探索在音频深度伪造检测中使用多模态大语言模型的可行性

Akanksha Chuchra[1], Shukesh Reddy[2],Sudeepta Mishra[1],Abhijit Das[2],Abhinav Dhall[31]
印度理工学院罗帕尔分校，印度[2] 比尔拉理工学院机器智能组，皮拉尼海得拉巴校区，印度[3] 莫纳什大学，墨尔本，澳大利亚

abhijit.das@hyderabad.bits-pilani.ac.in，abhinav.dhall@monash.edu

arXiv:2601.00777v1 [cs.SD] 2 Jan 2026

摘要

尽管视觉语言模型（*VLMs*）和多模态大语言模型（*MLLMs*）在检测图像和视频深度伪造方面展现出强大的泛化能力，但它们在音频深度伪造检测中的应用仍鲜有探索。本研究旨在探索多模态大语言模型用于音频深度伪造检测的潜力。我们将音频输入与一系列文本提示结合起来作为查询，以探究多模态大语言模型能否学习适用于音频深度伪造检测、跨模态的稳健表征。因此，我们尝试探索基于文本、内容丰富且采用问答形式并给出二元判定的提示。我们假设，这种特征引导推理有助于促进更深入的多模态理解，并实现面向音频深度伪造检测的稳健特征学习。我们在两种评估模式下评估了两种多模态大语言模型 *Qwen2-Audio-7B-Instruct* 和 *SALMONN*：（*a*）零样本和（*b*）微调。实验表明，将音频与多提示方法相结合，可能是推进音频深度伪造检测的一条可行路径。实验还显示，若不进行特定任务训练，模型表现不佳，且难以泛化至域外数据。然而，在仅有少量监督的情况下，它们在域内数据上表现良好，展现出用于音频深度伪造检测的潜力。

1. 引言

近年来，音频深度伪造的兴起已成为一个重大问题 [1, 2]。音频深度伪造是人工生成的语音，能够高度模仿人类的声音。语音生成技术的发展使得合成语音几乎无法与真实的人类语音区分 [1, 3]。这些伪造的音频片段可被用于传播虚假信息、冒充个人身份以及绕过基于语音的安全系统。因此，音频深度伪造检测已成为一个活跃的研究领域 [3, 4, 5]。

近期，人们提出了多种音频深度伪造检测方法，例如 AASIST [6] 和 RawNet2 [7]。这些方法依托端到端架构，并直接针对分类任务进行训练。另一类研究则采用两阶段策略：使用 Whisper [8], 、WavLM [9], 和 wav2vec 2.0 XLS-R [10]等预训练模型（PTM）作为特征提取器，再搭配轻量级的特定任务分类头。有趣的是，基于 PTM 的方法往往优于端到端模型，在检测深度伪造[11, 12, 13, 14]方面也展现出出色的性能。这主要归因于这些预训练模型以自监督方式在大规模数据集上进行训练，因此能够学习到丰富且稳健的语音表征，并在不同领域中实现良好的泛化。

大语言模型（LLM）已展现出强大的泛化能力和推理能力，能够应对广泛的任务 [15]。它们在多样化的文本语料库上训练，能够学习丰富的语义表征，并遵循指令执行文本生成、问答等任务，甚至无需或仅需少量特定任务训练 [16]，即可支持多模态应用。在大语言模型的基础上，多模态大语言模型（MLLM）进一步拓展了这些能力 [17]。它们使模型能够处理文本、音频和图像等多种模态并进行推理，为需要跨模态理解的任务开辟了新的可能性。此外，近年来，视觉语言模型（VLM）因具备理解和生成视觉与文本模态内容的能力而受到广泛关注 [18, 19, 20]。这些模型在多种视觉语言任务中展现出强劲表现，包括图像描述、视觉问答和多模态推理 [21, 22]。

与此同时，人们对利用多模态大语言模型进行媒体操

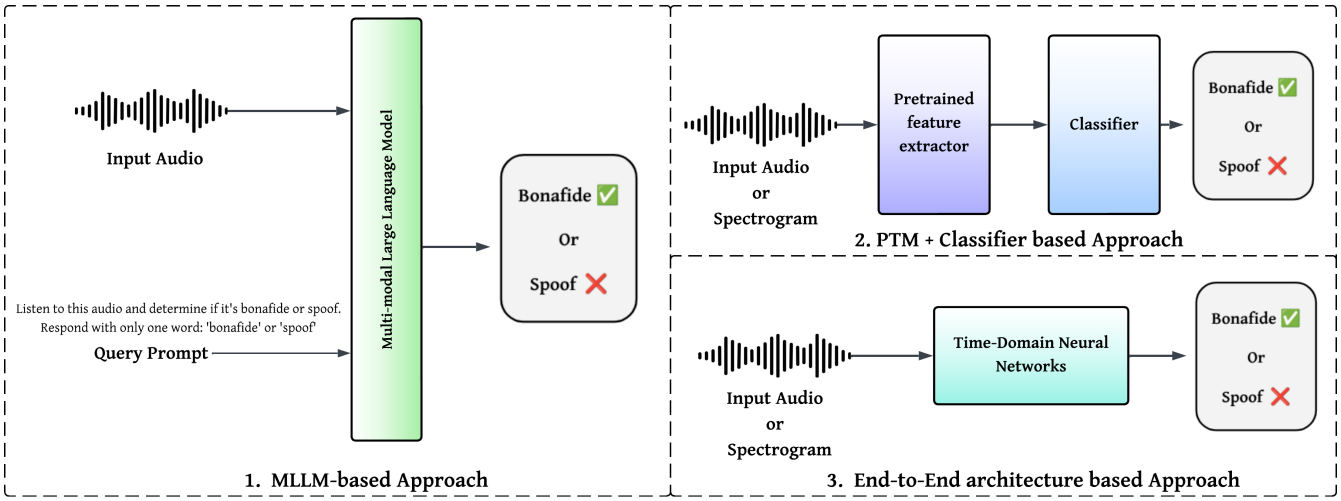

图 1. **左：**基于音频多模态大语言模型的音频深度伪造检测方法，将其表述为音频问答任务。**右：**传统方法依赖端到端架构，或使用预训练模型的特征提取器，再接分类器进行离散标签预测。

纵检测（如深度伪造检测）的兴趣日益浓厚，涉及文本、图像和音频等不同模态 [23]。尽管当前大多数研究聚焦于视觉内容，但人们正越来越多地研究视觉语言模型识别图像篡改和检测合成伪迹的潜力 [24, 25]。值得注意的是，一些视觉语言模型还展现出在零样本评估中检测视觉篡改和合成内容的能力，也就是说，无需针对特定任务进行额外训练 [26]。视觉语言模型在检测篡改方面取得的成功促使我们研究和探讨在音频深度伪造检测这一领域中使用基于音频的多模态大语言模型的可行性，并全面梳理文献中对此问题的处理方式（见图 1）。

