arXiv:2603.10420v1 [eess.AS] 11 Mar 2026

FireRedASR2S：业界领先的工业级一体化自动语音识别系统

Kaituo Xu、Yan Jia、Kai Huang、Junjie Chen、Wenpeng Li、Kun Liu、Feng-Long Xie、Xu Tang、Yao Hu Super Intelligence Team，小红书

摘要

我们推出了 FireRedASR2S，这是一套业界领先的工业级一体化自动语音识别（ASR）系统。它通过统一的处理流程集成了四个模块：自动语音识别、语音活动检测（VAD）、口语语言识别（LID）和标点预测（Punc）。在所评测的基准测试中，所有模块均达到当前最佳性能：

FireRedASR2：一个包含两个变体的自动语音识别模块：FireRedASR2-LLM（8B+参数）和 FireRedASR2-AED（1B+参数），支持普通话、汉语方言和口音、英语以及语码转换场景下的语音和歌唱转录。与 FireRedASR 相比，FireRedASR2 的识别准确率更高，方言和口音覆盖范围也更广。FireRedASR2-LLM 在 4 个公开普通话基准测试上的平均字符错误率为 2.89%，在 19 个公开汉语方言和口音基准测试上的平均字符错误率为 11.55%，优于包括豆包-ASR、Qwen3-ASR 和 Fun-ASR 在内的有竞争力的基线方法。

FireRedVAD：一个基于深度前馈序列记忆网络（DFSMN）的超轻量级模块（参数量为 0.6M ），支持流式语音活动检测、非流式语音活动检测和多标签语音活动检测（ mVAD）。在 FLEURS-VAD-102 基准测试中，其帧级 F1 达到 97.57%，AUC-ROC 达到 99.60%，性能优于 Silero-VAD、TEN-VAD、FunASR-VAD 和 WebRTC-VAD。

FireRedLID：一个支持 100+ 种语言以及20+汉语方言和口音的编码器-解码器语言识别模块。在 FLEURS（82 种语言）上，其话语级准确率达到 97.18%，性能优于 Whisper 和 SpeechBrain。

FireRedPunc：一个面向中文和英文的 BERT 风格标点预测模块。在多领域基准测试中，其平均 F1 达到 78.90%，优于 FunASR-Punc（62.77%）。

为推动语音处理研究，我们在 https://github.com/FireRedTeam/FireRedASR2S 发布了模型权重和代码。

1 引言

随着端到端建模、大规模训练以及大语言模型（LLM） [1–19] 的集成，自动语音识别（自动语音识别）取得了快速发展。然而，在现实场景中实际部署通常需要的不仅仅是一个独立自动语音识别模型。真实场景音频通常包含长篇录音、静音和非语音区域、背景音乐、歌唱、多语言语音，以及汉语方言和口音（下文简称“方言”）。为了在这些条件下提供可靠的转录文本，需要完整的处理流程，包括用于分段的语音活动检测（语音活动检测）、用于路由和标记的口语语言识别（语言识别），以及用于生成易读输出的标点预测（标点预测）。

在实践中，许多系统通过组装来自不同来源的模块构建而成（例如，独立的语音活动检测/语言识别/自动语音识别/标点预测工具包或云服务）。此类处理流程往往存在接口不一致、可复现性有限以及错误传播复杂等问题。此外，有些组件

依赖薄弱或间接的监督（例如，从自动语音识别强制对齐中训练语音活动检测），这可能会削弱系统在严苛声学条件下的鲁棒性。这些局限促使我们开发一种开源、工业级的一体化自动语音识别系统，兼具出色性能、清晰的模块化设计和全面的评测。

在这份技术报告中，我们介绍 **FireRedASR2S**，这是一套先进的一体化自动语音识别系统，集成了四个模块：用于自动语音识别的 **FireRedASR2**、用于语音活动检测和多标签语音活动检测（mVAD）的 **FireRedVAD**、用于多语言和方言语言识别的 **FireRedLID**，以及用于标点预测的 **FireRedPunc**。后缀 **2S** 表示第二代 FireRedASR，并扩展为一套一体化自动语音识别系**2**统。FireRedASR2S 提供从波形到结构化转录输出的统一处理流程，同时支持独立部署各个模块。FireRedASR2 基于我们此前的 FireRedASR [1]模型构建，架构改动极少。与 FireRedASR 相比，FireRedASR2 提高了识别准确率，并扩大了对更多汉语方言的覆盖范围，这主要得益于监督训练数据规模扩大至约 20 万小时，同时涵盖更丰富的领域、语言和方言。FireRedVAD 使用数千小时经人工高质量标注的声学事件数据进行训练，可在多种声学条件下实现可靠分段。FireRedLID 是一种基于编码器-解码器的 [20, 21]模型，以 FireRedASR2-AED 编码器为初始模型，并进行分层语言与方言预测。FireRedPunc 采用 BERT 风格编码器 [22]，以 LERT [23] 为初始模型，并使用大规模多领域中英文语料库进行训练。

我们的主要贡献如下：

- **一体化开源系统**：我们发布了一个集成式自动语音识别流水线，提供统一接口并支持模块化部署。
- **提升自动语音识别准确率并扩大方言覆盖范围**：在 FireRedASR 的基础上，Fir- eRedASR2 提升了识别准确率并扩大了对汉语方言的支持，在 24 个公开测试集上取得了优异成绩。
- **基于人工标注事件的稳健分段**：FireRedVAD 展现出优异的多语言语音活动检测性能，并采用高质量的人工标注事件数据进行训练，而非使用强制对齐派生监督。
- **分层多语言与方言语言识别**：FireRedLID 支持 100+ 种语言和 20+ 种汉语方言，并采用紧凑的双词元解码形式。
- **有效的标点预测**：FireRedPunc 在多领域中文和英语标点基准测试中取得了优异成绩。

本报告其余部分安排如下：第 2 节介绍系统概述。第 3 至第 6 节介绍各个模块。第 7 节报告评测结果。第 8 节讨论关键设计选择与局限性，第 9 节总结本报告。

2 FireRedASR2S：系统概述

FireRedASR2S 是一款工业级一体化自动语音识别系统，将四个模块——FireRedVAD（第 4 节）、FireRedLID（第 5 节）、FireRedASR2（第 3 节）和 FireRedPunc（第 6 节）——集成到统一的处理流程中。该系统采用模块化设计：每个组件均可独立使用，而默认配置则构成端到端转录流水线，可处理多种声学条件下的内容（语音、歌唱、音乐和非语音），以及多语言和汉语方言场景，并生成结构化输出，包括带标点文本、时间戳、置信度分数和语言标签。

处理流程：如图 1 所示，FireRedASR2S 通过四个阶段处理输入波形。首先，FireRedVAD 在原始音频时间轴上检测语音段，并滤除非语音区域，从而提升其处理长音频时的鲁棒性。其次，FireRedLID 为每个检测到的语音段预测语句级语言标签，并在适用时提供中文方言标签。第三，FireRedASR2 将每个语音段转录为文本，并返回 ASR置信度分数；使用 FireRedASR2-AED 时，还会提供词元级和词级时间戳。最后，FireRedPunc 为 ASR输出恢复标点，以提升可读性和下游使用效果。



图 1：FireRedASR2S 概览。输入波形依次经过 FireRedVAD（语音活动检测）、FireRedLID（语言识别）、FireRedASR2（自动语音识别）和 FireRedPunc（标点预测）处理，生成结构化转录输出，包括带标点文本、时间戳、置信度分数和语言标签。

