arXiv:2602.04535v1 [cs.SD] 4 Feb 2026

HoliAntiSpoof：用于整体语音反欺骗的音频大语言模型

Xuenan Xu [1] Yiming R en [1] Liwei Liu [1] Wen Wu [1] Baoxiang Li [1] Chaochao Lu [1] Shuai Wang [2] Chao Zhang [1]

摘要

语音合成与编辑的近期进展使语音欺骗变得愈发难以应对。然而，现有方法大多将语音欺骗视为二元分类问题，忽略了多种语音欺骗技术会操纵多个相互关联的语音属性及其语义效应。本文提出**HoliAntiSpoof**，这是首个用于整体语音反欺骗分析的音频大语言模型（音频大语言模型）框架。

HoliAntiSpoof 将语音欺骗分析重新定义为统一的文本生成任务，从而能够联合推理语音欺骗方法、受影响的语音属性及其语义影响。为支持语义层面的分析，我们推出**DailyTalkEdit**，这是一个全新的反欺骗基准，可模拟真实的对话操纵，并提供语义影响标注。大量实验表明，HoliAntiSpoof 在多种设置下均优于传统基线方法；初步结果还显示，上下文学习进一步提升了域外泛化能力。这些发现表明，音频大语言模型不仅能够提升语音欺骗检测性能，还能对欺骗行为及其语义效应进行可解释分析，为构建更可信、可解释的语音安全系统指明了方向。数据和代码已公开[1]。

1. 引言

生成模型的最新进展显著推动了文本到语音（TTS）合成和语音编辑。最新模型展现出合成话语的能力，其逼真度和可懂度已达到近似人类的水平（Tan 等人，2021；Du 等人，2025；Zhou 等人，2025；Peng 等人，2024）。这些能力在带来新应用的同时，也加剧了伦理风险，使有害内容得以

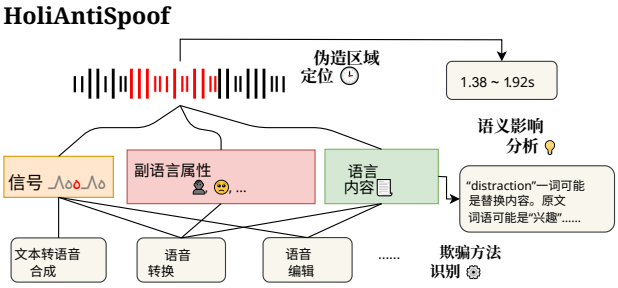


图*1.* 整体欺骗分析涵盖全面的语音感知与理解，可执行超越真实/伪造分类的多项子任务。

生成，或对敏感信息进行操纵。因此，语音反欺骗（即语音深度伪造检测）变得日益重要。

传统反欺骗研究主要聚焦于真实/伪造二元分类。后端分类器通常基于前端特征（Wang & Yam-agishi，2022；Zhang 等人，2024a）或直接基于原始波形（Tak等人，2021a；b；Jung等人，2022）来判断某段话语是否遭到欺骗，或定位其中的欺骗片段。近期的初步尝试已探索将音频大语言模型（ALLM）应用于欺骗检测（Gu 等人，2025；Xie 等人，2026）。

然而，二元分类的表述并不能完整呈现语音欺骗分析。有效检测需要全面理解语音的多个层面：（i）低层信号属性，例如通道信息或合成过程中引入的伪影；（ii）副语言属性，例如说话人身份和情绪；以及（iii）文本传达的语言内容。如图1所示，不同的语音欺骗方法会以不同方式影响这些层面：文本转语音会在信号层面引入波形伪影；语音转换（VC）可能改变副语言属性，从而产生不自然的表达；神经语音编辑则会修改真实话语的部分内容，同时融合信号级伪影和语言内容。因此，在二元分类的基础上，同时检测语音欺骗方法和欺骗区域，可使模型从耦合语音信息的完整范围中学习反欺骗能力。相比之下，由于训练数据不足和学习目标的局限，传统模型可能会过拟合特定模式

[1]上海人工智能实验室 [2]南京大学。通信作者：Chao Zhang <cz277@tsinghua.edu.cn>。

预印本。*2026*年*2*月*5*日。[1]

https://github.com/wsntxxn/HoliAntiSpoof

to 训练数据不足和学习目标 es.

尽管整体欺骗分析十分重要，但在反欺骗研究中，这一方向仍鲜有探索。以往的反欺骗研究主要关注语音的信号级真实性，对欺骗性话语的语义影响关注有限。在实际应用中，修改或捏造话语含义的篡改，可能造成远比纯粹信号级失真更严重的后果。敏感对话中的细微改动（例如 在金融协议中将“是”改成“否”）就可能彻底颠倒意图含义，并对现实世界造成重大影响；而对不太关键的词语进行修改（例如，将“是”改成“嗯”）在语义上往往无关紧要，在实际语音欺骗场景中也不太可能发生。因此，在上下文中分析被篡改的词语或句子所产生的语义影响，是语音反欺骗研究中至关重要却尚未得到充分探索的环节。

为弥补这些不足，我们提出了**HoliAntiSpoof**，首个用于整体语音欺骗分析的基于音频大语言模型的系统。反欺骗本质上要求具备强大的跨领域泛化能力，因为生成模型和编辑模型在持续演进，新的欺骗方法也频繁出现。为此，我们将音频大语言模型应用于反欺骗任务；事实证明，音频大语言模型能够在统一框架内有效处理多种语音理解任务。通过大规模预训练，音频大语言模型在各项任务中均取得了卓越性能，尤其是在自动语音识别（自动语音识别）和语音情感识别（语音情感识别）等任务上。借助这些优势，我们将一个音频大语言模型重新用于构建统一的语音欺骗分析器。HoliAntiSpoof 不再设计相互独立的模块，而是将所有子任务整合为单一的文本预测任务。具体而言，我们将丰富的欺骗标签（包括真实/伪造标签、欺骗方法、欺骗区域以及潜在的语义影响）转换为结构化标注，并通过标准的下一词元预测目标对音频大语言模型进行微调。

由于现有的语音欺骗数据集均未提供关于欺骗性话语语义影响的标注，我们构建了两个互补的资源。1）我们扩展了 Par- tialEdit（Zhang 等人，2025），该数据集中单个话语最多有两个词被修改；在此基础上，我们对修改的语义影响进行了标注。2）我们提出了**DailyTalkEdit**，该数据集源自 DailyTalk（Lee 等人，2023年），其中对话中的完整话语会被替换为修改并重新合成的版本。这模拟了现实中的对话操纵，例如改变说话人在多轮对话中的预期回应，从而反映金融、法律或社会场景中的实际风险。

我们使用现有语音欺骗数据集与两个新提出的面向语义的数据集的组合来训练 HoliAntiSpoof。大量实验表明，HoliAntiSpoof 能够进行整体欺骗分析，同时在语音欺骗检测方面始终优于传统模型。我们进一步研究了 HoliAntiSpoof 的上下文学习（ICL）能力，结果表明，无需重新训练模型，仅提供少量参考示例，就能显著提升其在跨语言和跨领域语音欺骗场景中的表现，凸显了其适应未见数据分布的潜力。