音频多模态大语言模型在大量多样化的音频-文本语料库上进行训练，因此能够学习复杂的语音模式，并响应指令式提示。近期，音频多模态大语言模型在各种语音相关任务中展现出令人瞩目的泛化能力 [27, 28]。本研究是较早探索利用多模态大语言模型进行音频深度伪造检测的工作之一。我们的研究不仅旨在评估现有音频多模态大语言模型能否胜任这一复杂的音频深度伪造检测任务，还旨在了解它们在不同推理条件下的表现及其局限性。因此，本文旨在探讨以下问题：

- 当前的多模态大语言模型能否有效用于音频深度伪造检测任务？
- 如何高效利用多模态大语言模型，从特征理解和决策准确率方面提升音频深度伪造检测？
- 基于多模态大语言模型的方法在多大程度上能够提升音频深度伪造检测在不同数据集和攻击类型之间的泛化能力？

2. 相关工作

MLLMs 的快速发展 [9, 20, 29, 30], 深度生成模型 [31], 和扩散模型 [32] 显著改变了合成音频创作领域。生成式音频建模的进步大幅降低了创建逼真合成语音的障碍，也因此给音频通信中的真实性、安全性和信任带来了重要问题。目前的研究通常遵循传统流水线模型，该模型将前端特征提取器 [9, 33, 8, 34, 35] 与后端分类器 [36, 37] 相结合；或者采用端到端模型，直接分析原始音频波形 [38, 39]。特征提取通过检测语音信号中的音频伪迹来识别区分性特征，而端到端模型则以原始形式处理音频数据，以捕捉细粒度细节，这些细节会直接影响音频深度伪造检测的性能。RawNet 2 [7]采用 Sinc层直接从波形中提取特征，而 RawGAT-ST [40] 则利用频谱和时间子图。Rawformer [41]将卷积层与 Transformer架构相结合，用于表示局部和全局伪迹。LFCC [42] 是广泛使用的手工设计特征，采用线性滤波器组，能够有效捕捉高频域中更多的频谱信息。然而，手工设计特征会受到人工表示约束所导致的偏差影响。为缓解这些约束，研究者提出了从深度神经网络中提取的深层特征。

其中，最具代表性的是预训练自监督语音模型，包括 Wav2vec2 [43], HuBERT [33], Whisper [8], BEATs [34],WavLM [9], 和 Data2Vec [35], 。LCNN [44]是一种常用的分类器，被认为是在包括 ASVspoof [45] 和 ADD 2022 [46]在内的各种竞赛中表现有效的基线模型。

文献 [47]提出了 Llama-AVSR，这是一种多模态大语言模型，结合预训练的音频和视频编码器，以及通过 LoRA 和轻量级投影器增强的静态大语言模型，可执行自动语音识别、视觉语音识别和视听语音识别。文献 [24]则利用 GPT-4V 进行媒体取证，重点分析视频内容并检测文本图像错位。他们提出直接处理视频，而非采用逐帧方法。尽管该方法效果良好，但受限于 GPT-4V 容易产生幻觉的倾向，因此需要人工监督。

多种音频语言模型，包括 AudioGPT[48], SpeechGPT [49], LTU [50], Qwen2-Audio [51], DESTA [52], 和 SALMONN [53], , 在语音识别 [54, 55], 、音频描述 [39, 56, 57], 和音频问答等任务中展现出出色的性能。然而，它们在检测伪造或篡改音频方面的应用仍鲜有探索。为弥补这一不足，我们将音频深度伪造检测建模为音频问答问题，以此探究多模态大语言模型在该领域的能力，并利用这些模型的推理和感知能力。

3. 所提方法

如前所述，传统的音频深度伪造检测采用针对离散标签预测进行优化的二元分类器。相比之下，多模态大语言模型基于文本和音频生成预测，因为它们经过训练，能够预测下一个词元。为此，针对音频深度伪造检测任务，我们将其重新定义为音频问答任务：模型根据输入音频和查询提示，输出真实语音或伪造语音（见图2）。我们研究了模型在两种评估模式下的性能：零样本模式，以及通过微调多模态大语言模型以提升任务特定性能的模式。为确保输出集中且一致，我们精心设计提示词，要求模型仅返回标签本身，避免添加任何解释或评论。

3.1. 问题定义

令 $\mathcal{M} = (\mathbf{x}^{(i)}_{\text{音频}}, \mathbf{x}^{(i)}_{\text{提示词}})_{i=1}^{N}$ 表示一个包含 N 对输入音频波形 $\mathbf{x}^{(i)}_{\text{音频}}$ 及其对应文本提示词 $\mathbf{x}^{(i)}_{\text{提示词}}$ 的数据集，这些数据经过处理

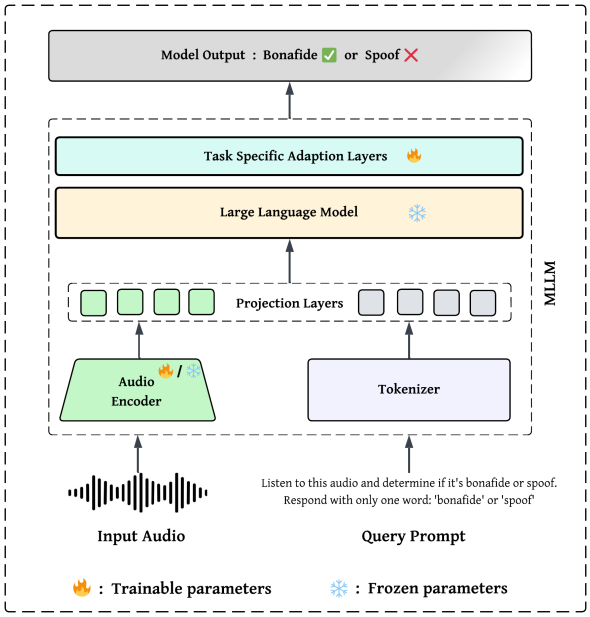

图2. 使用微调多模态大语言模型进行音频深度伪造检测的总体概览。微调采用低秩适配（LoRA），以最少的可训练参数实现高效适配。

并通过多模态流水线处理。对于每个实例，音频信号会先经过预训练音频编码器，再经过模态适配器，生成一系列音频标记 $\mathbf{A}_{\text{tokens}}$，如公式（1）所示：

$$\mathbf{A}_{\text{tokens}} = \text{Adapter}(\text{AudioEncoder}(\mathbf{x}_{\text{audio}})) \quad (1)$$

