在复杂指令任务中表现出色。人工评估结果进一步证实了生成语音的自然度和鲁棒性。

贡献总结如下：

- 我们提出了**FlexiVoice**，这是一种文本转语音系统，可实现零样本语音合成的灵活、精准风格控制。说话风格由可选的自然语言指令控制，音色则由可选的参考语音引导。该系统支持风格与音色的任意组合，实现灵活控制。

- 我们构建了大规模、多样化的自然语言指令语音数据集**FlexiVoice-Instruct**。这些标注依据语音属性，使用大语言模型（即 Deepseek-V3（Liu 等，2024a））生成。该数据集广泛涵盖类人指令，包括富有表现力的场景。

- 我们提出了一种专为可控文本转语音设计的新型课程学习方法——**渐进式后训练（PPT）**框架。通过策略性地安排偏好对齐与多目标优化的顺序，PPT 明确解决了风格-音色-内容冲突，实现了稳健的解耦，而标准训练范式无法做到这一点。

2 相关工作

可控语音合成可控语音合成研究主要沿着两个方向发展。（i）零样本文本转语音从简短的参考语音中克隆音色（Chen et al., 2024; Wang et al., 2024; Zhou 等，2025; Zhang et al., 2025年c），同时对说话风格进行部分控制。（ii）基于指令的文本转语音使用自然语言提示指定风格；例如，PromptTTS（Guo et al., 2023）和 PromptStyle（Liu et al., 2023）支持文本引导的风格控制，但控制范围有限；Parler-TTS（Lyth & King, 2024）则借助弱标签扩展条件控制能力，却不支持音色控制。较新的系统，如 VoxInstruct（Zhou et al., 2024b）、AudioBox（Vyas et al., 2023）、ControlSpeech（Ji et al., 2024b）和 CosyVoice2（Du 等，2024），将指令与参考语音相结合，但仍缺乏稳健的解耦能力和丰富多样的风格。这些局限表明，我们需要一个统一框架，能够处理多模态输入、根据指令定义的风格生成语音、保留音色，并明确解决解耦问题。

指令语音数据集 基于指令的文本转语音推动了将文本描述与语音相结合的数据集的发展。TextrolSpeech（Ji et al., 2024a）引入了包含五种风格因素的大规模提示语音对；Audiobox（Vyas et al., 2023）将这一范式扩展到多模态音频生成，但由于未公开发布，因而可及性有限。Parler-TTS（Lyth & King, 2024）将说话人/风格/条件的弱标签扩展至数万小时的数据规模。SpeechCraft（Jin et al., 2024）和 ParaSpeechCaps（Diwan 等，2025）通过自动字幕生成和多样的副语言属性，丰富了标注粒度。尽管如此，大多数语料库都来自同质化来源，并侧重于模板化描述，对自然、多样指令的覆盖仍然不足。我们的数据集旨在提供更高质量、更自然的标注，以加强指令遵循能力。

语音合成中的强化学习 近期，研究者开始探索利用强化学习提升语音合成的可控性。INTP（Zhang et al., 2025a）将直接偏好优化应用于具有挑战性的零样本案例，以提升可懂度。Emo-DPO（Gao et al., 2025）将偏好对齐扩展至情感控制。Vevo2（Zhang 等，2025b）采用多目标后训练，同时提升语音和歌唱的可懂度与韵律。这些工作凸显了基于强化学习的对齐在针对性提升文本转语音特定方面能力上的前景。与之不同，我们的方法采用渐进式课程，利用强化学习明确解决模态解耦和复杂指令遵循问题，从而实现零样本多模态文本转语音更广泛的可控性。

3 FlexiVoice 概述

FlexiVoice 基于与近期其他文本转语音系统类似的大语言模型构建（Du 等，2024；Zhou 等，2025）。在 FlexiVoice 中，语音分词器将语音转换为离散词元，作为大语言模型的输入和输出。大语言模型核心处理输入文本、自然语言指令和参考词元，以生成离散语音词元。生成的词元通过流匹配（Lipman 等，2023）转换为梅尔频谱特征，最后由声码器转换为波形音频。详细描述见附录 A.1。FlexiVoice 首先在大规模数据集上预训练为 **FlexiVoice-Base**，然后采用渐进式策略进行后训练。

3.1 预训练

FlexiVoice-Base 是 FlexiVoice 的预训练模型。它使用 Emilia（He 等，2024）以及附录 A.2 中列出的多样化指令语音数据集进行预训练。

Emilia（He 等，2024）是一个大规模、多语言且多样化的语音数据集，涵盖广泛的说话风格和场景，主要支持语音生成的基础能力。为进一步实现指令引导的文本转语音，我们构建了一个高质量、多样化的数据集 **FlexiVoice-Instruct**，其中包含适用于各种场景的自然语言指令。处理流程详见第 4 节。此外，我们还利用现有的指令-语音语料库（Diwan 等，2025）以及包含情绪、年龄和辩论等属性的专门资源，丰富预训练阶段；相关处理细节见附录 A.2。我们还纳入了 NVSpeech（Liao 等，2025），该数据集提供副语言标签并覆盖丰富的表达方式。这些资源共同构成了我们模型的预训练语料库。

遵循文本转语音系统的传统预训练策略（Du 等，2024；Zhou 等，2025），我们仅训练大语言模型核心，并冻结其他模块，且在预训练期间不引入参考语音。如图 3 所示，首先使用大语言模型输入模板对文本和指令进行格式化。配对的真实语音经冻结的语音分词器预处理为离散词元，并用于与生成的词元计算损失。为确保输入格式一致，对于来自 Emilia（He 等，2024）和 NVSpeech（Liao 等，2025）且缺少明确指令的数据，我们采用 请朗读以下文本 作为默认指令。

3.2 后训练

我们提出一种**渐进式后训练（PPT）**方案，以解耦说话风格与音色，并使 FlexiVoice 具备复杂指令遵循能力。该方案受到课程学习（Bengio 等，2009）的启发：课程学习从较简单的目标开始，逐步推进到更困难的目标，从而实现稳定优化并提升泛化能力。预训练后，FlexiVoice-Base 已具备扎实的零样本文本转语音能力，但在处理多模态输入和复杂指令时仍有困难。PPT 旨在提升 FlexiVoice-base 的泛化能力与性能。最终，这种渐进式课程将打造出强大的多模态指令文本转语音模型**FlexiVoice**。

PPT 分为三个阶段：

- **阶段一** 在带有显式标签的受控情绪中心任务中，对齐指令与参考语音的多模态可控性。
- **阶段二** 在与阶段一相同的场景中，解耦参考语音中的音色与风格，以及目标文本中的内容与风格，
- **阶段三**扩展到更模糊、也更难对齐的复杂现实指令。

3.2.1 阶段一：多模态可控性

阶段一主要采用直接偏好优化，使 FlexiVoice 能够同时处理风格指令和音色参考。前两个阶段针对情感指令，以降低任务复杂度并明确解决核心问题。我们将指令限制为如下模板：使用 {标签} 情绪来朗读，标签从 中性、快乐、愤怒、悲伤和惊讶中选择。

对于与情绪相关的任务，可以直接从语音情感识别数据集中获取成对偏好数据。我们使用情感语音数据集（ESD）（Zhou 等人，2021年），其中同一说话人会以不同情绪朗读相同的句子。遵循 Gao et al.（2025年）的方法，我们为每个数据点通过指令模板（所有模板均列于附录 A.5）指定一个目标情绪标签（例如，快乐），选择带有目标情绪的句子作为偏好样本，将情绪不同的相同句子（例如，愤怒）作为不偏好样本，并使用同一说话人的中性样本作为参考语音（附录 A.5 提供了一个样本对示例）。

DPO 可直接使模型的情绪输出与指令和参考语音保持一致，而无需显式的奖励模型（Rafailov 等人，2023年）。偏好数据集 $\mathcal{D}$ 由元组 $(x,\ y_w,\ y_l)$ 组成，其中 x 包括指令、文本和参考；y_w 是符合指令的“胜出”响应，而 y_l 是“失败响应”。DPO 损失定义如下：

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

其中，π_θ 是策略模型，π_{ref} 是参考模型，二者均以 **FlexiVoice-Base** 初始化。

3.2.2 阶段二：参考语音与目标文本的解耦

阶段二聚焦于 FlexiVoice 在语音参考和目标文本包含与指令相冲突的风格的场景中的解耦能力。经过 DPO训练后，模型在中性参考语音下能够良好地遵循情感指令，但当参考语音或文本本身带有情绪，且与指令中定义的目标情绪相冲突时，仍会受到干扰。为明确抑制这些影响，我们通过构建冲突训练场景（例如，快乐指令与悲伤参考语音）采用多目标 GRPO 形式。奖励 r_{ser}（定义如下）充当

风格约束：如果模型从参考语音或文本中泄露风格信息，就会受到惩罚。相反，r_{sv}充当音色约束，确保说话人身份得到保留。通过优化联合优势，模型被迫将这些因素解耦，以最大化总奖励。我们使用相同的指令模板，将 SER 数据集中的中性和情绪片段结合起来作为参考语音，并从 NCSSD（Liu 等，2024b）中采样文本（详情见附录 A.4）。

