信也科技杯

学术指导单位
中国计算机学会自然语言处理专业委员会(CCF-NLP)

合作院校
复旦大学自然语言处理实验室(FudanNLP)

赛事合作
第15届CCF自然语言处理与中文计算国际会议(NLPCC2026)

第十一届信也科技杯
全球AI算法大赛

LISTEN BEYOND 循声知意

2026 FINVOLUTION GLOBAL DATA SCIENCE COMPETITION

TURN-TAKING MODELING IN CONVERSATION CHALLENGE

信也科技杯

多模态跨注意力对齐的对话轮次预测

2026第十一届信也科技杯全球AI算法大赛

团队：和合台电

汇报人：李纯翰

CONTENT

01 团队介绍

02 赛题解读

03 方案介绍

04 感悟与致谢

Part 1

01 团队介绍

团队介绍

第十一届信也科技杯
全球AI算法大赛

RANK

01

和合台电团队

最终得分 **0.683575**

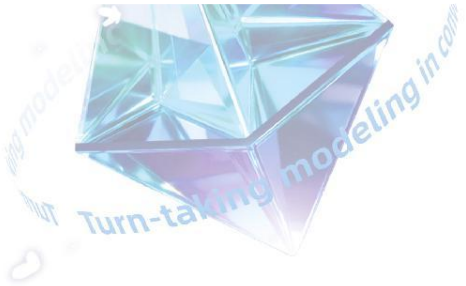
领先第二名 **0.007929**

第2名 → 第5名 **0.007566**

		排名变化	评价指标-F1
1	和合台电	0	0.683575
2	101010101010	+1	0.675646
3	jscc	-1	0.674916
4	LinkW	0	0.668153
5	带专小分队	0	0.66808

团队成员 王鹏 · 王新禹 · 王林梅 · 周洪青 · 李纯翰

竞赛履历 Kaggle Grandmaster丨6金3银 · 2025数字中国创新大赛丨二等奖
2024国家智慧教育平台创新活动赛题二、五丨第1名 · KDD2023 推荐任务丨第4名
浙江省能源科学创新大赛丨一等奖



Part 2

02 赛题解读

赛题解读 | 任务定义

赛题背景

对话轮次预测（**Turn-taking Prediction**）是智能对话系统中的经典任务，旨在预测对话参与者未来的发言状态，使机器能够判断何时倾听、何时接话，从“语音识别工具”进化为“对话参与者”，提升人机交互的自然度与流畅度。

输入数据 → 预测目标

历史上下文 | 5-30s 80 ms / chunk

双声道音频	声道A / B 两路波形
文本转写	[A] ..... [B] .....
历史标签	C · NA · T · BC · ...

未来窗口 | 2 s • 25 chunks

C T BC I NA

输出 5 个独立概率 | 评测指标 Macro-F1

标签	英文	含义
C	Continuation	当前说话人继续发言
T	Turn Change	话轮切换，另一方接话
BC	Backchannel	短反馈，不夺取话轮
I	Interruption	双方同时说话，争夺话轮
NA	Silence	双方均未说话

五类事件相互独立 | 未来 2 s 内出现一次即记为 1

赛题解读 | 核心难点

01 海量滑窗样本下的实验迭代效率

369	163 h	17 h
段长音频	训练音频总量	Baseline 单次耗时

369 段对话共约 163 小时，滑动窗口切分后样本量迅速膨胀；Baseline（单epoch双3090卡）完整运行一次约 17 小时，单次试错成本很高，实验迭代效率低。

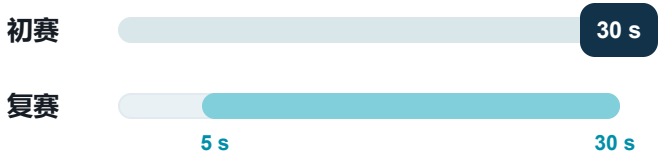
难点：有限资源下如何加速实验迭代，缩短“训练—验证—对比”周期

02 多模态+多时间尺度信息对齐与融合

音频特征	Whisper embedding	20 ms / 帧
历史标签	Embedding初始化训练	80 ms / chunk
文本序列	Qwen embedding	逐token