同时，输入提示会经过分词器处理，以获得一系列文本词元 $\mathbf{T}_{\text{tokens}}$，如公式（2）所定义：

$$\mathbf{T}_{\text{tokens}} = \text{Tokenizer}(\mathbf{x}_{\text{prompt}}) \quad (2)$$

随后，这两种模态——音频词元和文本词元——会一同输入语言模型（见图2），模型根据这两种信息来源生成文本输出 $\mathbf{y}_{\text{out}}$，如公式（3）所述：

$$\mathbf{y}_{\text{out}} = \text{LLM}(\mathbf{A}_{\text{tokens}}, \mathbf{T}_{\text{tokens}}) \quad (3)$$

3.2. 文本提示

我们设计并测试了各种文本提示，以引导多模态大语言模型完成音频深度伪造检测任务。文本提示充当模型的输入指令，塑造模型对随附音频的解读方式以及生成回复的方式。受近期视觉语言深度伪造取证研究的启发 [24, 58], , 我们的提示词

在问题表述方式、指令明确程度和上下文丰富度方面各不相同。

提示 #1（二元判断——直接式）

请听这段音频，并判断它是真实语音还是伪造语音。只需回复一个词：“真实语音”或“伪造语音”。

提示 #2（是/否）

这段音频是伪造语音吗？只回答“是”或“否”，无需解释。

提示 #3（上下文丰富的描述式）

真实的人类语音通常具有自然的韵律、情感一致性，且没有机器人般的质感或故障伪影。相比之下，虚假或欺骗语音可能会出现单调音调、节奏不自然、机器人般的质感，或咔嗒声、失真等音频故障。现在，请听给定的音频，并根据这些特征判断它是真实语音还是伪造语音。只需回复一个词：“伪造语音”或“真实语音”。

提示 #多重（双分类——直接式和富含上下文的描述式）

二元提示：聆听这段音频，并判断它是真实语音还是伪造语音。请只用一个词回答：“真实语音”或“伪造语音”。

富含上下文的描述性提示：真实的人类语音通常具有自然的韵律、情感一致性，且没有机器人般或故障伪影般的特征。相比之下，伪造语音可能会表现出单调的音调、不自然的语速、机器人般的质感，或咔嗒声、失真等音频故障。现在，请聆听给定的音频，并根据这些特征判断它是真实语音还是伪造语音。请只用一个词回答：“伪造语音”或“真实语音”。

从最简的二分类提示开始，（提示词 #1）直接要求模型在“真实语音”和“伪造语音”之间做出选择。提示词 #2 以是非问答的形式提出问题，采用略有不同的语言结构。提示词 #3 则加入丰富的上下文描述，通过提及单调音调、机器人般的质感或不自然的语速等典型听觉模式和伪影来引导模型，而这些特征通常与合成语音相关。提示词 #Multi 同时结合直接二分类提示和上下文丰富的描述性提示，使模型能够在不同引导程度下进行推理。这一递进过程使我们能够评估不同程度的具体性和上下文如何影响模型的表现及回答一致性。较简单的提示旨在反映普通用户提出直接问题的方式，并期待简短、直接的回答。相比之下，描述性提示旨在通过加入音调、节奏或故障伪影等具体语音特征，帮助模型更好地推理，从而判断音频内容是真实语音还是伪造语音。

3.3. 零样本评估

在此模式下，不经过任何特定任务训练或微调，直接对模型执行音频深度伪造检测任务的评估。目标是评估多模态大语言模型识别伪造语音的内在泛化能力，这仅依赖于其预训练多模态知识和指令遵循能力。此项评估有助于确定，在精心设计的提示词引导下，多模态大语言模型能否准确区分真实语音与合成语音。

3.4. 微调多模态大语言模型

我们分析中的第二种评估模式，是使用针对音频深度伪造检测任务定制的带标签数据集对多模态大语言模型进行微调。在这种监督设置下，模型使用真实语音和欺骗语音的样本进行训练，从而能够调整其内部表示，并在零样本评估中取得更好的表现。

微调多模态大语言模型等大规模模型可能需要大量计算资源和内存，带来重大挑战，尤其是在资源受限的环境中。为缓解这一问题，我们采用低秩适配（LoRA） [59], ，这是一种参数高效微调技术。LoRA 冻结原始权重，并在选定层中引入小型可训练低秩矩阵（通常位于注意力模块和前馈模块中），从而避免更新整个模型。这种设计在保持微调效果的同时，大幅减少了可训练参数的数量。形式上，LoRA 不更新完整的权重矩阵 $W \in \mathbb{R}^{d\times k}$，而是使用矩阵 $A \in \mathbb{R}^{d\times r}$ 和 $B \in \mathbb{R}^{r\times k}$引入低秩更新，其中 $r \ll \min(d, k)$。微调期间修改后的权重表示为：

$$W' = W + \Delta W = W + AB \tag{4}$$

在此， W 保持冻结，仅优化 A和 B 。完成微调过程后，我们采用与零样本评估相同的评估设置，对适配后的模型进行评估，以考察其性能。

4. 实验与结果

本节介绍所提出方法的实验细节。第 4.1 节说明用于评估我们模型的数据集。第 4.2 节介绍实现细节以及实验中使用的多模态大语言模型。第 4.3 节展示并分析我们用于检测音频深度伪造的方法的结果。

4.1. 数据集

我们的实验基于两个标准数据集：ASVspoof 2019 逻辑访问数据集和野外（ITW）数据集。**ASVspoof 19 LA：**ASVspoof 19 LA 数据集 [45]广泛用于音频深度伪造检测研究。然而，该数据集存在严重的类别不平衡（7355 个真实语音样本与 63882 个伪造语音样本），在训练集、开发集和评估集中的真实语音与伪造语音比例约为 1:9。为解决这一问题，我们为每个数据集划分构建了类别平衡的子集，分别记为 S_{train}、S_{dev} 和 S_{eval}。具体而言，我们纳入所有可用的真实样本，并随机抽取数量相同的伪造样本。此外，我们确保抽取的欺骗音频均匀涵盖全部 19 种攻击类型，以提高多样性，避免偏向任何特定攻击类型。我们使用 ASV19 作为 ASVspoof 2019 LA 完整数据集的简称。**ITW：**野外（ITW）数据集 [60]包含 31,779 个样本，其中有 19,963 个真实语音样本和 11,816 个伪造语音样本。由于该数据集的类别不平衡程度较轻，我们在评估时使用了整个数据集。ITW 具有多样的录音条件和伪造方法，因此提供了更贴近现实且更具挑战性的基准，非常适合用于评估我们模型的泛化能力。表 1 总结了我们实验中使用的 ASV19 子集划分和 ITW 数据集中的真实语音与伪造语音样本数量。