奖励定义如下：(1) $r_{ser} \in (0, 1)$，由 Emotion2vec-Large（Ma 等，2024）的情感识别结果给出的、符合所指示情绪的概率分数；(2) $r_{sv} \in \{0, 1\}$，由 CAM++ （Wang 等，2023）给出的说话人验证结果，用于确保音色一致性。遵循 Zhang 等（2025年b），采用多目标奖励时的优势为：

$$A^i_{emo} = \frac{r^i_{ser} - \text{mean}(r^i_{ser})}{\text{std}(r^i_{ser})} + \frac{r^i_{sv} - \text{mean}(r^i_{sv})}{\text{std}(r^i_{sv})}$$

其中，i 表示同一输入 x对应的 K 个候选项中第 i个生成结果。

3.2.3 阶段三：增强复杂指令遵循能力

最后一个阶段旨在提升模型对复杂现实指令的遵循能力，不再局限于情绪任务。由于在这种规模下难以获取成对偏好数据，我们直接采用 GRPO。与第一阶段可通过 SER 模型获得奖励不同，评估语音与开放式指令之间的一致性要困难得多。Huang 等人（2025年）表明，Gemini-2.5-Pro 能够提供与人类偏好一致的可靠判断，但将其用于 GRPO 的成本过高。因此，鉴于其出色的语音理解能力，我们采用开源的 Kimi-Audio-7B-Instruct（Ding 等，2025年）作为奖励模型。我们提示它针对生成语音是否符合指令输出是/否二元判断，并将其映射为奖励 $r_{llm} \in \{1, 0\}$。

在此阶段，我们仅使用指令和文本作为输入，舍弃参考音频，因为参考音频可能与开放式约束（例如性别）冲突，并导致训练不稳定。数据构建过程的详情见附录 A.4。

为缓解灾难性遗忘，我们在此阶段混入少量 S2-GRPO 数据。因此，最终的 GRPO 训练成为一项多任务、多目标优化。此阶段的单目标优势（左）和最终优势（右）如下：

$$A^i_{ins} = \frac{r^i_{llm} - \text{mean}(r^i_{llm})}{\text{std}(r^i_{llm})}, \qquad A^i = \begin{cases} A^i_{emo} & \text{for inputs in S2,} \\ A^i_{ins} & \text{for inputs in S3.} \end{cases}$$

4 FLEXIVOICE-INSTRUCT 数据集

4.1 概述

为了在预训练阶段赋予模型基本的指令遵循能力，必须构建一个涵盖多种风格和场景的大规模指令-语音数据集。因此，我们构建了一个高质量且多样化的数据集，总计 4,316 小时，并使用基于大语言模型的标注器处理与语音相关的文本元数据。这种方法能够高效生成自然的高级指令，更好地反映真实的人类使用模式。

4.2 主要流程

我们采用统一的处理策略，使用两个数据源：Emilia 和游戏配音 ：

Emilia（He 等，2024）是一个大规模、多样化的真实场景语音数据集。由于音频来源于视频平台和播客，其中一些条目包含视频标题和标签列表等源元数据。结合转录文本，这些线索使我们能够推断说话风格和场景语境，即使不进行直接的语音分析，也能生成开放式指令。我们使用 Deepseek-V3（Liu 等，2024a）进行指令标注。为过滤噪声样本（例如网址或相互冲突的元数据–转录文本对），我们提示大语言模型评估元数据对于风格与场景推断的信息价值，从而确保更高的数据质量。

为增强表现力和风格丰富度，我们还加入了来自两款热门游戏的游戏配音数据。这些数据的一个显著特点是说话风格与角色

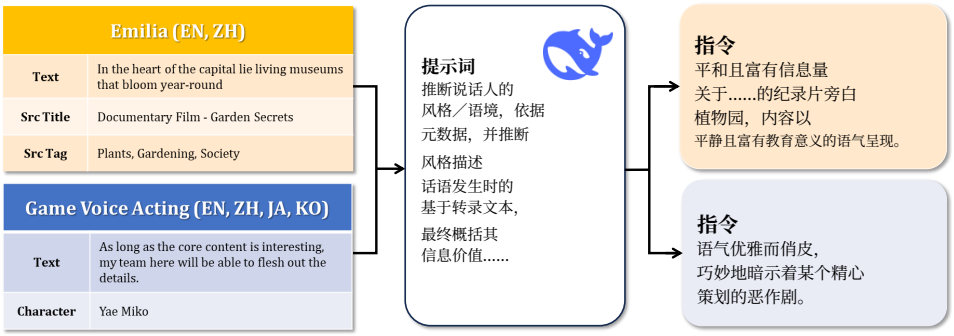

图2：FlexiVoice-Instruct 的处理流程和示例。对于两种数据来源，我们都使用语音转录文本及相关元信息提示大语言模型，生成自然、类人化的描述，作为预训练阶段的指令。

个性之间联系紧密。在大规模网络语料库上训练的大语言模型通常能够识别这些角色，并捕捉其标志性风格。因此，我们使用 Deepseek-V3（Liu 等，2024a），根据转录文本和角色名称生成指令。具体而言，我们首先提供完整的角色列表，让大语言模型识别其能够可靠识别的角色。对于已知角色，我们提示大语言模型根据角色名称推断其个性和说话风格，然后利用转录文本进一步完善描述。与 Emilia 一样，我们加入了信息价值检查，以过滤噪声标注。提示词见附录 A.3，整体处理流程和示例则见图 2。

5 实验

5.1 实验设置

基准测试 为评估模型的多模态可控性及其遵循复杂现实指令的能力，我们采用一个自建数据集和一个开源基准测试。针对多模态控制和解耦，我们定义了两种任务类型，每种各有两个难度级别，重点关注基于指令的情感语音合成，如表 1 所示。对于复杂指令遵循，我们采用 InstructTTSEval（Huang 等，2025）基准测试。

表 1：用于多模态控制和解耦评估的任务。包含两种任务，每种有两个难度级别，并为每种任务提供一个示例。

任务类型	难度	Text	参考	示例
单模态: 纯文本（TO）	Easy	中性	-	指令：使用**快乐**情绪来朗读。 文本：今天是星期一。 参考语音：-
	Hard	情感	-	指令：使用**快乐**情绪来朗读。 文本：外面在下雨，我真是**悲伤**。 参考语音：-
多模态: 文本和参考音频（TR）	Easy	中性	中性	指令：用**快乐**的情绪来表达。 文本：今天是星期一。 参考语音：[一种**中性**的声音]
	Hard	中性	情感	指令：用**快乐**的情绪来表达。 文本：今天是星期一。 参考语音：[一个**悲伤的**声音]

第一个评估集由 MEAD（Wang 等，2020）（英语）和 CSEMOTIONS（Tian 等，2025）（中文）构建而成。在 TR任务中，参考语音会从与真实标注相同的说话人的中性或冲突情绪片段中随机选取。对于 TO-hard，文本会替换为承载不同于目标情绪的情绪的句子，因此该情况下不存在真实标注音频。我们从两个数据集中各保留五种情绪类别，并分别为英语和中文随机抽取 500 个样本，以确保目标情绪覆盖均衡。

表2：不同任务和两种语言中的多模态可控性与解耦评估结果。

模型	以纯文本作为输入（TO）				以文本和参考语音作为输入（TR）						
	Easy		Hard		Easy			Hard			
	ACC-I↑	E-SIM↑	ACC-I↑	文本准确率↓	ACC-I↑	E-SIM↑	SV↑	ACC-I↑	参考语音准确率↓	E-SIM↑	SV↑
EN											
真实值	93.4	1.00	-	-	93.4	1.00	-	93.4	0.6	1.00	-
Parler-TTS	44.6	0.72	12.2	42.0	-	-	-	-	-	-	-
PromptStyple	43.8	0.70	14.0	33.6	-	-	-	-	-	-	-
PromptTTS	57.8	0.79	15.0	41.0	-	-	-	-	-	-	-
CosyVoice2	-	-	-	-	65.6	0.85	**99.8**	61.0	14.4	0.84	**99.8**
VoxInstruct	70.6	0.84	17.8	41.2	58.5	0.81	89.0	49.7	23.9	0.80	90.6
FlexiVoice-Base	72.4	0.83	39.4	30.6	58.8	0.81	99.2	48.8	32.2	0.78	99.4
FlexiVoice	**97.4**	**0.89**	**89.4**	**6.6**	**89.4**	**0.90**	91.0	**78.2**	**10.6**	**0.87**	95.8
ZH											
真实值	61.6	1.00	-	-	61.6	1.00	-	61.6	4.4	1.00	-
CosyVoice2	-	-	-	-	44.4	0.84	**99.8**	47.8	15.3	0.79	**100.0**
VoxInstruct	48.6	**0.76**	29.0	21.2	19.4	0.75	46.8	18.7	23.2	0.73	59.8
FlexiVoice-Base	78.4	**0.76**	66.8	14.2	25.2	0.78	99.6	22.4	38.0	0.74	99.2
FlexiVoice	**99.8**	0.72	**98.4**	**0.8**	**81.8**	**0.85**	98.8	**75.8**	**13.2**	**0.80**	98.4