音频、文本与历史标签经独立编码后分布于不同特征空间，且时间粒度不同。如何实现跨模态语义映射与时序对齐，并充分建模模态间的互补关系，是多模态融合的核心难点。

03 变长历史输入下的模型鲁棒性



初赛上下文长度固定为 30 s，复赛则在 [5, 30] s 范围内随机截断。模型不仅需要充分利用长时历史信息，还应降低对固定位置与完整上下文的依赖，在有限历史输入下保持稳定的表征能力与预测性能。

04 类别不均衡与多标签学习

标签	C	NA	T	I	BC
正样本比例	94.2%	65.6%	26.0%	14.2%	**3.6%**

五类极不均衡：C/NA 高频易学，I、BC稀疏且难

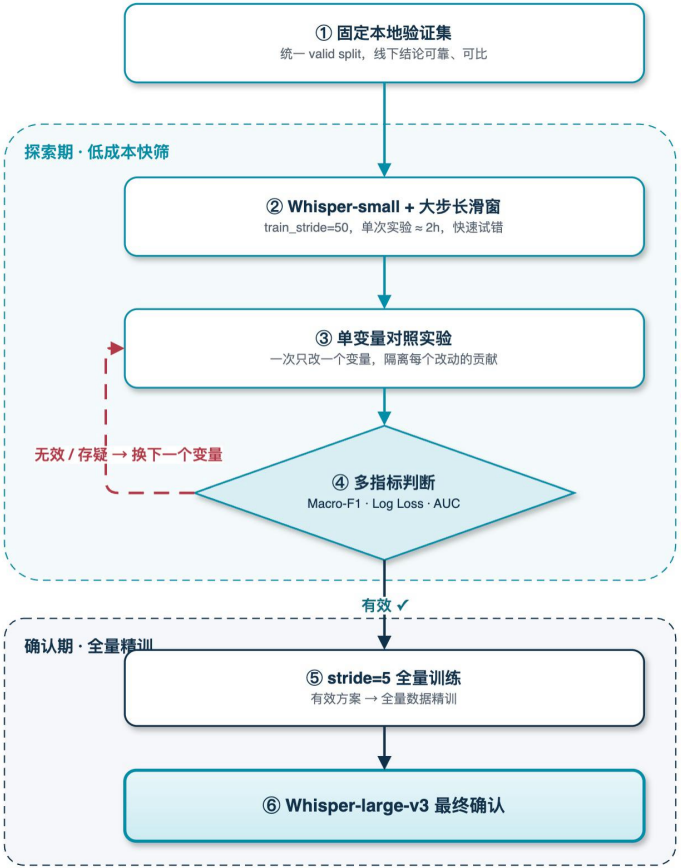
类别分布高度不均衡，且多标签联合学习存在性能此消彼长的问题。如何提升少数类与复杂交互事件的识别能力，并平衡各标签的学习效果，是核心难点。