表 1. 用于训练、验证和评估的数据集统计信息

数据集	#真实语音	#伪造语音	总计
S_{train}	2580	2580	5160
S_{dev}	2548	2548	5096
S_{eval}	7355	7355	14710
ITW	19963	11816	31779

4.2. 实验设置

我们选用两种近期最先进的多模态大语言模型进行评估：Qwen2-Audio [51]和 SALMONN [53]。Qwen2-Audio 由两个主要组件构成：音频编码器和大语言模型。音频编码器基于 Whisper-large-v3 模型初始化，而语言组件则基于 Qwen-7B。具体而言，在实验中我们使用 Qwen2-Audio-7B-Instruct 模型[1]，，该模型是聊天模型，属于 Qwen2-Audio模型系列。为便于简写，我们仅将该模型称为 Qwen2-Audio。SALMONN [53] 是一种多模态模型，将 Vi-cuna 大语言模型与两个音频编码器连接起来：用于语音的 Whisper 和用于通用音频的 BEATs。这些编码器处理输入音频，并将其输出传递给 Q-Former，由后者整合特征并将其转换为大语言模型能够理解的格式。具体而言，在实验中我们使用 SALMONN-13B [2] 变体。SALMONN 在各种语音和音频任务上表现良好，例如自动语音识别、翻译和情感识别。我们选择这些模型进行分析，是因为它们在各种任务和已建立的基准测试（如 Dynamic-SUPERB [27] 和 AIR-Bench-Chat [28]）上展现出强劲的性能。

4.3. 实现细节与评估指标

对于 LoRA微调，我们将 LoRA 秩设置为 8，缩放因子（alpha）设置为 32。应用 0.1 的丢弃率，并将 LoRA 集成到模型的查询投影层和值投影层中。我们使用监督微调对模型进行 10 个周期的微调，学习率为 1e-4。所有音频样本均重采样至 16 kHz。除重采样外，不进行其他额外的预处理。原始音频及其对应的提示直接输入模型。所有实验均在 NVIDIA-A100 GPU 上进行。在评估过程中，我们将模型生成的文本预测与真实标签直接进行比较。对于不符合预期格式的预测，例如超出范围的响应，将其视为未知值，并从指标计算中排除。在评估中，我们报告准确率和宏平均F1分数。准确率衡量两个类别中预测正确的总体比例。另一方面，宏平均F1分数分别计算每个类别的 F1 分数，然后取平均值，使真实语音类别和伪造语音类别获得相同的权重。

$$\text{mF1} = \frac{1}{C}\sum_{i=1}^{C}\text{F1}_i = \frac{1}{C}\sum_{i=1}^{C}\frac{2\cdot\text{Prec}_i\cdot\text{Rec}_i}{\text{Prec}_i+\text{Rec}_i} \tag{5}$$

其中，C 为类别数，F1_i、Prec_i和Rec_i 分别表示类别 i的F1分数、精确率和召回率。

4.4. 结果

本节展示了所评估模型在各种实验设置下的性能。我们分析了它们在不同提示词和数据集上的零样本与微调模式表现。表 2 总结了多模态大语言模型在 ASVspoof 19 评估子集上的零样本及微调设置下的性能。模型$_P$ 表示

[1]https://huggingface.co/Qwen/Qwen2-Audio-7B-Instruct

[2]https://huggingface.co/tsinghua-ee/SALMONN

使用特定提示词*p*进行微调的模型。我们考虑了三类提示词：直接式、描述式和多重。这里，*Dir* 指提示词#1，*Desc*指提示词#3，而*Multi*表示同时使用直接式和描述性提示进行微调。为分析模型性能对提示词变化的敏感度，我们在推理过程中使用多个提示词评估每个模型。表3展示了 ITW 数据集的结果，以检验跨数据集泛化能力。

提示	零样本评估	S_{eval}	
		ACC	mF1
提示词#1	Qwen2-Audio	0.34	0.28
	SALMONN	0.46	0.46
提示词#2	Qwen2-Audio	0.36	0.22
	SALMONN	0.50	0.33
提示#3	Qwen2-Audio	0.43	0.38
	SALMONN	0.45	0.32
微调模型			
提示词#1	Qwen2-Audio 方向	0.96	0.96
	Qwen2-Audio 描述	0.90	0.48
	Qwen2-Audio 多重	0.59	0.49
	SALMONN 方向	0.96	0.96
	SALMONN 描述	0.96	0.96
	SALMONN 多重	0.97	0.97
提示词#3	Qwen2-Audio 目录	0.95	0.95
	Qwen2-Audio 描述	0.91	0.90
	Qwen2-Audio 多重	0.58	0.47
	SALMONN 直接式	0.96	0.96
	SALMONN 描述式	0.97	0.97
	SALMONN 多重	**0.98**	**0.98**

表 2. 使用不同提示词微调模型后，在 S_{eval} 数据集上评估的性能比较。直接式、描述式和多重 分别表示使用直接提示、描述性提示以及两种提示相结合进行微调。结果分别报告了推理时使用直接提示和描述性提示的情况。模型名称的下标表示微调时使用的提示。

不同提示词下的性能：在不同提示词中，直接提示（即提示 #1）的平均准确率约为 50%，而描述性提示（提示 #3）在所有数据集和模型上取平均后，准确率约为 44.5%。这种较低的性能可能归因于描述性提示具有更长的上下文长度或更高的词元复杂度，而当前模型可能难以有效处理。总体而言，我们未观察到提示词之间存在一致的规律，这表明模型对提示措辞高度敏感。

微调与零样本性能。在零-

提示	零样本评估	ITW	
		ACC	mF1
提示词#1	Qwen2-Audio	0.66	0.54
	SALMONN	0.54	0.53
提示词#2	Qwen2-Audio	0.52	0.52
	SALMONN	0.52	0.51
提示词#3	Qwen2-Audio	0.51	0.44
	SALMONN	0.39	0.31
微调模型			
提示词#1	Qwen2-Audio 直接	0.37	0.27
	Qwen2-Audio 描述	0.36	0.26
	Qwen2-Audio 多重	0.59	0.49
	SALMONN 直接	0.58	0.57
	SALMONN 描述	0.57	0.56
	SALMONN 多重	0.63	0.59
提示词#3	Qwen2-Audio 方向	0.37	0.27
	Qwen2-Audio 描述	0.38	0.27
	Qwen2-Audio 多重	0.59	0.49
	SALMONN 方向	0.56	0.54
	SALMONN 描述式	0.59	0.58
	SALMONN 多重	**0.66**	**0.62**