第二个基准测试 InstructTTSEval（Huang 等，2025）是一个综合测试集，用于评估超越简单情绪的复杂指令遵循能力。它包含 1,000 个英语样本和 1,000 个中文样本，涵盖 12 种不同的语音属性：性别、音高、音质、清晰度、流畅度、语速、口音、年龄、音量、情绪、语气和个性。这些属性被封装在涵盖三个任务层级的指令中——声学参数规范（APS）、描述性风格指令（DSD）和角色扮演（RP）——从而能够严格评估模型处理多样且细粒度风格控制的灵活性。

指标 对于第一个评估集，我们使用 Emotion2vec-Large（Ma 等，2024）进行语音情感识别和情绪嵌入提取。我们报告相对于目标情绪的指令遵循准确率（**ACC-I**，↑）；在困难设置中，还报告与文本或参考语音中的冲突情绪相比的准确率（**ACC-T, ACC-R**，↓），数值越低越好。真实标注语音与生成语音的情绪嵌入之间的余弦相似度（**E-SIM**，↑）进一步反映了遵循情况。为验证 TR 任务中的音色保持情况，我们使用 CAM++（Wang 等，2023）评估说话人验证准确率（**SV**，↑）。我们还评估可懂度和感知质量：英语词错误率（**WER**，↓，使用 Whisper-Large-V3 进行自动语音识别）和中文字符错误率（**CER**，↓，使用 Paraformer-zh 进行自动语音识别）、语音质量平均意见分（**Q-MOS**，1–5，↑）以及比较平均意见分（**CMOS**，–2 至 2，↑）。主观评估设置的详细信息见附录 A.9。

对于复杂指令遵循，我们遵循 InstructTTSEval（Huang 等，2025），使用 Gemini 作为评判模型评估一致性，并报告宏平均准确率。

基线方法 我们将 FlexiVoice 与具有代表性的开源指令文本转语音系统进行比较，包括 Parler-TTS（Lyth & King，2024）、Prompt-Style（Liu 等人，2023）和 PromptTTS（Guo 等人，2023）的可复现变体（Ji et al.，2024b）、VoxInstruct（Zhou 等人，2024b），以及能够联合处理自然语言指令和参考语音的 CosyVoice2（Du 等，2024）。此外，我们还报告了 InstructTTSEval（Huang 等，2025）的结果，该评测涵盖上述开源系统和闭源商用模型（gemini-2.5-flash-preview-tts、gemini-2.5-pro-preview-tts、gpt-4o-mini-tts 和 Hume）；我们还纳入了 MiMo-Audio-7B-Instruct，这是一款近期推出、具有强大指令遵循能力的大型音频语言模型。

5.2 FlexiVoice 的解耦能力

从表2可以看出，FlexiVoice 在处理多模态输入，以及在英语和中文任务中分离内容、音色与风格方面表现出强大的能力。讨论从四个角度展开。

表3：第一个评估集中的词错误率分数和主观评估结果。

模型	纯文本作为输入（TO）			文本和参考语音作为输入（TR）					
	Easy			Easy			Hard		
	WER↓	Q-MOS↑	CMOS↑	WER↓	Q-MOS↑	CMOS↑	WER↓	Q-MOS↑	CMOS↑
EN									
真实值	4.50	3.16 ±0.07	0.00	4.50	3.50 ±0.13	0.00	4.50	4.26 ±0.22	0.00
CosyVoice2	-	-	-	**3.71**	3.50 ±0.36	-0.75 ±0.32	**3.60**	3.68 ±0.43	-0.88 ±0.21
VoxInstruct	7.29	3.02 ±0.34	-0.75 ±0.44	14.44	2.10 ±0.23	-1.50 ±0.19	12.61	2.66 ±0.63	-0.86 ±0.36
FlexiVoice-Base	**5.01**	3.72 ±0.14	-0.12 ±0.39	5.31	**3.90** ±0.26	-1.25 ±0.29	6.55	3.82 ±0.29	-0.56 ±0.31
FlexiVoice	5.99	**4.08** ±0.29	**0.91** ±0.20	5.23	3.62 ±0.25	**0.89** ±0.30	6.99	**4.06** ±0.35	**0.78** ±0.40
ZH									
真实值	4.55	3.78 ±0.14	0.00	4.55	3.38 ±0.14	0.00	4.55	3.92 ±0.07	0.00
CosyVoice2	-	-	-	**0.78**	3.54 ±0.27	-1.14 ±0.24	**0.89**	3.58 ±0.39	0.36 ±0.29
VoxInstruct	**3.37**	3.18 ±0.28	-1.60 ±0.30	10.04	2.62 ±0.28	-1.88 ±0.09	9.40	3.04 ±0.17	-1.62 ±0.19
FlexiVoice-Base	5.01	**4.10** ±0.26	**0.75** ±0.40	3.08	3.74 ±0.28	-1.50 ±0.23	5.63	3.66 ±0.06	-0.88 ±0.45
FlexiVoice	7.59	4.04 ±0.23	0.60 ±0.41	4.34	**3.79** ±0.26	**-0.36** ±0.24	7.02	**3.76** ±0.29	**0.57** ±0.44

多模态控制 比较这两项任务的简单集合，在指令的指导下，TO-easy 仅包含文本条件（模型使用随机音色），而 TR-easy 则将文本和参考语音作为多模态条件输入（模型使用参考语音的音色）。大多数基线方法无法同时支持这两者；即使是能够支持的基线方法以及 FlexiVoice-Base，表现也有较大差距，尤其是在中文任务中。经过渐进式后训练后，FlexiVoice 的性能大幅提升，在 TO-easy 上的英语和中文 ACC-I 分别达到 97.4% 和 99.8 %，在某些情况下甚至超过真实标注。在 TR-easy 上，FlexiVoice 也保持了较高的准确率（英语 89.4%，中文 81.8%），同时维持说话人一致性，凸显了其强大的多模态控制能力。

目标文本解耦 对于 TO 的两个难度等级，简单等级使用中性文本，而困难等级使用与指令所表达的情绪不同的文本。此设置旨在测试模型能否忽略文本中的冲突情绪线索。基线方法和 FlexiVoice-基础模型通常无法应对这一情况，表现为文本准确率较低、ACC-T 较高，这表明它们倾向于受文本所暗示的风格影响。相比之下，FlexiVoice 显著提升了解耦能力：在英语中，ACC-I 达到 89.4%，而 ACC-T 仅为 6.6%；在中文中，两者分别为 98.4% 和 0.8%。值得注意的是，中文设置下简单与困难等级之间的差距缩小至仅 1.4%，表明 FlexiVoice 能够出色地确保风格完全由指令控制。

参考语音解耦 对于ForTR，简单任务使用中性语音作为音色参考，而困难任务使用与指令不一致的带有强烈情绪的参考语音。大多数基线方法的准确率都大幅下降，且参考语音准确率较高，表明其受到参考语音风格的干扰。FlexiVoice 缓解了这一问题，在英语中取得了78.2%的ACC-I，在中文中取得了75.8%的ACC-I，同时参考语音准确率较低（英语为10.6%，中文为13.2%）。这表明，FlexiVoice能够有效地将参考音色与风格解耦，在遵循指令所定义情绪的同时保留音色。

风格控制与说话人验证之间的权衡 我们注意到，在某些场景下（例如表2中的TR任务），预训练的FlexiVoice-Base的说话人验证分数略高于最终的FlexiVoice。这是风格合成任务固有的预期现象。由于缺乏显式的解耦能力，FlexiVoice-Base模型倾向于直接克隆参考语音的韵律。虽然这会带来较高的说话人相似度，但无法遵循风格指令（其较低的ACC-I分数证明了这一点）。相比之下，FlexiVoice必须显著修改声学特征（例如音高和能量），以满足目标风格指令（例如将悲伤的参考语音转换为愉快的声音）。这些必要的韵律调整不可避免地会导致说话人嵌入的余弦相似度在数值上略有下降。重要的是，FlexiVoice在保持较高说话人验证分数的同时实现了更优的指令遵循，表明它能够在不损害说话人基本身份的情况下，成功优先保证风格控制。

可懂度与主观评估表 3 展示了可懂度和感知质量。与 FlexiVoice-Base 相比，FlexiVoice 的 WER/CER 略有上升。这一现

表 4：闭源和开源模型在两种语言上的 InstructTTSEval 复杂指令遵循评估结果。

模型	InstructTTSEval-英语				InstructTTSEval-中文			
	APS	**DSD**	**RP**	**Avg.**	**APS**	**DSD**	**RP**	**Avg.**
真实值	96.2	89.4	67.2	84.3	90.9	86.7	69.8	82.5
闭源								
Gemini-flash	92.3	93.8	80.1	88.7	88.2	90.9	77.3	85.4
Gemini-pro	87.6	86.0	67.2	80.3	89.0	90.1	75.5	84.8
GPT-4o-mini-TTS	76.4	74.3	54.8	68.5	54.9	52.3	46.0	51.1
Hume	83.0	75.3	54.3	71.1	-	-	-	-
开源								
ParlerTTS	60.0	45.9	31.2	45.7	-	-	-	-
PromptStyle	57.4	46.4	30.9	38.2	-	-	-	-
VoxInstruct	54.9	57.0	39.3	50.4	47.5	52.3	42.6	47.5
PromptTTS	64.3	47.2	31.4	47.6	-	-	-	-
MiMo-Audio-7B-Instruct	80.6	77.6	59.5	72.6	**75.7**	**74.3**	61.5	70.5
FlexiVoice-Base	63.6	75.0	60.6	66.4	56.7	59.1	59.5	58.4
FlexiVoice	**81.2**	**85.2**	**71.4**	**79.3**	71.0	71.8	**69.7**	**70.8**

