软件学报 ISSN 1000-9825, CODEN RUXUEW
2026,37(5):1903−1918 [doi: 10.13328/j.cnki.jos.007541] [CSTR: 32375.14.jos.007541]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

基于音频-语言模型的端到端说话人日志系统*

韦舒羽[1], 丘德来[2], 刘升平[2], 桑基韬[1]

[1](北京交通大学 计算机科学与技术学院, 北京 100044)

[2](云知声智能科技股份有限公司, 北京 100096)

通信作者: 桑基韬, E-mail: jtsang@bjtu.edu.cn

摘　要: 会议纪要、客服质检等应用对多说话人语音转写与归属判断的需求正日益增长. 随着近年来多模态大语言模型的迅速发展, 音频-语言模型因其能够同时理解音频信号与自然语言提示, 并在自回归解码框架中统一处理两种模态的能力, 天然契合这种“说话人日志”任务的需求, 为端到端多说话人音频转写提供了全新的思路. 提出一种基于音频-语言模型的端到端说话人日志系统, 通过两阶段训练策略实现语音识别能力与判断说话人归属能力的协同优化, 将音频-语言模型的能力泛化到具体的下游任务上. 训练的第 1 阶段采用监督微调 (SFT), 在标准交叉熵损失中引入“说话人损失”, 以加权的方式强化对稀疏说话人标签 token 的学习信号; 第 2 阶段使用了基于组相对策略优化 (GRPO) 算法的强化学习策略, 以联合指标 *cpCER* 与 *SA-CER* 设计奖励函数, 突破了监督学习的性能瓶颈. 在双说话人的场景下开展实验, 对比了热门开源工具 3D-Speaker、Diar Sortformer 和闭源的 AssemblyAI、Microsoft Azure 说话人日志 API, 并通过消融实验证明了训练方法的合理性, 随后将实验拓宽至四说话人场景. 结果表明, 两阶段的训练方法在双说话人环境中显著提升了模型的语音识别能力与判断说话人归属的能力, 而在四说话人场景中, 常规的监督微调已取得较大收益. 进一步讨论了大模型资源消耗、输入时长限制、跨域适应等问题, 提出了引入流式音频编码器、课程学习、拒绝采样策略等未来优化方向. 研究结果表明音频-语言模型在多说话人日志任务中具备显著潜力, 但亦需在复杂声学场景下完成更多技术突破.

关键词: 说话人日志; 音频-语言模型; 监督微调; 强化学习; GRPO 算法

中图法分类号: TP37

中文引用格式: 韦舒羽, 丘德来, 刘升平, 桑基韬. 基于音频-语言模型的端到端说话人日志系统. 软件学报, 2026, 37(5): 1903–1918. http://www.jos.org.cn/1000-9825/7541.htm

英文引用格式: Wei SY, Qiu DL, Liu SP, Sang JT. End-to-end Speaker Diarization System Based on Audio-language Model. Ruan Jian Xue Bao/Journal of Software, 2026, 37(5): 1903–1918 (in Chinese). http://www.jos.org.cn/1000-9825/7541.htm

End-to-end Speaker Diarization System Based on Audio-language Model

WEI Shu-Yu[1], QIU De-Lai[2], LIU Sheng-Ping[2], SANG Ji-Tao[1]

[1](School of Computer Science and Technology, Beijing Jiaotong University, Beijing 100044, China)

[2](Unisound AI Technology Co. Ltd., Beijing 100096, China)

Abstract: The demand for multi-speaker speech transcription and speaker attribution in applications such as meeting minutes and customer service quality inspection is increasing. Recent advances in multimodal large language models have given rise to audio-language models (ALMs) that can simultaneously interpret audio signals and natural-language prompts within a unified autoregressive decoding framework, making them a natural fit for the speaker diarization task and offering a fresh approach to end-to-end multi-speaker audio transcription. This study proposes an end-to-end speaker diarization system based on an ALM and achieves synergistic optimization of speech-

* 基金项目: 国家重点研发计划 (2023YFF1204100)
本文由“多媒体智能理解与生成”专题特约编辑孙立峰教授、闵巍庆副研究员、马占宇教授、蒋树强研究员、彭宇新教授、田丰研究员、黄庆明教授推荐.
收稿时间: 2025-05-25; 修改时间: 2025-07-11, 2025-08-30; 采用时间: 2025-09-05; jos 在线出版时间: 2025-09-23
CNKI 网络首发时间: 2025-12-11

recognition capability and speaker-attribution capability via a two-stage training strategy, thus generalizing the capability of ALMs to specific downstream tasks. In the first stage, supervised fine-tuning (SFT) introduces a “speaker loss” into the standard cross-entropy objective to weight and strengthen the learning signal for sparse speaker-label tokens. In the second stage, a reinforcement-learning scheme based on group relative policy optimization (GRPO) is employed, designing a reward function that jointly considers *cpCER* and *SA-CER* to break through the performance bottleneck of supervised learning. Experiments in a two-speaker setting compare with the open-source 3D-Speaker toolkit and the Diar Sortformer model, as well as the proprietary speaker diarization APIs from AssemblyAI and Microsoft Azure. Ablation studies are further conducted to validate the training methodology, and experiments are subsequently extended to a four-speaker scenario. Results demonstrate that the two-stage approach significantly improves both ASR and speaker-attribution performance in the two-speaker environment, whereas in the four-speaker setting, conventional SFT already yields substantial improvements. Challenges such as resource consumption, input-length limitations, and cross-domain adaptation are also discussed, and future enhancements are proposed, including streaming audio encoders, curriculum learning, and rejection-sampling strategies. Experimental results show that ALMs hold great promise for multi-speaker diarization tasks but require additional technical advances to handle more complex acoustic scenarios.

Key words: speaker diarization; audio-language model (ALM); supervised fine-tuning (SFT); reinforcement learning (RL); group relative policy optimization (GRPO) algorithm

随着人机交互和会议纪要等应用场景对语音转录需求的不断提高, 多说话人语音转写系统成为近年的研究热点. 说话人日志任务的目标是在包含多位说话者的音频中准确定位“谁在什么时间说了什么”. 该任务是经典的“鸡尾酒会问题”, 广泛应用于会议记录、实时字幕、多语种翻译等场景. 在技术上, 说话人日志要求将录音按照不同的说话人归属进行分段并建立索引, 即检测说话人边界并将同一说话人的片段聚合在一起, 同时确认说话人的归属. 与单纯的语音转录不同, 说话人日志需要输出带有说话人身份标记的文字, 形成带有说话人归属的转写结果.

如图 1 所示, 说话人日志系统接收一段包含多位说话人的音频作为输入, 输出文本形式的语音转录结果, 并附带说话人的标签, 指明每句话归属于哪位说话人. 为解决这类说话人日志任务, 系统不仅要完成高质量的语音转写, 还需准确区分每句话所属的说话人, 这对模型的声学建模与上下文理解能力提出了双重挑战. 传统的自动语音识别 (automatic speech recognition, ASR) 系统在单一说话人场景下已十分成熟, 但在多说话人的场景下, 环境噪声、重叠语音以及频繁的说话人切换等因素可能导致语音识别错误率和说话人归属错误率大幅上升.

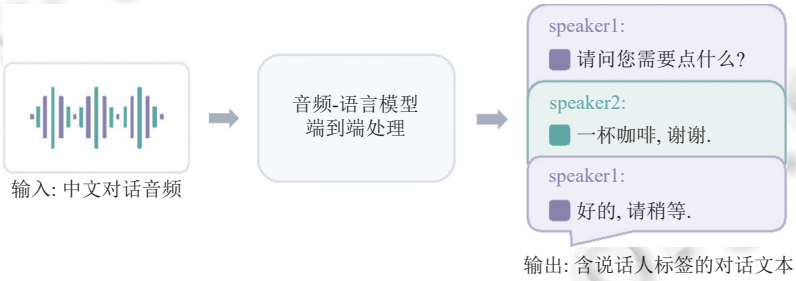

图 1　说话人日志任务示意图

近年来, 多模态大语言模型借助自回归解码架构, 将音频信号与自然语言提示联合建模, 为传统 ASR、翻译和问答等任务带来范式转变. 其中, 音频-语言模型 (audio-language model, ALM)[1]使用 MLP 层连接音频编码器与大语言模型, 赋予了大语言模型理解音频信息的能力, 代表性模型有 Qwen-Audio[2]、Step-Audio[3]等. 通过预训练阶段的监督学习进行模态对齐, 音频-语言模型实现了对语音、文本输入的统一映射. 经过少量指令式微调, 音频-语言模型即可迅速泛化至自动语音识别、文本翻译、语音问答、场景识别等多种任务, 具有优秀的下游任务泛化能力[4], 能够在零样本提示下完成语音转写、对话摘要与问答等多种任务.

音频-语言模型有着将多说话人音频直接映射成带有说话人标签的文本的潜力, 这避免了分离、聚类等模块化流水线的误差累积与工程复杂度; 得益于预训练阶段的大规模跨模态表示, 模型既能统一优化声学和语言信息, 又天然具备上下文感知与连贯生成能力; 更重要的是, 音频-语言模型不仅能高效完成说话人日志任务, 还可通过自然语言指令无缝扩展至摘要、问答等一系列后续处理, 一个模型便可同时兼容多项声学任务. 基于以上原因, 本研究提出基于音频-语言模型的说话人日志系统优化方案. 第 1 节的内容将更加详细地对比本研究相比现有方法

的差异与优势.