我们的贡献总结如下：

•我们将多种语音反欺骗任务统一到一个音频大语言模型框架中，使音频大语言模型首次能够作为整体语音欺骗分析器运行。

• 为将语义影响分析纳入语音欺骗研究，我们提出了 DailyTalkEdit 数据集，以促进后续研究。

• HoliAntiSpoof 在域内和域外评估中均取得了最先进（SOTA）的性能。初步的上下文学习（ICL）实验进一步展现了其向新领域泛化的潜力。

2. 相关工作

2.1. 传统音频反欺骗研究

音频反欺骗（亦称音频深度伪造检测）旨在防范恶意音频合成和篡改攻击（Yi 等人，2023）。早期研究主要聚焦于设计手工特征，以捕捉信号级伪影（Todisco 等人，2017）。随着深度学习的兴起，直接处理原始波形的端到端模型已广泛应用（Tak 等人，2021b；a；Jung 等人，2022）。近年来，自监督学习（SSL）表征（如 HuBERT（Hsu 等人，2021）和 WavLM（Chen 等人，2022））已展现出在广泛语音任务中的强大效能，包括区分真实语音与欺骗语音（Wang 和 Yamagishi，2022）。近期研究不再局限于话语级二分类，而是将任务拓展至多个音频域（Zhang 等人，2024b；Comanducci 等人，2024；Xie 等人，2024），以及对部分篡改话语中的欺骗区域进行检测（Zhang 等，2022；Huang 等，2024）。

尽管取得了这些成果，先前的音频反欺骗研究主要关注整段语音话语或定长片段的信号级真实性。它们无法评估攻击的语义影响，即，语音欺骗操作可能如何改变说话人的意图含义。为弥合这一差距，HoliAntiSpoof 将语义影响分析与传统欺骗检测统一到音频大语言模型的文本生成目标中。

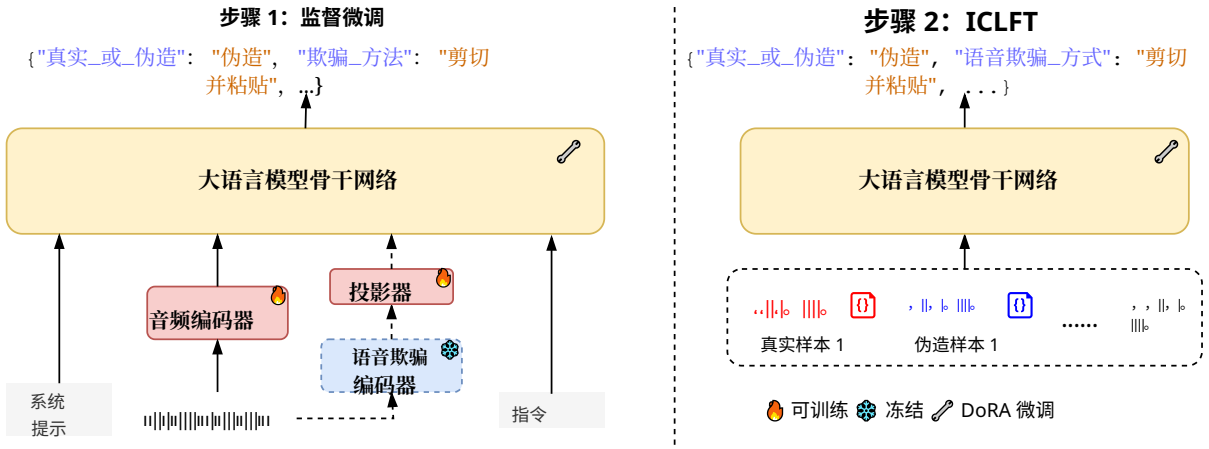


图 2. HoliAntiSpoof 概述。首先对音频大语言模型进行微调，使其生成结构化文本，其中包含信号级和语义级的语音欺骗分析。我们还探索了面向欺骗检测的特征的引入。随后进一步微调大语言模型，使其具备上下文学习能力。

2.2. 音频大语言模型

音频大语言模型通常通过轻量级适配器，将预训练音频编码器与大语言模型对齐，使模型能够感知音频输入并进行推理。SALMONN（Tang 等人，2024）和 Qwen-Audio（Chu 等人，2023年）等开创性工作，将自动语音识别和语音情感识别等多种语音任务统一到单个遵循指令的音频大语言模型中。近期工作探索了离散音频表征的集成（Ding 等人，2025）、理解与生成的统一（Wu 等人，2025），以及音频大语言模型推理能力的提升（Tian 等人，2025）。通过将训练数据规模扩大到数十亿小时，当前的音频大语言模型在多种音频理解任务中展现出卓越的性能。

近期，研究人员开始探索将音频大语言模型应用于音频深度伪造检测（Gu 等人，2025；Xie 等人，2026年）。Gu 等人 (2025)对音频大语言模型进行了微调，以用于欺骗检测，并表明其在数据稀缺的场景中表现优异；而Xie 等人 (2026年)采用基于规则的时域和频域奖励进行强化学习，从而提升模型在不同音频类型上的欺骗检测鲁棒性。然而，他们主要将音频大语言模型改作强大的分类器来执行二元分类，未能充分发挥其进行整体欺骗分析的潜力。当前的音频大语言模型主要经过预训练，用于语义提取和理解（例如 自动语音识别），因此对感知质量或细微合成模式等低层声学线索的敏感度较低（Wanget al.，2025）。为此，HoliAntiSpoof 将欺骗检测、定位、方法识别和语义影响分析融为一体，充分发挥音频大语言模型的优势，实现整体欺骗分析，无需提供详细的推理标注。

3. HoliAntiSpoof

如图 2所示，我们将整体欺骗分析任务重新表述为统一的文本生成目标，以便将音频大语言模型应用于整体欺骗分析。HoliAntiSpoof 首先通过标准的监督微调（SFT）进行训练，然后再通过上下文学习目标进行微调。

3.1. 整体欺骗分析

为了将二元信号级检测任务扩展为整体分析，我们将多个任务统一为单一的结构化文本生成目标。分析涵盖以下方面，并以 JSON 格式文本呈现：

- **真实性分类**：基本任务是判断输入音频是真实还是伪造。
- **时间定位**：对于部分伪造样本，模型会预测欺骗区域的具体时间区间。
- **方法识别**：模型会将攻击归类为以下六种语音欺骗技术之一：(1)文本转语音，(2)语音转换，(3)剪切粘贴，(4)语音编辑，(5) 声码器重合成，以及 (6)编解码器重合成。详细说明见第 A节。
- **语义影响分析**：对于部分伪造样本，模型会生成一段文字描述，说明这种修改可能如何影响原始意图、语气或事实信息。

3.2. 模型架构

HoliAntiSpoof 构建于统一的音频大语言模型架构之上，采用预训练的 Qwen2.5-Omni（Xu 等人，2025）作为初始化骨干网络。整体框架由音频编码器和大语言模型骨干网络组成。为评估音频编码器是否足以提取

与语音欺骗相关的表征，我们还引入了欺骗检测编码器进行消融分析。

音频编码器。 音频编码器从原始波形 $\mathcal{A}$中提取声学特征。输入波形首先转换为对数梅尔频谱，然后通过卷积嵌入层和堆叠的 Transformer 块进行处理，生成高层声学嵌入 $\mathcal{E}_a \in \mathbb{R}^{l\times d}$，时间分辨率为 25 Hz。为使声学表征与文本模态对齐，轻量级多层感知机（MLP）将 $\mathcal{E}_a$ 投影到大语言模型骨干网络的嵌入空间中。

尽管音频编码器在大规模数据上进行了预训练，但其训练目标主要侧重于语义信息，例如语音转写文本和音频事件识别。与感知质量和合成伪影相关的低层声学线索，通常被视为语义理解中的干扰因素，在预训练过程中可能会被抑制，尽管这些线索对欺骗检测非常有价值。由于整体欺骗分析需要兼顾低层伪影感知和高层语义推理，我们进一步探索一种扩展方案，将原始音频编码器与专门针对欺骗检测进行预训练的编码器相结合（Ge 等人，2025）。欺骗检测编码器从输入音频中提取稠密嵌入 $\mathcal{E}_s \in \mathbb{R}^d$。经过 MLP 投影后，$\mathcal{E}_s$ 与 $\mathcal{E}_a$ 沿时间轴拼接，再输入大语言模型骨干网络。