象与既有研究结果一致（Li 等，2023）：标准自动语音识别模型在面对表现力强或富有情绪的语音时，往往会因显著的韵律变化而出现性能下降。然而，这并不意味着人类听众感受到的可懂度降低。正如 Q-MOS 分数始终较高所表明的那样（例如，EN-TO-Easy 中为 4.08，而对比值为 3.72），FlexiVoice 在呈现更丰富的风格表达时，仍保持了更高的感知清晰度。此外，正向的 CMOS（最高达 +0.9）表明其整体表现力和指令遵循能力更佳，同时确保生成语音的质量不受影响。相比之下，基线方法通常获得负向 CMOS，这表明与真实值相比，人类判断对其偏好较低。这些结果证实，FlexiVoice 不仅实现了更优的解耦，还保持了接近真实标注的自然度和质量。

5.3 复杂指令遵循能力

InstructTTSEval（Huang 等，2025）定义了复杂指令遵循的三个层级。这些任务从低层级的声学控制到开放式风格生成和角色模仿，旨在测试指令遵循的不同方面。如表 4 所示，预训练模型 FlexiVoice-Base 已展现出较强的竞争力，尤其是在 DSD（75.0 英语）和 RP（60.6 英语）上表现突出，这得益于这些任务与我们构建的数据集中自然指令之间的契合。采用渐进式后训练策略后，FlexiVoice 在所有任务类型上均取得稳定提升。在英语任务上，其 APS 得分提高了 12 分以上（81.2 对比 63.6），RP 得分提高了近 11 分（71.4 对比 60.6），总体平均分达到 79.3，接近 Gemini-pro（80.3）。在中文任务上，FlexiVoice 的平均准确率达到 70.8，超过 MiMo-Audio-7B-Instruct（70.5），并缩小了与 Gemini 模型之间的差距。总体而言，FlexiVoice 的表现始终优于所有开源基线方法，并缩小了与闭源商业系统之间的差距，展现出在复杂真实场景中的稳健指令遵循能力。此外，附录 A.8 中的实证结果证实，FlexiVoice 在语速和音高等非情感属性的细粒度控制方面同样表现出色，在相关性分析中显著优于基线方法。

5.4 渐进式后训练的有效性

为严格验证所提出的渐进式后训练（PPT）方案的必要性，我们开展了全面的消融研究，对比不同的训练顺序和优化策略。结果汇总于表 5。

训练顺序的重要性 我们首先探究各阶段的顺序是否重要。如表 5 的第一部分所示，直接从复杂指令对齐（+阶段三）开始会导致性能不佳（InstructTTSEval 平均值为 72.3，而我们的方法为 79.3）。这表明，若缺少阶段一所提供的基础对齐，模型便难以针对阶段三中稀疏且高层次的奖励进行优化。此外，另一种顺序 +S3→S1→S2 在

表 5：FlexiVoice 在不同训练顺序和策略下的消融研究。我们比较了预训练模型（基础模型）、从指令 GRPO 开始的效果（先进行阶段三）、联合训练，以及我们提出的渐进式后训练（PPT）策略（S1 → S 2 → S3）。

训练策略	解耦集（英语）					InstructTTSEval（英语）			
	TO Easy	TO Hard	TR Easy	TR Hard	Avg.	APS	DSD	RP	Avg.
FlexiVoice-Base	72.4	39.4	58.8	48.8	54.9	63.6	75.0	60.6	66.4
替代训练顺序									
+ S3	72.8	59.0	48.6	38.2	54.7	73.3	78.8	64.7	72.3
+ S3 →阶段一	96.0	79.6	80.8	72.6	82.3	74.3	81.2	67.5	74.3
+ S3 →S1 → 阶段二	94.8	80.2	86.8	75.8	84.4	74.9	81.4	68.2	74.8
联合训练									
+S1 →S2 +阶段三（联合）	94.8	80.8	86.2	74.6	84.1	78.3	80.5	67.8	75.5
渐进式后训练（*PPT*）									
+ S1	96.0	81.4	83.3	72.2	83.2	63.4	79.3	64.2	69.0
+ S1 →阶段二	96.4	88.0	89.4	**80.2**	88.5	66.8	80.3	67.9	71.7
+S1 → S2 →阶段三（我们的方法）	**97.7**	**89.4**	**89.4**	78.2	**88.7**	**81.2**	**85.2**	**71.4**	**79.3**

复杂指令上的表现（74.8）明显低于我们的方法。这表明，在复杂指令训练之后再进行基础的直接偏好优化（阶段一），会导致模型灾难性遗忘在阶段三中习得的细粒度指令遵循能力。这些发现证实，阶段一是必要的"冷启动"基础，可在模型着手处理更高级的解耦和泛化任务之前，建立稳健的多模态响应能力。

渐进式训练与联合训练 另一个关键问题是，阶段二和阶段三是否可以同时优化，因为二者都采用 GRPO 算法。我们将渐进式方法与联合训练策略（+S1→S2+阶段三）进行比较，在该策略中，解耦目标与指令目标相结合。结果表明，联合训练在两项基准测试中的表现均逊于渐进式策略（解耦平均值为 84.1，对比 88.7；InstructTTSEval 平均值为 75.5，对比 79.3）。我们认为，这源于梯度之间的冲突：阶段二利用固定分类器施加严格约束，以抑制风格泄漏；而阶段三则通过基于 ALM 的奖励鼓励开放式风格泛化。同时优化这两者会造成干扰，使模型无法同时精通任一任务。

PPT 策略的累积增益 表 5 的最后一部分展示了我们提出的优化路径循序渐进的有效性。多模态直接偏好优化（+阶段一）首先显著提升了基础可控性（解耦平均值提高 +28.3）。随后，解耦 GRPO（+阶段二）显著增强了模型将风格与内容、音色分离的能力，使解耦平均值升至 88.5。最后，指令 GRPO（+阶段三）进一步提升了模型处理复杂真实场景指令的能力（InstructTTSEval 平均值提高 +7.6），同时保持了上一阶段所实现的高解耦能力。完整的 PPT 课程学习在所有指标上实现了最佳平衡，相较于基础模型取得了显著提升，也验证了渐进式设计是一条稳定且有效的优化路径。

6 结论

本文介绍 FlexiVoice，这是一种支持多模态控制的文本转语音系统，使用自然语言指令控制风格，并通过参考语音控制音色；该系统以高质量的指令—语音数据集和渐进式后训练（PPT）范式为支撑。PPT 首先通过以情绪为中心的直接偏好优化/GRPO，增强多模态可控性，并强化指令、文本和参考语音之间的解耦；随后借助基于 ALM 的奖励，扩展复杂指令遵循能力，从而实现稳定优化和广泛泛化。实验表明，该系统在解耦方面持续取得提升（例如，ACC-I 大幅提高，且文本/参考语音带来的干扰较小），并在 InstructTTSEval 上表现出色：FlexiVoice 超越了已有基线方法，缩小了与闭源系统之间的差距，同时在人工评估中保持了自然度和鲁棒性。

7 伦理声明

本研究探讨了基于多模态输入的可控文本转语音。指令—语音数据集由公开或经授权的资源构建，并经过筛选，以剔除冒犯性、带有偏见或有害的内容。研究未收集或处理任何个人或敏感用户数据。尽管语音合成技术的进步可能被滥用（例如生成具有欺骗性或有害的音频），本研究旨在为研究目的提升可控性、鲁棒性和透明度，并倡导负责任地使用所提出的方法。

8 可复现性声明

我们采取了多项措施，以确保本研究可复现。正文和附录中详细描述了模型架构、训练目标和评估方案。我们将发布指令—语音数据集、模型检查点以及所有训练和推理代码，以便他人复现并开展进一步研究。发布材料还将包括超参数设置、数据处理流程和评估脚本。