在音频-语言模型中, 阿里开源的 Qwen2-Audio-7B-Instruct 模型[5]因其强大的多模态理解能力和在多种音频任务上的优异表现, 成为本研究的理想基座. 然而, 我们发现, 尽管这类预训练模型具备强大的基础能力, 但其在特定、结构化任务上的指令遵循能力仍有不足. 例如, 当直接使用提示 (prompt) 引导其执行说话人日志任务时, 模型难以稳定地输出预设格式. 这表明了设计专门的微调策略的必要性, 从而将模型的通用能力“适配”到说话人日志这一具体的下游任务上.

因此, 本研究使用 Qwen2-Audio-7B-Instruct 模型作为基座模型, 通过设计阶段式的训练策略, 使模型适应说话人日志任务. 模型的训练使用音频-文本问答对数据, 将音频输入与说话人 prompt、文本输出统一建模为自回归模型的训练目标, 直接在对话级别进行联合训练, 这样就天然地兼顾了语音识别与说话人归属判别这两项能力需求. 本研究设计了两阶段的训练策略: 第 1 阶段使用监督学习 (supervised fine tuning, SFT), 在常规交叉熵损失的基础上额外引入了针对说话人标签的加权损失项, 在模型学习作答格式的同时强化模型的说话人分辨能力; 第 2 阶段则基于使用组相对策略优化 (group relative policy optimization, GRPO) 算法[6]的强化学习 (reinforcement learning, RL), 通过构造结合 *cpCER* (concatenated minimum permutation character error rate) 与 *SA-CER* (speaker attributed character error rate) 两个综合指标的奖励函数, 实现了对语音识别能力和说话人分辨能力的端到端协同优化, 突破 SFT 阶段的性能瓶颈. 本研究聚焦于中文多说话人场景, 首先在双说话人场景下验证了方法的有效性, 随后将场景扩展至四说话人.

综上所述, 本研究的主要贡献如下.

(1) 提出基于音频-语言模型的说话人日志系统. 本研究将音频-语言模型直接应用于说话人日志任务, 通过自回归解码器统一处理音频特征与文本输出, 实现了从原始音频到带有说话人标签的文字转录的端到端流程.

(2) 设计监督学习阶段的“说话人损失”加权策略. 针对说话人标签 token 在序列中占比稀少的问题, 本研究提出在监督学习阶段对该部分 logits 进行加权交叉熵优化, 有效地提升了模型对说话人归属的敏感度.

(3) 构建基于 GRPO 算法的强化学习优化方法. 在监督学习之后, 引入基于 GRPO 算法的强化学习阶段, 以 *cpCER* 和 *SA-CER* 为联合奖励, 直接对下游评测指标进行策略梯度优化, 突破了仅靠 SFT 难以逾越的性能瓶颈.

(4) 讨论模型的应用局限与改进方向. 本文分析了音频-语言模型在训练资源、输入时长限制与跨域泛化等方面的挑战, 并提出了可行的未来改进方向, 为后续在实际部署与更复杂场景下的进一步优化提供了参考.

1 相关工作

说话人日志任务的传统方法通常采用多步流水线: 先使用语音活动检测 (voice activity detection, VAD) 工具获得语音段落, 再提取说话人特征, 用聚类等方法将语音片段分组, 最后利用语音识别工具将语音转录为文本. 近年来, 深度学习极大地提升了嵌入提取和聚类的效果, 同时也催生了新的方法与思路. 本节将按照方法类型简单介绍说话人日志任务的代表性方法, 如表 1 所示.

表 1 说话人日志任务的现有方法梳理

方法类型	代表论文	方法特征	主要局限性
基于语音分离	von Neumann等人[7]、Morrone等人[8]、Gruttadauria等人[9]、Kalda等人[10]	先使用语音分离模型将混合语音分解到固定通道, 再分别处理各路信号	高度依赖分离质量; 通常需预设通道数; 对长时静默或重叠变化敏感
模块化流水线	Landini等人[11]、朱必松等人[12]、毛青青等人[13]	逐步执行VAD、分割、说话人嵌入提取、聚类、语音识别等模块	对重叠语音处理能力有限; 常需先验说话人数量; 各模块误差易累积
端到端SA-ASR	Kanda等人[14]、Li等人[15]、Kanda等人[16]	单模型并行输出文本与说话人标签, 无需后续聚类即可得出说话人划分	模型结构复杂、训练数据需求大; 对声学、语言标注一致性要求高
LLM后处理	Wang等人[17]、Efstathiadis等人[18]	将初步的转录文本和说话人标签输入LLM, 利用 LLM知识进行文本级纠错和优化	依赖模型初步输出的准确度; LLM推理开销大; 对不同系统的泛化性仍需验证

1.1 基于语音分离的方法

基于语音分离的方法先将混合的多说话人信号拆分为若干个无重叠的单说话人音轨, 然后分别对每一音轨应用常规的语音识别模型进行转录. 常用的分离模型包括 Conv-TasNet、dual-path RNN (DPRNN) 以及 TF-GridNet 等.

在 von Neumann 等人[7]提出的连续语音分离框架中, 混合语音被映射到预设数目的输出通道, 实现无重叠的单说话人流. Morrone 等人[8]提出的低延迟 SSGD 模型在电话通话上使用 DPRNN 进行实时分离, 然后对每个通道分别进行语音活动检测, 在 CALLHOME 数据集上达到了 11.1% 的 DER (diarization error rate). Gruttadauria 等人[9]提出了面向会议的在线分离-识别方案: 他们先用 Conv-TasNet 或 DPRNN 对会议录音产生 2–3 路分离信号, 再对每路信号进行语音活动检测, 并通过说话人嵌入进行递增聚类, 最终在 AMI 数据集上取得了新的最优效果. 另一个代表性工作 PixIT[10]联合训练了基于 PIT 的说话人分离和基于 MixIT 的无监督分离模型, 借助现有说话人聚类信息减少过分离, 提高了分离后的语音识别准确率.

总的来说, 基于语音分离的方法能有效处理重叠语音, 但高度依赖分离网络的质量, 一般需要预设输出通道数, 对快速轮转或长静默的说话人场景仍有挑战.

1.2 模块化流水线系统

模块化流水线始终是说话人日志系统的常用方法, 这类系统包括语音活动检测、声学分割、说话人特征提取、相似度评估或聚类、语音识别等多个阶段. 例如 Landini 等人[11]提出的 VBx 方法在 x-vector 序列上使用贝叶斯隐马尔可夫模型进行聚类, 在 CALLHOME、AMI、DIHARD 等基准测试中取得当时的最优结果; 近期, 朱必松等人[12]基于相邻语音片段的背景噪声往往相同这一现象, 提出一种新的基于时间分段和重组聚类的说话人日志方法; 毛青青等人[13]提出了一种基于图神经网络的方法, 优化了模块化流水线中“说话人嵌入提取”这一关键步骤, 通过结合局部和全局说话人信息并采用动态自适应的多原型学习模块, 在 AMI_SDM 和 CALLHOME 数据集上取得了显著优于基线系统的成绩.

流水线方法的优点是结构清晰、可针对各模块独立优化, 易于集成和升级; 但流水线系统的最终性能高度依赖于每个子模块与任务场景的匹配程度, 且模块间错误的累积同样会影响整个系统的性能. 不仅如此, 流水线系统对重叠语音的鲁棒性较弱, 而且一般需要提前知道或估计说话人数量, 否则可能需要额外后处理来合并聚类结果.

1.3 端到端说话人归属 ASR 系统

此方法使用区分说话人的语音识别 (speaker-attributed ASR, SA-ASR) 模型, 这类模型通过联合学习声学、语言与说话人建模, 端到端地输出带有说话人标签的转录文本, 即用一个统一模型同时输出转写文本和说话人标签.

Kanda 等人[14]提出的“Transcribe-to-Diarize”是典型的例子: 这个模型可以在模型内部状态中估计词语的起止时间, 从而实现对不限数量说话人的日志标注. Li 等人[15]提出的 SA-Paraformer 采用并行解码的 Paraformer 结构加速多说话人语音识别, 并引入“speaker-filling”策略和 inter-CTC 损失来减少说话人识别错误. 类似地, Kanda 等人[16]还提出了一种基于 Transformer 的端到端 SA-ASR, 在 LibriCSS 测试上实现了 11.9% 的 cpWER, 优于传统分离+聚类方案.

端到端 SA-ASR 的优点是可以联合训练语音识别和说话人归属任务, 但缺点是模型结构复杂、训练数据需求大, 对声学和语言标注的统一性要求高.

1.4 结合大语言模型的后处理

尽管上述方法在声学层面已经取得长足进展, 它们并没有在文本层面上进一步优化说话人归属和转写结果的准确性. 近期的部分研究聚焦于利用大语言模型对已有说话人日志结果进行文本级的后处理与纠错.

比如, Wang 等人[17]提出 DiarizationLM 框架, 将模型输出的音频转录文本和说话人标签作为文本提示输入微调后的 PaLM-2 模型, 生成优化后的带有说话人标签的转录文本. 实验结果表明微调后的大语言模型可以显著降低词级说话人错误率. Efstathiadis 等人[18]也研究了基于 LLM 的后处理, 他们针对不同语音识别系统的输出分别微调多个模型, 并通过集成策略提升了跨系统的稳健性.

这类方法的优点是无需改动原始模型即可提升结果, 能兼容任意系统; 其主要局限在于模型呈现的最终效果

与初始的转录文本质量高度相关, 泛化能力也有待进一步验证.

1.5 本研究相比现有方案的优势

基于音频-语言模型的端到端说话人日志系统, 在错误累积抑制、表示与泛化能力、内置语言建模和多任务交互扩展等方面, 相比现有方案具备如下几方面优势.