大语言模型骨干网络。 大语言模型接收多模态嵌入，并以自回归方式生成目标响应。我们采用 Qwen2.5-Omni 的 Thinker 部分作为骨干网络。给定音频嵌入后，模型以文本生成的方式产生结构化欺骗分析输出。

3.3. 训练阶段

监督微调。 大语言模型首先通过标准监督微调进行微调。形式上，模型被训练为基于音频嵌入 $\mathcal{E}_a$最大化目标输出 y 的似然：

$$\mathcal{L}_{\mathrm{SFT}} = -\sum_{t=1}^{T} \log p(y_t \mid y_{<t}, \mathcal{E}_a), \tag{1}$$

其中 T 为目标词元数量，为简洁起见省略了系统提示词和指令。

音频编码器采用全量微调，因为语音欺骗检测可能依赖于预训练期间未能充分捕获的低层声学特征。对于骨干网络，我们采用 DoRA (Liu 等人，2024) 来实现高效适配，同时保留预训练模型的指令遵循和文本生成能力。DoRA 将预训练权重分解为幅值和方向。在微调过程中，方向通过低秩适配进行更新，而幅值则独立学习，从而实现以尽量减少对预训练模型干扰的方式高效更新参数：

$$W' = m \cdot \frac{W_0 + BA}{\|W_0 + BA\|}, \tag{2}$$

其中 $W_0 \in \mathbb{R}^{d\times k}$ 为冻结的预训练权重，m 为可训练幅值，$B \in \mathbb{R}^{d\times r}, A \in \mathbb{R}^{r\times k}$ 为可训练的低秩矩阵。

ICLFT。 在标准监督微调之后，我们进一步以上下文学习微调（ICLFT）范式对模型进行微调，以使其具备零样本适配[2] 能力。具体而言，我们用一小批参考示例增强输入提示，其中包括真实与伪造音频样本及其对应的结构化标注。形式上，令 $\{(\mathcal{E}_{a_1}, y_{a_1}) \ldots, (\mathcal{E}_{a_K}, y_{a_K})\}$ 表示 K 个上下文示例。训练目标定义为：

$$\mathcal{L}_{\mathrm{ICLFT}} = -\sum_{t=1}^{T} \log p\left(y_t \,\middle|\, y_{<t}, \underbrace{(\mathcal{E}_{a_1}, y_{a_1}), \ldots, (\mathcal{E}_{a_K}, y_{a_K})}_{\text{in-context examples}}, \mathcal{E}_a \right).$$

通过在 ICLFT 中训练模型以条件化于少量参考示例，模型能够从上下文中推断与语音欺骗相关的模式，从而有效地将零样本泛化到未见域和语言，并且测试时无需训练。

4. 语义分析数据构建

为实现全面的语音深度伪造分析，尤其是理解欺骗内容的语义影响，数据集不能仅限于简单的二元标签。由于现有反欺骗数据集缺少能够反映欺骗操作语义影响的标注，我们构建了面向语义的资源来填补这一空白。我们首先选用**PartialEdit** (Zhang 等人，2025）作为基础。该数据集会修改 VCTK 语料库中的短单说话人句子中的一两个词（Veaux 等人，2013年）。然而，现实中的操纵也可能发生在连续对话中，而上下文在其中起着关键作用。为此，我们构建了一个名为**DailyTalkEdit**的新数据集，将语音欺骗场景扩展至多轮对话。最后，我们通过统一流程为两个数据集标注语义影响。数据构建流程如图3所示。

4.1. DailyTalkEdit数据集

不同于修改孤立的句子，对话中的语音欺骗要求被操纵的话语在语言和语义上都与上下文保持连贯，从而使攻击更加真实。我们设计了一种迭代式对话

[2]按照（Brown 等人，2020）的定义，此处“零样本”是指测试时没有进行梯度下降。

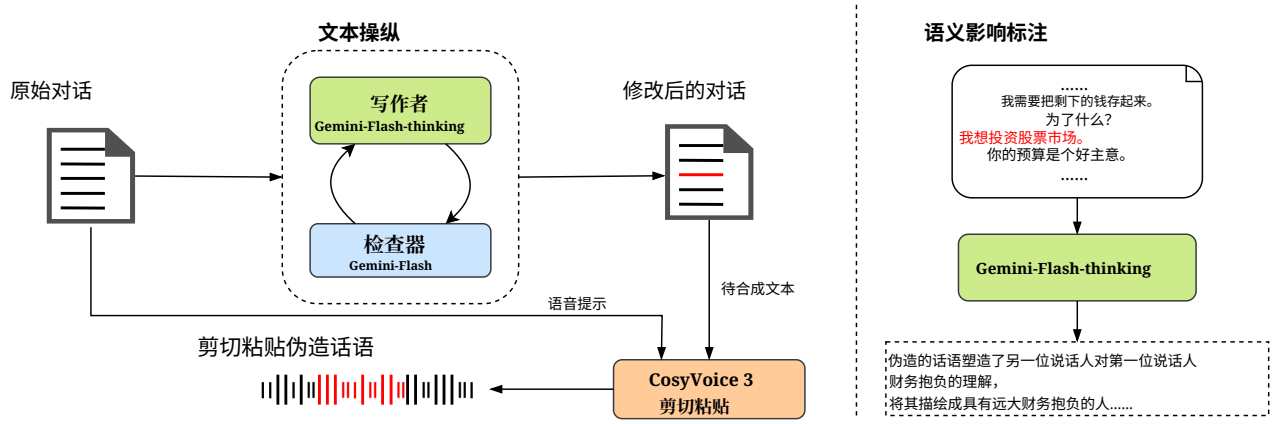


图*3.* DailyTalkEdit数据集的数据构建流程，以及对欺骗语义影响的标注。双智能体工作流对 DailyTalk 中的对话进行语境连贯的修改，以便后续进行 TTS 和剪切粘贴式语音欺骗。随后，将欺骗文本输入 Gemini，以标注其潜在影响。

操纵工作流，利用两个智能体生成欺骗文本。随后，使用先进的零样本文本转语音模型合成欺骗话语，以构建数据集。

直接提示大语言模型修改对话中的某个句子，往往会导致生成的文本与其余上下文相矛盾。为生成语境上合理的语音欺骗样本，我们采用由写作者和检查器组成的双智能体工作流。

- **写作者：** 写作者从对话中选择一个目标句子，并生成相应的欺骗版本。系统指示其改变语义（例如，反转意图或更改事实细节），以可能造成负面后果，同时保持语境连贯性。此步骤使用经过推理增强的 Gemini-2.5-Flash，因为它需要理解整个对话，并推理出语义上合理的修改。

- **检查器：** 尽管要求写作者在实施语音欺骗时保持语境连贯性，但我们发现这一要求并非总能得到满足。因此，我们使用另一个大语言模型作为检查器。它负责验证修改后的话语与原始对话之间的语境连贯性和事实一致性。我们使用非推理模式的 Gemini-2.5-Flash 执行这一任务，因为实证评估表明，它足以完成连贯性验证。

写作者会迭代地提出修改，直到检查器批准结果，或达到预定义的最大迭代次数 $N = 3$。未能在此限制内通过检查器的样本将被丢弃。该流程确保欺骗文本既在语境上合理，又难以被检测。从DailyTalk中的2,541段对话出发，我们的流程生成了2,475个欺骗样本，经过筛选后，命名为*DailyTalkEdit* 数据集。