结构化输出：FireRedASR2S 会返回结构化输出，其中包含：(1) 最终转录文本；(2) 句子级分段列表，其中包括起始/结束时间戳、识别文本、自动语音识别置信度，以及可选的语言标签及其置信度；(3) VAD分段结果。启用 ASR时间戳标注后，系统还可以利用标点预测生成句子级时间戳。所有时间戳均基于原始波形时间轴报告。
模块化：尽管 FireRedASR2S 采用端到端流水线设计，但每个模块都可以作为独立组件部署（例如，仅VAD分段、仅LID路由、对预分段音频进行自动语音识别，或对纯文本进行标点预测）。这种模块化设计支持灵活部署，并便于各组件独立迭代。

3 FireRedASR2：自动语音识别

我们在此总结 FireRedASR2 的关键组件，并着重介绍增量更新。有关 Conformer 和 Transformer 模块的详细规格、编码器-适配器-大语言模型训练流程以及优化策略，请参阅 FireRedASR技术报告 [1]。

FireRedASR2包含两个变体：FireRedASR2-AED和FireRedASR2-LLM。FireRedASR2-AED采用传统的基于注意力机制的编码器-解码器架构 [20, 21], , 而FireRedASR2-LLM则基于编码器-适配器-大语言模型架构 [1–4, 10, 18, 19]，借助大语言模型的能力实现自动语音识别。两个模型的输入特征处理和声学编码策略相似，但词元序列建模方法不同。FireRedASR2-AED还支持词元级时间戳和话语级置信度分数。词级时间戳通过对词元时间戳进行后处理获得（例如，将英语子词单元合并为单词）。

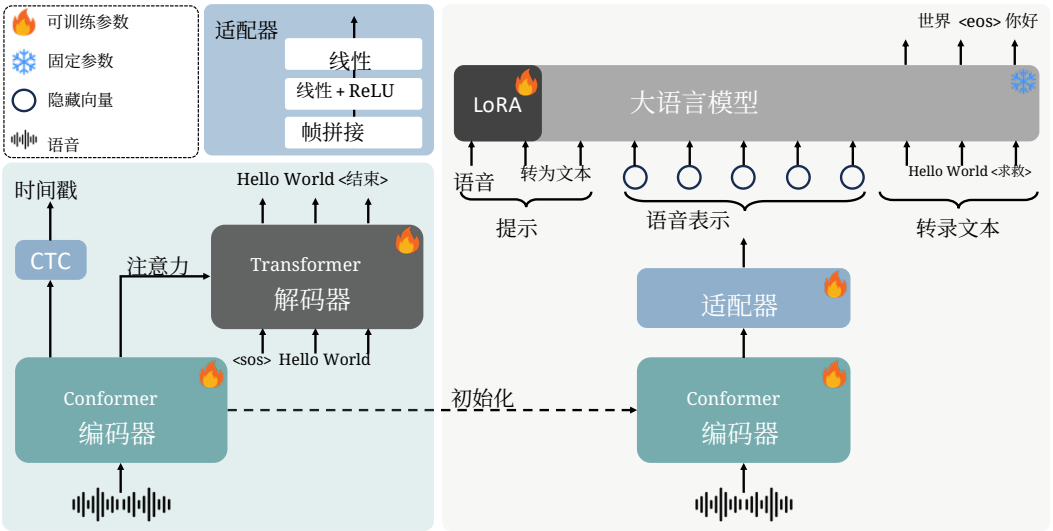


图2：FireRedASR2-AED（左下）、FireRedASR2-LLM（右侧）和适配器的架构。

3.1 FireRedASR2-AED：基于注意力的编码器-解码器自动语音识别模型

FireRedASR2-AED 的整体架构如图2（左下）所示。

架构：FireRedASR2-AED采用端到端自动语音识别架构，遵循FireRedASR-AED [1, 24, 25]的基于Conformer的编码器和基于Transformer的解码器设计。编码器首先通过卷积下采样降低帧率，随后由多个Conformer块堆叠而成。解码器是标准的基于Transformer的解码器，通过关注编码器状态，使用交叉熵目标函数生成词元序列。除非另有说明，架构超参数和训练方案均遵循FireRedASR-AED（L型模型） [1]。

训练数据：与 FireRedASR 相比，FireRedASR2 的主要更新是将监督训练数据从 7 万小时扩充至约 20 万小时。训练语料库涵盖普通话、英语、汉语方言、语码转换、语音和歌唱，以及非语音音频。我们认为，FireRedASR2 的性能提升和泛化能力增强，主要归功于这个规模更大、更多样化的训练语料库。

输入特征：我们使用从 25 毫秒窗口和 10 毫秒帧移中提取的 80 维对数梅尔滤波器组（FBank）特征，并进行全局均值方差归一化（倒谱均值方差归一化）。

分词：FireRedASR2-AED 采用混合分词策略：中文文本使用汉字，英语文本使用词元级字节对编码（BPE） [26] 词元。与 FireRedASR-AED 相比，FireRedASR2-AED 更新了词汇表大小，将其设为 8,667，以更好地覆盖多语言和方言场景。

3.1.1 基于解码器概率的置信度估计

FireRedASR2-AED 会返回一个话语级置信度分数，用于表示转录文本的可靠性。该分数根据解码器的词元概率得出。具体而言，我们提取束搜索生成的一优假设中各词元的后验概率（即 Softmax 输出），并排除特殊令牌。随后，将这些词元级概率汇总为单一序列级分数，通常采用有效词元概率的几何平均数。为提高其在实际部署中的可靠性，还可以使用启发式策略进一步调整这一原始汇总分数（例如，筛除统计离群值或进行置信度截断）。最后，该序列级分数可用于下游筛选、排序或用户界面显示。

3.1.2 用于时间戳的后处理 CTC 分支

FireRedASR2-AED 的一项重要更新是通过在编码器上附加一个 CTC [27]分支来支持时间戳。在基础 AED模型（Conformer 编码器 + Transformer解码器）完成训练后，我们在编码器输出上添加一个轻量级 CTC投影头，并在后处理中冻结编码器和解码器，仅使用标准 CTC 目标函数优化 CTC 分支。CTC 头采用线性投影，将编码器隐藏状态映射为 logits；其 CTC词汇表与 AED 词汇表相同，从而能够对 CTC后验概率与 AED解码词元进行强制对齐。这种设计在保留基础 AED模型识别准确率的同时，实现了基于对齐的时间戳预测。

推理时，我们首先使用 AED 解码器（束搜索）解码出令牌序列。随后，根据编码器状态计算帧级 CTC logits，并在 CTC logits 与 AED 解码得到的令牌序列之间执行 CTC 强制对齐（空白符 ID 设为 0）。最后，根据编码器下采样率，将帧级对齐转换为词元级起止时间。

对于最终系统输出，我们通过对词元时间戳进行后处理来提供词级时间戳。具体而言，我们将子词单元的时间戳合并为单词：将对应的 BPE 词元分组，并取合并后每个单词的最早起始时间和最晚结束时间。对于中文，我们将每个字符词元视为一个词单元。

3.2 FireRedASR2-LLM：基于编码器-适配器-大语言模型的自动语音识别模型

FireRedASR2-LLM 也是一种端到端自动语音识别模型，遵循 FireRedASR-LLM [1]的编码器-适配器-大语言模型框架。该模型由以下部分组成：（1）基于 Conformer 的音频编码器，用于将声学特征转换为高层表示；（2）轻量级适配器，用于将编码器输出映射到预训练文本大语言模型 [28], 的嵌入空间；以及（3）一个自回归

大语言模型，通过执行下一词元预测来生成转录文本。FireRedASR2-LLM 的整体架构如图2（右）所示。
FireRedASR2-LLM 使用与 FireRedASR2-AED 相同的训练数据、输入特征和处理方法。FireRedASR2-LLM 的编码器使用 FireRedASR2-AED 编码器的预训练权重进行初始化。