(1) 相比“语音分离+ASR”或“分段+聚类+ASR”的非端到端模块化系统, 音频-语言模型端到端地将音频输入映射至带有说话人标签的文本输出, 消除了中间模块误差累积的问题. 文献 [19] 表明, 在级联系统中, 早期阶段产生的错误会持续传递至后续模块, 并在最终结果中被放大, 使系统难以获得全局最优解; 而端到端方法因其统一的优化目标, 可有效降低这种误差传播带来的性能下降. 同时, 音频-语言模型的端到端特性也减少了各模块之间参数调优的开销, 简化了工程实施难度.

(2) 与现有端到端 SA-ASR 系统相比, 音频-语言模型提供了更通用的表示空间与更强大的任务迁移能力. 通过融合大规模预训练的文本与音频知识, 它能够在同一架构下同时处理不同人数、不同任务 (如转录、翻译、摘要等), 在跨任务泛化上具备优势. 这种统一的多模态表征不仅提高了判断说话人归属的准确性, 也增强了模型对复杂语义和上下文的理解能力.

(3) 音频-语言模型天然地内置语言模型的功能, 将语言建模与声学建模融为一体, 避免了现有系统在完成文本转录后再使用大语言模型进行后处理的繁琐流程. 音频-语言模型在生成阶段即可考虑全局语义一致性和说话人风格, 从而输出更加连贯、符合语言习惯的多说话人转录结果.

(4) 此外, 音频-语言模型具备广泛的交互和多任务扩展能力, 不仅能完成说话人日志的单一任务, 还能按照自然语言指令执行一系列下游操作 (如关键词提取、情感分析等). 现有的“专一型”模型仅限于输出“说话人标签+文本转录结果”, 而音频-语言模型则可在同一模型中通过不同的 prompt 实现定制化的人机交互, 真正做到了从“听懂”到“理解再处理”的转变. 这一优势为未来构建智慧会议助手或多模态语音助理奠定了基础.

2 训练方法

说话人日志任务可以拆解为两方面的能力需求: 其一是模型语音识别的能力, 其二是模型判断说话人归属的能力 (即为转录文本打上准确的说话人标签, 例如“speaker1”“speaker2”的能力). 本研究设计的两阶段训练框架围绕这两项能力如何提升、如何权衡展开.

模型的训练流程如图 2 所示. 在训练的第 1 阶段, 本研究使用监督学习对 Qwen2-Audio 进行指令微调, 让模型学会指定的作答格式, 并初步具备判别说话人的能力; 为了适应说话人日志任务, 损失函数采用加权交叉熵损失, 以强化模型对说话人标签 token 的学习, 从而平衡语音识别与说话人判别的能力.

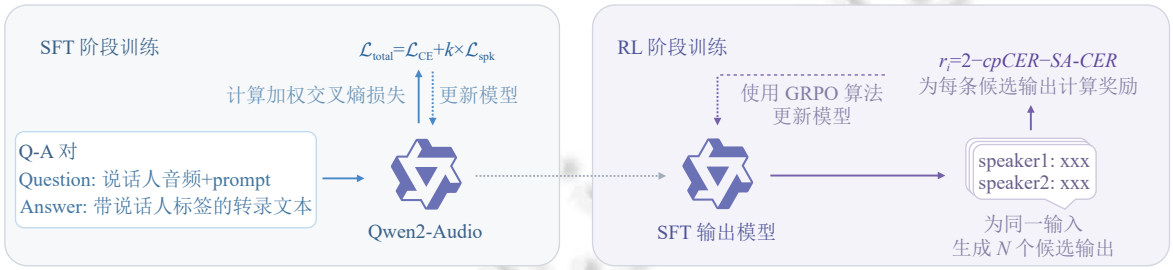

图 2 两阶段训练流程示意图

在随后的第 2 阶段训练中, 本研究在 SFT 模型的基础上引入基于 GRPO 算法的强化学习策略, 使用联合 *cpCER* 与 *SA-CER* 这两个综合指标的奖励函数, 实现模型语音识别能力与判别说话人归属能力的端到端协同提升.

2.1 第 1 阶段: 采用加权交叉熵损失的监督学习

如前文所述, Qwen2-Audio 模型在直接遵循复杂的日志任务指令时表现不佳. 因此, 在训练的第 1 阶段, 本研究基于指令式微调的思路构建训练数据, 旨在引导模型遵循“对输入音频进行多说话人区分, 并输出带有说话人标

签的转写”这一复杂指令, 并按照指定格式输出文本. 具体而言, 用于训练的问答对数据采用多种自然语言提示的形式, 例如“音频中有多位说话人在交谈, 请将每个人的发言区分开, 并为他们各自生成文字记录”, 并附上对应的使用标准作答格式的输出, 例如“speaker1: … speaker2: …”.

然而, 常规的交叉熵损失函数在计算损失时对所有 token 一视同仁, 缺乏对稀有的说话人标签 token 的额外关注. 说话人标签 token (数字“1”“2”…) 在整个文本序列中仅占极小比例, 例如, 对于一个正确的说话人序列“speaker1→2→1→3”, 若模型误识别为“speaker1→2→3→1”, 则序列中仅有两个 token 的归属发生错误, 而整个文本序列可能有上百个 token. 因此, 模型在输出错误的说话人顺序时, 受到的梯度惩罚可能不足以纠正说话人标签的归属. 标准交叉熵损失主要聚焦于词文本身的识别错误, 而对说话人标签置换的惩罚较弱, 更适合在不需要输出说话人标签的单说话人自动语音识别模型中充当损失函数, 不适用于多说话人场景.

为此, 本研究在标准交叉熵损失的基础上, 在真实标签序列中提取出所有代表说话人标签的 token 的位置. 根据模型生成的 logits 计算损失时, 对这些位置的 logits 额外计算它们与真实标签之间的交叉熵损失, 得到新的训练目标:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CE}} + k \times \mathcal{L}_{\text{spk}} \tag{1}$$

其中, $\mathcal{L}_{\text{CE}}$ 代表标准的交叉熵损失, $\mathcal{L}_{\text{spk}}$ 代表仅针对说话人标签 token 位置计算的“说话人损失”, 而 k 是一个用于调节两者权重的超参数.

这种加权策略与多任务学习中对不平衡类别进行加权的做法具有一致性. 例如, 在多语种或低资源 ASR 领域, 为提升少数语种的识别性能, Piñeiro-Martin 等人[20]通过语言权重交叉熵显著降低了低资源语言的错误率, 证明对稀有类别添加的额外惩罚能有效缓解类别不平衡问题. 同样, 在端到端说话人切换检测任务中, Khare 等人[21]针对“说话人发生改变”与“说话人未发生改变”类别的严重不平衡, 也采用了加权交叉熵的方法提高说话人切换检测准确率, 同样验证了加权策略对稀有标签学习的有效性. 此外, 多任务联合训练的经验也表明, 当各子任务的损失通过系数进行灵活调节时, 模型能够在保持主任务性能的同时兼顾辅助任务指标. 例如, Kumar 等人[22]将自动语音识别与说话人切换检测的损失线性组合后, 能在不额外增加模型复杂度的情况下实现任务间互补, 优化收敛速度与最终性能.

在训练过程中, 超参数 k 的取值既不宜过小, 以免说话人损失 $\mathcal{L}_{\text{spk}}$ 提供的监督信号过于弱; 也不宜过大, 以免任务过度偏向说话人标签预测而损害语音识别的精度. 本文在第 4.3 节“消融实验”中展示了超参数 k 对实验结果的影响, 并介绍了如何设定合适的 k 值. 合理调节超参数 k 后, 模型被期望获得更强的说话人归属判断的能力, 为后续强化学习阶段的进一步精细优化奠定了基础.

2.2 第 2 阶段: 基于 GRPO 算法的强化学习

在完成监督学习阶段的基础能力预热后, 本研究进一步引入基于强化学习的训练方法, 以突破模型在语音识别任务和判断说话人归属任务上的性能瓶颈. 近期研究[23]表明, 音频-语言模型如果仅靠监督微调, 在数据足够丰富时常会陷入性能平台期, 而基于策略梯度的 RL 方法则能够通过直接对离散指标进行优化, 进一步提升模型性能. 也有研究表明[24], 在语音识别领域, 通过基于人类反馈的强化学习 (reinforcement learning from human feedback, RLHF) 或直接的策略优化, 模型在适应病态语音或新领域时获得了显著增益, 表明 RL 对语音识别模型适配任务具有良好效果.

本研究选择了 GRPO 算法来替代直接偏好优化 (direct preference optimization, DPO)[25]、近端策略优化 (proximal policy optimization, PPO)[26]等强化学习算法. 多说话人日志任务的核心在于同时优化语音识别与说话人归属两大能力, 相较于传统的强化学习算法, GRPO 天然支持多响应生成与多指标联合奖励: 只需在奖励函数中完成对优化目标的定义, 训练时对生成的每条候选回复计算综合奖励, 就可以直接在策略梯度中隐式平衡各指标. 比起需要专门收集偏好数据的 DPO 流程, GRPO 无需额外组织正负样本对; 而与 PPO 相比, GRPO 以用户自定义的奖励函数替代了价值网络, 从而提供了一条更为简洁的训练路径.

GRPO 算法内部使用群体相对的方式计算优势, 如公式 (2) 所示, 定义响应的优势函数为:

$$A_{i,t} = \frac{r_i - mean(r)}{std(r)} \tag{2}$$

在每个训练批次中, GRPO 对同一输入生成多条响应. 在公式 (2) 中, r_i 代表对第 i 条候选输出得到的奖励分数, 而 r 是当前批次内所有候选输出的奖励集合 $\{r_i\}_{i=1}^{G}$. 该函数通过将单个奖励与组内平均奖励进行比较并用标准差归一化, 来计算其相对优势.