表3. 在野外（ITW）数据集上对预训练和微调的 Qwen2-Audio 与 SALMONN 模型进行跨领域评估，以考察其跨领域泛化能力。

少样本评估中，模型的表现差强人意，准确率仍接近随机猜测水平。我们发现，即使使用规模极小且均衡的带标签数据集进行微调，两个模型的性能也都显著提升，尤其是在 S_{eval} 数据集上。这凸显了两个模型的适应能力，并表明即使监督有限，也能显著提升其检测能力。微调模型取得的最佳分数以**粗体**标示，而零样本的最佳分数则在表 2、3 中加下划线。

按模型比较： 在这两种模型中，SALMONN 在大多数评估模式下始终优于 Qwen2-Audio。在零样本场景中，尤其是在 S_{eval} 上，SALMONN 表现更出色。此外，经过微调后，SALMONN 在默认和描述式这两种评估提示词下仍持续优于 Qwen2-Audio，展现出更强的泛化能力和适应性。

最先进水平比较表4展示了在 ASV19 数据集上与经典深度伪造语音检测方法的比较结果，同时列出了文献 [61] 报告的结果。比较的方法包括基于手工特征的方法，例如短时傅里叶变换（STFT）、常数Q变换（CQT）、线性滤波器（LF），以及 Rawformer、RawNet2、RawPC 和 RawGAT-

传统模型	ASV-19	
	ACC	mF1
卷积神经网络（短时傅里叶变换和低频）	0.88	0.90
循环神经网络（短时傅里叶变换和低频）	0.92	0.91
卷积循环神经网络（短时傅里叶变换和低频）	0.88	0.90
Swin Transformer（短时傅里叶变换和低频）	0.84	0.87
ConvNeXt-Tiny（短时傅里叶变换与低频）	0.88	0.90
SinC-CNN（原始音频）	0.84	0.87
Whisper+多层感知机（原始音频）	0.85	0.88
Speechbrain+多层感知机（原始音频）	0.77	0.81
Seamless+多层感知机（原始音频）	0.86	0.88
Pyannote+多层感知机（原始音频）	0.64	0.71
Whisper、ConvNeXt-Tiny（原始音频、短时傅里叶变换和低频）	0.86	0.88
Whisper、卷积神经网络（原始音频、短时傅里叶变换和低频）	0.87	0.89
Rawnet2	0.93	0.92
RawGAT-ST	0.97	0.93
Rawformer	0.98	0.99
所提出方法的最佳结果 **SALMONN 多重**	**0.98**	**0.98**

表4. 所提出的方法与经典方法在 ASV19数据集上的比较 [61]

传统模型	ITW	
	ACC	mF1
LCNN	0.65	0.63
LCNN-LSTM	0.66	0.62
Mesonet	0.53	0.53
ResNet18	0.49	0.46
CRNNSpoof	0.41	0.39
RawNet2	0.33	0.33
RawPC	0.45	0.43
RawGAT-ST	0.37	0.38
表现最佳的 SALMONN $_{\text{Multi}}$	**0.66**	**0.62**

表5. 不同传统模型与所提出模型在 ITW 数据集上的比较 [60]。

ST。表5展示了与 [60] 中传统模型在 ITW 数据集上的性能比较。通过比较可以发现，所提出的微调版本取得了与这些模型相当或更优的性能

文献中已有的经典方法。我们模型取得的最佳分数以**粗体**显示，而最佳 SOTA 结果则以下划线标出。

5. 讨论与挑战

尽管视觉语言模型在深度伪造取证领域取得了显著进展，但基于多模态大语言模型的音频深度伪造检测仍相对有限。与视觉语言模型相比，音频多模态大语言模型在语音相关任务上的表现相当，因此值得将其用于音频深度伪造检测。然而，音频模态本身的固有特性可能带来诸多挑战。音频数据本质上具有高维特征，包含密集的时间和频率信息，因此建模和有效解读更加复杂。图像和视频具有空间结构且可直接进行视觉解读，而音频信号则较为抽象，需要借助频谱图或学习得到的嵌入等专门变换，才能进行有意义的分析。

由于缺乏大规模、高质量且经过指令对齐的音频数据集，这一复杂性进一步加剧，从而限制了稳健音频多模态大语言模型的训练和基准测试。因此，针对音频多模态大语言模型的类似扩展可以包括在时频域中进行定位，识别音频信号中存在伪造语音伪迹的特定区域或特征。朝着这一层次的可解释性迈进，有望显著提升音频深度伪造检测中的信任和透明度。接下来，我们将讨论研究过程中遇到的显著挑战与局限性；我们相信，这些内容能够为这一新兴领域未来的研究工作提供指导和参考。随后，我们将重新审视引言中提出的关键研究问题，并介绍从实验分析中获得的见解。在此基础上，我们还将重点阐述研究期间发现的主要挑战与局限性；我们相信，这些内容能够为这一不断发展的领域的未来研究提供启示和指导。

问：多模态大语言模型能否有效应用于音频深度伪造检测任务?

如前所述，视觉语言模型在深度伪造媒体取证中的表现并不理想。这些成果引发了若干重要问题，例如将多模态大语言模型用于音频深度伪造检测。我们的研究结果表明，多模态大语言模型确实可以用于音频深度伪造检测。通过多提示输入训练的多模态大语言模型展现出了识别欺骗音频的潜力，表 2 和表 3 中的结果便是明证。

问：如何利用多模态大语言模型从特征理解和决策方面改进音频深度伪造检测?

准确率?

多模态大语言模型通过指令引导推理，而非静态的固定特征分类，为音频深度伪造检测提供了一种全新的思路。它们不再仅仅依赖手工设计的声学特征或学习得到的嵌入，也能够理解自然语言提示。这些提示词可以引导模型关注特定的音频特征，例如“机器人般的质感”、“单调音调”或“听觉异常”，这些特征往往与合成语音相关。这种灵活性有助于构建更具可解释性和适应性的检测流程。然而，我们的实验结果表明，当前的多模态大语言模型对深度伪造特有声学线索的内在理解仍然有限，尤其是在缺乏任务特定监督或微调的情况下。不过，经过任务特定微调并采用恰当的文本提示设计后，模型能够取得改进的结果，如表 2 和表 3 所示。

问：基于多模态大语言模型的方法在多大程度上能够提升音频深度伪造检测在不同数据集和攻击类型上的泛化能力?