FireRedASR-LLM 与 FireRedASR2-LLM 的主要区别在于，后者采用了第 3.1 节所述、扩充至 20 万小时的监督训练语料库。除此之外，两者的架构和训练策略保持一致。有关提示格式化、参数高效的大语言模型适配和解码配置等详情，请参阅 [1]。

3.3 与 FireRedASR 的差异总结

表1总结了 FireRedASR2 与 FireRedASR 的主要差异。总体而言，FireRedASR2 保留了 FireRedASR 中经验证的模型设计，同时通过采用规模更大、更多样化的训练语料库来提升泛化能力，并在 AED 变体中通过后置CTC分支实现时间戳生成。

表1：FireRedASR 到 FireRedASR2 的关键更新。

Item	FireRedASR	FireRedASR2
训练数据	~70千小时	~200千小时
词汇表大小（AED）	7,832	8,667
时间戳（AED）	不支持	支持

4 FireRedVAD：语音活动检测

FireRedVAD 为下游自动语音识别提供稳健的分段能力，适用于语音可能与歌唱、背景音乐及各种非语音声学事件共存的真实场景音频。许多工业级语音活动检测方案依赖自动语音识别强制对齐信号，且主要使用自动语音识别语料进行训练；与之不同，FireRedVAD 使用高质量的人工标注声学事件数据进行训练，因此能在复杂声学条件下实现更可靠的检测。

FireRedVAD 包含三种基于 DFSMN 的模型：(1) 用于离线分段的**非流式语音活动检测**模型，(2) 用于低延迟在线分段的**流式语音活动检测**模型，以及 (3) 用于声学事件识别的**非流式多标签语音活动检测 (mVAD)**模型。

4.1 任务与标签定义

多标签语音活动检测 (mVAD)：mVAD 被定义为一项帧级多标签分类任务，针对三种事件后验概率进行分类：语音、歌唱和音乐。mVAD 模型为每种事件输出独立的后验概率，并通过按事件后处理得到事件片段。

语音活动检测（VAD）：VAD 被定义为帧级二分类任务，用于预测有声与无声。我们将有声定义为语音和歌唱的并集，将无声定义为音乐、静音和噪声。此定义符合用户生成内容（UGC）场景中的典型自动语音识别用法，在此类场景中，歌唱片段通常会以与语音类似的方式进行处理。

4.2 训练数据

人工标注事件语料库：我们使用数千小时的人工标注声学事件数据训练 FireRedVAD。每条话语都标注了语音、歌唱和音乐的时间边界。不同于从自动语音识别强制对齐或弱分段启发式方法中获取 VAD 监督的常见做法，FireRedVAD 使用直接的人工标注。

mVAD与语音活动检测的监督：mVAD模型直接使用原始三分类标签。语音活动检测模型使用由同一标注空间派生的二分类标签，将语音和歌唱 映射为正类，并将音乐、静音和噪声 映射为负类。尽管

这些任务共享相关的本体，但mVAD与语音活动检测作为独立模型进行训练，并采用任务特定的目标与后处理准则。

4.3 模型架构

输入特征：FireRedVAD 使用与 FireRedASR2 相同的声学特征（第 3 节）。

DFSMN 主干网络：所有 FireRedVAD 模型均采用深度前馈序列记忆网络（DFSMN）[29, 30], ，该网络适用于帧级声学分类，兼具高效性。我们使用深度一维卷积和膨胀来实现 FSMN [31] 记忆块，以建模时间上下文，并结合残差连接实现稳定优化。

网络配置：我们使用 8 个 DFSMN 块，后接一个额外的前馈层。隐藏层大小为 256，投影层大小为 128。在时间上下文方面，我们采用回看阶 20、步长为 1。对于非流式语音活动检测和 mVAD，我们采用前瞻阶 20、步长为 1，以利用未来上下文改善离线分段。对于流式语音活动检测，我们采用前瞻阶 0，以确保因果推断。我们采用 0.05 的 dropout 率。

模型大小：得益于紧凑的设计，三种 FireRedVAD 模型都极其轻量，每种模型仅包含 ∼0.6M 个参数（以 float32 格式存储时约为 2.2 MB）。这种超轻量的占用可最大限度地减少内存和计算开销，使其非常适合在云服务器上进行大规模并发处理，也适用于资源受限环境下的边缘部署。

输出层：最终分类器通过线性投影将 DFSMN 状态映射为 logits。VAD 使用一维输出（有声与无声），而 mVAD 使用三维输出（语音、歌唱和音乐）。我们应用 Sigmoid激活以获得后验概率。

流式推理：为支持在线语音活动检测，流式模型会维护一个小型逐层缓存，用于存储 FSMN 回看记忆所需的固定长度历史信息。推理期间，模型会增量更新缓存并输出帧后验概率，无需重新处理过去的音频，从而实现低延迟、内存占用有上限的流式处理。

4.4 后处理与分段

DFSMN 模型会生成帧级后验概率，并通过确定性的后处理流程将其转换为时间片段。我们首先使用移动平均滤波平滑后验序列，然后应用概率阈值，得到帧级决策。为抑制局部声学波动导致的异常切换，有限状态后处理器会强制满足有声和静音的最短持续时间约束，从而提升离线和流式场景下的稳定性。还可以通过合并短间隙、延长边界以及拆分过长的语音段来优化片段，从而提升长音频及下游自动语音识别的鲁棒性。

对于 FormVAD，针对每种事件的后验概率流（语音、歌唱和音乐）分别独立应用相同的处理流程，并采用各事件专属的阈值，从而生成各事件对应的时间戳片段。非流式语音活动检测会输出一组带有起止时间戳的语音段；流式语音活动检测会输出增量式帧级决策以及语音起止事件；mVAD 会输出对应于语音、歌唱和音乐的各事件时间戳，以支持下游进行事件感知处理。

5 FireRedLID：分层式口语语言与方言识别

口语语言识别（语言识别） [32–35] 是集多功能自动语音识别系统中多语言及汉语方言语音处理的关键组成部分。在实际部署中，语言识别常用于将话语路由至特定语言的下游处理环节；语言识别中的错误可能会传递至自动语音识别解码和标点预测等后续模块。FireRedLID 旨在适应多样的声学条件，并支持多语言识别和细粒度中文方言识别。

5.1 模型与训练

架构：FireRedLID采用基于编码器-解码器的架构，包含一个Conformer编码器和一个Transformer解码器，其实现方式遵循我们基于注意力的编码器-解码器自动语音识别模型的风格。给定一个

输入语音，编码器会生成声学表示，解码器则生成一个表示语言识别结果的短令牌序列。

初始化：Conformer 编码器基于预训练的 FireRedASR2-AED 编码器（第 3.1 节）初始化，以利用其大规模自动语音识别表征学习能力。语言识别解码器则采用随机初始化，并从头开始训练。

输入特征：FireRedLID 使用与 FireRedASR2 相同的声学特征（第 3 节）。
训练数据：FireRedLID 使用约 20 万小时的多语言语音进行训练，涵盖100+ 种语言，包括普通话和 20+ 种汉语方言。数据经过筛选，涵盖多种领域和声学条件，以提升泛化能力。

训练目标：FireRedLID 使用教师强制，采用标准的序列到序列交叉熵目标进行训练。

5.2 层次化标签空间与解码

两级标签：FireRedLID 将语言识别建模为两级层次结构。第一级预测语言（例如 zh、en、ja、ko 等）。当预测语言为中文（zh）时，模型还会预测中文方言标签（例如普通话、粤语、吴语、闽语、湘语等）。这种设计反映了自然的标签结构，并通过基于粗粒度语言判定进行方言预测，提高了方言识别的稳定性。
短序列令牌预测：我们将分层语言识别建模为短序列生成任务，最大解码长度为 2。实际运行中，模型会先输出语言令牌，通常还会在生成 <eos> 之前输出第二个令牌，表示汉语方言。对于非中文话语，解码器通常在预测语言令牌后输出 <eos> 并终止。这种建模方式使标签序列保持紧凑，并且与扁平标签空间相比，减少了歧义。

