arXiv:2603.05231v1 [cs.SD] 5 Mar 2026

通过结合音频-文本语义奖励的测试时强化学习提升自动语音识别鲁棒性

Linghan Fang[1,2],**Tianxin Xie**[1],**Li Liu**[1*]香港科技大学（广州），中国广州[2]慕尼黑工业大学，德国慕尼黑 linghan.fang@tum.de, txie151@connect.hkust-gz.edu.cn,avrillliu@hkust-gz.edu.cn

摘要

近年来，自动语音识别（ASR）系统（例如 Whisper）在准确率方面取得了显著提升，但对于真实世界未见数据（存在较大分布偏移的数据）仍然高度敏感，包括嘈杂环境和多样口音。为解决这一问题，测试时自适应（TTA）展现出提升模型在推理时间适应能力的巨大潜力，且无需真实标签；现有测试时自适应方法通常依赖伪标签或熵最小化。然而，将模型置信度作为学习信号可能会强化高置信度错误，从而导致确认偏差，削弱自适应效果。为克服这些局限，我们提出了**ASR-TRA**，一种受因果干预启发的新型**测**试时**强**化**自**适应框架。具体而言，我们的方法引入了可学习的解码器提示，并利用温度控制的随机解码生成多样化的转写候选。这些候选项由奖励模型进行评分，该模型衡量音频-文本语义对齐程度；随后，通过强化学习利用所得反馈更新模型参数和提示参数。在带有合成噪声的 LibriSpeech 数据集和 L2 Arctic 带口音的英语数据集上开展的大量实验表明，与现有测试时自适应基线相比，我们的方法在保持更低延迟的同时实现了更高的准确率。消融研究进一步证实了结合音频和语言奖励的有效性，凸显了我们方法更强的稳定性和可解释性。总体而言，我们的方法为在充满挑战的真实世界条件下部署自动语音识别系统提供了一种实用且鲁棒的解决方案。

代码— https://github.com/fangcq/ASR-TRA

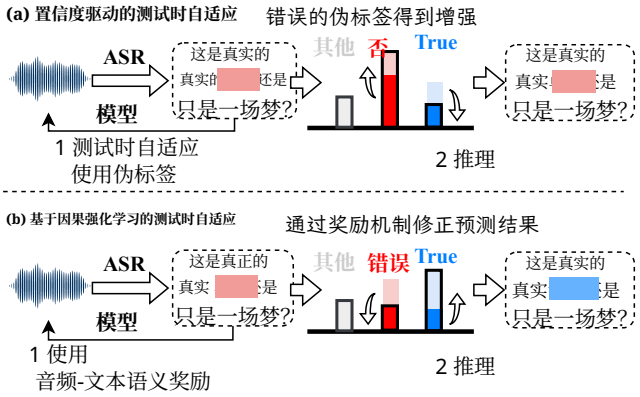

图 1：**噪声条件下测试时自适应策略概览**。（a）基于置信度的自适应会强化错误但高概率的预测（例如，*word*），从而导致持续性错误。（b）我们的奖励引导自适应方法 **ASR-TRA** 通过强化学习，倾向于选择语义准确的替代项（例如，*world*），即使其初始置信度较低。

引言

近年来，自动语音识别（ASR）取得的进展得益于自监督学习和大规模弱监督训练领域的突破（Baevski 等，2020；Hsu 等，2021；Schneider 等，2019），这些突破使模型能够学习丰富的声学和语言表征—从海量未标注或弱标注语音数据中学习表征。借助这些技术，wav2vec 2.0（Baevski 等，2020）和 Whisper（Radford 等，2023）显著提升了各种领域中的转录文本准确率和泛化能力。

尽管取得了这些改进，但在边缘设备、在线流式处理以及资源受限的小模型场景等现实应用中部署自动语音识别仍然面临挑战。在实际应用中，轻量级自动语音识别系统经常会遇到严重的分布外（OOD）条件（Hendrycks 和 Dietterich，2019），包括背景噪声、浓重口音和地区方言；这些因素在训练数据中的代表性不足，并会导致显著的领域偏移（Ben-David 等，2007）。提高鲁棒性的传统方法通常依赖离线再训练或有监督领域自适应，例如多条件训练或噪声增强（Ko 等，2015；Li、Deng 和 Gong，2020），但由于测试时无法获得标注数据，这些方法难以实施。

在相关文献中，测试时自适应（TTA）已成为一种很有前景的方法，无需额外训练或有标签监督即可提升自动语音识别的鲁棒性（Wang

*通讯作者。版权所有 © 2026，美国人工智能促进协会（www.aaai.org）。保留所有权利。

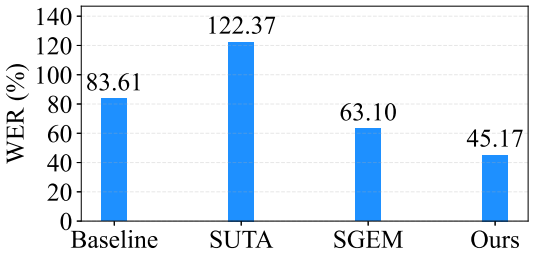

图2：我们在 LibriSpeech test-other 中经高斯噪声污染的置信度最高的 100 个样本上评估词错误率。虽然 Whisper-Tiny 的基线性能在噪声下有所下降，但 SUTA 等启发式方法由于过度自信而表现进一步恶化。相比之下，我们的方法取得了更低的词错误率。（越低越好。）