我们的研究发现，尽管微调后的多模态大语言模型在域内评估（S_{eval}）中表现出色，如表 2 所示，但其对域外数据（如 ITW 数据集）的泛化能力仍然有限，与表5中报告的结果相当。尽管模型经过多种欺骗攻击类型的训练，但往往会对训练分布过拟合，并表现出偏差，经常将大多数 ITW 样本标记为伪造语音。我们认为，提升泛化能力可能不仅需要调整超参数或扩大数据集规模。少样本学习和提示工程等技术，尤其是使用解释性提示或思维链提示，能够帮助模型关注稳健且与领域无关的音频模式。这些策略或许可以提升模型适应陌生输入和多样化欺骗场景的能力。

音频多模态大语言模型在通用音频深度伪造检测等复杂任务中的表现有限，这也可归因于用自然语言描述音频这一根本性挑战。在视觉语言模型训练中，图像描述通常语义丰富，涵盖场景元素、物体属性、情感线索和上下文信息。这些多样而详尽的标注有助于模型建立牢固的跨模态关联。相比之下，描述音频，尤其是合成或经过操纵的语音，往往不那么直观，而且常常缺乏能够捕捉其细微声学细节的词汇或结构。因此，音频多模态大语言模型难以形成同样深入的语义表征，限制了其在复杂下游任务中的效用。

此外，传统的深度伪造检测方法依赖输出类别概率的二元分类器，因此等错误率和曲线下面积等评估指标自然适用。然而，使用多模态大语言模型时，问题的表述转变为音频问答任务。在这种设置下，模型根据输入音频和提示生成响应，通过预测词元而非类别概率估计来产生输出。这一根本差异使其与传统分类器的直接比较意义不大。此外，多模态大语言模型容易产生幻觉，有时生成听起来颇具说服力、实际上却不正确或无关的响应。这些因素表明，我们需要重新审视并可能重新设计评估指标，使其更能体现多模态大语言模型在深度伪造检测任务中的生成特性。

6. 结论与未来工作

本研究初步探索了最先进的多模态大语言模型在音频深度伪造检测中的适用性。我们评估了两种近期模型 Qwen2-Audio-7B-Instruct 和 SALMONN，以了解它们识别合成语音的效果。我们的研究涵盖两种评估模式：零样本评估，即在不进行任何特定任务训练的情况下评估模型；以及微调，即利用带标签数据调整模型。在零样本评估中，两种模型都难以可靠地区分真实语音和欺骗音频，准确率往往接近随机猜测水平（50%）。在极小的带标签数据集上进行微调后，模型在域内数据上的检测表现有所提升。然而，它们对更具挑战性、更多样的数据集（如 In-the-Wild 数据集）的泛化能力仍然有限。未来，我们计划开发专为音频深度伪造检测量身定制的专用多模态大语言模型架构。

除了提升原始检测性能外，我们还计划探索利用多模态大语言模型提升音频深度伪造检测的可解释性。鉴于这些模型能够生成自然语言输出，我们可以利用它们为预测结果提供可解释的理由。这一方向目前鲜有研究，有望深入揭示模型的决策过程，从而提升音频深度伪造检测的透明度。此外，现有研究大多假设整段音频要么是真实的，要么是伪造的。一个很有前景的方向是探索使用多模态大语言模型进行局部深度伪造定位，即仅对部分片段进行篡改。通过精心设计提示词，多模态大语言模型有望识别这些区域，并以人类可读的方式解释故障伪影或不自然的韵律等异常。

致谢

本研究由 IDEAS、TiH、ISI、加尔各答、DST、印度政府资助，项目名称为“媒体通用篡改检测（GTDM）”，项目编号为 OO/ISI/IDEAS-TIH/2023-24/86。

参考文献

[1] B. Zhang, H. Cui, V. Nguyen 和 M. T. Whitty，“音频深度伪造检测：取得了哪些进展，未来又将如何发展”，*Sensors*，第 25 卷，第 7 期，第 1989 页，2025年。

[2] J. Yi, C. Wang, J. Tao, X. Zhang, C. Y. Zhang, 和 Y. Zhao,《音频深度伪造检测：一项综述》，2023年。

[3] J. Yi, C. Y. Zhang, J. Tao, C. Wang, X. Yan, Y. Ren, H. Gu 和 J. Zhou，“Add 2023：面向真实场景的音频深度伪造检测与分析”，载于 *CoRR*，2024年。

[4] T. M. Wani, S. A. A. Qadri, F. A. Wani 和 I. Amerini，《探寻欺骗的声音景观：关于音频深度伪造生成、检测及未来展望的综合综述》，*NOWPublishers*，2024年。

[5] Z. Zhang, W. Hao, A. Sankoh, W. Lin, E. Mendiola-Ortiz, J. Yang 和 C. Mao，“我能听见你：用于深度伪造音频检测的选择性鲁棒训练”，载于国际学习表征会议（*ICLR*），2025年。[6] J.-w. Jung, H.-S. Heo, H. Tak, H.-j. Shim, J. S. Chung, B.-J. Lee, H.-J. Yu 和 N. Evans，“AASIST：使用集成时频图注意力网络的音频反欺骗”，载于 *ICASSP 2022-2022 IEEE*国际声学、语音与信号处理会议（*ICASSP*），第6367–6371页，IEEE，2022年。[7]H. Tak, J. Patino, M. Todisco, A. Nautsch, N. Evans 和 A. Larcher，“基于RawNet2的端到端反欺骗”，载于*ICASSP 2021-2021 IEEE*国际声学、语音与信号处理会议（*ICASSP*），第6369–6373页，IEEE，2021年。[8] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey 和 I. Sutskever，“通过大规模弱监督实现鲁棒语音识别”，2022年。[9] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Qian, Y. Qian, J. Wu, M. Zeng 和 F. Wei，“WavLM：用于全栈语音处理的大规模自监督预训练”，2021年。[10] A. Babu, C. Wang, A. Tjandra, K. Lakhotia, Q. Xu, N. Goyal, K. Singh, P. Von Platen, Y. Saraf, J. Pino，等，“XLS-R：大规模自监督跨语言语音表征学习”，*arXiv*预印本 *arXiv:2111.09296*，2021年。[11] O. C. Phukan, G. S. Kashyap, A. B. Buduru 和 R. Sharma，“异质性优于同质性：探究用于音频深度伪造检测的多语言语音预训练模型”，*arXiv*预印本 *arXiv:2404.00809*，2024年。[12] Y. Guo, H. Huang, X. Chen, H. Zhao 和 Y. Wang，“基于自监督WavLM和多重融合注意力分类器的音频深度伪造检测”，载于*ICASSP 2024-2024 IEEE*国际声学、语音与信号处理会议（*ICASSP*），第12702–12706页，IEEE，2024年。[13] P. Kawa, M. Plata, M. Czuba, P. Szyma´nski 和 P. Syga，“利用Whisper特征改进深度伪造检测”，*arXiv*预印本 *arXiv:2306.01428*，2023年。