解码与置信度：推理期间，我们使用束搜索解码标签令牌序列。由于输出长度最多为 2 个令牌，解码开销可以忽略不计。我们报告最优假设，并将解码后标签令牌（不包括 <sos> 和 <eos> 等特殊令牌）的后验概率均值作为话语级置信度。

5.3 支持的语言和方言

标签覆盖范围：FireRedLID 支持 100+种语言和 20+种汉语方言。我们使用简洁的语言代码（例如 zh、en、ja、ko）表示语言，并将 20+ 种汉语方言划分为 8 个不同的地理或语言方言簇（例如普通话、粤语、吴语、闽语等）。附录 B 列出了支持的语言和方言的完整清单。

6 FireRedPunc：标点预测

FireRedPunc 为自动语音识别转录文本添加标点，以提高可读性和下游可用性（例如字幕显示和机器翻译）。它面向开放领域自动语音识别输出，支持中文和英语标点预测。

架构：FireRedPunc 采用 BERT 风格编码器 [22]和词元级分类头。给定一个输入令牌序列，模型会为每个词元预测一个标点标签，指示应在该词元后插入的标点符号。我们使用预训练的 LERT 检查点 [23], 初始化编码器。LERT 是一种受语言学原理启发的 BERT 变体，我们会针对标点预测对其进行微调。

标点集：我们采用精简的五分类标点集，包含无标点以及四种标点符号：，。？！。在实现中，中文文本使用中文全角标点符号。这种设计涵盖了自动语音识别应用中最常见的标点符号，同时保持分类器简单稳定。

训练数据：FireRedPunc 使用带有标点标注的大规模多领域文本语料库进行训练。训练数据包含约 18.57B 个汉字和 2.20B 个英语单词，涵盖不同领域和写作风格，以提升模型对类似自动语音识别的输入的泛化能力。

训练目标：我们使用标准的词元级交叉熵目标函数训练 FireRedPunc。

推理：在推理时，我们使用与预训练 LERT 编码器相同的分词器对 ASR 输出进行分词，并应用模型获取词元级标点标签。最终的带标点文本通过将预测的标点符号插入原始文本序列中生成。

7 评测

本节通过报告自动语音识别、语音活动检测、语言识别和标点预测各模块的结果，在公开基准上评测 FireRedASR2S。各模块均单独评测，以避免上游或下游组件带来的混杂影响。除非另有说明，所有结果均在非流式设置下获得。

7.1 FireRedASR2评测

我们在 24 个公开测试集上评测 FireRedASR2，这些测试集涵盖普通话、汉语方言和歌唱歌词识别。FireRedASR2 包含两个变体：FireRedASR2-LLM（8B+ 个参数）和 FireRedASR2-AED（1B+ 个参数），分别代表准确率与效率权衡中的不同取舍。

指标：中文采用字符错误率（CER，%）进行评估，数值越低越好。对于汇总结果，我们报告：(1) Avg-All-24：全部 24 个测试集的平均字符错误率；(2) Avg-Mandarin-4：4 个普通话测试集的平均字符错误率；(3) Avg-Dialect-19：19 个汉语方言测试集的平均字符错误率；(4) Sing-1：歌唱歌词测试集（opencpop）上的字符错误率。所有平均字符错误率均按测试集进行宏平均（每个测试集权重相等）。

测试集：Avg-Mandarin-4 包含 AISHELL-1 测试集 [36] (AISHELL-1)、AISHELL-2 iOS 测试集[37] (AISHELL-2)，以及 WenetSpeech 的 [38]互联网/会议领域（ws-net/ws-meeting）。Avg-Dialect-19 包含 KeSpeech [39]，以及从 WenetSpeech-Yue [40],、WenetSpeech-Chuan [41], 和 MagicData [42] 整理的方言测试集（完整列表见附录A）。Sing-1 使用 opencpop [43] 进行歌唱歌词识别。

基线方法：我们将其与表现强劲的商业和开源基线方法进行比较：(1) 豆包-ASR（商业 API） [4, 44], (2) Qwen3-ASR（开源检查点） [2], (3) Fun-ASR（商业 API） [3, 45], (4) Fun-ASR-Nano（开源检查点；表现不如 Fun-ASR API；结果见附录） [3]。需要强调的是，基于 API 的基线方法可能会因服务器端更新而随时间变化，也可能包含专有系统级组件。我们报告其结果是为了提供实用的参考，而非作为严格可复现的基线。各测试集的完整结果见附录A。

表2：FireRedASR2-LLM（FRASR2-LLM）、FireRedASR2-AED（FRASR2-AED）与其他大型自动语音识别基线方法在公开自动语音识别测试集上的字符错误率（CER%）对比。

测试集\ 模型	FRASR2-LLM	FRASR2-AED	豆包-ASR	Qwen3-ASR	Fun-ASR
Avg-All-24	**9.67**	9.80	12.98	10.12	10.92
Avg-Mandarin-4	**2.89**	3.05	3.69	3.76	4.16
Avg-Dialect-19	**11.55**	11.67	15.39	11.85	12.76
Sing-1	**1.12**	1.17	4.36	2.57	3.05
AISHELL-1	0.64	0.57	1.52	1.48	1.64
AISHELL-2	2.15	2.51	2.77	2.71	2.38
ws-net	4.44	4.57	5.73	4.97	6.85
ws-meeting	4.32	4.53	4.74	5.88	5.78

API基线系统： 豆包-ASR（volc.seedasr.auc）于2026年2月上旬进行评测，Fun-ASR于2025年11月下旬进行评测。API结果可能会因服务器端更新而随时间变化，也可能包含专有组件。为确保比较公平，在API提供相应选项时，我们会在输出中关闭逆文本正则化和标点预测，并采用各API提供的默认语音活动检测配置。**数据重叠：** 我们的自动语音识别训练数据不包含任何来自MagicData的汉语方言或带口音语音数据；所有MagicData方言数据集仅用于评测。

结果与分析：表2汇总了主要结果。FireRedASR2-LLM在所有汇总指标上均取得最佳整体准确率：普通话（Avg- Mandarin-4）的平均字符错误率为2.89%，汉语方言（Avg-Dialect-19）为11.55%，Avg-All-24为9.67%。FireRedASR2在歌唱歌词识别方面也表现出色：在opencpop上，FireRedASR2-LLM的字符错误率为1.12%。FireRedASR2-AED以更小的模型尺寸取得了有竞争力的准确率，为实际部署提供了更均衡的选择。各测试集的详细结果见附录A。

7.2 FireRedVAD评测

任务与标签定义：我们在二元语音活动检测任务（有声与无声）上评测 FireRedVAD。该评测侧重于语音是否存在，并与自动语音识别分段中典型的语音活动检测用法保持一致。

指标：我们报告 AUC-ROC、F1分数、虚警率（FAR）和漏检率（MR）。AUC-ROC 不依赖阈值，而 F1 值、虚警率和漏检率取决于决策阈值。

测试集：我们在 FLEURS-VAD-102 上进行评测，这是一个覆盖 102 种语言的多语言语音活动检测基准。该基准通过从 FLEURS测试集的每种语言中抽取约 100 个音频文件，并手动标注二元语音活动检测标签构建而成，共包含 9,443 个音频文件。我们将发布 FLEURS-VAD-102 及其标注协议，以促进可复现研究。