参考文献

Marah Abdin、Jyoti Aneja、Hany Awadalla、Ahmed Awadallah、Ammar Ahmad Awan、Nguyen Bach、Amit Bahree、Arash Bakhtiari、Jianmin Bao、Harkirat Behl、Alon Benhaim、Misha Bilenko、Johan Bjorck、S´ebastien Bubeck、Martin Cai、Qin Cai、Vishrav Chaudhary、Dong Chen、Dong- dong Chen、Weizhu Chen、Yen-Chun Chen、Yi-Ling Chen、Hao Cheng、Parul Chopra、Xiyang Dai、Matthew Dixon、Ronen Eldan、Victor Fragoso、Jianfeng Gao、Mei Gao、Min Gao、Amit Garg、Allie Del Giorno、Abhishek Goswami、Suriya Gunasekar、Emman Haider、Junheng Hao、Russell J. Hewett、Wenxiang Hu、Jamie Huynh、Dan Iter、Sam Ade Jacobs、Mojan Javaheripi、Xin Jin、Nikos Karampatziakis、Piero Kauffmann、Mahoud Khademi、Dongwoo Kim、Young Jin Kim、Lev Kurilenko、James R. Lee、Yin Tat Lee、Yuanzhi Li、Yunsheng Li、Chen Liang、Lars Liden、Xihui Lin、Zeqi Lin、Ce Liu、Liyuan Liu、Mengchen Liu、Weishung Liu、Xiaodong Liu、Chong Luo、Piyush Madan、Ali Mahmoudzadeh、David Majercak、Matt Mazzola、Caio C´esar Teodoro Mendes、ArindamMitra、HardikModi、Anh Nguyen、Brandon Norick、Barun Patra、Daniel Perez- Becker、Thomas Portet、Reid Pryzant、Heyang Qin、Marko Radmilac、Liliang Ren、Gustavo de Rosa、Corby Rosset、Sambudha Roy、Olatunji Ruwase、Olli Saarikivi、Amin Saied、Adil Salim、Michael Santacroce、Shital Shah、Ning Shang、Hiteshi Sharma、Yelong Shen、Swadheen Shukla、Xia Song、Masahiro Tanaka、Andrea Tupini、Praneetha Vaddamanu、Chunyu Wang、Guanhua Wang、Lijuan Wang、Shuohang Wang、Xin Wang、Yu Wang、Rachel Ward、Wen Wen、Philipp Witte、Haiping Wu、Xiaoxia Wu、Michael Wyatt、Bin Xiao、Can Xu、Jiahang Xu、Weijian Xu、Ji- long Xue、Sonali Yadav、Fan Yang、Jianwei Yang、Yifan Yang、Ziyi Yang、Donghan Yu、Lu Yuan、Chenruidong Zhang、Cyril Zhang、Jianwen Zhang、Li Lyna Zhang、Yi Zhang、Yue Zhang、Yunan Zhang 和 Xiren Zhou。Phi-3 技术报告：一款可在手机本地运行的高性能语言模型，2024年。网址：https://arxiv.org/abs/2404.14219。

Yoshua Bengio、J´erˆome Louradour、Ronan Collobert 和 Jason Weston。课程学习。载于第*26*届国际机器学习年会，第 41–48 页，2009年。

Carlos Busso、Murtaza Bulut、Chi-Chun Lee、Abe Kazemzadeh、Emily Mower、Samuel Kim、Jean- nette NChang、Sungbok Lee 和 Shrikanth S Narayanan。Iemocap：交互式情感双人动作捕捉数据库。语言资源与评测，42(4)：335–359，2008年。

Houwei Cao、David G Cooper、Michael K Keutmann、Ruben C Gur、Ani Nenkova 和 Ragini Verma。Crema-d：众包情感多模态演员数据集。*IEEE* 情感计算汇刊，5(4)：377–390，2014年。

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* 信号处理选题期刊，16(6):1505– 1518，2022年。

Yushen Chen、Zhikang Niu、Ziyang Ma、Keqi Deng、Chunhui Wang、Jian Zhao、Kai Yu 和 Xie Chen。F5-tts：一位通过流匹配生成流畅且忠实语音的童话讲述者。*arXiv* 预印本 *arXiv:2410.06885*，2024年。

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。

Anuj Diwan、Zhisheng Zheng、David Harwath 和 Eunsol Choi。扩展丰富的风格提示文本转语音数据集，2025年。网址：https://arxiv.org/abs/2503.04713。

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。

Xiaoxue Gao、Chen Zhang、Yiming Chen、Huayun Zhang 和 Nancy F Chen。Emo-DPO：通过直接偏好优化实现可控的情感语音合成。载于 *ICASSP 2025-2025*年*IEEE*声学、语音与信号处理国际会议（*ICASSP*），第 1–5 页。IEEE，2025年。

Zhifang Guo、Yichong Leng、Yihan Wu、Sheng Zhao 和 Xu Tan。PromptTTS：通过文本描述实现可控的文本转语音。载于 *ICASSP 2023-2023*年*IEEE*声学、语音与信号处理国际会议（*ICASSP*），第 1–5 页。IEEE，2023年。

Haorui He、Zengqiang Shang、Chaoren Wang、Xuyuan Li、Yicheng Gu、Hua Hua、Liwei Liu、Chen Yang、Jiaqi Li、Peiyang Shi 等。Emilia：用于大规模语音生成的广泛、多语言且多样化的语音数据集。载于 *2024*年*IEEE*口语语言技术研讨会（*SLT*），第 885–890 页。IEEE，2024年。

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的基准测试。*arXiv* 预印本 *arXiv:2506.16381*，2025年。

Yiqiao Huang、Yuancheng Wang、Jiaqi Li、Haotian Guo、Haorui He、Shunsi Zhang 和 Zhizheng Wu。Debatts：零样本文本转语音合成辩论。*arXiv* 预印本 *arXiv:2411.06540*，2024年。

Shengpeng Ji、Jialong Zuo、Minghui Fang、Ziyue Jiang、Feiyang Chen、Xinyu Duan、Baoxing Huai 和 Zhou Zhao。Textrolspeech：一种用于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 *ICASSP 2024-2024 IEEE*声学、语音与信号处理国际会议 *(ICASSP)*，第 10301–10305 页。IEEE，2024a。

Shengpeng Ji、Jialong Zuo、Wen Wang、Minghui Fang、Siqi Zheng、Qian Chen、Ziyue Jiang、Hai Huang、Zehan Wang、Xize Cheng等。ControlSpeech：借助解耦编解码器实现零样本说话人克隆与零样本语言风格控制的同步进行。*arXiv*预印本 *arXiv:2406.01205*，2024b。

Zeyu Jin、Jia Jia、Qixin Wang、Kehan Li、Shuoyi Zhou、Songtao Zhou、Xiaoyu Qin和Zhiyong Wu。SpeechCraft：带有自然语言描述的细粒度富有表现力的语音数据集。在第*32*届*ACM*多媒体国际会议论文集中，第1255–1264页，2024年。

Jiaqi Li、Xiaolong Lin、Zhekai Li、Shixi Huang、Yuancheng Wang、Chaoren Wang、Zhenpeng Zhan和Zhizheng Wu。DualCodec：用于语音生成的低帧率语义增强神经音频编解码器。*arXiv*预印本 *arXiv:2505.13000*，2025年。

Yuanchao Li、Zeyu Zhao、Ondrej Klejch、Peter Bell 和 Catherine Lai。自动语音识别与情感语音：对语音与情感识别相互影响的词级研究。载于 *Interspeech2023*，第 1449–1453 页。国际语音通信协会，2023年。

Huan Liao、Qinke Ni、Yuancheng Wang、Yiheng Lu、Haoyue Zhan、Pengyuan Xie、Qiang Zhang 和 Zhizheng Wu。Nvspeech：一种通过副语言发声实现类人语音建模的集成式可扩展流程。*arXiv* 预印本 *arXiv:2508.04195*，2025年。

Yaron Lipman、Ricky TQ Chen、Heli Ben-Hamu、Maximilian Nickel 和 Matt Le。用于生成建模的流匹配。载于 第*11*届国际学习表征会议（*ICLR 2023*），2023年。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*，2024a。Guanghou Liu、Yongmao Zhang、Yi Lei、Yunlin Chen、Rui Wang、Lei Xie 和 Zhifei Li。Prompt-style：使用自然语言描述实现可控的文本转语音风格迁移。载于 *Interspeech 2023* 论文集，第4888–4892页，2023年。Rui Liu、Yifan Hu、Yi Ren、Xiang Yin 和 Haizhou Li。生成式富有表现力的对话语音合成。载于第*32*届 *ACM*多媒体国际会议论文集，第4187–4196页，2024b。Steven R Livingstone 和 Frank A Russo。Ryerson情感语音与歌曲视听数据库（RAVDESS）：一个关于北美英语中面部和声音表达的动态多模态数据集。*PLOS ONE*，13(5):e0196391，2018年。R. Lotfian 和 C. Busso。通过从现有播客录音中检索情感语音，构建自然的情感均衡语音语料库。*IEEE* 情感计算汇刊，10(4):471–483，2019年10月至12月。doi：10.1109/TAFFC.2017.2736999。Dan Lyth 和 Simon King。利用合成标注进行高保真文本转语音的自然语言引导，2024年。Ziyang Ma、Zhisheng Zheng、Jiaxin Ye、Jinchao Li、Zhifu Gao、ShiLiang Zhang 和 Xie Chen。emotion2vec：用于语音情感表征的自监督预训练。载于 计算语言学协会 *ACL 2024* 论文集，第15747–15760页，2024年。Kari Ali Noriy、Xiaosong Yang 和 Jian Jun Zhang。EMNS/IMZ/Corpus：用于游戏、电视和图像小说叙事讲述的情感单说话人数据集。*arXiv* 预印本 *arXiv:2305.13137*，2023年。Soujanya Poria、Devamanyu Hazarika、Navonil Majumder、Gautam Naik、Erik Cambria 和 Rada Mihalcea。MELD：用于对话情感识别的多模态多方数据集。载于第*57*届计算语言学协会年会论文集，第527–536页，2019年。Rafael Rafailov、Archit Sharma、Eric Mitchell、Christopher D Manning、Stefano Ermon 和 Chelsea Finn。直接偏好优化：你的语言模型其实是一个奖励模型。神经信息处理系统进展，36:53728–53741，2023年。Hubert Siuzdak。Vocos：弥合时域与基于傅里叶变换的神经声码器之间的差距，实现高质量音频合成。*arXiv* 预印本 *arXiv:2306.00814*，2023年。Zhiyuan Tang、Dong Wang、Yanguang Xu、Jianwei Sun、Xiaoning Lei、Shuaijiang Zhao、Cheng Wen、Xingjun Tan、Chuandong Xie、Shuran Zhou、Rui Yan、Chenjia Lv、Yang Han、Wei Zou 和 Xiangang Li。KeSpeech：普通话及其八种次方言的开源语音数据集。载于第三十五届神经信息处理系统会议数据集与基准赛道（第二轮），2021年。Fengping Tian、Chenyang Lyu、Xuanfan Ni、Haoqin Sun、Qingjuan Li、Zhiqiang Qian、Haijun Li、Longyue Wang、Zhao Xu、Weihua Luo 等。Marco-Voice 技术报告。*arXiv* 预印本 *arXiv:2508.02038*，2025年。Apoorv Vyas、Bowen Shi、Matthew Le、Andros Tjandra、Yi-Chiao Wu、Baishan Guo、Jiemin Zhang、Xinyue Zhang、Robert Adkins、William Ngan 等。Audiobox：使用自然语言提示实现统一的音频生成。*arXiv* 预印本 *arXiv:2312.15821*，2023年。Hui Wang、Siqi Zheng、Yafeng Chen、Luyao Cheng 和 Qian Chen。Cam++：一种使用上下文感知掩蔽的快速高效说话人验证网络。载于 *INTERSPEECH*，2023年。