等人 2021；Lin、Li 和 Lee 2022；Kim 等人 2023）。TTA 使模型能够动态适应测试时的输入，从而有助于缓解训练条件与部署条件之间的不匹配。然而，将当前的 TTA 方法直接应用于自动语音识别系统时，仍面临以下两个主要挑战（见图 1）。**挑战 1：**尽管 Whisper 是一种强大的编码器-解码器模型，但它缺乏针对 Whisper 的适应机制（Radford 等人 2023）。**挑战 2：**许多现有方法依赖启发式优化策略，并大量使用模型自行生成的伪标签（Wang 等人 2021；Lin、Li 和 Lee 2022；Lee 2013；Sohn 等人 2020），这类方法在声学环境多样时可能速度慢、稳定性差，并且在模型不确定时容易导致错误累积。如图2所示，SUTA（Lin、Li 和 Lee 2022）等基于置信度的方法在噪声环境下处理高置信度样本时，甚至可能表现更差。这凸显了置信度与真实准确率之间的不匹配，也正是这一点促使我们提出本方法（详情见实验部分）。

为应对这些挑战，我们提出了一种新方法 **ASR-TRA**（**自动语音识别**中的**测**试时**强**化学习**适**应），该方法采用强化学习（RL）的视角（Williams 1992；Ziegler 等，2019），并将测试时自适应视为不确定性下的奖励驱动决策过程。ASR-TRA 基于两个关键思路解决这些局限。**思路 1：**为利用 Whisper 等编码器-解码器架构的结构，我们通过可学习的解码器提示引入专门的适应路径，从而实现针对自动语音识别的高效、低开销测试时优化。**思路 2：**我们不再依赖脆弱的启发式方法或自我确认的伪标签，而是引入外部语义奖励，以更稳健地指导适应过程。这避免了错误累积，并能在没有真实标签监督的情况下实现更稳定的更新。

为实现这两个关键构想，ASR-TRA 首先使用结构因果模型（SCM）（Pearl 2009）对适应过程进行结构化建模，将研究动机与具体实现衔接起来。该 SCM 包含括四个关键变量：编码后的音频特征 A、可学习的解码器提示 P、生成的转录文本 Y以及奖励 R。有向边捕捉了因果依

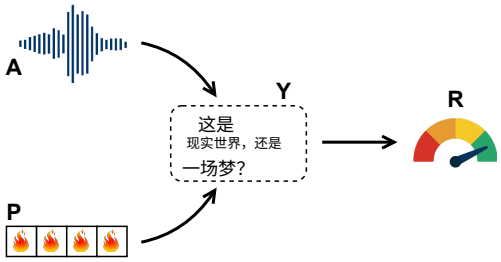

图 3：结构因果模型（SCM）示意图。节点A、 P、 Y和 R分别表示音频特征、可学习提示、转录输出和奖励，因果流向为 A， $P \rightarrow Y \rightarrow R$（Pearl 2009）。

赖关系（见图3）。

该模型不仅阐明了我们框架中的适应过程如何展开，还指导设计了三个具体组件，以在测试时实现奖励驱动的学习和基于提示的适应：1）**提示注入（对应构想 1）：**在解码器输入序列开头插入一个可学习向量 P，使 P与词元嵌入一同处理，并通过因果干预直接影响解码过程。2）**候选生成与评估（对应构想 2）：**调整采样温度，以生成多个不同的转录假设。随后使用 CLAP（对比语言-音频预训练）（Elizalde 等，2023）评估每个候选结果，并计算音频-文本相似度分数，将其作为序列级奖励来指导适应。3）**参数更新：**采用策略梯度强化学习算法反向传播 CLAP 奖励，并联合更新提示参数 P和模型权重，使后续输出朝着更高奖励的方向调整。

总之，所提出的 ASR-TRA 方法有三项主要贡献：

- 我们将测试时自适应建模为由音频-文本奖励模型 CLAP 引导的强化学习过程，从而缓解启发式伪标签或基于置信度的方法所导致的错误累积。
- 我们为 Whisper 设计了特定的结构因果模型（SCM），其中可学习的解码器提示词用于调节解码过程。结合策略梯度更新和 CLAP 奖励，该框架为自动语音识别提供了一种有理论依据且轻量的测试时自适应方法。
- 在噪声与口音语音基准上的实验表明，该方法在速度和识别准确率方面均持续优于以往的测试时自适应方法。

相关工作

自动语音识别。近年来，自动语音识别取得的进展主要得益于在多样化音频-文本语料库上训练的大规模神经网络模型。wav2vec 2.0（Baevski 等，2020）利用自监督对比预训练，从未标注音频中学习稳健的语音表征。Whisper（Radford 等，2023）基于 Transformer

编码器-解码器架构，通过在 68 万小时的弱标注数据上训练，实现了出色的多语言转录效果。尽管这些模型表现令人印象深刻，但它们仍容易受到分布偏移的影响，例如环境噪声、说话人口音或自发语音，这凸显了对自适应机制的需求（Hendrycks and Dietterich 2019；Ben-David 等，2007），以便模型能够很好地适应真实世界的条件。

深度学习中的因果性。因果推理为提升模型鲁棒性与可解释性提供了一个有原则的框架。经典表述（Pearl 2009）将干预定义为通过主动改变模型变量来研究下游效应的机制。近期工作探索了因果表征学习（Sch¨olkopf 等人 2021；Mitrovic 等人 2021），旨在解耦生成性变异因素，以在领域偏移下获得更强的鲁棒性。在深度学习中，因果干预可通过架构修改或受控采样方案实现，并已被证明有助于模型更可靠地适应分布变化。