奖励函数的设计是使用 GRPO 算法进行强化学习训练的核心, 也是方法创新之一, 可以定义多个奖励函数来灵活地衡量不同的评价维度. 本研究将转录文本的质量、说话人归属判断的准确性、格式规范性等多种指标融入奖励中. 在本任务中, 如公式 (3) 所示, 每条候选输出的奖励 r_i 由联合评价指标构成:

$$r_i = 2 - cpCER - SA\text{-}CER \tag{3}$$

其中, *cpCER* 为拼接最小置换字错误率, *SA-CER* 为说话人归属字错误率, 二者均为越低越好. 该奖励函数旨在同时优化这两个负向指标. *cpCER*、*SA-CER* 是语音识别与说话人归属错误的综合性指标, 能同时反映模型的语音识别能力与说话人归属能力, 与说话人日志任务的优化目标高度契合. 本文第 3.2 节“评价指标”会同实验的测试指标一起, 简要介绍这两个指标.

在策略优化阶段, GRPO 的目标是最大化批次级别的期望优势, 通过最小化以下目标:

$$J_{\text{GRPO}}(\theta) = \mathbb{E}\left[\min\left(\rho_t(\theta)A_t, clip\left(\rho_t(\theta), 1-\epsilon, 1+\epsilon\right)A_t\right)\right] - \beta KL\left[\pi_\theta \| \pi_{\text{ref}}\right] \tag{4}$$

其中, $\rho_t(\theta)$ 为策略比值, 用于衡量当前需要优化的模型策略 π_θ 与固定的参考模型策略 π_{ref} (在本研究中 π_{ref} 固定为 SFT 阶段结束后的模型) 之间的差异; A_t 是 t 时刻的优势函数; ϵ 是用于裁剪策略比值的超参数, 以限制更新步长; β 是 KL 散度项的权重系数, 以防止策略崩溃. GRPO 简化了价值估计并提高了多样化响应的利用效率.

在训练实施时, 本研究以经过加权 SFT 训练后的模型为基础继续进行强化学习训练, 结合 GRPO 目标与 KL 正则项, 保持与监督学习阶段优化目标的一致性; 同时, 通过 GRPO 算法的小批量多响应机制收集组内奖励, 将 *cpCER* 与 *SA-CER* 的组内差异转化为增强学习信号, 在降低语音识别错误的同时进一步压低说话人归属的错误. 整个训练流程借鉴了 RL 强化 LLM 的成功经验[27], 并融合了端到端多模态对话系统的优化策略, 以期模型在中文多说话人日志场景中获得最佳综合性能.

3 实验设置

3.1 基座模型与对比基线

本研究选用 Qwen2-Audio-7B-Instruct 作为端到端说话人日志系统的基座模型. 音频-语言模型 Qwen2-Audio-7B 基于 Qwen2-7B 大语言模型, 它在预训练阶段加入音频编码器, 将 16 kHz 的语音特征映射到与语言模型共享的表示空间. 这样, 整个模型既保留了大语言模型的自然语言理解能力, 又具备音频编码器带来的声学建模能力, 因此音频-语言模型无需再外挂额外的语音识别模块或声道分离模块, 就可以在自回归解码器中直接生成带有说话人标签的文本响应.

为了评估本方法的提升幅度, 本研究选择开源社区 ModelScope 的“说话人日志”板块中下载量最高的模型, 即“3D-Speaker”系列中的说话人日志模型 (后文简称为 3D-Speaker 模型) 作为对比的基线. 截至 2025 年 4 月, 此模型的下载量达 390k 次, 被各团队广泛使用. 3D-Speaker 模型由语音端点检测模型、说话人确认模型、说话人转换点定位模型、语音识别模型这 4 个组件串联而成, 提供了完整的说话人日志流水线, 在多个公开基准上展现了稳健的 DER 表现.

同时, 我们引入了 NVIDIA NeMo 工具集中的 Diar Sortformer (4spk, V1)[28]模型作为另一重要的开源基线. Sortformer 是一种时域端到端的说话人日志模型, 输出带说话人标签的起止时间戳. 这种方法虽然在日志任务本身上是端到端的, 但仍需一个独立的 ASR 模块来完成最终的文本转录. 因此, 在本研究的基线实现中, 我们将 Sortformer 的输出结果与 Paraformer-large[29]中文离线 ASR 模型相结合, 构成一个完整的“时域端到端+ASR”流水线. 这一基线代表了另一条先进的、与本研究不同的技术路线.

此外, 为了更全面地评估本研究方法的性能, 本研究还引入了两个业界领先的商业解决方案作为对比基线: AssemblyAI 和 Microsoft Azure 的多说话人转录服务, 它们代表了当前工业界成熟的说话人日志技术水平. 由于其实现细节并未公开, 本研究将其作为“黑盒”模型, 仅对其最终输出结果进行评估.

值得注意的是, 由于本研究所用基座模型 Qwen2-Audio 存在 30 s 的音频输入时长限制, 故本研究中的所有训练和测试样本均被裁剪或控制在 30 s 以内. 这一约束对不同类型的模型影响是不同的. 对于本研究提出的端到端模型, 短音频是其固有的处理单元. 然而, 对于 3D-Speaker 等基于分段和聚类的流水线模型, 其性能往往受益于更长的音频输入. 在长音频中, 模型可以观察到更丰富的说话人信息和更稳定的声学特征, 从而做出更鲁棒的聚类判断. 因此, 在“中文语境”和“30 s 输入时长”的双重限制下, 基线模型的性能可能受到了限制, 未能完全发挥其在长时程任务上的优势. 这使得我们的端到端方法在当前实验设置下, 相比基线模型, 可能表现出比在长音频场景下更大的相对优势.

3.2 评价指标

本研究涉及以下评价指标.

(1) *CER* (character error rate): 语音识别任务的常见指标, 衡量模型在文本转录层面的准确性, 定义为替换、删除、插入错误之和与参考文本总字符数之比.

(2) *cpCER*: 将各说话人对应的转录结果按说话顺序拼接, 遍历所有可能的说话人标签置换组合, 取最低 *CER*, 联合反映语音识别错误与说话人归属错误.

(3) $\Delta cp = cpCER - CER$: 度量说话人归属错误带来的额外字符错误率增量.

(4) *SA-CER*: 确定说话人标签映射的情况下, 分别计算每位说话人的 *CER* 并汇总, 评估词级与说话人级匹配的整体表现.

(5) $\Delta SA = SA\text{-}CER - CER$: 与$\Delta cp$ 的意义相似, 反映说话人映射的不准确对语音识别造成的附加错误.

其中, Δcp 与ΔSA 主要用于衡量说话人归属的准确性, *CER* 则专注于语音识别质量; *cpCER* 与 *SA-CER* 提供了两者的联合视角. 这些指标数值越小, 代表模型能力越优秀. 本研究使用Δcp、ΔSA、*CER* 这 3 个指标作为衡量模型性能的度量, 而综合性指标 *cpCER* 与 *SA-CER* 则被应用于强化学习训练阶段的奖励函数.

3.3 数据集的构建

本研究围绕双说话人和四说话人两种场景, 构建了包含真实对话与人工合成对话的训练集, 并从相同数据源中另外构建了相应的测试集. 真实场景下, 带有文本转录标注的中文多人对话数据集规模有限, 而中文的单人语音识别数据集更易于寻找. 只要单人的语音识别数据集带有说话人标签, 就可以通过句子间组合拼接的方式构建多人对话数据集, 且说话人的数量、对话的长度、语速等属性均可控. 通过这种人工合成数据的方法, 可以快速构建大规模 (但可能缺少上下文语义关联) 的多人对话数据集. 本研究使用了这种数据生成策略, 将人工合成的音频 (超长的音频通过快进的方式保证符合 30 s 的输入时长限制) 结合真实场景下截取的音频一并纳入训练集和测试集中. 实验选用的训练数据集详见表 2、表 3.

表 2 双说话人场景 SFT 阶段训练集

数据集	种类	数量	备注
MagicData-RAMC	1–2人, 真实场景	10000	GRPO时随机抽取5000条
THCHS-30	1–2人, 人工场景	8000	选取1–2位说话人的数据

表 3 四说话人场景 SFT 阶段训练集

数据集	种类	数量	备注
MagicData-RAMC	1–2人, 真实场景	24000	GRPO时, 在所有数据集中随机抽取8000条
AliMeeting	1–4人, 真实场景	24800	
AISHELL-4	1–4人, 真实场景	15000	
THCHS-30	1–4人, 人工场景	15000	

下面简单介绍各数据集. 真实场景的数据集来源如下.

(1) MagicData-RAMC: 由 663 名说话人通过手机采集的 180 h 高质量对话语音, 覆盖 15 个领域并提供人工校对的精确标注, 包括发言时间戳与说话人标识.

(2) AliMeeting: 源自 ICASSP 2022 会议转写挑战的 118.75 h 真实会议录音, 包含 2–4 人的自然交谈场景, 既有远端八通道麦克风阵列, 也有主体头戴麦克风记录, 覆盖多种室内对话的声学场景. 本研究使用 AliMeeting 的近场子集 (单通道音频).

(3) AISHELL-4: 录制自 211 场真实录制的中文会议, 每场 4–8 人, 合计 120 h, 采用圆形八通道麦克风阵列, 提供准确的语音起止时间戳与说话人信息标注, 真实反映了环境噪声、语音重叠与说话人快速切换等复杂现象.