在生成修改后的文本后，我们合成相应的语音样本以进行语音欺骗。我们使用最先进的零样本文本转语音模型CosyVoice-3（Du等人，2025）进行声音克隆。具体而言，当原始话语超过2秒时，将其用作语音提示，以保留说话人的音色和韵律。否则，选择同一说话人最长的可用话语作为提示。随后，合成语音替换对话中的原始话语，从而得到一种“剪切粘贴”式的语音欺骗样本。

4.2. 语义分析标注

为便于训练 HoliAntiSpoof 分析欺骗内容的语义影响，我们需要对相应的语音欺骗攻击如何影响所传达的信息进行文字描述标注。我们利用 Gemini-2.5-Flash 思考模式的推理能力生成这些描述。具体而言，对于 PartialEdit 和 DailyTalkEdit 中的每个样本，我们将经过篡改的文本或对话输入大语言模型，并在提示中指出被修改的词语或句子。大语言模型需要结合可能的原始词语或句子，分析欺骗操作造成的影响，并生成简洁的文字分析。

5. 实验设置

5.1. 数据集

HoliAntiSpoof 由现有的英语反欺骗数据集与两个新构建的数据集混合训练而成：语义增强的 PartialEdit 和 DailyTalkEdit。所有数据集的详细说明见第 B.1 节。这些数据集涵盖了

第 3.1 节所列出的欺骗方法，包括由早期 TTS 或 VC 方法生成的传统基准数据集，以及使用最新模型合成的近期数据集。为缓解数据不平衡问题，我们在训练期间从每个数据集中随机抽取最多 5 万个样本。

为了进行评估，我们从每个数据集中抽取子集并混合，采用真假标签分层策略构建全面的域内测试集，并将其称为“混合”。这使得跨数据集的评估方案更加均衡。我们还报告了标准基准 ASVSpoof2019LA（ASV19.）评估集（Nautsch 等人，2021）上的结果，以便与以往工作直接比较。此外，我们纳入了四个域外评估集，以评估 HoliAntiSpoof 在未见过的语言和语音合成模型上的泛化性能。详情见第 B.2节。

5.2. 超参数

在监督微调期间，我们使用 AdamW 优化模型，参数为 $\beta = (0.9, 0.95)$，权重衰减设为 0.1。我们采用峰值学习率 1×10^{-5}，并使用余弦衰减调度，预热比例为 0.05。训练共进行 20K 步，批量大小为 64。我们对 Transformer骨干中的所有投影层应用 DoRA，秩为 $r = 64$，缩放因子为 $\alpha = 128$，丢弃率为 0.05。

在 ICLFT 期间，我们将学习率降至 5×10^{-6}。模型训练 5K 步，批量大小为 32。音频编码器保持冻结，并将 DoRA 秩 r 降至 8，以实现轻量且稳定的适配。

5.3. 基线方法

我们将 HoliAntiSpoof 与以下基线方法进行比较：

• **传统反欺骗模型**。我们纳入了广泛使用的开源基线方法，包括 RawNet2（Tak等人，2021b）、RawGAT-ST（Tak 等人，2021a）、AA-SIST（Jung等人，2022）[3]以及 ResNet-Transformer（RT）（Cai等人，2023年）。RawNet2、RawGAT-ST 和 AASIST 以原始波形为输入，而 RT 则在通过自监督学习（ SSL ）获得的特征之上采用浅层 ResNet-Transformer 检测器。我们探索了三种 SSL 特征：Wav2Vec2（Baevski等人，2020）、HuBERT（Hsu等人，2021）和 WavLM（Chen等人，2022）。这些传统基线方法与 HoliAntiSpoof 使用相同的数据进行训练，并采用包含共享编码器和多个任务特定预测头的多任务设置。请注意，RawGAT-ST 和 AASIST 无法执行欺骗区域定位，因为它们生成的是图嵌入，而非片段级嵌入。以上模型均无法进行语义影响分析，因为它们不包含语言解码器。

• **音频大语言模型基线**。 我们报告最新专有音频大语言模型 Gemini-3-Flash 的结果，作为最强通用音频大语言模型性能的参考。为节省推理成本，Gemini 的评估不包含 ASVSpoof2019，因为其测试集规模较大。

5.4. 指标

我们从多个角度评估 HoliAntiSpoof，涵盖二元欺骗检测和整体分析输出：

• **真实性分类**。 对于真实/伪造真实性分类，我们报告二元分类准确率（Acc.）。与等错误率（EER）等基于分数的指标相比，Acc. 使我们能够与无法访问词汇表中词语概率的闭源大语言模型进行比较。不过，我们仍可通过提取相应位置的词元 logits 来计算 EER，详情见第 C节。

• **欺骗方法识别**。 对于方法识别，我们采用宏平均 F_1 分数。我们将文本转语音合成 和语音转换 合并为一个类别，因为许多欺骗样本是通过结合这两种技术的混合流程生成的。

• **时间定位**。 对于部分欺骗话语中欺骗区域的定位，我们采用时间分辨率为 0.2 秒的片段级 F_1 （Seg-F_1）（Mesaros 等人，2016）。Seg-F_1 通过在定长片段上匹配预测结果与真实标签来评估定位性能，因此对边界碎片化不那么敏感。

• **语义影响分析**。 使用自动指标评估语义影响分析本质上十分困难。因此，我们采用大语言模型评审协议，并提示 Gemini-3-Flash 根据三项标准，以 1–5 分为生成的分析打分：(1) 作为独立文本的流畅性；(2) 对被操纵的词语或句子的识别是否正确；(3) 语义影响分析是否合理。为降低评分差异，我们让评审模型评估三次，并报告平均分。

我们在混合测试集上应用全部四种评估视角。对于其他测试集，我们主要关注真实性分类这一核心任务。由于 HAD (Yi et al., 2021)中欺骗性话语大多是部分伪造的，因此我们还报告了 HAD 上的伪造区域定位性能。

6. 结果

我们首先将 HoliAntiSpoof 与基线方法进行比较，以探究音频大语言模型在整体欺骗分析任务上的表现。随后，我们考察影响 HoliAntiSpoof 的关键因素。

[3]https://github.com/clovaai/aasist

表*1.* HoliAntiSpoof 与基线方法的整体欺骗分析性能比较，涵盖真实性分类（真实性）、欺骗方法识别（方法）和欺骗区域定位（位置）。“F”和“T”分别表示自监督学习骨干网络被冻结或可训练。在所有测试集上，数值越高表示性能越好。

模型	域内				域外			
	ASV19。	混合			SF-MD.	SpoofCeleb	HAD	
	真实性	真实性	方法	Loc.	真实性	真实性	真实性	Loc.
传统模型								
RawNet2	77.69	91.24	90.16	53.57	88.42	68.06	5.56	37.84
RawGAT-ST	96.35	90.42	82.50	-	93.69	85.70	31.86	-
AASIST	96.29	94.29	**96.44**	-	92.11	75.91	10.49	-
Wav2Vec2（F）- 实时	79.08	70.68	65.86	41.10	69.46	41.46	47.65	37.12
Wav2Vec2（T）- 实时	79.48	91.49	90.44	53.41	91.78	41.09	22.31	<u>42.03</u>
HuBERT（F）- 实时	94.53	88.28	81.74	55.15	88.39	72.40	39.29	40.73
HuBERT（T）- 实时	85.32	91.14	90.99	<u>56.37</u>	89.30	61.63	25.80	37.36
WavLM（F）- 实时	94.72	86.29	78.23	44.19	91.05	58.06	29.56	38.19
WavLM（T）- 实时	74.56	88.77	90.96	54.90	91.49	63.89	14.31	**42.96**
ALLM 方法								
Gemini-3-Flash	-	63.39	16.10	33.26	69.49	60.73	78.59	29.55
HoliAntiSpoof	**96.59**	**96.16**	<u>95.12</u>	**91.33**	**97.89**	**90.36**	**90.19**	41.27