自动语音识别中的测试时自适应。测试时自适应使模型能够在推理过程中仅使用无标注测试数据进行调整。在早期方法中，Tent（Wang 等人 2021）提出了一种情景式测试时自适应框架，通过最小化熵并冻结大部分参数，使模型适应每个测试样本。这启发了一系列旨在实现快速、轻量级领域自适应的情景式测试时自适应方法。在自动语音识别领域，SUTA（Lin、Li 和 Lee 2022）通过从高置信度预测生成伪标签来自适应Whisper式模型，而SGEM（Kim 等人 2023）则最小化序列级熵以正则化输出分布。然而，这些方法严重依赖熵或置信度等模型内部信号，在严重分布偏移下这些信号可能变得不可靠。此外，它们缺乏纳入模型先验信念之外的外部反馈的机制。

用于测试时自适应的强化学习。强化学习为基于延迟反馈优化决策提供了一个通用框架（Sutton 和 Barto，2018），并已广泛应用于机器人学（Levine 等，2016）、对话（Li 等，2016）和语言模型对齐（Ouyang 等，2022；Ziegler 等，2019）。近年来，强化学习也开始应用于测试时自适应。例如，RLCF（Zhao 等，2024）学习奖励驱动策略，以实现分类器自适应。较新的工作（如 BiTTA（Lee 等，2025））则利用二元反馈（正确/错误）来引导情节式自适应。尽管这些方法前景可期，但它们大多应用于分类或导航任务，这些任务中的动作空间和反馈信号相对简单。相比之下，它们在自动语音识别等结构化序列建模问题中的应用仍然有限。

ASR-TRA 的定位。虽然此前的自动语音识别测试时自适应方法（如 SUTA 和 SGEM）依赖置信度和熵等内部启发式指标，但它们在分布偏移下往往缺乏鲁棒性，并可能陷入伪标签反馈循环。近期基于强化学习的方法，例如 BiTTA 展示了奖励信号的效用，但这类方法专为分类或导航任务设计，尚未适用于自动语音识别中的结构化序列建模。ASR-TRA 通过在基于 Whisper 的框架中结合因果干预与外部奖励引导的自适应，弥合了这一差距。具体而言，我们将可学习的解码器提示词引入为因果变量，并利用语义反馈（Elizalde 等，2023）对其进行优化，从而实现鲁棒、可解释且高效的自适应，而无需依赖内部确定性估计。

方法

我们提出了一种面向 Whisper 的测试时自适应框架 ASR-TRA，将因果推理与强化学习相结合。其核心思路是将 Whisper 解码器中的提示注入视为一种因果干预，并将解码过程视为生成反事实假设的过程。我们通过奖励与 CLAP 预测结果更一致的生成内容，在推理时间优化模型，而无需依赖真实转录文本。这使模型能够针对未见过的声学条件进行动态、无标签的自适应。

Whisper 架构

输入语音信号首先经过预处理，并转换为对数梅尔频谱图 $s \in \mathbb{R}^{F\times T}$，其中包含 F 个频率频点和 T 个时间帧（Logan 2000；McFee 等，2015）。Whisper 是一种基于Transformer的编码器—解码器模型（Vaswani 等，2017；Radford 等，2023），遵循自动语音识别（ASR）中广泛使用的自回归序列到序列范式（Chan 等，2016）。音频编码器 Enc(·) 生成一个隐藏表示：

$$h = \mathrm{Enc}(s), \tag{1}$$

文本解码器则通过对输出概率进行条件词元分布建模，生成输出序列 $\hat{y} = (y_1, y_2, \ldots\ldots, y_N)$：

$$P(y_t \mid y_{<t}, h) = \mathrm{Dec}(y_{<t}, h), \tag{2}$$

其中 $y_{<t}$ 表示先前生成的词元。实际应用中，解码过程可采用贪心选择、束搜索或随机采样，并通过温度缩放进行调节（Holtzman 等，2020）。

将提示注入视为因果干预

我们引入一个可学习的提示向量 $p \in \mathbb{R}^{L\times d}$，其中$L$ 表示提示词元的数量，d 表示解码器嵌入维度。软提示会前置到解码器的输入嵌入中，直接拼接在 <bos> 词元的嵌入之前。在生成过程中，解码器会同时关注该提示和编码器输出，从而让提示引导每次预测，而非完全依赖从头开始的自回归解码。

由于提示词在每个步骤都对解码器的注意力可见，因此它可以直接塑造隐藏状态，从而影响所有后续 token 的预测（Li 和 Liang，2021；Lester、Al-Rfou 和 Constant，2021）。

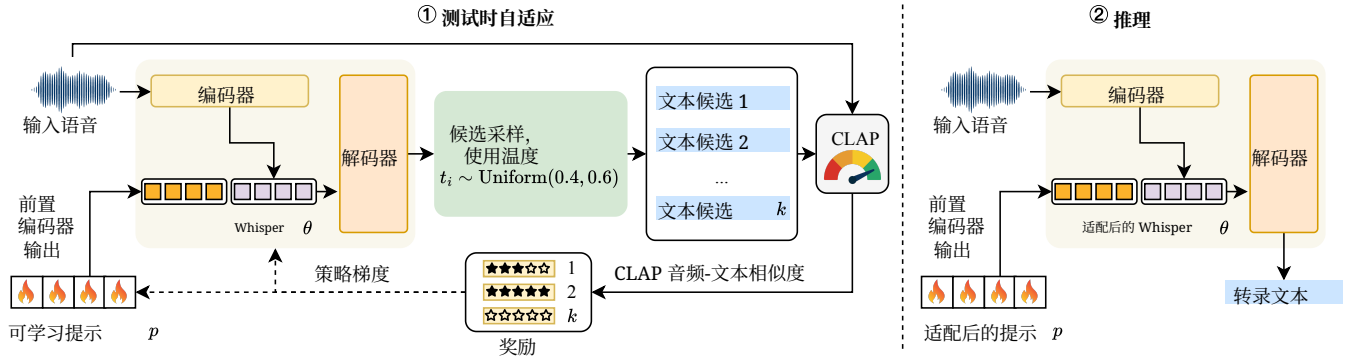

图4：Whisper 的测试时自适应。首先根据输入的梅尔频谱图解码得到基线转录文本；随后以不同温度采样软提示词变体，由 CLAP 对它们进行评分，并汇总奖励以更新模型，从而无需标签即可实时纠正。