[14] Y. Xiao and R. K. Das, “XLS-R-mamba: A dual-column bidirectional state space model for spoofing attack detection”，*IEEE* 信号处理快报，2025年。[15] J. Yang、H. Jin、R. Tang、X. Han、Q. Feng、H. Jiang、S. Zhong、B. Yin 和 X. Hu，“在实践中发挥大语言模型的力量：关于 ChatGPT 及其他模型的调查”，*ACM* 数据知识发现汇刊，第 18 卷，第 6 期，第 1–32 页，2024年。[16] T. Kojima、S. S. Gu、M. Reid、Y. Matsuo 和 Y. Iwasawa，“大语言模型是零样本推理器”，神经信息处理系统进展，第 35 卷，第 22199–22213 页，2022年。[17] D. Zhang、Y. Yu、J. Dong、C. Li、D. Su、C. Chu 和 D. Yu，“多模态大语言模型：近期进展”，*arXiv* 预印本 *arXiv:2401.13601*，2024年。[18] S. Wu、H. Fei、L. Qu、W. Ji 和 T.-S. Chua，“Next-GPT：任意模态到任意模态的多模态大语言模型”，发表于第四十一届国际机器学习会议，2024年。[19] G. Team、R. Anil、S. Borgeaud、J.-B. Alayrac、J. Yu、R. Soricut、J. Schalkwyk、A. M. Dai、A. Hauth、K. Millican、等，“Gemini：一系列高能力多模态模型”，*arXiv* 预印本 *arXiv:2312.11805*，2023年。[20] H. Zhang、X. Li 和 L. Bing，“Video-LLaMA：一种用于视频理解的指令调优音视频语言模型”，*arXiv* 预印本 *arXiv:2306.02858*，2023年。[21] J. Zhang、J. Huang、S. Jin 和 S. Lu，“用于视觉任务的视觉语言模型：综述”，*IEEE* 模式分析与机器智能汇刊，2024年。[22] A. Ghosh、A. Acharya、S. Saha、V. Jain 和 A. Chadha，“探索视觉语言模型的前沿：当前方法与未来方向综述”，*arXiv* 预印本 *arXiv:2404.07214*，2024年。[23] Y. Zou、P. Li、Z. Li、H. Huang、X. Cui、X. Liu、C. Zhang 和 R. He，“人工智能生成媒体检测综述：从非多模态大语言模型到多模态大语言模型”，*arXiv* 预印本 *arXiv:2502.05240*，2025年。[24] S. Jia、R. Lyu、K. Zhao、Y. Chen、Z. Yan、Y. Ju、C. Hu、X. Li、B. Wu 和 S. Lyu，“ChatGPT 能检测深度伪造吗？关于使用多模态大语言模型进行媒体取证的研究”，发表于 *IEEE/CVF* 计算机视觉与模式识别会议论文集，第 4324–4333 页，2024年。[25] P. Yu、J. Fei、H. Gao、X. Feng、Z. Xia 和 C. H. Chang，“释放视觉语言模型在可泛化、可解释深度伪造检测方面的能力”，*arXiv* 电子印本，第 arXiv–2503 页，2025年。[26] Y.-M. Chang、C. Yeh、W.-C. Chiu 和 N. Yu，“Antifake-prompt：提示调优视觉语言模型是伪造图像检测器”，*arXiv* 预印本 *arXiv:2310.17419*，2023年。[27] C.-y. Huang、K.-H. Lu、S.-H. Wang、C.-Y. Hsiao、C.-Y. Kuan、H. Wu、S. Arora、K.-W. Chang、J. Shi、Y. Peng、等，“Dynamic-SUPERB：迈向动态、协作且全面的语音指令微调基准”，发表于*ICASSP 2024—2024 IEEE* 国际声学、语音与信号处理会议（*ICASSP*），第 12136–12140 页，IEEE，2024年。

[28] Q. Yang、J. Xu、W. Liu、Y. Chu、Z. Jiang、X. Zhou、Y. Leng、Y. Lv、Z. Zhao、C. Zhou、等，“Air-bench：通过生成式理解对大型音频语言模型进行基准测试”，*arXiv* 预印本 *arXiv:2402.07729*，2024年。

[29] Qwen《GGT-1 技术报告》，2024年。[30] DeepSeek-AI，《DeepSeek-V3 技术报告》，2025年。[31] Z. Zheng、X. Peng、T. Yang、C. Shen、S. Li、H. Liu、Y. Zhou、T. Li 和 Y. You，《Open-Sora：让所有人都能进行高效视频制作》，2024年。[32] R. Rombach、A. Blattmann、D. Lorenz、P. Esser 和 B. Ommer，《使用潜在扩散模型进行高分辨率图像合成》，2021年。[33]W.-N. Hsu、B. Bolte、Y.-H. H. Tsai、K. Lakhotia、R. Salakhutdinov 和 A. Mohamed，《HuBERT：通过隐藏单元掩码预测进行自监督语音表征学习》，2021年。[34] S. Chen、Y. Wu、C. Wang、S. Liu、D. Tompkins、Z. Chen 和 F. Wei，《BEATs：使用声学分词器进行音频预训练》，2022年。[35] A. Baevski、W.-N. Hsu、Q. Xu、A. Babu、J. Gu 和 M. Auli，《Data2Vec：面向语音、视觉和语言的通用自监督学习框架》，2022年。[36] F. Chollet，《Xception：使用深度可分离卷积的深度学习》，2017年。[37] M. Tan 和 Q. V. Le，《EfficientNet：重新思考卷积神经网络的模型缩放》，2020年。[38] Y. Chen、J. Yi、J. Xue、C. Wang、X. Zhang、S. Dong、S. Zeng、J. Tao、L. Zhao 和 C. Fan，《RawBMamba：用于音频深度伪造检测的端到端双向状态空间模型》，*arXiv* 预印本 *arXiv:2406.06086*，2024年。[39] X. Liu、M. Liu、L. Wang、K. A. Lee、H. Zhang 和 J. Dang，《利用与位置相关的局部—全局依赖关系进行合成语音检测》，载于 *ICASSP 2023-2023 IEEE* 声学、语音与信号处理国际会议（*ICASSP*），第 1–5 页，IEEE，2023年。[40] H. Tak、J. weon Jung、J. Patino、M. Kamble、M. Todisco 和 N. Evans，《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*2021* 年自动说话人验证与伪造语音对策挑战赛会议录，第 1–8 页，2021年。[41] W. Xu、X. Dong、L. Ma、A. B. J. Teoh 和 Z. Lin，《RawFormer：用于低光照原始图像增强的高效视觉 Transformer》，*IEEE* 信号处理快报，第 29 卷，第 2677–2681 页，2022年。[42] X. Zhou、D. Garcia-Romero、R. Duraiswami、C. Espy-Wilson 和 S. Shamma，《用于说话人识别的线性频率倒谱系数与梅尔频率倒谱系数比较》，载于*2011 IEEE* 自动语音识别与理解研讨会，第 559–564 页，2011年。[43] A. Baevski、Y. Zhou、A. Mohamed 和 M. Auli，《wav2vec 2.0：语音表征自监督学习框架》，神经信息处理系统进展，第 33 卷，第 12449–12460 页，2020年。