(4) SeniorTalk: 专为 75 岁及以上超高龄老年人设计的中文对话数据集, 包含 55.53 h 的自然对话录音, 涵盖 202 名参与者, 具有丰富的方言和老年人特有的音色特征.

人工场景的数据集来源如下.

THCHS-30: 由清华大学发布的免费中文语音数据库, 包含 30 h 朗读式语音和说话人信息的标注. 本研究将数据集中单句语音互相拼接以构造多轮对话, 通过快进的方式控制音频的总长度不超过 Qwen2-Audio 的 30 s 输入长度限制, 作为人工合成的补充数据集.

本研究构造的数据集包含至多 6 轮对话 (5 次说话人切换), 并涵盖了 1–4 位说话人在至多 6 轮对话中可能产生的所有说话人排列顺序, 尽可能地提升场景丰富度.

在双说话人与四说话人两种情境下, 监督学习的训练集均由上述真实数据与合成数据两部分组成. 对于双人场景下的强化学习步骤, 因监督学习后模型在双人人工场景已无性能的优化空间, 实验只在真实场景数据中采样 5 000 条数据作为训练集; 对 4 人场景下的强化学习, 在整个监督学习训练集中随机采样 8 000 条作为训练集. 测试集则分别从各数据集中另外随机采样 50 条 (4 人场景) 或 100 条 (双人场景) 样本.

3.4 超参数设置与硬件需求

两个训练阶段均使用 ModelScope 的 swift 训练框架训练 Qwen2-Audio 模型, 训练使用 4 张 NVIDIA A800 80 GB 显卡. 训练的两个阶段均冻结音频编码器, 仅对 Qwen2-Audio 的语言模型部分进行全参数微调, 以降低显存占用并专注提升解码质量. 为确保训练效率与收敛稳定性, 训练过程中的关键超参数设置如表 4、表 5 所示. 在双说话人场景下, 说话人损失 $\mathcal{L}_{\text{spk}}$ 的权重 k 设置为 4, 四说话人场景下设置为 $k=1$. 双说话人场景训练耗时约 10 h, 四说话人场景训练耗时约 30 h.

表 4 SFT 阶段主要超参数设置

超参数	数值
train_epoch	1
per_device_train_batch_size	1
learning_rate	1E–4
warmup_ratio	0.05
gradient_accumulation_steps	16

表 5 GRPO 阶段主要超参数设置

超参数	数值
train_epoch	1
per_device_train_batch_size	4
learning_rate	1E–5
warmup_ratio	0.05
num_generations	4
temperature	0.9
gradient_accumulation_steps	1

4 实验结果与分析

本节首先对双说话人场景对所提两阶段训练方法进行了系统验证, 通过与 3D-Speaker 基线的对比实验, 展示了监督微调及加权损失对 ASR 与说话人归属性能的影响, 并进一步引入 GRPO 强化学习以突破模型能力瓶颈; 随后, 本研究尝试将训练方法推广至四说话人的场景, 再次评估其在各种对话环境中的表现差异, 并通过消融实验探讨各环节对整体性能的贡献. 本节表中 Average 指代Δcp 与ΔSA 的均值.

4.1 双人场景测试

● 发现 1: 如表 6 所示, 在 MagicData-RAMC 测试集 (真实场景) 中通过监督学习与后续的 GRPO 强化学习, 采用一位、两位说话人数据训练的“双人模型”在语音识别 (*CER*) 和判断说话人归属 (Δ*cp*、Δ*SA*) 两项能力上均优于所有基线模型. 说话人权重的引入能提升归属准确度, 以可接受的 *CER* 回升为代价; GRPO 强化学习阶段同步提升两项性能.

表 6 MagicData-RAMC 测试集 (双人真实场景) (%)

模型	Δ*cp*↓	Δ*SA*↓	Average↓	*CER*↓	方法类型
3D-Speaker	8.66	27.58	18.12	14.88	流水线-开源
Sortformer	7.68	25.80	16.74	25.27	时域端到端-开源
AssemblyAI	17.61	28.51	23.06	20.22	API-闭源
Microsoft	8.68	22.46	15.57	14.36	API-闭源
双人SFT	2.83	5.81	4.32	10.98	ALM-SFT
双人加权SFT	2.57	5.63	4.10	12.29	ALM-加权SFT
+双人GRPO	**1.94**	**3.55**	**2.75**	**10.61**	ALM-加权SFT+RL
4人SFT	2.82	4.91	3.87	13.64	ALM-SFT
4人加权SFT	2.55	4.51	3.53	13.91	ALM-加权SFT
+4人GRPO	**2.26**	**4.26**	**3.26**	**13.36**	ALM-加权SFT+RL

分析: 即便在监督微调中不加入“说话人损失”, 模型的语音识别能力与判断说话人归属的能力已经较为优秀, 高于基线模型; 引入说话人损失后, 尽管代表语音识别错误的 *CER* 相比常规 SFT 模型略增, 但加权策略为稀疏的说话人标签提供了强化信号, 使Δ*cp* 与Δ*SA* 进一步下降, 实现了优化目标之间的权衡; 最终经过 GRPO 强化学习阶段, 模型不仅在Δ*cp* 与Δ*SA* 达到了最低水平, 其 *CER* 也恢复至与未加权 SFT 模型相当的水平, 甚至略有提升, 这表明策略梯度优化在突破监督学习瓶颈方面具有明显优势.

● 发现 2: 如表 7 所示, 在 THCHS-30 测试集 (人工场景) 中, 不引入说话人权重的监督学习模型在说话人归属相关的指标Δ*cp* 与Δ*SA* 上均达到 0.00, *CER* 也几乎为 0. 这表明模型在该数据集上已达到性能上限. 而后续的加权 SFT 和 GRPO 阶段反而导致了性能的轻微下降.

表 7 THCHS-30 测试集 (双人人工场景) (%)

模型	Δ*cp*↓	Δ*SA*↓	Average↓	*CER*↓	方法类型
3D-Speaker	2.74	4.03	3.39	9.47	流水线-开源
Sortformer	1.83	4.24	3.04	18.64	时域端到端-开源
AssemblyAI	6.50	7.32	6.91	15.03	API-闭源
Microsoft	9.94	13.54	11.74	8.42	API-闭源
双人SFT	**0.00**	**0.00**	**0.00**	**0.03**	ALM-SFT
双人加权SFT	0.26	0.26	0.26	0.33	ALM-加权SFT
+双人GRPO	0.54	0.54	0.54	0.68	ALM-加权SFT+RL
4人SFT	11.27	66.43	38.85	**0.00**	ALM-SFT
4人加权SFT	**1.68**	4.95	3.32	0.12	ALM-加权SFT
+4人GRPO	**1.68**	**4.04**	**2.86**	0.12	ALM-加权SFT+RL

分析: 这一“零错误”现象并非由评估方式导致, 而是在该特定数据集上模型能力与任务难度共同作用的结果. THCHS-30 人工数据由纯净的单人朗读语音拼接而成, 完全不含真实场景中的背景噪声和语音重叠, 这极大地降低了声学分离和语音识别的难度. Qwen2-Audio 面对如此理想化的任务游刃有余, 因此标准 SFT 足以使其达到“性能天花板”. 在这种近乎饱和的状态下, 任何对辅助任务 (如强化说话人标签学习) 的额外干预, 都可能引入不必要的梯度噪声, 从而轻微损害已经接近完美的整体表现, 这与多任务学习中的经验相符. 这一结果也反衬出在 MagicData-RAMC 等真实场景数据集上进行测试的必要性.

● 发现 3: 如表 6、表 7 所示, 在双人场景中, 采用 1–4 位说话人数据训练的“4 人模型”在测试结果上同样符合训练方法设计的预期. 从最终结果来看, 模型的表现相比“双人模型”略差, 但仍相当接近.

分析: 经过加权 SFT 的 4 人模型在判别说话人的能力上优于不加权 SFT 的模型, 语音识别能力相比不加权 SFT 的模型稍差; 经过 GRPO 强化学习阶段后, 两项能力均获得进一步提升, 这些现象和双人模型的测试结果展示了相同的趋势, 符合训练方法的设计思路. 相比双人模型, 4 人模型在双人测试集上的表现略差, 这归因于训练-测试分布的不匹配. 在强化学习阶段, 本研究对模型进行了以 4 人对话为主的联合奖励优化, 而在双人对话测试中, 其优化策略可能对多说话人交互模式产生了过拟合, 从而降低了模型在双人场景下对对话轮次与说话人切换的敏感度, 导致性能的略微下降. 这与强化学习中常见的“分布外泛化 (out-of-distribution, OOD)”[30,31]问题相契合: 当 RL 策略在测试时遭遇与训练环境显著不同的输入分布, 会因价值估计或策略执行的偏移而导致性能退化.

● 发现 4: 如表 7 所示, 在 THCHS-30 测试集 (人工场景) 中, 未使用加权交叉熵的监督学习模型尽管实现了近乎完美的语音识别 (*CER*=0), 却因对第 3、第 4 说话人的大量误检导致 Δ*cp* 和Δ*SA* 暴增; 引入说话人加权损失后, Δ*cp* 与Δ*SA* 大幅下降至 1.68% 和 4.95%, 且经 GRPO 后Δ*SA* 进一步优化至 4.04%, 同样符合预期.

分析: 人工合成数据的对话结构简单且无环境噪声和重叠语音, 常规模型微调足以优化文本生成, 但对说话人归属缺乏区分信号, 错误地将训练数据中分布较大的第 3、第 4 说话人标签直接应用于只有两位说话人的场景, 导致模型识别出多余的说话人. 说话人加权损失通过专门放大稀有的说话人标签 token 的误差, 提供了额外的判别监督, 有效缓解了此问题. GRPO 阶段的指标提升也进一步印证了强化学习方法的适用性.