关键在于，我们将 p 视为对生成过程的 因果干预，而不只是一个额外条件。从形式上看，解码器的预测可写为

$$y_t = \text{Dec}(h, y_{<t}, \text{do}(p)), \tag{3}$$

其中，do(p) 遵循 Pearl 的 do 演算（Pearl 2009），表示将 p 设为外部变量，而不是从观测到的输入 s 中推断得出。这种干预不会修改声学输入，却会扰动内部生成动态，使模型能够在相同观测下探索其他假设。

反事实采样与奖励评估

为了在固定输入和提示词下探索多样化的输出轨迹，我们采用随机解码，以温度参数 $T > 0$对词元进行采样（Holtzman et al. 2020）：

$$P_T(y_t \mid y_{<t}, h, p) \propto \exp\left(\frac{\log P(y_t \mid y_{<t}, h, p)}{T}\right). \tag{4}$$

较高的温度 T 会使词元分布趋于平坦并促进多样性，而 $T \to 0$则接近贪心解码。通过反复采样，我们得到 K 个候选转录文本 $\{\hat{y}^{(1)}, \ldots, \hat{y}^{(K)}\}$，每一个都代表一个反事实假设，即在相同音频和提示词条件下的一条合理替代转录轨迹。

随后使用奖励函数 $R(\hat{y}^{(i)}) \in \mathbb{R}$ 对每个采样假设进行评估，以量化其与输入之间的语义对齐。在我们的实现中，我们采用CLAP 音频文本相似度模型（Elizalde et al. 2023）作为主要奖励，该模型计算输入语音的音频嵌入与所生成转录文本的文本嵌入之间的余弦相似度：

$$r^{(i)} = R(\hat{y}^{(i)}). \tag{5}$$

我们还将其他预训练语言模型（LM）（Ziegler et al. 2019；Ouyang et al. 2022）的分数用作消融研究（表3）中的补充信号。

这种标量反馈可作为转录质量的代理指标，并实现无标签优化。由此产生的奖励会被汇总，用于更新提示向量 p 和模型参数，并通过策略梯度目标（Williams 1992；Sutton 和 Barto 2018）使模型倾向于在相同条件下获得更高奖励的生成结果。

通过强化学习进行优化

为了在线适配提示向量 p 和模型参数，我们构建了一个强化学习目标，明确鼓励生成由奖励函数衡量的更高语义质量的结果。我们将以提示为条件的 Whisper 视为输出序列 $\hat{y}$ 上的随机策略 $\pi_p(\hat{y})$，并旨在最大化期望奖励：

$$\mathcal{J}(p) = \mathbb{E}_{\hat{y} \sim \pi_p}[R(\hat{y})]. \tag{6}$$

这一设置与策略梯度方法（Sutton 和 Barto 2018）自然契合。我们采用经典的 REINFORCE 算法（Williams 1992）来估计 $\mathcal{J}(p)$ 的梯度。对于一批 N 个采样假设 $\{\hat{y}^{(i)}\}_{i=1}^{N}$ 及其对应奖励 $r^{(i)}$ 和对数概率 $\log P(\hat{y}^{(i)})$，我们引入基线以降低梯度方差。该基线即为该批次的平均奖励：

$$\bar{r} = \frac{1}{N} \sum_{i=1}^{N} r^{(i)}. \tag{7}$$

利用该基线，强化学习目标相对于可学习提示 p 和 Whisper 参数 θ 的梯度可估计如下：

$$\nabla_{\theta,p} \mathcal{L} = -\frac{1}{N} \sum_{i=1}^{N} \nabla_{\theta,p} \log P(\hat{y}^{(i)}) \cdot (r^{(i)} - \bar{r}), \tag{8}$$

从而提高生成高奖励结果的可能性，同时惩罚低奖励的替代方案。

在测试时，我们针对每个输入样本独立进行优化，利用少量经奖励评估的候选项来更新提示词和模型参数。当前样本的自适应与预测完成后，模型参数会恢复到初始状态，再处理下一个样本，从而确保更新不会在整个测试集上累积。

算法概述

我们在**算法 1**中总结了我们的测试时自适应方法 ASR-TRA，展示了如何将提示干预、反事实采样和强化学习更新整合到统一的在线“自适应与预测”循环中。

算法 1：**ASR-TRA**

要求：输入梅尔频谱图 s、Whisper 模型 θ、CLAP 奖励模型、样本数量 n、提示词嵌入 p1：初始化 Whisper 模型 θ，设置温度 $t = 0$2：生成确定性输出： $y_0 \leftarrow$ Whisper($s; \theta, t$ =0) 3：计算奖励： $r_0 \leftarrow$ CLAP(y_0, s) 4：将随机提示词 p插入 Whisper 解码器 5：对于 $i = 1, \ldots, n$，执行 6：将温度采样为 $t_i \sim$ Uniform(0.4, 0.6) 7：通过干预生成输出：$y_i \leftarrow$ Whisper($s; \theta$, do(p) t_i) 8：计算奖励： $r_i \leftarrow$ CLAP(y_i, s) 9：结束循环 10：计算优势：adv$_i \leftarrow r_i -$mean($\{r_j\}_{j=0}^{n}$) 11：计算策略梯度损失：$\mathcal{L} \leftarrow -\sum_i$ adv$_i \cdot$ log $P_\theta(y_i|s, p)$ 12：更新模型参数和提示词：$\theta \leftarrow \theta - \eta_1 \nabla_\theta \mathcal{L}$, $p \leftarrow p - \eta_2 \nabla_p \mathcal{L}$13：生成适配后的输出： $y \leftarrow$ Whisper(s; θ, p) 14：将 Whisper 模型 θ 和提示词 p恢复到原始状态 15：返回适配后的转录文本 y