Part 3

03 方案介绍

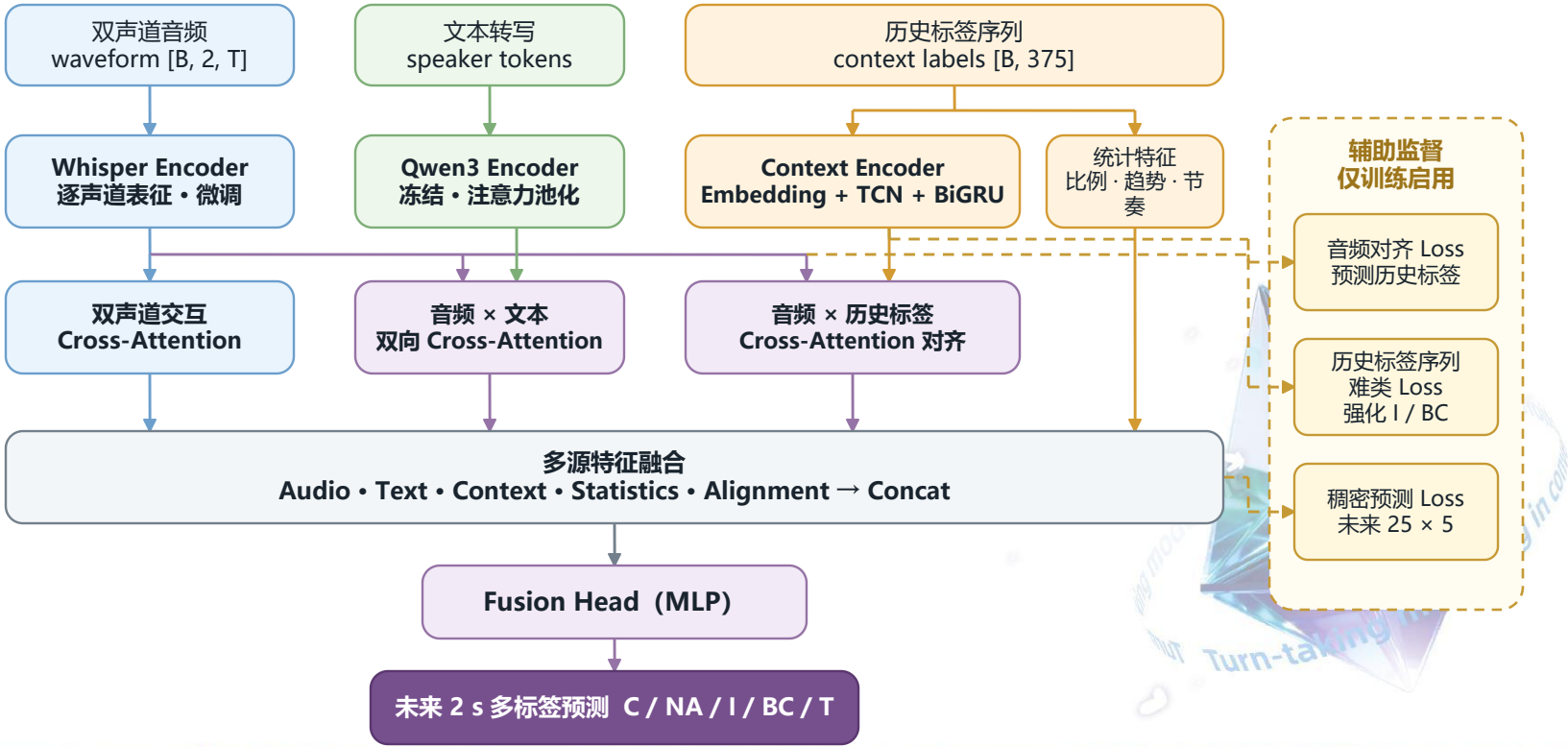
实验框架：低成本快筛，有效才全量

- **统一实验口径**：固定数据划分，单变量对照，准确评估各项改动的贡献
- **双验证集评估**：30s 等长(对齐初赛) + 5–30s 变长(对齐复赛)，双口径检验模型鲁棒性
- **两阶段推进**：使用 Whisper-small 与大步长滑窗快速筛选方案（约 2 h/次）；仅将稳定有效的改动升级Whisper-large-v3 全量训练
- **数据管线优化**：将 WAV 离线转换为 NPY，并通过 mmap 按需读取，避免滑窗样本重复解码，每个 epoch 节省约 20 min