综上所述, 双说话人场景下的实验结果验证了两阶段训练策略的合理性: 监督学习阶段通过加权损失为说话人归属提供强化信号, GRPO 强化学习阶段则借助联合奖励函数突破监督学习瓶颈, 实现对语音识别与说话人归属能力的全面优化. 相比 3 个基线模型, 训练后的 Qwen2-Audio 模型不仅在指标上大幅领先, 其转录的语句也具有更好的可读性, 这归功于音频-语言模型出色的下游任务泛化能力和 Qwen2 系列大语言模型基座优秀的自然语言交互能力.

4.2 4 人场景测试

● 发现 5: 如表 8、表 9 所示, 在四说话人真实会议场景中, 不加权的监督学习模型在判别说话人的能力上反而优于加权模型与进一步经过强化学习的模型. 在此场景下, 我们提出的加权交叉熵策略不仅未能带来预期的能力进步, 反而在指标上出现负提升.

表 8 AliMeeting 测试集 (4 人真实场景) (%)

模型	Δ*cp*↓	Δ*SA*↓	Average↓	*CER*↓
3D-Speaker	7.84	26.24	17.04	41.87
Sortformer	14.15	18.54	16.35	71.92
AssemblyAI	11.03	22.53	16.78	45.58
Microsoft	9.17	29.47	19.32	42.74
4人SFT	**1.99**	**12.34**	**7.17**	38.56
4人加权SFT	3.88	15.60	9.74	36.41
+4人GRPO	2.61	14.69	8.65	**34.74**

表 9 AISHELL-4 测试集 (4 人真实场景) (%)

模型	Δ*cp*↓	Δ*SA*↓	Average↓	*CER*↓
3D-Speaker	3.34	19.71	11.53	59.56
Sortformer	15.28	21.10	18.19	57.23
AssemblyAI	13.73	27.26	20.50	50.66
Microsoft	5.61	15.73	10.67	51.24
4人SFT	**0.97**	**13.41**	**7.19**	7.17
4人加权SFT	4.24	14.60	9.42	7.63
+4人GRPO	2.73	14.09	8.41	**6.75**

分析: 此现象揭示了当前方法在处理高复杂度声学场景时的局限性. 从双人到 4 人, 任务难度并非线性增加, 多人会议录音中急剧增多的语音重叠、更相似的说话人音色以及更复杂的轮转模式, 都对模型的声学建模能力提出极大的挑战. 尽管常规 SFT 模型相比基线仍有较大优势, 但部分输出已出现较为明显的说话人归属混乱现象. 在这种高压环境下, 我们为双人场景设计的加权策略反而成为梯度噪声源, 干扰了模型对核心声学信息的学习, 造成了严重的多任务负迁移.

对更多说话人场景的适应性推断: 基于 4 人场景的实验结果, 我们可以合理推断, 直接将本方法应用于更多说话人 (如 5 人以上) 的场景是困难的. 任务复杂度的进一步提升可能彻底压垮当前仍较为粗糙的训练框架. 未来的研究既需要探索全新的范式, 也要聚焦于对现有框架的细致优化. 例如, 只精细地依靠音频模态信息可能已不足以

解决问题, 可以使用引入视觉模态信息 (如唇动、人脸位置) 的图文音全模态大模型 (omnimodal large language model, OLLM); 亦或在当前框架下更精细地设计训练集, 并结合课程学习 (curriculum learning) 等策略, 让模型从易到难逐步学习. 这在本文第 5.3 节和第 5.5 节中进行了更详细的讨论.

4.3 消融实验

本节通过 3 组消融实验, 在双人说话人场景下探讨了第 1 阶段监督学习中说话人权重 k 的作用, 以及第 2 阶段强化学习对监督学习模型性能瓶颈的突破, 最终在分布外数据上验证了训练方法的合理性. 通过对比不同设置下的 Δcp、ΔSA 与 *CER*, 消融实验验证了两阶段训练方法对模型能力的塑造效果, 并揭示了背后的机理与权衡.

(1) 说话人权重 k 产生的影响

在表 10 中, 当说话人损失权重 k 从 0 (不加权) 逐步增加至 4 时, 衡量说话人归属能力的 Δcp 与 ΔSA 呈现持续下降趋势, 而反映语音识别能力的 *CER* 则出现轻微上升; 当 k 超过 4 时, Δcp 与 ΔSA 的下降趋势开始逆转, 不再进一步降低.

表 10 消融实验: 说话人权重 k 的影响 (%)

模型	Δcp↓	ΔSA↓	Average↓	*CER*↓
SFT, k=0	2.83	5.81	4.32	**10.98**
SFT, k=1	3.40	7.28	5.34	12.66
SFT, k=2	3.07	7.00	5.04	11.93
SFT, k=3	2.97	6.21	4.59	11.47
SFT, k=4	**2.57**	**5.63**	**4.10**	12.29
SFT, k=5	2.83	6.23	4.53	11.47

实验数据显示, 结果与方法设计的初始预期相符: 在序列到序列任务中, 少数类别 (说话人标签 token) 因出现频率低, 往往被标准交叉熵淡化, 难以获得足够的梯度信号. 引入加权交叉熵, 通过对指定类别赋予更高的损失权重, 可以有效缓解类别不平衡问题, 提升模型对重要但稀疏信息的敏感度. 具体而言, 本研究从 $k = 1$ 开始, 尝试将说话人标签 token 的梯度放大, 与多标签分类中为少数类加权的常用做法一致, 可获得对稀有但关键标签更强的优化信号. 随着权重 k 增加到 4, 模型的 Δcp 从 2.83% 降至 2.57%, ΔSA 从 5.81% 降至 5.63%, 表明说话人归属能力的提升; 同时, *CER* 从 10.98% 上升至 12.29%, 反映了将部分学习能力偏向说话人分类带来了语音识别能力的轻微退让. 当 k 进一步增至 5, 归属指标未能继续下降, 说明过度加权会导致模型对文本流畅性关注不足. 综上所述, 本实验揭示了加权策略在语音识别能力与说话人归属能力之间的性能权衡.

(2) GRPO 突破监督学习性能瓶颈

如表 11 所示, 在监督学习模型 ($k = 4$) 的基础上, 继续使用 MagicData-RAMC 数据集进行监督微调 (抽取 5000 条样本作为训练集) 难以使模型的性能获得进一步提升, 甚至在语音识别能力上略有下降; 相比之下, 使用相同的训练集进行 GRPO 强化学习后, Δcp、ΔSA 与 *CER* 均获得显著的下降, 分别达 1.94%、3.55% 与 10.61%.

表 11 消融实验: GRPO 训练策略的有效性 (%)

模型	Δcp↓	ΔSA↓	Average↓	*CER*↓
双人SFT	2.57	5.63	4.10	12.29
双人SFT+继续SFT训练	2.32	5.48	3.90	14.80
双人SFT+GRPO	**1.94**	**3.55**	**2.75**	**10.61**

由于监督学习目标为最大化似然, 对非可微指标 (如 *cpCER*、*SA-CER*) 难以直接优化, 常在数据丰富后陷入性能瓶颈. 策略梯度方法则通过直接优化下游评价标准, 能够突破监督学习的性能极限. 在消融实验中, 继续进行监督学习未能带来联合指标的有效提升, 这反映了监督目标与评测指标之间存在一定的目标错配; 而 GRPO 在奖励函数的直接指导下, 实现了对语音识别与说话人归属两项能力的协同优化, 证明了第 2 阶段强化学习设计的必要性与有效性.

(3) 两阶段训练方法在分布外数据的泛化能力

为验证两阶段训练策略在分布外数据上的泛化能力, 本研究在 SeniorTalk 数据集上进行了消融实验. SeniorTalk

数据集与训练双人模型所使用的 THCHS-30 合成数据和 MagicData-RAMC 真实场景数据在说话人年龄、音色特征和方言分布等方面存在显著差异, 因而可作为评估模型在分布外场景下表现的理想测试集.

如表 12 所示, 从衡量模型区分说话人能力的 Average 指标来看, 经过监督学习训练的音频-语言模型在识别说话人的能力上已优于基线, 且引入说话人权重和 GRPO 训练均进一步提升了性能, 表明模型习得的区分说话人能力在分布外数据上依旧有效. 在 *CER* 指标上, Microsoft 基线表现最佳, 说明数据分布的变化影响了音频-语言模型文本转录的能力. 尽管如此, 引入说话人权重的 SFT 模型在说话人分离能力提升的同时, *CER* 略有妥协, 而随后的 GRPO 训练在提升说话人分离能力的同时, 同步改善了 *CER*, 这些规律都符合训练策略设计时的初始设想, 进一步证明了两阶段训练方法的合理性.

表 12 消融实验: 分布外测试集上的性能比较 (%)

模型	Δ*cp*↓	Δ*SA*↓	Average↓	*CER*↓
3D-Speaker	7.89	22.94	15.42	30.82
Sortformer	9.97	21.49	15.73	42.37
AssemblyAI	15.25	24.13	19.69	35.25
Microsoft	8.72	25.01	16.87	**29.56**
双人SFT	7.07	15.66	11.37	37.74
双人加权SFT	7.06	**14.80**	10.93	40.44
+双人GRPO	**6.56**	14.95	**10.76**	39.68

5 讨 论

本研究提出了基于 Qwen2-Audio 的端到端说话人日志系统, 并通过两阶段训练策略实现了语音识别与说话人归属的联合优化. 尽管实验结果表明该方法相比 3D-Speaker 基线具有显著优势, 但在实际应用与模型设计层面仍存在诸多值得讨论的问题. 本文设计的实验只覆盖了不超过 4 位说话人的场景, 但真实应用中还存在更多人的混合对话、跨设备收音及多人快速切换等极端场景, 还需要更大规模、多样化的标注数据进行深入验证. 本节针对模型与训练方法的缺陷以及未来的优化方向进行了更进一步的讨论.