实验

实验设置

我们在轻量级 Whisper-Tiny 模型（Radford 等，2023）上评估 ASR-TRA。该模型约含 3900 万个参数，由于计算成本低，已广泛应用于真实环境中的自动语音识别任务。其紧凑的尺寸使其非常适合部署在对延迟敏感或资源受限的环境中，例如设备端转录或流式自动语音识别。然而，Whisper-Tiny 仍然对分布偏移高度敏感，在声学变化下难以维持性能。

为实现快速且有针对性的适配，我们插入一个长度为四的可学习解码器提示词，仅增加 $4 \times d$ =1，536 个参数，其中 $d = 384$ 表示解码器嵌入维度。适配过程中，解码器温度从均匀范围 $[0.4, 0.6]$内随机采样，Whisper-Tiny 并行生成总计 4 个候选项，以鼓励产生多样化的假设。模型参数根据数据复杂度，以 η_1 的学习率在 10^{-6} 至 10^{-5} 的范围内更新；提示参数则以更大的学习率 $100\times$ η_2更新，这与提示微调文献（Li 和 Liang，2021；Lester、Al-Rfou 和 Constant，2021）的发现一致。所有实验均在单台 NVIDIA RTX 6000 Ada GPU 上进行。

数据集与评估指标

我们考虑两种具有挑战性的分布偏移场景，用于测试模型对声学和语言变化的鲁棒性：

1) LibriSpeech test-other 集上的环境噪声。为评估模型在声学干扰下的鲁棒性，我们在 LibriSpeech test-other 集（Panayotov 等人，2015）中加入从 MS-SNSD 语料库（Reddy 等人，2019）采样的背景噪声。我们采用 10 dB 信噪比下的八种加性噪声：空调、机场广播、喋喋声、复印机、咀嚼声、邻居、关门声和打字声，遵循既有鲁棒性基准（Hendrycks 和 Dietterich，2019；Kim 等人，2023）。为近似真实环境，每条语音都加入每种噪声类型中随机采样的一个实例。

2) L2-Arctic 非母语口音语音。 为评估对说话人及发音差异的敏感性，我们使用 L2-Arctic 数据集（Zhao 等，2018），其中包含来自六种不同母语背景说话人的英语语音。该条件引入了 Whisper 在预训练期间未曾遇到的严重口音偏移，从而对其跨说话人泛化能力的评估构成了严峻挑战。

这两种设置共同评估了我们的方法对声学领域偏移的适应能力，而这是现实世界自动语音识别部署中的一项关键挑战（Wang 等人 2021；Kim 等人 2023）。

评估指标。 我们在本文的所有实验中报告词错误率（WER），遵循自动语音识别测试时自适应的标准做法（Lin、Li 和 Lee 2022；Kim 等人 2023）。词错误率衡量将模型的转录文本转换为真实标注所需的最少词级插入、删除和替换次数，并以参考转录文本中的词数进行归一化。较低的词错误率表示更高的转录准确率。我们还以秒为单位报告延迟，以量化所提出的测试时自适应过程引入的开销。

与最先进方法的比较

我们在两种主要的分布偏移类型下评估了我们的方法：环境噪声和说话人口音变化。结果汇总于表 1 和表 2 中，遵循了此前针对自动语音识别的测试时自适应研究（Lin、Li 和 Lee，2022；Kim 等人，2023；Wang 等人，2021）。

噪声鲁棒性。我们将所提出的方法与三种基线方法进行比较：（1）未经适配的原始 Whisper-Tiny 模型（基线）；（2）SUTA（Lin、Li 和 Lee，2022）；（3）SGEM（Kim 等人，2023）。表 1 报告了在 LibriSpeech test-other（Panayotov 等人，2015）上应用的八种 MS-SNSD 噪声条件（Reddy 等人，2019）下的词错误率（WER）和推理延迟。我们的方法取得了最低的平均词错误率（28.64%）和延迟（0.720 秒），始终优于基线方法。在机场广播和多人嘈杂声等高熵噪声条件下，改进尤为显著，因为这些条件下的声学变异性较高。虽然在邻居条件下 SUTA 略胜于我们的方法，但我们的方法在推理时仍显著更快。这与相关研究中的发现一致

数据集	基线 词错误率（%）↓	+ SUTA 词错误率（%）↓	+ SUTA 延迟（秒）	+ SGEM 词错误率（%）↓	+ SGEM 延迟（秒）	+ASR-TRA（我们的方法） 词错误率（%）↓	+ASR-TRA（我们的方法） 延迟（秒）
AC	26.54	24.02	1.683	23.76	0.730	**22.39**	**0.678**
AA	40.22	37.69	1.770	36.66	0.759	**35.66**	**0.670**
BA	36.85	40.12	1.643	33.92	0.710	**31.39**	**0.679**
CM	39.40	43.52	1.583	35.02	0.691	**34.72**	**0.631**
MU	29.98	34.12	1.645	34.40	0.705	**25.88**	**0.584**
NB	40.77	**33.75**	1.847	36.13	**0.756**	35.81	0.833
SD	23.12	22.91	1.656	**19.50**	0.909	22.34	**0.801**
TP	24.81	22.05	1.699	22.40	0.941	**20.91**	**0.886**
Mean	32.71	32.27	1.690	30.22	0.775	**28.64**	**0.720**