方案介绍 | 整体架构

三模态独立编码，经交互对齐后融合；训练期引入辅助监督强化关键表征。



方案介绍｜编码模块与跨模态对齐

01 音频编码与双声道交互

- **音频预训练模型：**验证集评测决定 ☑Whisper模型 ☒Wav2vec2 ☒WavLM
- **双声道Cross-Attn[线下+~0.004] vs 平均声道:** 采用**VAP论文[1]**中思想，Whisper模型分别编码两个声道，采用**共享权重**的双向Cross-Attention，实现双声道表征的充分交互与协同对齐，避免平均声道造成的说话人信息损失
- **共享权重动机：**双声道本身同源，配合双声道翻转增强，共享权重保证一致性。
- **近端聚焦：**保留序列末端帧数(最近6s)，聚焦最近声学信息，也顺带压缩了序列长度、降低开销
- **Whisper微调：**主干冻结，仅微调最后 1 层transformer层与末端 LayerNorm

```python
class CrossAttentionBlock(nn.Module):
    """A single bidirectional cross-attention layer between two streams.

    Each stream attends to the other (pre-norm + residual), followed by a
    feed-forward block. Same spirit as the cross-attention transformer used in
    VAP to let the two speakers' audio channels interact.
    """

    def __init__(self, dim: int, num_heads: int, dropout: float):
        super().__init__()
        # The two channels are symmetric (and swapped by channel-flip augmentation),
        # so both directions share the same norm / attention / feed-forward weights.
        self.norm1 = nn.LayerNorm(dim)
        self.attn = nn.MultiheadAttention(dim, num_heads, dropout=dropout, batch_first=True)
        self.norm2 = nn.LayerNorm(dim)
        self.ff = nn.Sequential(
            nn.Linear(dim, dim * 2), nn.GELU(), nn.Dropout(dropout), nn.Linear(dim * 2, dim),
        )
        self.dropout = nn.Dropout(dropout)

    def forward(self, a: torch.Tensor, b: torch.Tensor):
        # a, b: [B, T, D]; query from one stream, key/value from the other.
        an, bn = self.norm1(a), self.norm1(b)
        a = a + self.dropout(self.attn(an, bn, bn, need_weights=False)[0])
        b = b + self.dropout(self.attn(bn, an, an, need_weights=False)[0])
        a = a + self.dropout(self.ff(self.norm2(a)))
        b = b + self.dropout(self.ff(self.norm2(b)))
        return a, b
```

02 文本编码与音频—文本对齐

- **文本预训练模型：** ☑Qwen3-0.6B ☒Qwen3-4B（没有显著变化）
- **近期语义聚合:** 保留最近 30% 的有效 Token，并通过 Padding Mask 与注意力池化
- **音频<->文本对齐[线下+~0.0015]：** 独立参数的双向 Cross-Attention：音频 → 文本聚合语义线索，文本 → 音频回查对应声学证据。

[1] Ekstedt E, Skantze G. Voice Activity Projection: Self-supervised Learning of Turn-taking Events[C]//Proc. Interspeech. 2022: 5190–5194.