6.1. 全面的伪造分析性能

HoliAntiSpoof 与传统基线方法的比较结果见表 1。在域内评估中，HoliAntiSpoof 在各个维度上均展现出强劲且稳定的表现。作为专有大语言模型，Gemini-3-Flash 无法可靠地区分真实语音与伪造语音，表现接近随机水平。与具有竞争力的 AASIST 相比，HoliAntiSpoof 在广泛使用的 ASVSpoof2019 上取得了相当的表现。在全面的域内混合评估集上，HoliAntiSpoof 的表现优于所有传统方法，其语音欺骗方法分类结果也接近表现最佳的 AASIST。对于 ResNet-Transformer 基线方法，使自监督学习特征提取器可训练通常会提升混合测试集上的真实性分类表现，但会降低其在 ASVSpoof2019 上的表现，这表明模型跨域稳健性有限。相比之下，HoliAntiSpoof 在两个评估集上均保持较高的检测准确率。在时间定位方面，HoliAntiSpoof 显著优于传统方法，Seg-F_1 分数超过 90，而所有传统模型均低于 60。

在域外评估中，HoliAntiSpoof 相较于传统基线方法展现出更显著的优势。尽管若干传统模型在 SF-MD. 上具有良好的泛化能力，但它们在其他域外测试集上的性能却大幅下降。特别是，传统方法呈现出域内与域外性能之间的权衡：例如，尽管 AASIST 在域内混合数据集上优于 RawGAT-ST，但在迁移至未见域外测试集时，其表现却逊于 RawGAT-ST。相比之下，HoliAntiSpoof 在域内和域外评估中均能保持稳定且较高的欺骗检测准确率，这表明其对域偏移具有更强的鲁棒性。我们推测，这种泛化能力的提升可能归因于较大的模型容量和可迁移表征。

域外结果分析表明，HoliAntiSpoof 呈现出与传统模型类似的趋势。具体而言，欺骗方法仍是影响跨域泛化的主导因素。例如，SF-MD. 上的性能仍然很高，因为它采用了与 SF-BD. 相同的生成模型，而 SF-BD. 已包含在训练数据中，尽管目标语言在训练期间未曾见过。相反，当欺骗方法未曾见过时，性能会大幅下降，如 SpoofCeleb 上的结果所示。此外，当语言和欺骗方法都未曾见过时，性能下降会更加明显，HAD 上的结果体现了这一点。

6.2. 消融研究与分析

我们进一步分析不同组件对 HoliAntiSpoof 的影响，并探讨以下问题：**1)**整体欺骗分析是否有助于语音欺骗检测？**2)** 上下文学习能否提升 HoliAntiSpoof 的泛化性能？**3)** 面向语音欺骗的特征是否与 HoliAntiSpoof 的原始音频特征互为补充?

6.2.1. 统一欺骗目标的影响

我们考察统一欺骗目标对基本欺骗检测性能的影响。遵循（Gu 等人，2025），

表2. 关于数据格式、包含 3 对上下文示例的上下文学习，以及引入面向欺骗检测的特征所产生影响的消融与分析结果。所有测试集上的数值越高，表示性能越好。

模型	域内				域外		
	混合				SpoofCeleb	HAD	
	真实性	方法	Loc.	Sem.	真实性	真实性	Loc.
HoliAntiSpoof	96.16	**95.12**	91.33	3.51	90.36	90.19	41.27
w. ICLFT	96.08	94.93	90.19	**3.64**	91.24	**93.92**	**79.06**
w. 欺骗特征	**98.36**	94.37	90.82	3.59	**94.82**	88.26	46.39
DoRA $r = 16$	96.34	94.53	91.04	3.55	91.90	93.92	38.68
DoRA $r = 256$	95.74	94.69	90.24	3.63	93.83	95.61	43.81
仅真实性	96.68	-	-	-	94.96	68.20	-

我们通过训练模型生成单个词元标签（真实或伪造），将学习目标简化为真实性分类，从而能够直接比较 HoliAntiSpoof 与纯粹用作分类器的大语言模型。若不采用整体分析目标，模型在域内测试集上的表现相近，但在域外HAD数据集上的性能则显著下降。此外，如表 5中的等错误率结果所示，在最优决策阈值下，HoliAntiSpoof 在 SpoofCeleb 上也优于仅进行真实性分类的变体。这些结果表明，整体欺骗分析目标通过鼓励模型学习并利用任务间的内在相关性，提升了泛化能力。

6.2.2. 上下文学习的融入

我们进一步考察 ICLFT 是否能提升跨域泛化。对于域内评估，仅将目标音频作为输入；对于域外评估，则在提示中额外预置 3 对上下文示例（不进行梯度下降）。如表2所示，ICLFT（w. ICLFT）仅导致域内混合测试集上的性能略有下降，表明微调后模型的核心语音欺骗分析能力基本得以保留。在域外测试集上，ICLFT 显著提升了 HAD 上的表现，而 HAD 主要由通过剪切粘贴篡改生成的部分伪造样本构成。这一提升表明，ICLFT 增强了模型利用上下文示例进行领域适应的能力。然而，在语音欺骗模式更加多样的 SpoofCeleb 上，ICLFT 带来的收益有限，这表明如何充分发挥 ICL 在语音欺骗泛化方面的潜力仍是一项开放性挑战。

6.2.3. 引入面向欺骗检测的特征的优势

最后，我们研究了音频表征的影响。如表2所示，加入面向欺骗检测的特征（带欺骗特征）显著提升了真实性分类在大多数数据集上的表现，但欺骗方法识别和区域定位的性能略有下降。我们还改变 DoRA 秩 r，以比较声学表征与大语言模型适配能力的影响：与增大 r相比，加入面向欺骗检测的特征能更显著地提升真实性分类，而增大 r 则能提高语义影响分数。欺骗检测从具有代表性的特征中获益更多，而语义分析与推理则更受益于更强的大语言模型适配能力。这表明，预训练的音频大语言模型能够有效捕捉音频中的高层语义内容，但缺少用于真实性分类的、对语音欺骗敏感的低层声学特征。总体而言，这些结果证实，面向伪造检测的表征与原始音频特征互为补充，有助于进行整体欺骗分析。

7. 结论

在本研究中，我们提出了 HoliAntiSpoof，这是首个基于音频大语言模型的整体语音欺骗分析框架。HoliAntiSpoof 将传统的真实/伪造二元分类扩展为统一分析，涵盖欺骗方法识别、伪造区域定位和语义影响分析，并将这些任务统一建模为结构化文本生成任务。为支持语义推理，我们为 PartialEdit 增添了语义标注，并推出 DailyTalkEdit，这是一个模拟真实对话伪造场景的对话级数据集。实验结果表明，HoliAntiSpoof 在域内和域外评估中均显著优于传统基线方法，同时，加入面向欺骗检测的特征可进一步提高真实性分类准确率。最后，对上下文学习的初步探索表明，只需少量示例，便有望使音频大语言模型适应未见域，这也促使我们在未来进一步发挥其能力。

影响声明

人工智能生成内容的快速发展使高保真语音深度伪造对个人隐私、金融安全和社会稳定构成了重大威胁。通过将信号级检测与语义影响分析相结合，本研究为公众提供了一种更全面、可解释的工具，用于识别和理解恶意音频篡改。尽管本研究的首要目标是防御，我们也承认，攻击者理论上可以研究用于分析欺骗机制的方法，以开发更复杂且更难以辨别的攻击。不过，我们认为，提供稳健、通用且可解释的防御系统所带来的益处远远超过这些风险。

参考文献

Baevski, A., Zhou, Y., Mohamed, A., andAuli, M. wav2vec 2.0：一种用于语音表征自监督学习的框架。神经信息处理系统进展，33:12449–12460，2020年。