表 1：MS-SNSD 的八种噪声条件下（信噪比 =10 dB），Whisper-Tiny 的词错误率（%）和推理延迟（秒）。每行的最佳值以**粗体**显示。

设置	基线	SUTA	SGEM	ASR-TRA（我们的方法）
阿拉伯语	32.74	35.94	35.19	**22.92**
普通话	28.03	27.68	28.03	**25.14**
印地语	14.62	14.32	**12.61**	14.18
韩语	13.42	13.56	**12.12**	13.76
西班牙语	42.64	41.25	41.78	**39.44**
越南语	61.21	62.78	58.68	**53.84**
Mean	32.11	32.59	31.40	**28.21**

表 2：具有不同第一语言（L1）背景的英语语音说话者在 L2-Arctic 数据集上的词错误率（%）。每行的最佳结果以**粗体**显示。

奖励引导适配领域的研究表明，增加多样性有助于在强扰动下稳定学习（Lee 等，2025）。

口音鲁棒性。 我们进一步使用 L2-Arctic 数据集（Zhao 等，2018）评估模型对说话者差异和发音差异的鲁棒性。该数据集包含来自六种不同第一语言（L1）背景的非英语母语者。表 2 显示，我们的方法取得了最佳平均词错误率（28.21%），并在阿拉伯语和越南语等更具挑战性的 L1 群体上实现了显著提升。这些结果表明，我们的适应框架能够很好地泛化至多样化的语音系统，缩小母语与非母语语音识别之间的差距。

总体而言，这些发现表明，我们的方法增强了 Whisper 对声学和语言分布偏移的鲁棒性，同时保持较低的推理延迟。与基于熵最小化的方法（SUTA、SGEM）相比，我们的奖励驱动适应在准确率和效率之间实现了更好的平衡，这对于实际的自动语音识别部署至关重要。

消融研究

为更好地了解各个组件的作用，我们在添加了信噪比为 10 dB 高斯噪声的 LibriSpeech test-other 数据集上开展消融研究。除 Whisper-Tiny 外，我们还纳入了 Whisper-Base （参数量分别为 74M 和 39M），以考察模型规模的影响。它在噪声环境下具有更好的识别能力，同时仍能保持高效推理。我们系统地调整了适应框架中的三个设计维度：

- **提示调优**：在解码过程中是否注入可学习的解码器提示。
- **模型微调**：测试时适配过程中是否更新模型参数。
- **奖励建模**：是否使用奖励反馈，以及反馈来自 CLAP（音频-文本）、预训练大语言模型（文本-文本；我们使用 DeepSeek V3（Guo 等，2024）），还是两者结合。

如表3所示，所有适配配置的表现均优于未适配的 Whisper 基线，凸显了提示调优和奖励反馈的有效性。由此可得出几项关键观察：

提示调优的有效性。 比较第二行（仅微调，使用 CLAP 奖励）和第四行（微调 +提示调优，使用 CLAP 奖励），可以看到词错误率显著降低（Whisper-Tiny 从 40.49%→36.65%），而推理延迟仅略有增加。这表明，解码器提示为参数更新提供了互补的适配能力，从而提升了与奖励信号的对齐程度。

奖励建模。 将 CLAP 用作奖励模型，能够显著降低词错误率，且几乎不增加延迟。加入基于大语言模型的反馈可进一步提升效果（例如，Whisper-Base 达到 0.3342→0.3024），但会增加 7–9× 的延迟开销。CLAP+大语言模型混合奖励在整体词错误率方面表现最佳。我们还发现，CLAP 分数与真实词错误率呈负相关（斯皮尔曼 $\rho = -0.431$），这印证了其作为语义奖励的可靠性。

效率考量。 我们还探索了基于 LoRA 的参数高效自适应（Hu等人（2021）），但观察到，与全量微调相比，加速效果微乎其微，因为大部分运行时间都耗费在自回归解码和奖励计算上。因此，我们不再将 LoRA 纳入后续比较，而是着重权衡奖励复杂度与推理效率。

配置			Whisper-Tiny		Whisper-Base	
微调	提示词	奖励	词错误率(%) ↓	延迟（秒）	词错误率(%) ↓	延迟（秒）
N	N	N	45.06	–	40.25	–
Y	N	CLAP	40.49	0.486	35.97	0.573
N	Y	CLAP	42.75	0.472	37.45	0.552
Y	Y	CLAP	36.65	0.489	33.42	0.580
Y	Y	LLM	36.35	4.191	**30.24**	4.355
Y	Y	CLAP +大语言模型	**35.49**	4.365	30.28	4.367

表3：在带噪声的 LibriSpeech test-other 测试集上，不同适配配置的准确率–延迟权衡消融研究。每行启用提示调优、参数微调和奖励反馈（CLAP、大语言模型或两者兼有）的一种组合。报告 Whisper-Tiny 和 Whisper-Base 的词错误率（↓）以及平均推理延迟（秒/条语音）。

高置信度样本子集评估

我们进一步分析了 LibriSpeech test-other 集（Panayotov 等人 2015）中的一个子集，该子集包含在加性高斯噪声下模型置信度最高的 100 个样本。有趣的是，这些高置信度样本的基线词错误率高达 **83.61%**，比整个测试集的词错误率更差。这一反直觉的结果表明，Whisper-Tiny 往往对错误预测表现出很高的置信度；我们将这一现象称为盲目置信，即在分布偏移下，模型的内部确信程度无法反映实际的转录准确率。