Kaisiyuan Wang、Qianyi Wu、Linsen Song、Zhuoqian Yang、Wayne Wu、Chen Qian、Ran He、Yu Qiao 和 Chen Change Loy。Mead：用于情感说话人脸生成的大规模音视数据集。载于*ECCV*，2020年。

Yuancheng Wang、Haoyue Zhan、Liwei Liu、Ruihong Zeng、Haotian Guo、Jiachen Zheng、Qiang Zhang、Xueyao Zhang、Shunsi Zhang 和 Zhizheng Wu。Maskgct：采用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。

Xueyao Zhang、Yuancheng Wang、Chaoren Wang、Ziniu Li、Zhuo Chen 和 Zhizheng Wu。通过偏好对齐提升多样化领域中的零样本文本转语音可懂度。*arXiv* 预印本 *arXiv:2505.04113*，2025年a。

Xueyao Zhang、Junan Zhang、Yuancheng Wang、Chaoren Wang、Yuanzhe Chen、Dongya Jia、Zhuo Chen 和 Zhizheng Wu。Vevo2：通过统一的韵律学习，打通可控语音与歌声生成。*arXiv* 预印本 *arXiv:2508.16332*，2025年b。

Xueyao Zhang、Xiaohui Zhang、Kainan Peng、Zhenyu Tang、Vimal Manohar、Yingru Liu、Jeff Hwang、Dangna Li、Yuhao Wang、Julian Chan 等。Vevo：通过自监督解耦实现可控的零样本声音模仿。*arXiv* 预印本 *arXiv:2502.07243*，2025年c。

Guanlong Zhao、Sinem Sonsaat、Alif Silpachai、Ivana Lucic、Evgeny Chukharev-Hudilainen、John Levis 和 Ricardo Gutierrez-Osuna。L2-ARCTIC：非英语母语者语音语料库。载于*Interspeech*论文集，第2783–2787页，2018年。doi：10.21437/Interspeech.2018-1110。网址：http://dx.doi.org/10.21437/Interspeech.2018-1110。

Jinming Zhao、Tenggan Zhang、Jingwen Hu、Yuchen Liu、Qin Jin、Xinchao Wang 和 Haizhou Li。M3ED：多模态多场景多标签情感对话数据库。载于 第*60*届计算语言学协会年会论文集（第*1*卷：长篇论文），第5699–5710页，2022年。

Jiaming Zhou、Shiyao Wang、Shiwan Zhao、Jiabei He、Haoqin Sun、Hui Wang、Cheng Liu、Aobo Kong、Yujie Guo 和 Yong Qin。ChildMandarin：面向3至5岁幼儿的综合普通话语音数据集。*arXiv*预印本 *arXiv:2409.18584*，2024a。

Kun Zhou、Berrak Sisman、Rui Liu 和 Haizhou Li。基于新情感语音数据集的语音转换中已见与未见情感风格迁移。载于 *ICASSP 2021—2021 IEEE*声学、语音与信号处理国际会议（*ICASSP*），第920–924页。IEEE，2021年。

Siyi Zhou、Yiquan Zhou、Yi He、Xun Zhou、Jinchao Wang、Wei Deng 和Jingchen Shu。IndexTTS2：情感表现力强、时长可控的自回归零样本文本转语音领域的一项突破。*arXiv*预印本 *arXiv:2506.21619*，2025年。

Yixuan Zhou、Xiaoyu Qin、Zeyu Jin、Shuoyi Zhou、Shun Lei、Songtao Zhou、Zhiyong Wu 和 Jia Jia。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。载于第*32*届 *ACM*多媒体国际会议论文集，第554–563页，2024b。

A 附录

A.1 模型结构

如第 3 节所示，我们的模型主要包含两个阶段：自回归大语言模型和流匹配。在第一阶段，模型接收文本、指令和参考语音作为输入。文本和指令按照大语言模型的输入模板进行格式化，并将参考语音转录文本拼接在文本之前。我们使用从 Dual- Codec（Li 等，2025年）中提取的语义编码，以离散形式表示系统中的语音。因此，参考语音首先会通过 DualCodec 转换为离散词元，然后将格式化后的文本和指令经大语言模型编码器处理所得的词元拼接到前面。它们共同作为自回归大语言模型的输入。考虑到 Phi- 3.5-mini-instruct（Abdin 等，2024年）适用于多模态任务，我们在此采用该模型。我们首先扩展

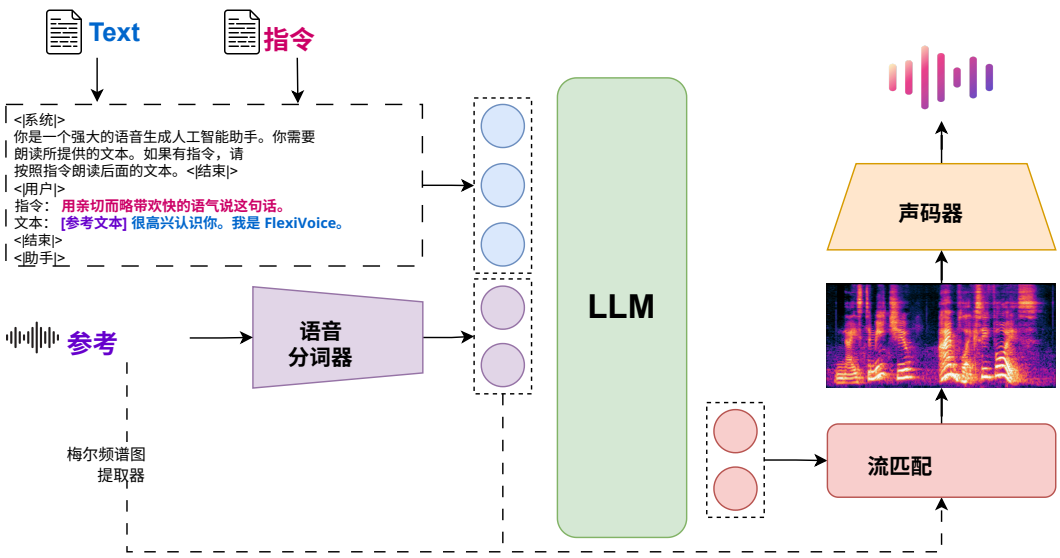

图 3：FlexiVoice 的完整结构。

其词汇量（与 DualCodec 的词汇量相同，为 16384），然后使用文本大语言模型的参数作为预训练和后训练的初始状态，如本文工作主体部分所示。

主流文本转语音工作（Du 等，2024；Zhou 等，2025；Zhang 等，2025年a）选择流匹配（Lipman 等，2023）作为解码器，因为它能够高质量地重建语音细节。在本文的第二阶段，我们采用在 Emilia（He 等，2024）上训练的流匹配模块，以参考语音代码为条件，将生成的代码转换为梅尔频谱。最后，通过声码器（本例中使用 Vocos（Siuzdak，2023））将梅尔频谱转换为目标音频。模型结构如图 3 所示。