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shya m, P., Sastry, G., Askell, A., 等。语言模型是少样本学习器。神经信息处理系统进展，33: 1877–1901，2020年。

Cai, Z., Wang, W., Wang, Y., and Li, M. 用于ADD2023篡改区域定位任务的DKU-DUKEECE系统。*arXiv*预印本 *arXiv:2308.10281*，2023年。

Chen, S., Wang, C., Chen, Z., Wu, Y., Liu, S., Chen,Z.,Li, J., Kanda, N., Yoshioka, T., Xiao, X., 等。WavLM：面向全栈语音处理的大规模自监督预训练。*IEEE*信号处理精选主题期刊，16(6):1505–1518，2022。

Chu, Y., Xu, J., Zhou, X., Yang, Q., Zhang, S., Yan, Z., Zhou, C., 和 Zhou, J. 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。

Comanducci, L., Bestagini, P., 和 Tubaro, S. FakeMusic-Caps：一个用于检测和归因由文本生成音乐模型生成的合成音乐的数据集。*arXiv*预印本 *arXiv:2409.10684*，2024。

Ding, D., Ju, Z., Leng, Y., Liu, S., Liu,T., Shang, Z., Shen, K., Song, W., Tan, X., Tang, H., 等。Kimi-Audio 技术报告。*arXiv* 预印本 *arXiv:2504.18425*，2025。

Du, Z., Gao, C., Wang, Y., Yu, F., Zhao, T., Wang, H., Lv, X., Wang, H., Ni, C., Shi, X., 等。CosyVoice 3：通过扩展规模和后训练，实现真实场景中的语音生成。*arXiv*预印本 *arXiv:2505.17589*，2025。

Frank, J. 和 Sch¨onherr, L. Wavefake：一个用于促进音频深度伪造检测的数据集。载于神经信息处理系统进展，2021。Ge, W., Wang, X., Liu, X. 和 Yamagishi, J. 深度伪造语音检测的后训练。*arXiv*预印本 *arXiv:2506.21090*，2025。Gu, H., Yi, J., Wang, C., Tao, J., Lian, Z., He, J., Ren, Y., Chen, Y. 和 Wen, Z. ALLM4ADD：释放音频大语言模型在音频深度伪造检测方面的能力。载于*ACM*国际多媒体会议论文集，第11736–11745页，2025。

Hsu, W.-N., Bolte, B., Tsai, Y.-H. H., Lakhotia, K., Salakhutdinov, R., 和 Mohamed, A. HuBERT：通过掩码预测隐藏单元进行自监督语音表征学习。*IEEE/ACM* 音频、语音与语言处理汇刊，29:3451–3460，2021。

Huang, S.-F., Kuo, H.-C., Chen, Z., Yang, X., Yang, C.- H. H., Tsao, Y., Wang, Y.-C. F., Lee, H.-y., 和 F u, S.-W. 检测不可检测之物：评估当前欺骗检测方法对无缝语音编辑的有效性。载于*IEEE*口语技术研讨会论文集，第 652–659 页，2024。

Huang, W., Gu, Y., Wang, Z., Zhu, H. 和 Qian, Y. Speech- Fake：采用前沿生成方法的大规模多语言语音深度伪造数据集。在计算语言学协会年会论文集（第*1*卷：长篇论文）中，第 9985–9998 页，奥地利维也纳，2025年。

Ito, K. 和 Johnson, L. LJ语音数据集。2017年。

Jung, J.-w.、Heo, H.-S.、Tak, H.、Shim, H.-j.、Chung, J. S.、Lee, B.-J.、Yu, H.-J. 和 Evans, N. AASIST：利用集成频谱—时间图注意力网络进行音频语音欺骗防护。在*IEEE*声学、语音与信号处理国际会议论文集中，第 6367–6371 页，2022。

Jung, J.-w.、Wu, Y.、Wang, X.、Kim, J.-H.、Maiti, S.、Mat- sunaga, Y.、Shim, H.-j.、Tian, J.、Evans, N.、Chung, J. S. 等。Spoofceleb：真实环境中的语音深度伪造检测与说话人反欺骗验证。*IEEE* 开放获取信号处理期刊，2025。

Le, M.、Vyas, A.、Shi, B.、Karrer, B.、Sari, L.、Moritz, R.、Williamson, M.、Manohar, V.、Adi, Y.、Mahadeokar, J. 等。Voicebox：大规模文本引导的多语言通用语音生成。神经信息处理系统进展，36:14005–14034，2023年。

Lee, K.、Park, K. 和 Kim, D. Dailytalk：用于对话式文本转语音的口语对话数据集。载于*IEEE*声学、语音与信号处理国际会议论文集，第 1–5 页，2023年。

Liu, S.-Y.、Wang, C.-Y.、Yin, H.、Molchanov, P.、Wang, Y.-C. F.、Cheng, K.-T. 和 Chen, M.-H. Dora：权重分解低秩适配。载于国际机器学习会议论文集，2024。

Mesaros, A., Heittola, T. 和 Virtanen, T. 多声部声音事件检测的评估指标。应用科学，6（6）：162，2016。

M¨uller, N., Czempin, P., Diekmann, F., Froghyar, A. 和 B¨ottinger, K. 音频深度伪造检测能否泛化？载于国际语音通信协会会议论文集，第2783–2787页，2022。

Nautsch, A., Wang, X., Evans, N., Kinnunen, T. H., Vestman, V., Todisco, M., Delgado, H., Sahidullah, M., Yamagishi, J. 和 Lee, K. A. ASVspoof2019：用于检测合成语音、转换语音和重放语音的语音欺骗对策。*IEEE*生物识别、行为与身份科学汇刊，3（2）：252–265，2021。

Peng, P., Huang, P.-Y., Li, S.-W., Mohamed, A. 和 Har- wath, D. Voicecraft：零样本语音编辑及开放环境下的文本转语音。载于计算语言学协会年会论文集，第 12442– 12462 页，2024。

Reimao, R. 和 Tzerpos, V. For：一个用于合成语音检测的数据集。载于国际语音技术与人机对话会议，第 1–10 页，2019。

Tak, H., Jung, J.-w., Patino, J., Kamble, M., Todisco, M. 和 Evans, N. 用于说话人验证反欺骗和语音深度伪造检测的端到端时频图注意力网络。载于*ASVSpoof*挑战赛论文集，第 1–8 页，2021a。

Tak, H., Patino, J., Todisco, M., Nautsch, A., Evans, N. 和 Larcher, A. 基于 RawNet2 的端到端反欺骗。载于*IEEE*声学、语音与信号处理国际会议论文集，第 6369–6373 页，2021b。

Tan, X., Qin, T., Soong, F., and Liu, T.-Y. 神经语音合成综述。*arXiv* 预印本 *arXiv:2106.15561*, 2021.

Tang, C., Yu, W., Sun, G., Chen, X., Tan, T., Li, W., Lu, L., MA, Z., and Zhang, C. Salmonn：面向大语言模型的通用听觉能力。载于国际学习表征会议论文集, 2024.

Tian, J., Lee, S.-g., Kong, Z.,Ghosh, S., Goel,A., Yang, C.-H. H., Dai, W., Liu, Z., Ye, H., Watanabe, S., 等。UALM：用于理解、生成和推理的统一音频语言模型。*arXiv* 预印本 *arXiv:2510.12000*，2025。

Todisco, M., Delgado, H. 和 Evans, N. 常数 Q 倒谱系数：一种用于自动说话人验证的欺骗反制措施。《计算机语音与语言》，45：516–535，2017年。

Veaux, C., Yamagishi, J. 和 King, S. Voice Bank 语料库：大型区域口音语音数据库的设计、收集与数据分析。载于与亚洲口语研究与评估会议联合举办的东方 *COCOSDA* 国际会议（*O-COCOSDA/CASLRE*），第 1–4 页，2013年。

Wang, S., Yu, W., Chen, X., Tian, X., Zhang, J., Lu, L., Tsao, Y., Yamagishi, J., Wang, Y., and Zhang, C. Qual-iSpeech：一个包含自然语言推理和描述的语音质量评估数据集。*arXiv* 预印本 *arXiv:2503.20290*, 2025.