[44] X. Wu, R. He, Z. Sun, 和 T. Tan，《采用噪声标签进行深度人脸表征的轻量级 CNN》，*IEEE*信息取证与安全汇刊，第 13 卷，第 11 期，页 2884–2896，2018年。[45] X. Wang, J. Yamagishi, M. Todisco, H. Delgado, A. Nautsch, N. Evans, M. Sahidullah, V. Vestman, T. Kinnunen, K. A. Lee，等，《ASVspoof 2019：一个大规模合成、转换和重放语音公共数据库》，《计算机语音与语言》，第 64 卷，文章编号 101114，2020年。[46] J. Yi, R. Fu, J. Tao, S. Nie, H. Ma, C. Wang, T. Wang, Z. Tian, Y. Bai, C. Fan, S. Liang, S. Wang, S. Zhang, X. Yan, L. Xu, Z. Wen, 和 H. Li，《ADD 2022：首届音频深度合成检测挑战赛》，载于*ICASSP 2022 - 2022 IEEE*国际声学、语音与信号处理会议（*ICASSP*），页 9216–9220，2022年。[47] U. Cappellazzo, M. Kim, H. Chen, P. Ma, S. Petridis, D. Falavigna, A. Brutti, 和 M. Pantic，《大语言模型是强大的视听语音识别学习器》，载于*ICASSP 2025 - 2025 IEEE*国际声学、语音与信号处理会议（*ICASSP*），页 1 5，2025 –。[48] R. Huang, M. Li, D. Yang, J. Shi, X. Chang, Z. Ye, Y. Wu, Z. Hong, J. Huang, J. Liu, Y. Ren, Z. Zhao, 和 S. Watan- abe，《AudioGPT：理解并生成语音、音乐、声音和说话头像》，2023年。[49] D. Zhang, S. Li, X. Zhang, J. Zhan, P. Wang, Y. Zhou, 和 X. Qiu，《SpeechGPT：赋予大语言模型内在的跨模态对话能力》，2023年。[50] Y. Gong, A. H. Liu, H. Luo, L. Karlinsky, 和 J. Glass，《音频与语音联合理解》，载于*2023*年 *IEEE* 自动语音识别与理解研讨会（*ASRU*），2023年。[51] Y. Chu, J. Xu, Q. Yang, H. Wei, X. Wei, Z. Guo, Y. Leng, Y. Lv, J. He, J. Lin, C. Zhou, 和 J. Zhou，《Qwen2-Audio 技术报告》，*arXiv*预印本 *arXiv:2407.10759*，2024年。[52] K.-H. Lu, Z. Chen, S.-W. Fu, H. Huang, B. Ginsburg, Y.- C. F. Wang, 和 H. yi Lee，《DESTA：通过描述式语音文本对齐增强语音语言模型》，载于*Interspeech 2024*，页 4159–4163，2024年。[53] C. Tang, W. Yu, G. Sun, X. Chen, T. Tan, W. Li, L. Lu, Z. MA, 和 C. Zhang，《SALMONN：面向大语言模型的通用听觉能力》，载于第十二届国际学习表征会议，2024年。[54] Y. Li, X. Wang, S. Cao, Y. Zhang, L. Ma, 和 L. Xie，《基于转录提示的高效音频大语言模型，用于鲁棒语音识别》，*arXiv*预印本 *arXiv:2408.09491*，2024年。[55] Y. Fathullah, C. Wu, E. Lakomkin, J. Jia, Y. Shangguan, K. Li, J. Guo, W. Xiong, J. Mahadeokar, O. Kalinli，等，《通过语音识别能力提示大语言模型》，载于*ICASSP 2024-2024 IEEE*国际声学、语音与信号处理会议（*ICASSP*），页 13351–13355，IEEE，2024年。

[56] C. Tang, W. Yu, G. Sun, X. Chen, T. Tan, W. Li, L. Lu, Z. Ma, 和 C. Zhang, 《拓展大语言模型以实现语音与音频描述》, 载于*ICASSP 2024-2024 IEEE*国际声学、语音与信号处理会议（*ICASSP*）, 页 11236–11240, IEEE, 2024年。[57] W. Shan, Y. Li, Y. Zhang, Y. Luo, C. Xu, X. Zhao, L. Meng, Y. Lu, M. Zhang, H. Yang, 等, 《利用提示感知的音频编码器混合增强语音大语言模型》, *arXiv*预印本 *arXiv:2502.15178*, 2025年。[58] S. A. Shahzad, A. Hashmi, Y.-T. Peng, Y. Tsao, 和 H.-M. Wang, 《ChatGPT 在视听深度伪造检测方面表现如何：ChatGPT、人工智能模型与人类感知的比较研究》, *arXiv*预印本 *arXiv:2411.09266*, 2024年。[59] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, 等, 《LoRA：大语言模型的低秩适配。》, *ICLR*, 第 1 卷, 第 2 期, 页 3, 2022年。[60] N. M¨uller, P. Czempin, F. Dieckmann, A. Froghyar, 和 K. B¨ottinger, 《音频深度伪造检测能否泛化？》, *arXiv*预印本 *arXiv:2203.16263*, 2022年。[61] L. Pham, P. Lam, D. Tran, H. Tang, T. Nguyen, A. Schindler, F. Skopik, A. Polonsky, 和 H. C. Vu, 《深度伪造语音检测综合综述与批判性分析》, *ComputerScience Review*, 第 57 卷, 文章编号 100757, 2025年。