帧设置：所有语音活动检测指标均在帧级计算。我们使用 25 毫秒分析窗和 10 毫秒帧移，与 FireRedASR2 使用的特征提取设置一致。

基线方法与工作点：我们将其与广泛使用的开源语音活动检测系统进行比较，包括 Silero-VAD [46], TEN-VAD [47], FunASR-VAD [30], 和 WebRTC-VAD [48]。对于依赖阈值的指标（F1值/虚警率/漏报率），我们对所有神经网络语音活动检测模型使用固定的 0.5 后验阈值，以确保工作点一致；在开发集上调整阈值可能会进一步提升各系统的 F1值。

表3：FLEURS-VAD-102 上帧级语音活动检测性能。AUC-ROC 和 F1值越高越好；虚警率和漏报率越低越好。

指标 \模型	FireRedVAD	Silero-VAD	TEN-VAD	FunASR-VAD	WebRTC-VAD
AUC-ROC (%) ↑	**99.60**	97.99	97.81	–	–
F1值 (%) ↑	**97.57**	95.95	95.19	90.91	52.30
虚警率 (%) ↓	**2.69**	9.41	15.47	44.03	2.83
漏检率 (%) ↓	3.62	3.95	2.95	0.42	64.15

由于 FunASR-VAD 和 WebRTC-VAD 不会输出与阈值无关的评估所需的连续后验概率，因此未报告其 AUC-ROC。

结果与分析：如表3所示，FireRedVAD 在多语言语音活动检测方面表现出色，AUC-ROC 达到 99.60%，F1值达到 97.57%，优于所有对比基线方法。值得注意的是，FireRedVAD 以极小的参数规模（∼0.6M）实现了当前最佳性能，展现出准确率与效率之间的出色平衡，适用于实际工业处理流程。FireRedVAD 在保持较低漏检率（3.62%）的同时，虚警率也较低（2.69%），说明其工作点均衡，适合下游分段。我们注意到，部分基线方法（如 FunASR-VAD）虽然漏检率很低，但代价是虚警率显著偏高，这在实际部署中可能导致过度分段，并增加不必要的下游自动语音识别计算。

7.3 FireRedLID 评测

测试集：我们在多语言和汉语方言语言识别基准测试上评测 FireRedLID。对于多语言语言识别，我们报告了 FLEURS [49] 测试集（82 种语言）和 CommonVoice[50] 测试集（74 种语言）上的结果。对于汉语方言识别，我们通过直接合并 KeSpeech 和 MagicData 的测试样本，构建了一个综合基准测试，涵盖 10+ 种汉语方言。

指标：我们报告话语级语言识别准确率（%）。越高越好。

基线方法：我们与 Whisper [51] 语言识别、SpeechBrain 语言识别模型 [52],以及 Dolphin [53]进行比较。

表 4：公开测试集上的话语级语言识别准确率（%）。数值越高越好。

测试集 \ 模型	FireRedLID	Whisper	SpeechBrain	Dolphin
FLEURS 测试集	**97.18**	79.41	92.91	–
CommonVoice 测试集	**92.07**	80.81	78.75	–
汉语方言	**88.47**	–	–	69.01

结果与分析：表 4 显示，FireRedLID 在多语言和中文方言语言识别任务中均表现出色。在 FLEURS 上，FireRedLID 的准确率达到 97.18%，显著优于 Whisper，且表现超过 SpeechBrain。在 CommonVoice 上，FireRedLID 的准确率也优于所有参与比较的系统。在综合中文方言基准上，FireRedLID 的准确率达到 88.47%，证明了层次化标签建模对于细粒度中文方言识别的有效性。

7.4 FireRedPunc 评测

测试集：我们在内部多领域中文和英语标点预测基准测试上评测 FireRedPunc。中文基准测试包含 88,644 个句子，英语基准测试包含 28,641 个句子。我们将发布中文和英语标点基准测试，以促进可复现研究。

标签集和指标：FireRedPunc 从一个紧凑的标点标签集（空格/无标点、逗号、句号、问号和感叹号；见第 6 节）中预测标点标签。评测时，我们报告基于标点标签计算的精确率/召回率/F1值（%）。根据我们的评测协议，Overall 分数是在所有非空格标点标签上计算的微平均精确率/召回率/F1值。由于感叹号在评测数据中的出现频率极低，我们在报告的 Overall 分数中排除感叹号。

Bas基线：我们将其与广泛使用的标点模型 FunASR-Punc（CT-Transformer） [30 进行比较].

表 5：内部测试集上的标点预测结果（精确率/召回率/F1值，单位为 %）。数值越高越好。

测试集 \ 模型	FireRedPunc	FunASR-Punc
多领域中文	**82.84 / 83.08 / 82.96**	77.27 / 74.03 / 75.62
多领域英语	**78.40 / 71.57 / 74.83**	55.79 / 45.15 / 49.91
平均 F1	**78.90**	62.77

结果与分析：如表 5 所示，FireRedPunc 在中文和英语基准测试中均持续优于基线。具体而言，FireRedPunc 在中文上的 F1值达到 82.96%，在英语上达到 74.83%，平均 F1 为 78.90%。英语上的显著提升表明，我们采用 LERT 初始化的 BERT 风格编码器和大规模多领域训练数据，能够有效地为类似 ASR 的文本进行标点预测。

8 讨论

我们将讨论 FireRedASR2S 的关键设计选择和实际应用考量。

系统设计：模块化与统一接口：FireRedASR2S 采用模块化处理流程设计，由语音活动检测、语言识别、自动语音识别和标点预测组成。这种设计简化了部署和维护，支持各组件独立迭代，并且与临时整合异构模块相比，提高了可复现性。

通过数据扩展提升自动语音识别准确率和方言覆盖范围：FireRedASR2基本保留了FireRedASR中经过验证的模型架构，重点将监督训练数据扩展至约20万小时，并覆盖更广泛的场景。在普通话

基准测试中持续取得的改进，以及在方言测试集上的出色表现表明，扩大监督数据多样性是提升识别准确率并增强对多样化汉语方言场景泛化能力的主要驱动力。

用于分段的人工标注事件监督：与基于自动语音识别强制对齐派生监督训练的语音活动检测模型相比，FireRedVAD 在数千小时人工标注的声学事件数据上进行训练。这种显式事件监督提升了在多样声学条件下的鲁棒性，并同时支持语音活动检测与 mVAD 的使用场景。

面向语言与汉语方言的层级语言识别：FireRedLID 将语言识别建模为带有层级标签的短序列生成任务，先预测语言，再在中文条件下预测方言。与扁平标签空间相比，该建模方式更契合标签结构并降低歧义，同时保持推理高效。

9 结论

我们介绍了 FireRedASR2S，这是一套先进的工业级一体化语音识别系统，集成了自动语音识别、语音活动检测、语言识别和标点预测模块。在 FireRedASR 的基础上，FireRedASR2 提升了识别准确率，并将覆盖范围扩展到更广泛的汉语方言，同时提供两种变体：基于大语言模型的模型（8B+ 参数）以实现最高准确率，以及基于 AED 的模型（1B+参数）以实现均衡的准确率与效率的权衡。FireRedVAD 提供稳健的分段，并在多语言语音活动检测上取得出色表现。FireRedLID 以高准确率执行层级语言与汉语方言识别。FireRedPunc 为中文和英语恢复标点，并在多领域基准上取得出色表现。我们发布模型权重与代码，以促进研究与实际部署。未来工作将聚焦于进一步提升性能，并扩展对更多语言的支持。