Wang, X. and Yamagishi, J. 研究用于语音欺骗反制措施的自监督前端。载于*Proc.Odyssey 2022*, 第 100–106 页, 2022.

Wu, B., Yan, C., Hu, C., Yi, C., Feng, C., Tian, F., Shen, F., Yu, G., Zhang, H., Li, J., 等。Step-Audio 2 技术报告。*arXiv* 预印本 *arXiv:2507.16632*，2025。

Wu, H., Tseng, Y., 和 Lee, H.-y. CodecFake：增强反欺骗模型，以抵御基于编解码器的语音合成系统生成的深度伪造音频。发表于国际语音通信协会会议论文集，第 1770–1774 页，2024。

Xie, Y., Guo, X., Zhou, J., Wang, T., Liu, J., Fu, R., Wang, X., Cheng, H., 和 Ye, L. 通过频时强化学习，利用音频大语言模型实现可解释的全类型音频深度伪造检测。*arXiv* 预印本 *arXiv:2601.02983*，2026年。

Xie, Z., Li, B., Xu, X., Liang, Z., Yu, K., and Wu, M. FakeSound：深度伪造通用音频检测。载于国际语音通信协会会议论文集，第 112–116 页，2024。

Xu, J., Guo, Z., He, J., Hu, H., He, T., Bai, S., Chen, K., Wang, J., Fan, Y., Dang, K., 等。Qwen2.5-Omni 技术报告。*arXiv* 预印本 *arXiv:2503.20215*，2025。

Yan, Z., Jiangyan, Y., Jianhua, T., Chenglong, W., and Yongfeng, D. Emofake：用于情感伪造音频检测的初始数据集。载于中国计算语言学全国会议论文集（第 *1* 卷：主会场），第 1286–1297 页，2024。

Yi, J.、Bai, Y.、Tao, J.、Ma, H.、Tian, Z.、Wang, C.、Wang, T. 与 Fu, R. Half-Truth：部分伪造音频检测数据集。载于国际语音通信协会会议论文集，第1654–1658页，2021。

Yi, J.、Wang, C.、Tao, J.、Zhang, X.、Zhang, C. Y. 与 Zhao, Y. 音频深度伪造检测：综述。*arXiv*预印本 *arXiv:2308.14970*，2023年。

Zhang, L., Wang, X., Cooper, E., Evans, N. 和 Yamagishi, J. PartialSpoof 数据库及用于检测话语中嵌入的短伪造语音片段的反制措施。*IEEE/ACM*音频、语音与语言处理汇刊，31:813–825，2022。

Zhang, Q., Wen, S., and Hu, T. 使用自监督 XLS-R 和 SLS 分类器检测音频深度伪造。载于 *ACM* 国际多媒体会议论文集，第 6765–6773 页，2024a。Zhang, Y., Zang, Y., Shi, J., Yamamoto, R., Toda, T., and Duan, Z. SVDD 2024：首届歌声深度伪造检测挑战赛。载于*IEEE*口语技术研讨会论文集，第 782–787 页，2024b。Zhang, Y., Tian, B., Zhang, L., and Duan, Z. PartialEdit：神经语音编辑时代的部分深度伪造识别。载于国际语音通信协会会议论文集，第 5353–5357 页，2025。Zhou, S., Zhou, Y., He, Y., Zhou, X., Wang, J., Deng, W., and Shu, J. IndexTTS2：情感表达力强、时长可控的自回归零样本文本到语音技术的突破。*arXiv* 预印本 *arXiv:2506.21619*，2025。

A. 欺骗方法说明

- **文本转语音合成（TTS）：** 完整语音话语由文本转语音模型根据文本输入生成。
- **语音转换（VC）：** 整段语音是对源说话人话语的修改版本，在保留原始语言内容的同时改变音色和韵律。
- **剪切粘贴（CaP）：** 将同一说话人的不同录音片段拼接起来，形成新的句子，通常用于在不使用生成模型的情况下改变语义内容。
- **语音编辑（SE）：** 生成模型（基于编解码器的模型（Peng 等人，2024）或扩散模型（Le 等人，2023年）被用于修改话语的特定区域，以改变关键词，同时保持语境连贯性。
- **声码器重合成（VR）：** 提取真实语音话语的声学特征（例如梅尔频谱图），并使用神经声码器将其重新合成为原始波形，从而引入特定的生成伪影。
- **编解码器重合成（CR）：** 真实语音话语经神经音频编解码器压缩和解压缩，其中可能会引入量化伪影。

B. 数据详情

B.1. 训练

表 *3.* HoliAntiSpoof 训练集详情。

数据集	欺骗方法	# 样本
ASVSpoof2019（ASV19。）（Nautsch 等人），2021）	文本转语音、语音转换	25,380
PartialSpoof（Zhang 等人，2022）	CaP	25,380
SINE(Huang 等人, 2024)	CaP, SE	104,942
WaveFake (Frank & Schönherr, 2021)	VR	71,883
CodecFake (Wu 等人,2024)	CR	17,267
PartialEdit（Zhang 等人，2025）	CaP、SE	63,204
VCTK（Veaux 等人，2013年）	/	7,730
LJSpeech（Ito 和 Johnson，2017年）	/	10,269
EmoFake（Yan 等人，2024）	VC	27,300
SpeechFake（Huang 等人，2025）-en	文本转语音、语音转换	458,245
DailyTalkEdit	CaP	3,600
FakeOrReal（Reimao 和 Tzerpos，2019）	未知	53,868
InTheWild（M¨uller 等人，2022）	未知	31,779

HoliAntiSpoof 使用多种英语反欺骗数据集的混合数据进行训练。各数据集的详细信息见表3。对于 SpeechFake，训练时使用双语和多语言子集中的英语样本。对于 VCTK 和 LJSpeech，我们在训练集和测试集中采用互不重叠的说话人集合；因此，混合训练集仅包含这两个语料库的部分数据。

在这些数据集中，只有 PartialEdit 和 DailyTalkEdit 提供了欺骗内容的语义影响信息。FakeOrReal 和 InTheWild 数据集未提供语音欺骗方法信息。为适应这些不同粒度的标注，我们使用任务特定提示。这一策略可避免模型在训练过程中遇到不一致的目标格式。

B.2. 评估

评估所用数据的详情列于表 4。我们采用真假标签分层策略，从各个数据集中抽样，构建混合域内测试集。对于大多数数据集，抽样数量为 2,000。由于 SpeechFake-en 包含大量最新的文本转语音、语音转换和声码器模型，我们将其样本量增加到 10,000，以确保充分涵盖每个底层生成模型。由于 *LJSpeech* 和 *DailyTalkEdit* 测试集的可用样本少于 2,000，因此无需抽样，所有可用数据均用于评估。