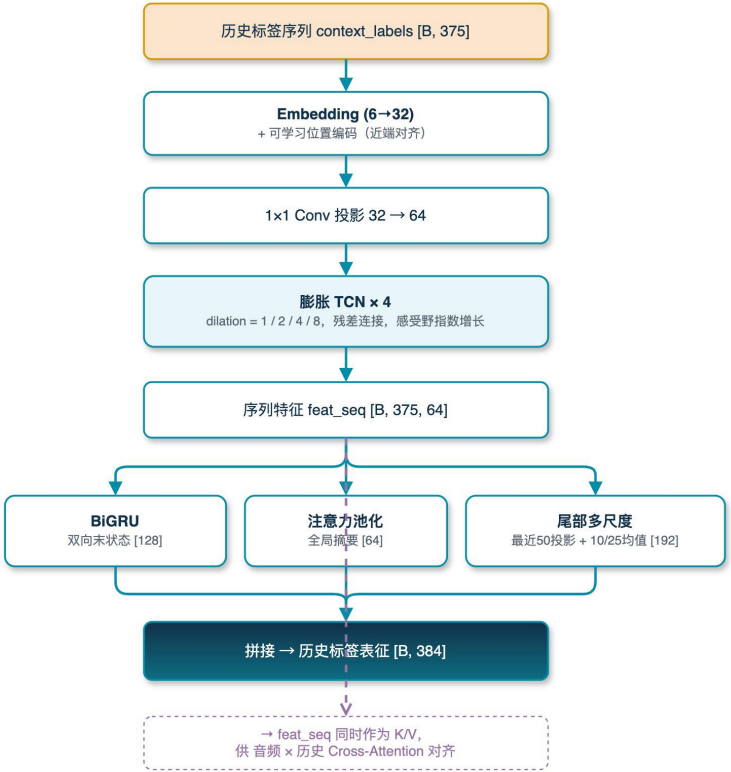
方案介绍｜编码模块与跨模态对齐

03 历史标签序列编码与音频—历史标签序列对齐[线下+～0.0030]

- **标签序列嵌入**：将历史标签映射为连续向量，并加入可学习的位置编码，保留事件顺序及相对预测时刻的位置信息
- **多尺度时序建模**：通过 1×1 Conv 将特征投影至 64 维，再采用四层残差膨胀 TCN（Dilation = 1/2/4/8）逐层扩大感受野，捕捉长时间范围内的话轮转换规律。
- **全局—局部信息聚合**：融合 BiGRU 的有序状态、注意力池化的全局摘要，以及最近 10/25/50 个 Chunk 的多尺度尾部特征，同时建模整体对话节奏与临近预测时刻的状态变化。
- **音频—历史标签序列对齐**：以音频序列为 Query、历史标签序列为 Key/Value，通过 Cross-Attention 从标签轨迹中检索与当前声学状态相关的话轮信息。

04 历史标签序列统计特征[线下+～0.0005]

- **特征扩展**：在 Baseline 19 维特征基础上扩展至 72 维，显式补充话轮先验
- **多尺度分布**：统计不同时间窗口及全局的类别占比，兼顾短期变化与长期分布趋势
- **状态与节奏**：提取最近标签、当前状态持续时长、状态切换标记与切换率，并通过短长期占比差刻画话轮变化趋势。



方案介绍 | 辅助Loss

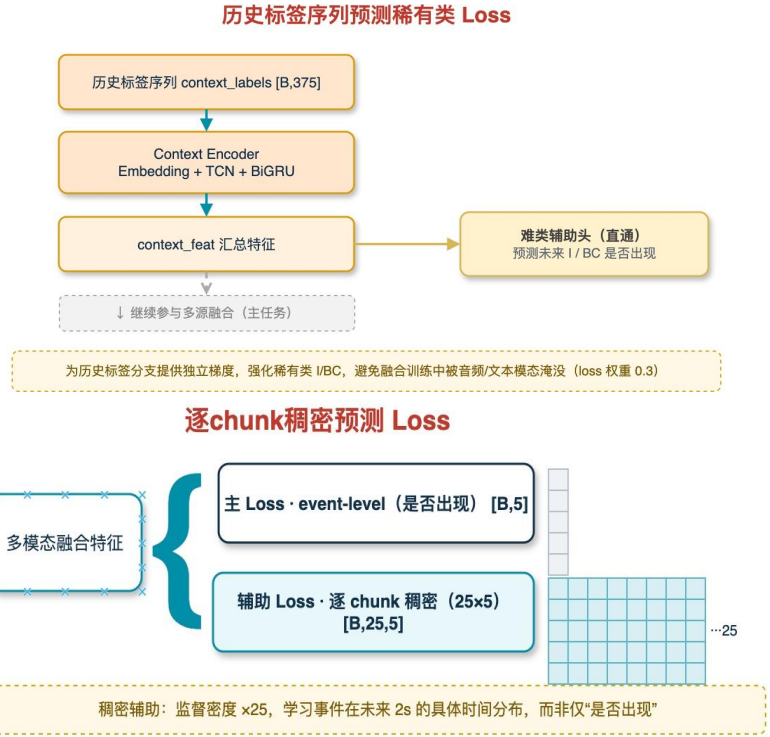
主任务为未来 2 s 的五标签 event-level BCE

01 历史标签序列预测稀有类Loss[线下+～0.0010]