参考文献

[1] Kai-Tuo Xu、Feng-Long Xie、Xu Tang 和 Yao Hu。FireRedASR：从编码器-解码器到大语言模型集成的开源工业级普通话语音识别模型。*arXiv* 预印本 *arXiv:2501.14350*，2025年。[2] Xian Shi、Xiong Wang、Zhifang Guo、Yongqi Wang、Pei Zhang、Xinyu Zhang、Zishan Guo、Hongkun Hao、Yu Xi、Baosong Yang 等。Qwen3-ASR 技术报告。*arXiv* 预印本 *arXiv:2601.21337*，2026年。[3] Keyu An、Yanni Chen、Zhigao Chen、Chong Deng、Zhihao Du、Changfeng Gao、Zhifu Gao、Bo Gong、Xiangang Li、Yabin Li 等。Fun-ASR 技术报告。*arXiv* 预印本 *arXiv:2509.12508*，2025年。[4] Seed-ASR（2024）。Seed-ASR：利用基于大语言模型的语音识别理解多样化语音与上下文。*arXiv* 预印本 *arXiv:2407.04675*，2024年。[5] Wenjie Tian、Bingshen Mu、Guobin Ma、Xuelong Geng、Zhixian Zhao 和 Lei Xie。DLLM-ASR：一种更快速的基于扩散大语言模型的语音识别框架。*arXiv* 预印本 *arXiv:2601.17902*，2026年。[6] Bingshen Mu、Yiwen Shao、Kun Wei、Dong Yu 和 Lei Xie。基于大语言模型的自动语音识别的高效扩展。*arXiv* 预印本 *arXiv:2508.04096*，2025年。[7] Keyu An、Qian Chen、Chong Deng、Zhihao Du、Changfeng Gao、Zhifu Gao、Yue Gu、Ting He、Hangrui Hu、Kai Hu 等。FunAudioLLM：用于人类与大语言模型自然交互的语音理解与生成基础模型。*arXiv* 预印本 *arXiv:2407.04051*，2024年。[8] Yunfei Chu、Jin Xu、Qian Yang、Haojie Wei、Xipin Wei、Zhifang Guo、Yichong Leng、Yuan-jun Lv、Jinzheng He、Junyang Lin 等。Qwen2-Audio 技术报告。*arXiv* 预印本 *arXiv:2407.10759*，2024年。[9] Yunfei Chu、Jin Xu、Xiaohuan Zhou、Qian Yang、Shiliang Zhang、Zhijie Yan、Chang Zhou 和 Jingren Zhou。Qwen-Audio：通过统一的大规模音频-语言模型推进通用音频理解。*arXiv* 预印本 *arXiv:2311.07919*，2023年。

[10] Jian Wu、Yashesh Gaur、Zhuo Chen、Long Zhou、Yimeng Zhu、Tianrui Wang、Jinyu Li、Shujie Liu、Bo Ren、Linquan Liu 等。关于用于语音转文本和大语言模型集成的仅解码器架构。载于自动语音识别与理解研讨会（*ASRU*），第 1–8 页。IEEE，2023年。[11] Paul K Rubenstein、Chulayuth Asawaroengchai、Duc Dung Nguyen、Ankur Bapna、Zalán Borsos、Félix de Chaumont Quitry、Peter Chen、Dalia El Badawy、Wei Han、Eugene Kharitonov 等。Audiopalm：一种能够说话和聆听的大语言模型。*arXiv* 预印本 *arXiv:2306.12925*，2023年。[12] Yuang Li、Yu Wu、Jinyu Li 和 Shujie Liu。提示大语言模型实现语音识别中的零样本领域自适应。载于自动语音识别与理解研讨会（*ASRU*），第 1–8 页。IEEE，2023年。[13] Mingqiu Wang、Wei Han、Izhak Shafran、Zelin Wu、Chung-Cheng Chiu、Yuan Cao、Nanxin Chen、Yu Zhang、Hagen Soltau、Paul K Rubenstein 等。SLM：弥合语音与文本基础模型之间的细微差距。载于自动语音识别与理解研讨会（*ASRU*），第 1–8 页。IEEE，2023年。[14] Jing Pan、Jian Wu、Yashesh Gaur、Sunit Sivasankaran、Zhuo Chen、Shujie Liu 和 Jinyu Li。COSMIC：面向语音上下文学习的数据高效指令微调。*arXiv* 预印本 *arXiv:2311.02248*，2023年。[15] Wenyi Yu、Changli Tang、Guangzhi Sun、Xianzhao Chen、Tian Tan、Wei Li、Lu Lu、Zejun Ma 和 Chao Zhang。连接语音编码器与大语言模型以实现自动语音识别。载于国际声学、语音与信号处理会议（*ICASSP*），第 12637–12641 页。IEEE，2024年。[16] Zhehuai Chen、He Huang、Andrei Andrusenko、Oleksii Hrinchuk、Krishna C Puvvada、Jason Li、Subhankar Ghosh、Jagadeesh Balam 和 Boris Ginsburg。SALM：通过上下文学习实现语音识别与翻译的语音增强语言模型。载于国际声学、语音与信号处理会议（*ICASSP*），第 13521–13525 页。IEEE，2024年。[17] Egor Lakomkin、Chunyang Wu、Yassir Fathullah、Ozlem Kalinli、Michael L Seltzer 和 Christian Fuegen。利用大语言模型实现端到端语音识别上下文化。载于国际声学、语音与信号处理会议（*ICASSP*），第 12406–12410 页。IEEE，2024年。[18] Xuelong Geng、Tianyi Xu、Kun Wei、Bingshen Mu、Hongfei Xue、He Wang、Yangze Li、Pengcheng Guo、Yuhang Dai、Longhao Li 等。揭示基于大语言模型的自动语音识别在中文开源数据集上的潜力。载于第*14*届中文口语语言处理国际研讨会（*ISCSLP*），第 26–30 页。IEEE，2024年。[19] Ziyang Ma、Guanrou Yang、Yifan Yang、Zhifu Gao、Jiaming Wang、Zhihao Du、Fan Yu、Qian Chen、Siqi Zheng、Shiliang Zhang 等。一种具备强大自动语音识别能力的大语言模型的极其简单方法。*arXiv* 预印本 *arXiv:2402.08846*，2024年。[20] Dzmitry Bahdanau、Jan Chorowski、Dmitriy Serdyuk、Philemon Brakel 和 Yoshua Bengio。端到端基于注意力的大词汇量语音识别。载于国际声学、语音与信号处理会议（*ICASSP*），第 4945–4949 页。IEEE，2016年。[21] William Chan、Navdeep Jaitly、Quoc Le 和 Oriol Vinyals。听、关注与拼写：一种用于大词汇量会话语音识别的神经网络。载于国际声学、语音与信号处理会议（*ICASSP*），第 4960–4964 页。IEEE，2016年。[22] Jacob Devlin、Ming-Wei Chang、Kenton Lee 和 Kristina Toutanova。BERT：用于语言理解的深度双向变换器预训练。载于*2019*年北美计算语言学协会会议：人类语言技术论文集（*NAACL-HLT*），第 4171–4186 页。计算语言学协会，2019年。