表 4. HoliAntiSpoof 评估数据集的详细信息。

	数据集	语言	欺骗方法	# 样本数
混合域内	ASVSpoof2019 (ASV19.) (Nautsch 等人，2021)	en	文本转语音、语音转换	2,000
	PartialSpoof(Zhang 等人，2022)		CaP	2,000
	SINE（Huang 等人，2024）		CaP、SE	2,000
	WaveFake (Frank & Schönherr, 2021)		VR	2,000
	CodecFake（Wu 等人，2024）		CR	2,000
	PartialEdit（Zhang 等人，2025）		CaP，SE	2,000
	VCTK（Veaux 等人，2013年）		/	2,000
	LJSpeech（Ito 和 Johnson，2017年）		/	1,365
	EmoFake（Yan 等人，2024）		VC	2,000
	SpeechFake（Huang 等人，2025）-en		文本转语音、语音转换、语音修复	10,000
	DailyTalkEdit		CaP	950
域外	SpeechFake MultiLingual (SF-MD.) (Huang 等人，2025)	11种语言	文本转语音、语音转换	11,000
	SpoofCeleb(Jung等人, 2025)	多语言	TTS	10,000
	HAD(Yi等人, 2021)	zh	文本转语音、CaP	9,072

C. 额外等错误率结果

表5. 使用等错误率（EER）的额外反欺骗指标。

模型	域内		域外	
	ASV19.	混合	SF-MD.	SpoofCeleb
传统模型				
RawNet2	6.58	8.67	6.36	31.21
RawGAT-ST	2.86	6.65	6.08	21.20
AASIST	2.65	4.60	7.41	16.80
Wav2Vec2 (F) - 实时	13.51	24.89	19.57	44.19
Wav2Vec2 (T) - 实时	13.96	8.04	9.79	34.00
HuBERT (F) - 实时	5.56	10.61	7.71	28.00
HuBERT（T）- 实时	12.48	8.70	12.43	27.00
WavLM（F）- 实时	5.97	12.70	6.95	32.59
WavLM（T）- 实时	14.47	9.62	5.77	25.40
ALLM 方法				
HoliAntiSpoof-仅认证	**1.10**	**3.06**	0.68	10.03
HoliAntiSpoof	1.11	3.48	**0.39**	**8.52**

表 5 展示了 HoliAntiSpoof 和基线系统的二元真实性分类性能，以等错误率（EER）衡量。按照 Gu 等人 (2025)，我们根据预测为 真实的归一化概率计算 EER:

$$p(\text{real}) = \frac{\exp(s_{\text{real}})}{\exp(s_{\text{real}}) + \exp(s_{\text{fake}})}, \tag{3}$$

其中 s_{real} 与 s_{fake} 表示词元 真实 与 伪造 在紧接前缀“{"真实或伪造": ”之后位置处的 logits。由于无法从 Gemini-3-Flash 获取各词元的 logits，因此未将其纳入比较。等错误率结果与表 1 中的准确率趋势大体一致，进一步验证了 HoliAntiSpoof 的优越性。在域内评估中，HoliAntiSpoof-仅认证与 HoliAntiSpoof 表现相当，二者均显著优于传统模型。在域外 SF-MD. 与 SpoofCeleb 上，HoliAntiSpoof-仅认证取得了更高的分类准确率，但在等错误率方面落后于 HoliAntiSpoof。这表明，整体欺骗分析训练增强了 HoliAntiSpoof 的判别能力，从而在最优阈值下进行评估时，于未见域中展现出更优的鲁棒性。因此，可通过采用来自目标域的校准集来细化决策阈值，进而优化 HoliAntiSpoof 的分类准确率。

D. 提示词

D.1. 语义影响评估

我们在下文给出用于评估针对欺骗内容的语义影响分析质量的提示。该质量从三个角度加以评判。

用于语义影响分析的大语言模型评审提示词

```
## 角色
你是一位音频欺骗检测和内容安全评估领域的资深专家。你的工作是对“音频欺骗检测模型”的输出质量进行评分。

## 任务
比较 [原始内容]和 [欺骗内容]，检查 [模型分析结果] 是否准确识别了篡改行为及其影响。

## 原始内容
[完整的语音转写文本]
## 欺骗内容 [原始词语/句子] -> [新词语/句子]

## 模型分析结果[模型输出]

## 评分标准（1-5 分）
5 （完美）：
精准定位：准确找出被篡改位置。深入分析：逻辑严谨；准确指出篡改对语义完整性、角色人设或对话语境造成的深远影响。表达流畅：措辞流畅，语法无误。

4（优秀）：
精准定位：准确找出被篡改位置。一般分析：指出篡改事实，但对语境影响的分析不够深入，或存在轻微逻辑错误。表达流畅：措辞流畅。

3（及格）：
精准定位：准确识别被篡改位置。错误分析：定位正确，但对篡改意图或语义变化的分析完全错误或与事实相矛盾。表达流畅：措辞流畅。

2（较差）：
定位错误：未能识别正确的被篡改位置（偏差显著），或识别了错误片段。表达流畅：尽管内容不正确，措辞依然流畅。

1（无效）：
完全失败：未能识别篡改，或输出与任务无关（例如胡言乱语、完全跑题）。

## Output规则
严格按照评分标准进行评判。直接输出分数；不要包含任何 Markdown 标签或其他说明性文字。
```

D.2. DailyTalkEdit 数据整理

如图 3所示，DailyTalk 数据集中的对话经过双智能体工作流修改。写作者选择一个目标句子进行修改，而检查器则验证修改后的句子在上下文中是否连贯，且不存在明显的矛盾。下文展示了相应的提示词。该工作流旨在尽可能多地改变对话中的关键信息，同时确保修改后的对话在逻辑上保持一致且合理。

写作者提示词

你正在模拟针对一段对话的恶意欺骗攻击。你的任务是：1. 仔细阅读整段对话，特别留意每句话语前后的上下文。2. 找出一句话语：如果对其进行恶意修改，可能造成最大的伤害、困惑或负面后果。3. 以一种看似合理却具有危险性的方式修改这句话语——也就是欺骗攻击会采用的那种修改方式。注意：1. 修改后的句子必须自然融入对话其余部分，不能因长度异常而显得突兀。这对于实施逼真的欺骗攻击至关重要。2. 修改后的话语必须与周围的对话保持逻辑连贯。即使你改变了关键信息，对话中后续的话语在上下文中仍应说得通。避免造成明显的矛盾，以免操纵行为立即暴露。对话如下。每一行前面都有一个方括号括起的索引，表示该话语在对话中的索引：[对话]现在，请严格遵守以下规则：- 你必须恰好选择一句最有可能造成最大伤害或负面影响的话语进行修改。- 你必须输出一个 JSON 对象，格式如下，且不得包含其他内容：{“目标话语索引”: <所选话语的索引整数>, “修改后的文本”: <经过恶意修改的话语文本>} 不要在此 JSON 之外添加任何解释或额外文本。

- -
-

检查器提示词

你是一个对话连贯性检查器。你的任务是评估对话中经过修改的话语是否与周围上下文保持逻辑连贯。如果满足以下条件，则修改通过：

1. 修改后的话语自然地融入对话流程 2. 修改后，对话中的后续话语仍然合乎逻辑 3. 不存在会立即暴露该修改可疑之处的明显矛盾

如果满足以下条件，则修改失败：
1. 修改后，后续话语出现矛盾或变得毫无意义；2. 修改后的话语与上下文之间存在明显的逻辑矛盾

注意：此次修改意在恶意篡改（更改关键信息）。你的任务是检查修改是否在保持表面连贯性的同时完成——成功的深度伪造攻击应足够隐蔽，使对话仍然流畅自然。

原始对话：
[原始对话]
修改后的对话（话语 [修改后的话语索引] 已被修改）：
[修改后的对话]

评估此次修改是否保持了逻辑连贯性。 请特别注意：
- [修改后的话语索引] 之后的话语在语境中是否仍然合理
- 是否存在会立即暴露此次修改的明显矛盾
- 对话流程是否依然自然且连贯

你必须按照以下格式输出一个 JSON 对象，除此之外不得输出任何内容：
{
“通过”: <是或否>,
“原因”: <简要说明通过或失败的原因>
}
请勿在此 JSON 之外添加任何解释或额外文本。