A.2 预训练的数据处理

表6：预训练阶段使用的指令语音数据集。

来源	时长（小时）	描述
ParaSpeechCaps（Diwan 等，2025）	2847	开源的风格提示语音数据
ChildMandarin（Zhou 等人，2024a）	41	儿童语音片段
Debatts（Huang 等人，2024年）	67	带有辩论风格的语音片段
情绪数据集*	117	带有不同情绪的语音片段
KeSpeech（Tang 等，2021）	1541	具有不同方言、年龄和性别的语音片段
L2-ARCTIC（Zhao 等，2018）	27	非母语英语语音语料库
NVSpeech（Liao 等，2025年）	775	带副语言标签的语音数据
FlexiVoice-Instruct	4316	带有高级指令的富有表现力的语音

* 包括 CREMA-D（Cao 等人，2014年）、EMNS（Noriy 等人，2023年）、ESD（Zhou 等人，2021年）、IEMOCAP（Busso 等人，2008年）、M3ED（Zhao 等人，2022年）、MELD（Poria 等人，2019年）、MSP-Podcast（Lotfian 与 Busso，2019年）和 RAVDESS（Livingstone 与 Russo，2018年）。

对于 ParaSpeechCaps（Diwan 等，2025），由于其自底向上的数据标注流程，部分语音条目同时包含详细的声学特征词典和摘要描述。对于这些数据点，我们将特征词典拼接起来，并随机采样描述。对于其他数据，我们将其描述用作指令。

对于 ChildMandarin（Zhou 等，2024a）、Detatts（Huang 等，2024）和情绪数据集，由于它们都是单标签数据集（男孩、辩论场景、情绪标签），我们预先生成了若干条指令

模板，并使用 Deepseek-V3（Liu 等，2024a）生成模板内容（例如：“用 {标签}所指的声音说话”、“想象自己身处辩论场景中”、“表达 {标签}所指的情绪”）。然后，对于每个语音数据点，我们随机选择一个模板，并填入相应的标签。
KeSpeech（Tang 等，2021）和 L2-ARCTIC（Zhao 等，2018）都是多标签数据集。例如，KeSpeech 中的每段语音都有三个属性：年龄、性别和口音。对于整个数据集中的每个属性三元组，我们使用 Deepseek-V3（Liu 等，2024a）生成三条中文指令和三条英文指令，然后将这些指令随机应用于整个数据集。

对于带有副语言标签的语音语料库 NVSpeech，我们不应用任何指令，而是使用其标签来扩充核心大语言模型模块的词汇量，以确保能够进行富有表现力的副语言生成。预训练阶段的所有指令-语音配对数据集均列于表6。

A.3 指令数据构建中的提示词

Emilia 的数据来源

角色与任务
你是一位多语言文本分析专家，负责为用户的指令式文本转语音任务生成所需的语音描述。你的核心职责是：**根据文本和元信息（包括但不限于场景、风格、情绪等）生成语音描述**

详细描述 -
元信息：包括视频标题以及音频来源视频的标签列表。如果标题无关（例如网址、乱码等），请忽略标题。

- 可以根据标签和标题推断场景和说话风格，但若两者与文本内容冲突，应优先以文本内容为准。

- 用自然的人类语言指令简洁描述语音。- 使用口语化、生动的表达方式。变换句式，避免写得像套用模板。

- 根据文本和元信息，判断能否生成有效或丰富的语音描述，并对信息价值（高／中／低）进行分类。

输入／输出规范I
nput Struct
ure:{"语音转录文本": "要分析的语音文本", "元信息": {"标题": [源视频的标题] ,"标签": [源视频的标签列表] }}输出要求：{"描述": "对语音风格的自然语言描述", "价值": "信息价值（高／中／低）"}

示例
示例输入：{"语音转录文本": "这场比赛目前显示马超拿下四次击杀，金币接近6000。", "元信息": {"标题": "《王者荣耀》KPL秋季赛总决赛", "标签": "游戏，电子竞技，《王者荣耀》，LGD，KPL，DYG"}}示例输出：{"描述": "专业且富有激情的电子竞技解说，表达激动而热烈。", "价值": "高"}

核心约束
- 元信息与文本内容冲突时，以文本内容为准。- 禁止毫无根据的推测；禁止描述任何不确定的细节。- 示例输出仅供参考；不要刻板地照搬其措辞。

游戏配音的数据来源

角色与任务
你是一名多语言文本分析专家，负责为用户的指令式文本转语音任务生成所需的语音描述。你的核心职责是：**根据文本和元信息（包括但不限于场景、风格、情绪等）生成语音描述**

详细描述 –
Main Info 指的是 [GameName]，中某个特定角色，其对话即为语音转录文本。- 根据角色名称推断其说话风格，并输出语音描述。- 当说话风格与文本表达的风格冲突时，以文本为准。- 描述中不得包含具体角色名称。- 使用口语化、生动的措辞。变换句式，避免显得像套用模板。- 根据文本和元信息判断能否生成有效或丰富的语音描述，并将信息价值分为高／中／低。

输入／输出规范 Input Structure:{"语音转录文本": "语音转录文本（来自 [GameName]角色的对话）", "元信息": "来源信息（[GameName]中的角色名称）"}输出要求：{"描述": "对语音风格的自然语言描述", "值": "信息价值（高/中/低）"}

示例
示例输入：{"语音转录文本": "只要内容有趣，这些专家就能帮忙增添趣味。", "元信息": "八重神子"}示例输出：{"描述": "她的语气中带着一丝俏皮的优雅，仿佛早已看透对方的心思，却选择含蓄地暗示出来。", "值": "高"}

核心约束
- 当元信息与文本冲突时，以文本内容为准。- 禁止毫无根据的猜测；禁止描述任何不确定的细节。- 示例输出仅供参考；不要刻板地照搬其措辞。

A.4 GRPO 训练的数据构建

解耦任务训练集由 NCSSD（Liu 等，2024b）的录音子集构成，包含约 20,000 条中文和 10,000 条英语对话语音样本。我们选择该数据集，主要是因为其转录文本与日常对话十分接近，而且在大多数情况下隐含着情绪倾向。我们为每个数据点随机分配情绪标签，从而生成两类数据：指令与转录文本中的情绪一致的数据，以及二者情绪相冲突的数据。这种方法有助于引导模型学习在文本中解耦内容与风格的能力。同时，我们以 90% 和 10% 的概率，分别从预训练数据的情绪数据集（6）中为每个数据点随机分配中性和情感语音作为参考语音。这使模型能够学习从参考语音中解耦音色与风格。我们使用一半的中文数据和全部英语数据，构成约 20,000 个均衡的数据点，作为解耦阶段的 GRPO 训练集。

复杂指令 为了提升模型在复杂指令遵循任务中的表现，我们构建了丰富多样的 GRPO 训练数据集。具体而言，对于每种语言（中文和英语），我们首先从

预训练数据中随机抽取 1,000 条现有的指令-文本输入，以确保优化稳定性。接下来，我们提示 Deepseek-V3（Liu 等，2024a）在三种不同配置下随机生成 6,000 条指令：（1）以词典格式生成详细且全面的声学特征描述；（2）生成包含 3/4/5/6 个明确声学特征的自然语言描述；（3）根据人类使用模式，为任意场景自由生成指令。总共生成了 14,000 条指令-文本配对输入，作为第二阶段的 GRPO 训练数据。

A.5 情感控制 DPO 详情

指令模板

- 用 {label}情绪说出这个句子。
- 用 {label} 语气说这句话。
- 以 {label}方式说出这个句子。
- 用 {label}情绪表达这段文本。
- 使用 {label}声音。
- 用 {label} 感受表达这个句子。
- 用 {label}语调朗读。
- 用 {label} 情感朗诵。
- 以 {label}风格发声。
- {标签}（仅使用一个情绪标签）
- 用{标签}的情感说出这句话。
- 以{标签}的情绪表达这句话。
- 带着{标签}的情感说出这段文本。
- 使用{标签}的情感。
- 以{标签}的心情表达这句话。
- 用{标签}的情感色彩说出这句文本。
- 带着{标签}的感情说出这句话。
- 用{标签}的情感基调传达此句。
- 以{label}的情绪状态说出这句话。
- {label}

一组配对样本（情绪=愤怒，说话人 ID=0013）

{”提示文本”：”门口的第十一章。”，”提示音频路径”：”情感语音数据集/**0013**/中性/0013 000267.wav”，”目标文本”：”你还在等什么？伙计。”，”指令”：”用愤怒的语气说这句话。”，”选定音频路径”：”情感语音数据集/**0013**/愤怒/0013 000697.wav” ”拒绝音频路径”：”情感语音数据集/**0013**/悲伤/0013 001397.wav”}

A.6 解耦 GRPO 中的奖励选择

在解耦 GRPO 中，我们首先使用 SER 信号作为奖励之一，以控制情感合成是否正确。对于输入的参考语音，我们还需要确保生成结果与参考语音来自同一说话人。这里有两种选择：说话人验证信号（0 或 1）和说话人相似度信号（取值介于 0 和 1 之间），分别使用 CAM++ (Wang 等，2023) 和 WavLM-large-finetuned（Chen 等人，2022）。基于经过直接偏好优化的模型，我们使用 SER奖励和两种独立的说话人相关奖励训练多目标 GRPO，共训练四个训练轮次，然后在解耦评估集的两个 TR 难度级别上进行测试，如图 4 所示。