5.1 训练效率与资源权衡

现有端到端多模态音频–语言模型的参数量往往以亿为单位, 如 Qwen2-Audio 拥有约 70 亿参数, 在推理与微调阶段的计算资源开销不可小觑. 相比之下, 3D-Speaker 基线由多个相对轻量的模块 (语音端点检测、说话人确认、说话人转换点定位、语音识别) 组成, 训练和使用时对资源的需求更低. 然而, 3D-Speaker 的最终效果高度依赖于每个模块各自的性能, 需要针对不同场景的需求单独地优化每个子模块, 所以各模块总的工程复杂度与调试成本可能较高. 对于这一点, Qwen2-Audio 通过统一的单模型架构直接端到端地实现了整个任务的闭环, 简化了任务管道并避免了多模块接口的误差累积, 尽管需要更强大的硬件资源作为支撑. 所以, 在保证端到端简洁性的前提下, 通过模型压缩、蒸馏或高效的微调技术 (如 LoRA、 QLoRA)、推理框架来降低音频-语言模型的资源需求, 仍亟待进一步探索.

5.2 输入时长限制与长序列处理问题

Qwen2-Audio 使用了基于 Whisper-large-v3 模型的音频编码器, 它对音频输入有 30 s 的时长限制, 以避免过长的输入序列导致 $O(n^2)$ 复杂度的庞大注意力机制开销与内存溢出风险. 音频长度的限制可以通过音频截断拼接或对音频加速的方式绕过, 但音频截断拼接的方法难以通过算法的设计来保证转录文本的跨段语义衔接与说话人连续性的正确恢复, 音频加速的方法则会不同程度地削弱模型的能力, 所以这两种方法都不能从根本上解决问题. 相比之下, 基于分段+聚类的流水线 (如 3D-Speaker) 对整体时长的限制更弱, 且聚类算法在更长时长的样本上可能表现更稳健, 所以在通用性上, 基于分段-聚类的流水线说话人日志系统仍存在一定的优势.

对于上述限制, 一种思路是更换更强的音频编码器, 让音频-语言模型能接收分钟级甚至小时级的音频输入. 比如, 可以引入流式 Transformer (streaming Transformer) 结构的音频编码器[32], 它被专门设计用于处理长序列, 通

过分块处理和在线解码, 在输入音频的同时同步生成结果, 非常适合长音频输入, 在语音识别领域已被广泛应用.

5.3 基于训练数据的简单优化策略

本研究通过加权 SFT 和 GRPO 强化学习两阶段的训练, 在双人说话人场景下获得了稳健的性能提升. 然而, 实验结果也显示加权 SFT 的训练策略在四说话人的复杂场景下出现性能退化. 在面对高语速、语句重叠、环境噪声大和说话人数多样化的真实会议场景时, 训练的收敛不够稳定, 甚至出现了“负提升”.

一种基于训练数据的简单优化方法是课程学习策略, 提前为每条数据打上难度标签, 让模型先使用简单数据 (双人、清晰录音) 完成训练, 再逐渐过渡到难度较大的数据 (多人、重叠、噪声), 或许能缓解这一问题. 同时, 还可以引入“拒绝采样 (rejection sampling)”的训练数据筛选流程: 先对原始训练集中的各样本按模型当前表现或样本难度打分, 剔除数据集中难度太小或难度太大的样本, 仅保留中等难度的数据作为新训练集. 通过拒绝采样的数据筛选流程, 模型在训练时可以专注于具有挑战性但可学习的样本.

5.4 任务专一化带来的能力退化及应对策略

本研究通过定向针对说话人日志任务的监督微调和强化学习训练, 显著提升了模型在多说话人转写与归属判别中的性能, 但这种针对某一个任务的专项训练会不可避免地侵蚀模型原本的多模态能力, 带来“灾难性遗忘”: 单任务微调通常会导致模型在新任务上表现优异的同时, 其固有的部分通用能力出现退化, 也会削弱模型在未见任务或新领域上的泛化能力. 具体到音频-语言模型, 经由说话人日志任务专项训练后, 它在常规模式下对自然语言指令 (如音频翻译或摘要生成) 的理解和执行是否仍能保持原有水平, 尚未获得系统验证; 若模型在这些任务上出现显著的性能下滑, 便会严重限制其作为通用音频-语言助理的应用场景.

针对这一问题, 未来工作需在尽量不损害模型基础能力的前提下, 让模型学习说话人日志任务. 首先, 可尝试多任务微调, 即在 SFT 阶段同时引入其他音频-语言子任务 (如语音翻译、音频问答), 以保持模型对各类指令的响应能力. 另一种简单的方法是采用 LoRA (low-rank adaptation) 等轻量化微调的方法, 只对特定模块插入小规模的可训练参数, 而冻结大部分预训练权重, 已有研究[33]使用此类轻量化的方式, 发现轻量化微调的思路可显著降低遗忘风险, 同时降低训练成本.

5.5 扩展到全模态大语言模型

2024 年以来, 以 GPT-4o、Qwen2.5-Omni 为代表的全模态大语言模型迅速崛起, 这类模型一般使用不同的编码器分别编码视觉、听觉模态的输入信息, 并通过共享自回归解码器 (多为大语言模型) 生成高质量的文本或语音输出. 相比音频-语言模型, OLLM 不仅能够处理文本和音频, 还支持图像与视频输入, 实现了视觉、听觉、文本这 3 个模态的统一感知. 它的应用场景十分广阔, 从智能家居助手到自动驾驶汽车, 再到虚拟现实环境中的交互体验, 它们都可以提供更加自然和直观的服务, 正如一个理想的多模态模型应当能够理解用户的语音指令, 同时分析背景图像, 据此作出恰当的回应.

OLLM 的多模态处理能力对于说话人日志任务尤为重要: 在视频会议或监控场景下, 模型可以利用人脸位置、口型运动及场景背景为说话人归属提供额外线索, 从而弥补纯音频信息在高重叠语音和嘈杂环境中的不足. 这意味着未来的端到端说话人日志系统不仅可以接收音频提示, 还能接收屏幕截帧或摄像头画面, 让模型根据视觉中人物位置与动作进一步提高区分说话人的准确率. 比如在一次多人会议中, 当多位与会者相继发言时, 模型可以通过视觉信号判断谁在说话, 从而减少音频重叠、环境噪音导致的说话人归属错误. 与音频-语言模型相似, 以大语言模型为基础的 OLLM 同样具备优秀的自然语言理解能力, 可以响应用户的个性化需求, 兼容多个任务. 所以随着硬件算力与模型架构的持续进步, OLLM 可能会成为下一个推动说话人日志系统发展的重要技术方向.

6 结 论

本研究设计了一种以音频-语言模型为基座训练得到的端到端说话人日志系统, 摒弃传统模块化流水线, 通过监督微调与加权交叉熵损失强化说话人分类信号, 再借助 GRPO 强化学习直接优化联合指标, 实现了语音识别与说话人归属能力的提升. 在各测试场景中, 经过训练的音频-语言模型相较基线模型具有显著优势. 消融实验进一

步展示了“说话人损失”对两个训练目标的权衡, 以及 GRPO 阶段起到的突破监督学习瓶颈的关键作用.

然而, 在复杂会议录音和跨场景测试中, 加权微调相比朴素的监督学习反而出现性能退化, 揭示了模型在应对复杂声学场景时的不足, 以及其对训练-测试数据分布偏移的敏感性. 本文讨论了可能的改进方向, 包括模型压缩与量化以降低资源消耗、使用支持流式输入的音频编码器扩展输入音频时长限制、采用课程学习和拒绝采样等方法提升训练稳定性与跨域泛化等.

总之, 本研究为音频-语言模型在说话人日志任务上的应用提供了系统化思路和实践经验, 验证了方案的可行性, 同时也指出了未来需突破的技术难点, 为后续研究提供了参考.

References

[1] Rubenstein PK, Asawaroengchai C, Nguyen DD, *et al*. AudioPaLM: A large language model that can speak and listen. arXiv:2306.12925, 2023.

[2] Chu YF, Xu J, Zhou XH, Yang Q, Zhang SL, Yan ZJ, Zhou C, Zhou JR. Qwen-Audio: Advancing universal audio understanding via unified large-scale audio-language models. arXiv:2311.07919, 2023.

[3] Step-Audio Team. Step-Audio: Unified understanding and generation in intelligent speech interaction. arXiv:2502.11946, 2025.

[4] Peng J, Wang YC, Li BH, Guo YW, Wang HK, Fang YG, Xi Y, Li HY, Li X, Zhang K, Wang S, Yu K. A survey on speech large language models for understanding. arXiv:2410.18908, 2025.

[5] Chu YF, Xu J, Yang Q, Wei HJ, Wei XP, Guo ZF, Leng YC, Lv YJ, He JZ, Lin JY, Zhou C, Zhou JR. Qwen2-Audio technical report. arXiv:2407.10759, 2024.

[6] Shao ZH, Wang PY, Zhu QH, Xu RX, Song JX, Bi X, Zhang HW, Zhang MC, Li YK, Wu Y, Guo DY. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. arXiv:2402.03300, 2024.