如图2所示，SUTA（Lin、Li 和 Lee，2022）高度依赖基于置信度的熵最小化，其性能大幅下降至**122.37%**，加剧了置信度与准确率之间的不一致。SGEM（Kim 等人，2023）引入了序列级不确定性，表现较好，达到**63.00%**，但仍未能完全解决盲目置信的问题。

我们的方法 ASR-TRA 在这一具有挑战性的子集上取得了最低的词错误率**45.17%**，与基线相比，错误减少了近一半。这种鲁棒性源于我们的适配不依赖模型置信度或熵等内部信号，而是利用外部奖励模型（例如 CLAP 和大语言模型）评估转录质量，因此即使模型盲目自信，也能修正预测。通过将适配决策与模型自身的不确定性估计解耦，我们的方法减轻了盲目置信带来的影响。

这些结果表明，在分布偏移的情况下，Whisper-Tiny 的内部置信度指标并不可靠；面对误导性输入时，依赖置信度的方法可能会变得脆弱。奖励引导的自适应提供了更可靠的标准，使自适应信号与模型自身的不确定性估计相互分离。

讨论

我们的实验表明，将因果推理与强化学习相结合，能够实现基于 Whisper 的自动语音识别的有效测试时自适应。与熵最小化（Wang 等人 2021；Lin、Li 和 Lee 2022）和伪标签（Lee 2013；Sohn 等人 2020）等由置信度驱动的基线方法不同，我们的因果干预框架避免了不可靠的置信度信号导致的错误放大，并提供了更清晰的可解释性；置信度子集分析证实了这一点：SUTA 等方法往往会加剧错误。此外，加入互补的外部奖励，能进一步提升模型对声学和语言偏移的鲁棒性：CLAP 提供快速、稳定的音频指导，而 DeepSeek V3 则能提供更精准的语义反馈，但计算成本更高。这些组件共同凸显了这样一种优势：在分布偏移下实现可靠的自适应，与其单纯依赖内部置信度估计，不如利用外部线索。

仍有两项局限。首先，我们的奖励模型存在固有局限：在混合设置下，基于大语言模型的奖励会显著增加推理延迟；而 CLAP 目前主要支持英语音频—文本相似度评估，限制了多语言评估。其次，该框架侧重于单条话语自适应；将其扩展至流式或对话式自动语音识别，有望在实际应用中实现持续且具备上下文感知能力的鲁棒性。此类设置还可能支持隐式少样本学习：跨话语随时间积累的反馈能够提供监督，其作用类似于少样本提示。

结论

我们提出了一种因果强化学习框架 ASR-TRA，可通过三个轻量级阶段在推理过程中自适应 Whisper。自适应首先通过注入可学习的解码器提示，在解码轨迹上实施显式因果干预。为探索不同方案，模型通过温度控制采样生成多样化的转录文本，再由 CLAP 评分并提供奖励。这些奖励驱动策略梯度更新，对提示词和模型参数进行微调，逐步引导模型生成语义更优的输出。

与现有测试时自适应方法相比，该闭环在仅增加有限延迟的同时，持续降低了噪声和带口音语音的词错误率。该方法不依赖模型内部置信度，而是采用外部的模态对齐奖励信号，因此在分布偏移情况下仍具有可解释性和鲁棒性。更广泛地说，将测试时自适应视为由奖励驱动的因果过程，为实用的端侧或低资源语音识别提供了一个有前景的方向，也为语音识别与下游多模态或对话系统的深度集成指明了一条自然路径。

致谢

本研究得到中国国家自然科学基金（编号：62471420）、广东省基础与应用基础研究基金（2025A1515012296）以及2025腾讯人工智能实验室犀牛鸟计划的资助。

参考文献

Baevski, A.; Zhou, Y.; Mohamed, A.; and Auli, M. 2020. wav2vec 2.0：语音表征自监督学习框架。神经信息处理系统进展, 33: 12449–12460. Ben-David, S.; Blitzer, J.; Crammer, K.; Pereira, F.; et al. 2007. 域适配表征分析。载于神经信息处理系统进展（*NeurIPS*）, 137–144. Chan, W.; Jaitly, N.; Le, Q.; and Vinyals, O. 2016. 聆听、关注与拼写。载于*IEEE*声学、语音与信号处理国际会议论文集（*ICASSP*）, 4960–4964. Elizalde, B.; Kumar, P.; Zhan, J.; Shi, H.; et al. 2023. CLAP：从大规模弱监督数据中学习音频-文本联合嵌入。*IEEE*声学、语音与信号处理国际会议论文集（*ICASSP*）。Guo, Z.; Li, X.; Wang, H.; Zhang, L.; Liu, R.; Zhao, Z.; Wu, Y.; et al. 2024. DeepSeek-V3：通过混合专家和强化学习扩展开源语言模型。*arXiv*预印本 *arXiv:2401.06066*。Hendrycks, D.; and Dietterich, T. 2019. 神经网络对常见损坏和扰动的鲁棒性基准测试。载于国际学习表征会议（*ICLR*）。Holtzman, A.; Buys, J.; Du, L.; Forbes, M.; and Choi, Y. 2020. 神经文本退化的奇特案例。载于国际学习表征会议（*ICLR*）。Hsu, W.-N.; Bolte, B.; Tsai, Y.-H. H.; Lakhotia, K.; Salakhutdinov, R.; and Mohamed, A. 2021. HuBERT：通过隐藏单元的掩码预测进行自监督语音表征学习。*IEEE/ACM*音频、语音与语言处理汇刊, 29: 3451–3460. Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, L.; Wang, L.; and Li, W. 2021. LoRA：大语言模型的低秩适配。*arXiv*预印本 *arXiv:2106.09685*。Kim, C.; Park, J.; Shim, H.; and Yang, E. 2023. SGEM：通过序列级广义熵最小化实现自动语音识别的测试时自适应。*arXiv*预印本 *arXiv:2306.01981*。Ko, T.; Peddinti, V.; Povey, D.; and Khudanpur, S. 2015. 用于语音识别的音频增强。载于*Interspeech*论文集, 3586–3589. Lee, D.-H. 2013. 伪标签：一种简单高效的深度神经网络半监督学习方法。载于*ICML*表征学习挑战研讨会。Lee, T.; Chottananurak, S.; Kim, J.; Shin, J.; Gong, T.; and Lee, S. 2025. 采用二元反馈的测试时自适应。*arXiv*预印本 *arXiv:2505.18514*。Lester, B.; Al-Rfou, R.; 与 Constant, N. 2021. 规模的力量：面向预训练语言模型的参数高效自适应。载于*EMNLP* 会议论文集。