- **输入->输出:** 仅历史标签序列表征，独立预测主任务中稀有类别(I/BC)，不与音频、文本特征融合
- **作用：** 将监督梯度直接传递至历史标签编码器，避免该分支在多模态联合训练中被强势模态抑制；同时通过类别加权增强对 I / BC 的学习
- **动机：** **论文[1]**指出不同模态的学习与过拟合速度存在差异，统一端到端训练可能导致部分模态学习不足，因此引入独立辅助监督强化历史标签分支

02 逐chunk稠密预测Loss[线下+～0.0010]

- **输入->输出:** 基于多模态融合表征，逐 Chunk 预测未来 25 个时间步（共 2 s）中的五类话轮事件
- **作用：** 主任务仅判断事件是否在未来窗口内出现，稠密监督进一步保留事件发生的时间位置，使单个样本由一次窗口级监督扩展为 25 个时间步的监督信号
- **动机：** 论文[2]中采用该Loss作为主Loss



[1] Wang W, Tran D, Feiszli M. What Makes Training Multi-modal Classification Networks Hard?[C]//CVPR. 2020: 12695–12705.
[2] Ekstedt E, Skantze G. Voice Activity Projection: Self-supervised Learning of Turn-taking Events[C]//Proc. Interspeech. 2022: 5190–5194.

03 当前音频—同期标签对齐 Loss[线下+～0.0020]

• **输入->输出:** 区别于主任务预测未来 2 s，该分支仅使用音频表征，逐 Chunk 预测输入末端最近 2 s 的已知历史标签

• **作用：**通过音频表征逐 Chunk 重建同期标签序列，构造“音频→话轮状态”的两模态生成式对齐约束，使声学特征显式编码发言持续、话轮切换、打断与反馈等交互状态。

• **有趣的发现：**仅微调 Whisper 并不能稳定提升效果；加入该 Loss 后，Whisper 微调才获得增益。可能是因为主任务的未来窗口级监督较弱、梯度链路较长，解冻后容易出现表征漂移；该辅助Loss提供了更直接的梯度，从而约束预训练特征向话轮预测任务有效适配。

方案介绍 | 训练/推理策略

01 训练策略

- **分层学习率**：解冻的 Whisper 末层用小的学习率（whisper-small: 2e-6，whisper-large-v3: 1.5e-6），其余可训练模块用主学习率1e-4量级，避免微调破坏预训练表征。
- **稀有类别加权**：多标签加权BCE Loss, pos_weight={C:1.0, NA:1.0, I:2.0, BC:5.0, T:1.0}
- **数据增强**：以 0.5 概率做变长[5,30]s历史截断（模拟复赛长度分布）、以 0.5 概率做双声道翻转
- **稳定性技巧**：EMA（ema_decay=0.999）、AMP 混合精度加速、梯度裁剪(max_grad_norm=1.0)、cosine_schedule 学习率 + warmup(step=100)
- **分布式**：双卡 3090 DDP 训练。训练时间：small 分支约 10h，large 分支约24h。