结果表明，将说话人验证作为奖励信号的模型取得了显著提升，而使用说话人相似度的模型性能基本不变，甚至有所下降。这是因为计算说话人嵌入的模型（此处为 WavLM）并非只依赖语音音色，还会利用音色之外的其他声学特征，例如音高和情绪。当人类以不同情绪说话时，音高、语速和音量等声学属性必然会有所不同。因此，富有表现力的情感表达会带来较高的 SER 奖励分数，却产生较低的说话人相似度信号。因此，这一现象可以解释为……之间的冲突

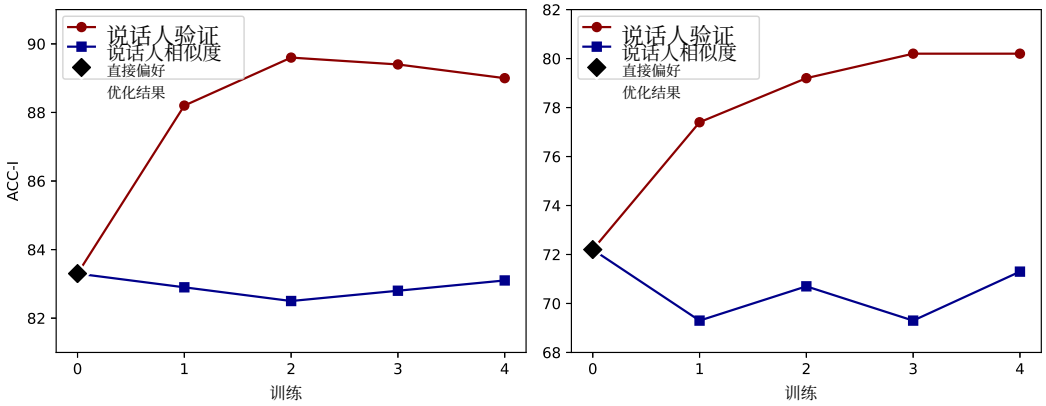

图 4：在解耦评估集的 TR-easy（左）和 TR-hard（右）任务上，解耦 GRPO 中两种奖励信号的结果对比。

SER奖励与说话人相似度奖励之间的冲突，阻碍了多目标 GRPO 任务的正常优化过程。

说话人验证信号是说话人相似度的一种较宽松形式，能够在确保说话人身份的同时，适应生成更具表现力的情感语音的模型。因此，我们选择说话人验证作为阶段一 GRPO 阶段的最终奖励信号之一。

A.7 奖励模型的验证

对于第3阶段（S3）的复杂指令遵循，我们需要一个可靠的奖励模型来评估开放式生成。尽管 Gemini-2.5-Pro 在 InstructTTSEval 基准测试中担任黄金标准评判器，且与人类偏好高度一致（Huang 等，2025），但由于在线采样期间推理成本高昂、延迟较高，直接将其用作 GRPO 的奖励模型并不现实。因此，我们提出使用 Kimi-Audio-7B-Instruct 作为经济高效的替代模型。

为验证这一替换方案的可靠性，我们评估了 Kimi-Audio-7B 与 Gemini-2.5-Pro 之间的一致性。我们在 InstructTTSEval数据集上进行评估，使用训练时相同的提示词，并通过 Macro-F1分数衡量两者判断的一致程度。如表 7 所示，在英语和中文的所有任务类型中，Kimi-Audio-7B 与 Gemini 均达到了较高的一致性（Macro-F1 > 0.60）。这些结果表明，Kimi-Audio-7B 能有效捕捉较大模型的偏好模式，为优化复杂指令遵循提供有效且高效的信号。

表 7：在 InstructTTSEval 上，我们的替代模型（Kimi-Audio-7B）与黄金标准评判器（Gemini-2.5-Pro）之间的一致性评估（Macro-F1）。较高的一致性验证了使用 Kimi-Audio 作为奖励模型的有效性。

指标	英语（EN）			中文（ZH）		
	APS	**DSD**	**RP**	**APS**	**DSD**	**RP**
Macro-F1	0.62	0.62	0.65	0.60	0.64	0.63

A.8 非情感风格控制评估

虽然我们的主要实验聚焦于情感表达和复杂指令遵循，但要证明能够灵活控制风格，就必须在情绪之外，对更细粒度的韵律属性进行验证。为此，我们针对**语速** 和**音高** 控制开展了专项评估。

表 8：语速和音高控制的斯皮尔曼相关系数评估。

模型	语速相关性		音高相关性	
	EN	ZH	EN	ZH
ParlerTTS	0.61	0.22	0.68	0.79
PromptTTS	0.24	-	0.12	-
PromptStyle	0.05	-	0.24	-
VoxInstruct	0.75	0.49	0.31	0.15
FlexiVoice-Base	0.62	0.47	0.66	0.52
FlexiVoice	**0.86**	**0.78**	**0.91**	**0.90**

我们构建了一个测试集，其中包含 100 个英语样本和 100 个中文样本。对于每个样本，我们根据对应于三个不同等级的指令合成语音：

- **语速：** 快，正常，慢（例如，指令：请快速朗读这句话。）
- **音高：** 高、正常、低 （例如，指令：用高音调的声音说话。）

为量化可控性，我们计算了指令级别（映射为 1、2、3）与生成音频中提取的声学特征之间的**斯皮尔曼相关系数** 。具体而言，我们使用音素时长（每秒音素数）衡量语速，使用基频中位数衡量音高。相关性越高，表明模型越能准确地遵循指令的梯度变化。

如表 8 所示，FlexiVoice 在两种语言和各项属性上均取得了最高的相关性得分。值得注意的是，它显著优于基线模型和预训练的 FlexiVoice-基础模型，表明我们的渐进式后训练（PPT）策略能够有效增强对细粒度物理参数的控制，以及对抽象情感风格的控制，即使我们没有针对语速或音高进行任何定向优化。

A.9 主观评估配置

主观评估由一组获得报酬的参与者完成，他们都具备扎实的语音与音频专业知识。我们选取了两个任务支持相对全面的基线模型、我们的预训练模型以及 FlexiVoice 进行主观评估。对于每种语言中的每项任务，我们从每个模型和真实标注中随机选取五组配对样本。为确保可靠性，每个音频样本均由至少三名不同人员分别进行主观评估。

Q-MOS 请参与者从清晰度、自然度以及是否存在失真或伪影等方面，对每个生成语音样本的整体质量进行五分制评分，但不考虑风格指令和音色参考。各分数的含义如下：

- **5** - 语音自然、清晰且悦耳。没有明显的伪影或失真，与专业录制的人类语音相当。
- **4–** 语音总体自然且清晰易懂，只有轻微瑕疵或偶尔出现的伪影，不会影响理解或聆听舒适度。
- **3–** 语音清晰可懂，但存在一些中等程度的问题，例如轻微失真、不自然的韵律或轻微背景噪声。质量尚可，但明显低于高标准的人类录音。
- **2 –** 由于存在明显的伪影、失真或不自然的表达方式，语音较难理解。质量问题显著影响聆听体验。
- **1 –**语音基本无法理解或极不自然，存在严重的伪影或失真，导致难以进行评估。

CMOS 在此任务中，要求参与者在给定目标情绪的条件下，对比两个音频样本（一个为真实标注样本，另一个由模型生成）。主要关注情感表达的丰富度和自然度，这反映了模型的指令遵循能力和解耦能力。

补充评估规则如下：（1）未提供参考语音时，忽略音色一致性。（2）提供参考语音时，音色相似度是次要标准（只要两个样本听起来像同一位说话人即可）。（3）应忽略轻微的发音错误或噪声；主要比较情绪的准确性和表现力。评分量表定义如下：（请注意，成对音频示例的顺序是随机的。）

- $+$**2** - 音频 B 在传达目标情绪方面明显优于音频 A。
- $+$**1** – 音频 B 在情感表达方面略优于音频 A。
- **0** – 两个样本在情感丰富度和自然度方面相当。
- $-$**1** – 音频 A 在情感表达方面略逊于音频 B。
- $-$**2** – 音频 A 在情感表达方面远不如音频 B。

A.10 可复现性细节

为了便于复现我们的结果并开展未来研究，我们提供了渐进式后训练（PPT）方案的计算资源和训练配置详细说明。FlexiVoice 基于 Phi-3.5-mini-instruct 骨干模型构建，参数量约为 38 亿。所有训练阶段均在 $8\times$ NVIDIA A800（80GB）GPU 上进行。后训练流程的总训练时间约为 3.5 天。

具体而言，S1（多模态 DPO）阶段作为高效的对齐基础，约 2 小时即可完成。我们以 1×10^{-5} 的学习率训练模型 3 个训练轮次，并使用 $\beta = 0.1$的 KL 惩罚系数。S2（解耦 GRPO）阶段通过在线采样来实现解耦，约需 36 小时。此阶段采用 $G = 8$的组大小，以 1×10^{-5} 的学习率训练 2 个训练轮次，并使用 $\beta = 0.1$。最后，S3（指令 GRPO）阶段采用 $G = 6$ 的组大小，以适应复杂指令遵循，并运行 2 个训练轮次。由于基于 ALM 的奖励计算较为复杂，此阶段约需 42 小时。