Levine, S.; Finn, C.; Darrell, T.; and Abbeel, P. 2016. 深度视觉运动策略的端到端训练。载于《机器学习研究杂志》, 1, 1334–1373. Li, J.; Deng, L.; and Gong, Y. 2020. 鲁棒自动语音识别的多条件训练。载于《鲁棒自动语音识别：通往实际应用的桥梁》, 43–72. 学术出版社。Li, J.; Monroe, W.; Ritter, A.; Galley, M.; Gao, J.; and Jurafsky, D. 2016. 用于对话生成的深度强化学习。载于*EMNLP*, 1192–1202. Li, X. L.; and Liang, P. 2021. 前缀调优：优化用于生成的连续提示。载于*ACL*论文集。Lin, G.-T.; Li, S.-W.; and Lee, H.-y. 2022. 聆听、自适应、更优词错误率：用于自动语音识别的无源单句测试时自适应。*arXiv*预印本 *arXiv:2203.14222*。Logan, B. 2000. 用于语音与说话人识别的梅尔频率倒谱系数。载于*IEEE*声学、语音与信号处理国际会议论文集（*ICASSP*）。McFee, B.; Raffel, C.; Liang, D. P.; Ellis, D. P.; McVicar, M.; Battenberg, E.; and Nieto, O. 2015. librosa：用于音频与音乐信号分析的 Python 库。载于第*14*届*Python*科学会议论文集, 18–25. Mitrovic, J.; McWilliams, B.; Walker, J.; Buesing, L.; and Blundell, C. 2021. 通过不变因果机制进行表征学习。载于国际学习表征会议（*ICLR*）。Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; et al. 2022. 通过人类反馈训练语言模型遵循指令。神经信息处理系统进展, 35: 27730–27744. Panayotov, V.; Chen, G.; Povey, D.; and Khudanpur, S. 2015. LibriSpeech：基于公版有声读物的自动语音识别语料库。载于*IEEE*声学、语音与信号处理国际会议论文集（*ICASSP*）, 5206–5210. Pearl, J. 2009. 因果性：模型、推理与推断。剑桥大学出版社。Radford, A.; Kim, J. W.; Xu, T.; Jeong, G.; et al. 2023. 通过大规模弱监督实现鲁棒语音识别。*arXiv*预印本 *arXiv:2212.04356*。Reddy, C. K.; Dubey, H.; Garg, V.; Cheng, R.; Cutler, R.; and Gehrke, J. 2019. 微软可扩展噪声语音数据集（MS-SNSD）。技术报告，微软。微软技术报告。Schneider, S.; Baevski, A.; Collobert, R.; and Auli, M. 2019. wav2vec：用于语音识别的无监督预训练。载于*Interspeech*, 3465–3469.

Sch¨olkopf, B.; Locatello, F.; Bauer, S.; Ke, N. R.; Kalchbrenner, N.; Goyal, A.; and Bengio, Y. 2021. 迈向因果表征学习。*IEEE* 汇刊, 109(5): 612–634. Sohn, K.; Berthelot, D.; Li, C.-L.; Zhang, Z.; Carlini, N.; et al. 2020. FixMatch：通过一致性和置信度简化半监督学习。神经信息处理系统进展, 33: 596–608. Sutton, R. S.; and Barto, A. G. 2018. 强化学习：导论。麻省理工学院出版社，第2版。Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. 注意力即一切。载于神经信息处理系统进展（*NeurIPS*）。Wang, D.; Zhu, E.; Yue, X.; and Gonzalez, J. E. 2021. Tent：通过熵最小化实现完全测试时自适应。载于国际学习表征会议（*ICLR*）。Williams, R. J. 1992. 用于联结主义强化学习的简单统计梯度跟随算法。载于*Machine Learning*, 3–4, 229–256. Zhao, G.; Hirst, D.; Povey, D.; and Khudanpur, S. 2018. L2-ARCTIC：非母语英语语音语料库。载于*IEEE*声学、语音与信号处理国际会议论文集（*ICASSP*）, 5204–5208. Zhao, S.; Wang, X.; Zhu, L.; and Yang, Y. 2024. 使用 CLIP 奖励进行测试时自适应，以提升视觉语言模型的零样本泛化能力。载于国际学习表征会议（*ICLR*）。Ziegler, D. M.; Stiennon, N.; Wu, J.; Brown, T.; et al. 2019. 基于人类偏好微调语言模型。*arXiv*预印本 *arXiv:1909.08593*。