02 推理策略

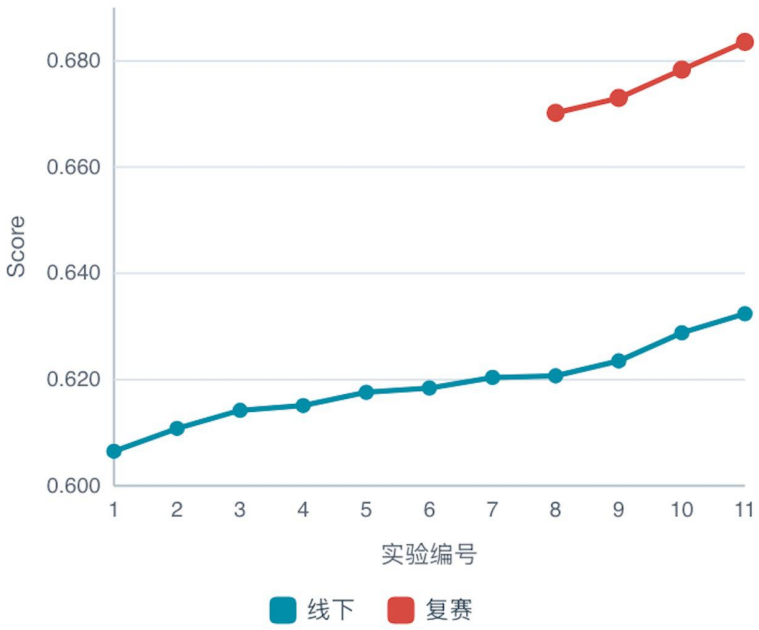
- **双声道翻转TTA：**每个模型对测试集跑两遍（原始输入 + 双声道翻转），sigmoid 概率取平均，降低方差。
- **简单加权模型融合：**pred = 0.3 * whisper-small模型 + 0.7 * whisper-large模型

方案介绍 | 上分过程

实验迭代记录

#	模型改进	线下	复赛
1	Baseline（whisper-small，stride=50）	0.6065	—
2	双声道 Cross-Attention	0.6108	—
3	历史标签 TCN+GRU+统计特征	0.6142	—
4	历史标签序列预测稀有类 Loss	0.6151	—
5	音频–文本 + 音频–历史对齐	0.6176	—
6	逐chunk稠密预测 Loss	0.6184	—
7	纯音频历史标签预测 Loss + whisper微调	0.6204	—
8	#5：全量训练（stride=5）	0.6207	0.6702
9	#7：全量训练（stride=5）	0.6235	~0.6730
10	#9：切换 Whisper-large	0.6288	0.67835
11	**Small#7 + Large#9 模型融合**	**0.6324**	**0.683575**

趋势折线图



方案介绍 | 模型实用性说明

模型得分与参数量

模型编号	音频骨干	参数量	线下成绩	复赛成绩	备注
618	whisper-small	0.69B	0.6207	0.6702	初赛模型, AMP 推理仅约 2 min
647	whisper-small	0.69B	0.6235	未单独提交，预估 ≈ 0.673	复赛模型
657	whisper-large-v3	1.24B	0.6288	0.678	复赛模型
融合（0.3·647 + 0.7·657）	—	1.93B	0.6324	0.683575（复赛第一名）	

• 647 和 657 分别为 Whisper-small 与 Whisper-large-v3 的复赛最优模型。其中，647 以初赛最优模型 618 为基础，进一步引入两个辅助loss及 Whisper 微调。

•融合参数量为两模型相加（1.24 + 0.69 = 1.93B）；每个模型各做一次双声道翻转 TTA（同一模型推理 2 次），等效推理计算量约 1.93 × 2 = 3.86B，远小于比赛要求的8B上限

Part 4

04 感悟与致谢

感悟与致谢

比赛感悟

- Trust your local validation：相信并坚持可靠的本地验证结论。
- 大模型与 AI Coding 的发展正在打破不同领域之间的知识壁垒，显著提升知识获取与技术落地的效率。在 AI Coding 时代，代码实现不再是主要瓶颈，真正决定方案上限的，是对问题的理解、创新的思路与高质量的 Idea。

致谢

感谢信也科技与赛事组委会提供高质量的赛题与数据，以及比赛期间专业、高效的组织协调与技术支持。从第一届到第十一届，信也科技始终坚持连接前沿技术与真实业务场景，为参赛者搭建了宝贵的实践交流平台，也为人工智能领域的技术创新与人才成长持续注入动力。

信也科技杯

THANKS