[7] von Neumann T, Boeddeker C, Cord-Landwehr T, Delcroix M, Haeb-Umbach R. Meeting recognition with continuous speech separation and transcription-supported diarization. In: Proc. of the 2024 IEEE Int’l Conf. on Acoustics, Speech, and Signal Processing Workshops. Seoul: IEEE, 2024. 775–779. [doi: 10.1109/ICASSPW62465.2024.10625894]

[8] Morrone G, Cornell S, Raj D, Serafini L, Zovato E, Brutti A, Squartini S. Low-latency speech separation guided diarization for telephone conversations. In: Proc. of the 2022 IEEE Spoken Language Technology Workshop (SLT). Doha: IEEE, 2022. 641–646. [doi: 10.1109/SLT54892.2023.10023280]

[9] Gruttadauria E, Fontaine M, Essid S. Online speaker diarization of meetings guided by speech separation. In: Proc. of the 2024 IEEE Int’l Conf. on Acoustics, Speech and Signal Processing (ICASSP). Seoul: IEEE, 2024. 11356–11360. [doi: 10.1109/ICASSP48485.2024.10447682]

[10] Kalda J, Pagés C, Marxer R, Alumäe T, Bredin H. PixIT: Joint training of speaker diarization and speech separation from real-world multi-speaker recordings. In: Proc. of the 2024 Speaker and Language Recognition Workshop (Odyssey 2024). Quebec City, 2024. 115–122. [doi: 10.21437/odyssey.2024-17]

[11] Landini F, Profant J, Diez M, Burget L. Bayesian HMM clustering of x-vector sequences (VBx) in speaker diarization: Theory, implementation and analysis on standard tasks. Computer Speech & Language, 2022, 71: 101254. [doi: 10.1016/j.csl.2021.101254]

[12] Zhu BS, Mao QR, Gao LJ, Shen YX. Temporal-segment-and-regroup clustering for speaker diarization. Application Research of Computers, 2024, 41(9): 2649–2654 (in Chinese with English abstract). [doi: 10.19734/j.issn.1001-3695.2024.01.0017]

[13] Mao QQ, Jia HJ, Zhu BS. Multi-prototype driven graph neural network for speaker diarization. Application Research of Computers, 2025, 42(6): 1778–1783 (in Chinese with English abstract). [doi: 10.19734/j.issn.1001-3695.2024.11.0458]

[14] Kanda N, Xiao X, Gaur Y, Wang XF, Meng Z, Chen Z, Yoshioka T. Transcribe-to-diarize: Neural speaker diarization for unlimited number of speakers using end-to-end speaker-attributed ASR. In: Proc. of the 2022 IEEE Int’l Conf. on Acoustics, Speech and Signal Processing (ICASSP). Singapore: IEEE, 2022. 8082–8086. [doi: 10.1109/ICASSP43922.2022.9746225]

[15] Li YZ, Yu F, Liang YH, Guo PC, Shi MH, Du ZH, Zhang SL, Xie L. Sa-Paraformer: Non-autoregressive end-to-end speaker-attributed ASR. In: Proc. of the 2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). Taipei: IEEE, 2023. 1–7. [doi: 10.1109/ASRU57964.2023.10389762]

[16] Kanda N, Ye GL, Gaur Y, Wang XF, Meng Z, Chen Z, Yoshioka T. End-to-end speaker-attributed ASR with Transformer. arXiv: 2104.02128, 2021.

[17] Wang Q, Huang YL, Zhao GL, Clark E, Xia W, Liao H. DiarizationLM: Speaker diarization post-processing with large language models. In: Proc. of the 25th Annual Conf. of the Int’l Speech Communication Association. Kos, 2024. 3754–3758. [doi: 10.21437/Interspeech.2024-209]

[18] Efstathiadis G, Yadav V, Abbas A. LLM-based speaker diarization correction: A generalizable approach. Speech Communication, 2025,

170: 103224. [doi: 10.1016/j.specom.2025.103224]
[19] Yang M, Kanda N, Wang XF, Chen JK, Wang PD, Xue J, Li JY, Yoshioka T. DiariST: Streaming speech translation with speaker diarization. In: Proc. of the 2024 IEEE Int'l Conf. on Acoustics, Speech and Signal Processing (ICASSP). Seoul: IEEE, 2024. 10866–10870. [doi: 10.1109/ICASSP48485.2024.10446050]
[20] Piñeiro-Martín A, García-Mateo C, Docío-Fernández L, López-Pérez MDC, Rehm G. Weighted cross-entropy for low-resource languages in multilingual speech recognition. In: Proc. of the 25th Annual Conf. of the Int'l Speech Communication Association. Kos, 2024. 1235–1239. [doi: 10.21437/Interspeech.2024-734]
[21] Khare A, Han EJ, Yang YG, Stolcke A. ASR-aware end-to-end neural diarization. In: Proc. of the 2022 IEEE Int'l Conf. on Acoustics, Speech and Signal Processing (ICASSP). Singapore: IEEE, 2022. 8092–8096. [doi: 10.1109/ICASSP43922.2022.9746964]
[22] Kumar S, Madikeri S, Nigmatulina I, Villatoro-Tello E, Motlicek P, Pandia K, Dubagunta SP, Ganapathiraju A. Multitask speech recognition and speaker change detection for unknown number of speakers. In: Proc. of the 2024 IEEE Int'l Conf. on Acoustics, Speech and Signal Processing (ICASSP). Seoul: IEEE, 2024. 12592–12596. [doi: 10.1109/ICASSP48485.2024.10446130]
[23] Li G, Liu JZ, Dinkel H, Niu YD, Zhang JB, Luan J. Reinforcement learning outperforms supervised fine-tuning: A case study on audio question answering. arXiv:2503.11197, 2025.
[24] Nagpal C, Venugopalan S, Tobin J, Ladewig M, Heller K, Tomanek K. Speech recognition with LLMs adapted to disordered speech using reinforcement learning. In: Proc. of the 2025 IEEE Int'l Conf. on Acoustics, Speech and Signal Processing (ICASSP). Hyderabad: IEEE, 2025. 1–5. [doi: 10.1109/ICASSP49660.2025.10888006]
[25] Rafailov R, Sharma A, Mitchell E, Ermon S, Manning CD, Finn C. Direct preference optimization: Your language model is secretly a reward model. In: Proc. of the 37th Int'l Conf. on Neural Information Processing Systems. New Orleans: Curran Associates Inc., 2023. 53728–53741.
[26] Schulman J, Wolski F, Dhariwal P, Radford A, Klimov O. Proximal policy optimization algorithms. arXiv:1707.06347, 2017.
[27] Wang SH, Zhang SY, Zhang J, Hu RY, Li XY, Zhang TW, Li JW, Wu F, Wang GY, Hovy E. Reinforcement learning enhanced LLMs: A survey. arXiv:2412.10400, 2025.
[28] Park T, Medennikov I, Dhawan K, Wang WQ, Huang H, Koluguri NR, Puvvada KC, Balam J, Ginsburg B. Sortformer: A novel approach for permutation-resolved speaker supervision in speech-to-text systems. arXiv:2409.06656, 2025.
[29] Gao ZF, Zhang SL, McLoughlin I, Yan ZJ. Paraformer: Fast and accurate parallel Transformer for non-autoregressive end-to-end speech recognition. arXiv:2206.08317, 2023.
[30] Mao YX, Wang Q, Chen C, Qu Y, Ji XY. Offline reinforcement learning with OOD state correction and OOD action suppression. In: Proc. of the 38th Int'l Conf. on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2025. 93568–93601.
[31] Haider T, Roscher K, da Roza FS, Günnemann S. Out-of-distribution detection for reinforcement learning agents with probabilistic dynamics models. In: Proc. of the 2023 Int'l Conf. on Autonomous Agents and Multiagent Systems. London: Int'l Foundation for Autonomous Agents and Multiagent Systems, 2023. 851–859.
[32] Gulati A, Qin J, Chiu CC, Parmar N, Zhang Y, Yu JH, Han W, Wang SB, Zhang ZD, Wu YH, Pang RM. Conformer: Convolution-augmented Transformer for speech recognition. In: Proc. of the 21st Annual Conf. of the Int'l Speech Communication Association. Shanghai, 2020. 5036–5040.
[33] Zhong SS, Gao SH, Huang ZZ, Wen WS, Zitnik M, Zhou P. MoExtend: Tuning new experts for modality and task extension. In: Proc. of the 62nd Annual Meeting of the Association for Computational Linguistics (Vol. 4: Student Research Workshop). Bangkok: ACL, 2024. 494–505.

附中文参考文献

[12] 朱必松, 毛启容, 高利剑, 沈雅馨. 基于时间分段和重组聚类的说话人日志方法. 计算机应用研究, 2024, 41(9): 2649–2654. [doi: 10.19734/j.issn.1001-3695.2024.01.0017]
[13] 毛青青, 贾洪杰, 朱必松. 面向说话人日志的多原型驱动图神经网络方法. 计算机应用研究, 2025, 42(6): 1778–1783. [doi: 10.19734/j.issn.1001-3695.2024.11.0458]

作者简介

韦舒羽, 博士生, 主要研究领域为多模态大模型.
丘德来, 硕士, 主要研究领域为多模态大模型, 自然语言处理, 智能问答.
刘升平, 博士, CCF 专业会员, 主要研究领域为大语言模型, 自然语言处理, 知识图谱.
桑基韬, 博士, 教授, 博士生导师, CCF 高级会员, 主要研究领域为多模态智能, 可信与对齐, 推理模型, AI Agent.