[23] Yiming Cui、Wanxiang Che、Shijin Wang 和 Ting Liu。LERT：一种受语言学启发的预训练语言模型。*arXiv* 预印本 *arXiv:2211.05344*，2022年。[24] Anmol Gulati、James Qin、Chung-Cheng Chiu、Niki Parmar、Yu Zhang、Jiahui Yu、Wei Han、Shibo Wang、Zhengdong Zhang、Yonghui Wu 等。Conformer：用于语音识别的卷积增强型 Transformer。*arXiv* 预印本 *arXiv:2005.08100*，2020年。[25] A Vaswani。Attention is all you need。神经信息处理系统进展，2017年。[26] Rico Sennrich。使用子词单元对罕见词进行神经机器翻译。*arXiv* 预印本 *arXiv:1508.07909*，2015年。[27] Alex Graves、Santiago Fernández、Faustino Gomez 和 Jürgen Schmidhuber。连接时序分类：使用循环神经网络为未分段序列数据标注。载于第*23*届国际机器学习会议论文集，第369–376页，2006年。[28] Qwen团队。Qwen2 技术报告。*arXiv* 预印本 *arXiv:2407.10671*，2024年。[29] Shiliang Zhang、Ming Lei、Zhijie Yan 和 Lirong Dai。面向大词汇量连续语音识别的 Deep-fsmn。载于国际声学、语音与信号处理会议（*ICASSP*），第5869–5873页。IEEE，2018年。[30] Zhifu Gao、Zerui Li、Jiaming Wang、Haoneng Luo、Xian Shi、Mengzhe Chen、Yabin Li、Lingyun Zuo、Zhihao Du、Zhangyu Xiao 和 Shiliang Zhang。FunASR：基础端到端语音识别工具包。载于*INTERSPEECH*，2023年。[31] Shiliang Zhang、Hui Jiang、Shifu Xiong、Si Wei 和 Li-Rong Dai。面向大词汇量连续语音识别的紧凑型前馈序列记忆网络。载于*Interspeech*，第3389–3393页，2016年。[32] Jörgen Valk 和 Tanel Alumäe。Voxlingua107：一个用于口语语言识别的数据集。载于*2021*年 *IEEE*口语语言技术研讨会（*SLT*），第652–658页。IEEE，2021年。[33] Irshad Ahmad Thukroo、Rumaan Bashir 和 Kaiser J Giri。口语语言识别中的深度学习技术综述。多媒体工具与应用，81(22)：32593–32624，2022年。[34] Adal A Alashban、Mustafa A Qamhan、Ali H Meftah 和 Yousef A Alotaibi。使用卷积循环神经网络的口语语言识别系统。应用科学，12(18)：9181，2022年。[35] Douglas O'Shaughnessy。口语语言识别：过去与当前研究趋势概述。语音通信，167：103167，2025年。[36] Hui Bu、Jiayu Du、Xingyu Na、Bengu Wu 和 Hao Zheng。AISHELL-1：一个开源普通话语音语料库及语音识别基线。载于国际语音数据库与语音输入/输出系统及评估协调委员会东方分会第*20*届会议（*O-COCOSDA*），第1–5页。IEEE，2017年。[37] Jiayu Du、Xingyu Na、Xuechen Liu 和 Hui Bu。AISHELL-2：将普通话自动语音识别研究拓展至工业级规模。*arXiv* 预印本 *arXiv:1808.10583*，2018年。[38] Binbin Zhang、Hang Lv、Pengcheng Guo、Qijie Shao、Chao Yang、Lei Xie、Xin Xu、Hui Bu、Xiaoyu Chen、Chenchen Zeng 等。WenetSpeech：一个包含 10000+ 小时语音数据的多领域普通话语料库。载于国际声学、语音与信号处理会议（*ICASSP*），第6182–6186页。IEEE，2022年。[39] Zhiyuan Tang、Dong Wang、Yanguang Xu、Jianwei Sun、Xiaoning Lei、Shuaijiang Zhao、Cheng Wen、Xingjun Tan、Chuandong Xie、Shuran Zhou 等。KeSpeech：一个包含普通话及其八种次方言的开源语音数据集。载于神经信息处理系统会议数据集与基准赛道第三十五届会议（第二轮），2021年。

[40] Longhao Li、Zhao Guo、Hongjie Chen、Yuhang Dai、Ziyu Zhang、Hongfei Xue、Tianlun Zuo、Chengyou Wang、Shuiyuan Wang、Jie Li 等。WenetSpeech-Yue：一个带有多维标注的大规模粤语语音语料库。*arXiv* 预印本 *arXiv:2509.03959*，2025年。[41] Yuhang Dai、Ziyu Zhang、Shuai Wang、Longhao Li、Zhao Guo、Tianlun Zuo、Shuiyuan Wang、Hongfei Xue、Chengyou Wang、Qing Wang 等。WenetSpeech-Chuan：一个带有丰富标注、用于方言语音处理的大规模四川话语料库。*arXiv* 预印本 *arXiv:2509.18004*，2025年。[42] Magic Data。开源自动语音识别语料库。https://magichub.com/datasets，2026年。[43] Yu Wang、Xinsheng Wang、Pengcheng Zhu、Jie Wu、Hanzhao Li、Heyang Xue、Yongmao Zhang、Lei Xie 和 Mengxiao Bi。opencpop：一个用于歌声合成的高质量开源中文流行歌曲语料库。*arXiv* 预印本 *arXiv:2201.07429*，2022年。[44]火山引擎。豆包-ASR。https://www.volcengine.com/docs/6561/ 1354868，2026年。[45]阿里云。Fun-ASR。https://help.aliyun.com/zh/model-studio/ recording-file-recognition，2026年。[46] Silero 团队。Silero 语音活动检测器：预训练的企业级语音活动检测器（语音活动检测）、数字检测器和语言分类器。https://github.com/snakers4/silero-vad，2024年。[47] TEN 团队。TEN 语音活动检测器：一种低延迟、轻量且高性能的流式语音活动检测器（语音活动检测）。https://github.com/TEN-framework/ten-vad.git，2025年。[48] Google。WebRTC：面向 Web 的实时通信。https://webrtc.org，2026年。[49] Alexis Conneau、Min Ma、Simran Khanuja、Yu Zhang、Vera Axelrod、Siddharth Dalmia、Jason Riesa、Clara Rivera 和 Ankur Bapna。FLEURS：通用语音表征的少样本学习评测。*arXiv* 预印本 *arXiv:2205.12446*，2022年。[50] Rosana Ardila、Megan Branson、Kelly Davis、Michael Kohler、Josh Meyer、Michael Henretty、Reuben Morais、Lindsay Saunders、Francis Tyers 和 Gregor Weber。Common Voice：一个大规模多语言语音语料库。载于第十二届语言资源与评测会议论文集，第4218–4222页，2020年。[51] Alec Radford、Jong Wook Kim、Tao Xu、Greg Brockman、Christine McLeavey 和 Ilya Sutskever。通过大规模弱监督实现鲁棒语音识别。载于国际机器学习会议，第28492–28518页。PMLR，2023年。[52] Mirco Ravanelli、Titouan Parcollet、Peter Plantinga、Aku Rouhe、Samuele Cornell、Loren Lugosch、Cem Subakan、Nauman Dawalatabad、Abdelwahab Heba、Jianyuan Zhong、Ju-Chieh Chou、Sung-Lin Yeh、Szu-Wei Fu、Chien-Feng Liao、Elena Rastorgueva、François Grondin、William Aris、Hwidong Na、Yan Gao、Renato De Mori 和 Yoshua Bengio。SpeechBrain：通用语音工具包，2021年。arXiv:2106.04624。[53] Yangyang Meng、Jinpeng Li、Guodong Lin、Yu Pu、Guanbo Wang、Hu Du、Zhiming Shao、Yukai Huang、Ke Li 和 Wei-Qiang Zhang。Dolphin：面向东部语言的大规模自动语音识别模型。*arXiv* 预印本 *arXiv:2503.20212*，2025年。

附录

公开测试集上的详细自动语音识别结果

本附录报告了第7.1节中使用的全部24个公开测试集上各测试集的字符错误率（%）。为完整起见，我们纳入了 FunAudioLLM 发布的开源检查点 Fun-ASR-Nano。

表6：FireRedASR2-LLM（FRASR2-LLM）、FireRedASR2-AED（FRASR2-AED）与其他大型自动语音识别基线方法在公开自动语音识别测试集上的字符错误率（%）比较。

测试集 \ 模型	FRASR2-LLM	FRASR2-AED	豆包-ASR	Qwen3-ASR	Fun-ASR	Fun-Nano
Avg-Mandarin-4	**2.89**	3.05	3.69	3.76	4.16	4.55
Avg-Dialect-19	**11.55**	11.67	15.39	11.85	12.76	15.07
Avg-All-24	**9.67**	9.80	12.98	10.12	10.92	12.81
AISHELL-1	0.64	0.57	1.52	1.48	1.64	1.96
AISHELL-2	2.15	2.51	2.77	2.71	2.38	3.02
ws-net	4.44	4.57	5.73	4.97	6.85	6.93
ws-meeting	4.32	4.53	4.74	5.88	5.78	6.29
KeSpeech	3.08	3.60	5.38	5.10	5.36	7.66
WenetSpeech粤语短音频	5.14	5.15	10.51	5.82	7.34	8.82
WenetSpeech粤语长音频	8.71	8.54	11.39	8.85	10.14	11.36
WenetSpeech四川话简单集	10.90	10.60	11.33	11.99	12.46	14.05
WenetSpeech四川话困难集	20.71	21.35	20.77	21.63	22.49	25.32
MagicData重口音	7.42	7.43	7.69	8.02	9.13	9.97
MagicData粤语对话	12.23	11.66	26.25	9.76	33.71	15.68
MagicData粤语日常	3.61	3.35	12.82	3.66	2.69	5.67
MagicData粤语车载	4.50	4.83	8.66	4.28	6.00	7.04
MagicData四川话对话	13.18	13.07	11.77	14.35	14.01	17.11
MagicData四川话日常	4.90	5.17	3.90	4.93	3.98	5.95
MagicData上海话对话	28.70	27.02	45.15	29.77	25.49	37.08
MagicData上海话日常	24.94	24.18	44.06	23.93	12.55	28.77
吴语	7.15	7.14	7.70	7.57	10.63	10.56
MagicData 普通话会话语音	10.20	10.65	9.83	9.55	10.85	13.09
MagicData 普通话日常语音	5.80	6.26	5.77	5.88	6.29	8.18
武汉话	9.60	10.81	9.94	10.22	4.34	8.70
天津话	15.45	15.30	15.79	16.16	19.27	22.03
长沙话	23.18	25.64	23.76	23.70	25.66	29.23
opencpop	1.12	1.17	4.36	2.57	3.05	2.95

缩写： ws 表示 WenetSpeech；md 表示 MagicData；conv 表示会话；daily 表示日常用语；Fun-Nano 表示 Fun-ASR-Nano-2512。

API 基线系统： 豆包-ASR (volc.seedasr.auc) 于 2026 年 2 月上旬进行了评测，Fun-ASR 于 2025 年 11 月下旬进行了评测。API 结果可能会因服务器端更新而随时间变化，也可能包含专有组件。为确保公平比较，在 API 输出支持相关选项时，我们均关闭了逆文本正则化和标点预测，并使用各 API 提供的默认语音活动检测配置。

数据重叠： 我们的自动语音识别训练数据不包含来自 MagicData 的任何汉语方言或带口音语音数据；所有 MagicData 方言数据集仅用于评测。Fun-ASR API 可能受益于专有训练数据，这或许能解释其在某些方言子集（例如 MagicData 上海话和武汉话测试集）上的优势。

B FireRedLID 标签列表

B.1 完整语言代码列表

B.2 完整汉语方言代码列表

表7：FireRedLID支持的完整语言代码列表。

代码	英语名称	中文名称	代码	英语名称	中文名称
zh	中文	中文	en	英语	英语
es	西班牙语	西班牙语	fr	法语	法语
ja	日语	日语	ko	韩语	韩语
ru	俄语	俄语	de	德语	德语
pt	葡萄牙语	葡萄牙语	ar	阿拉伯语	阿拉伯语
ab	阿布哈兹语	阿布哈兹语	af	南非荷兰语	南非荷兰语
am	阿姆哈拉语	阿姆哈拉语	as	阿萨姆语	阿萨姆语
az	阿塞拜疆语	阿塞拜疆语	ba	巴什基尔语	巴什基尔语
be	白俄罗斯语	白俄罗斯语	bg	保加利亚语	保加利亚语
bn	孟加拉语	孟加拉语	br	布列塔尼语	布列塔尼语
bs	波斯尼亚语	波斯尼亚语	ca	加泰罗尼亚语	加泰罗尼亚语
ceb	宿务语	宿务语	cs	捷克语	捷克语
cy	威尔士语	威尔士语	da	丹麦语	丹麦语
el	希腊语	希腊语	eo	世界语	世界语
et	爱沙尼亚语	爱沙尼亚语	eu	巴斯克语	巴斯克语
fa	波斯语	波斯语	fi	芬兰语	芬兰语
fo	法罗语	法罗语	gl	加利西亚语	加利西亚语
gn	瓜拉尼语	瓜拉尼语	gu	古吉拉特语	古吉拉特语
gv	马恩语	马恩语	ha	豪萨语	豪萨语
haw	夏威夷语	夏威夷语	hi	印地语	印地语
hr	克罗地亚语	克罗地亚语	ht	海地克里奥尔语	海地克里奥尔语
hu	匈牙利语	匈牙利语	hy	亚美尼亚语	亚美尼亚语
ia	国际语	国际语	id	印度尼西亚语	印度尼西亚语
is	冰岛语	冰岛语	it	意大利语	意大利语
iw	希伯来语	希伯来语	jw	爪哇语	爪哇语
ka	格鲁吉亚语	格鲁吉亚语	kk	哈萨克语	哈萨克语
km	高棉语	高棉语	kn	卡纳达语	卡纳达语
la	拉丁语	拉丁语	lb	卢森堡语	
ln	林加拉语	林加拉语	lo	老挝语	老挝语
lt	立陶宛语	立陶宛语	lv	拉脱维亚语	拉脱维亚语
mg	马尔加什语	马尔加什语	mi	毛利语	毛利语
mk	马其顿语	马其顿语	ml	马拉雅拉姆语	马拉雅拉姆语
mn	蒙古语	蒙古语	mr	马拉地语	马拉地语
ms	马来语	马来语	mt	马耳他语	马耳他语
my	缅甸语	缅甸语	ne	尼泊尔语	尼泊尔语
nl	荷兰语	荷兰语	no	挪威语	挪威语
oc	奥克语	奥克语	pa	旁遮普语	旁遮普语
pl	波兰语	波兰语	ps	普什图语	普什图语
ro	罗马尼亚语	罗马尼亚语	sd	信德语	信德语
si	僧伽罗语	僧伽罗语	sk	斯洛伐克语	斯洛伐克语
sl	斯洛文尼亚语	斯洛文尼亚语	so	索马里语	索马里语
sq	阿尔巴尼亚语	阿尔巴尼亚语	sr	塞尔维亚语	塞尔维亚语
sv	瑞典语	瑞典语	sw	斯瓦希里语	斯瓦希里语
ta	泰米尔语	泰米尔语	te	泰卢固语	泰卢固语
th	泰语	泰语	tr	土耳其语	土耳其语
uk	乌克兰语	乌克兰语	ur	乌尔都语	乌尔都语
uz	乌兹别克语	乌兹别克语	vi	越南语	越南语
yi	意第绪语	意第绪语	yo	约鲁巴语	约鲁巴语

表8：FireRedLID 支持的汉语方言代码完整列表。

Code	英文名称	中文名称
普通话	中文（普通话）	中文(普通话)
yue	中文（粤语：广东/香港）	中文(粤语:广东/香港)
wu	中文（吴语：上海/吴语片区）	中文(吴语:上海/吴语片区)
min	中文（闽语：福建）	中文（闽语：福建）
北方	中文（官话-北方：山东/甘肃/宁夏/河北/山西/辽宁/陕西）	中文(官话-北方:山东/甘肃/宁夏/河北/山西/辽宁/陕西)
西南	中文（官话-西南：四川/云南/贵州/湖北/重庆）	中文（官话-西南：四川/云南/贵州/湖北/重庆）
湘语	中文（湘语：湖南）	中文（湘语：湖南）
bo	藏语（中文语境）	中文（藏语）