arXiv:2601.02983v1 [cs.SD] 6 Jan 2026

通过频率–时间强化学习和音频大语言模型实现可解释的全类型音频深度伪造检测

Yuankun Xie[1,2], XiaoxuanGuo[1,2],Jiayi Zhou[2],Tao Wang[2],Jian Liu[2]*, Ruibo Fu[3], XiaopengWang[3],Haonan Cheng[1],Long Ye[1]*[1]中国传媒大学，北京，中国[2]蚂蚁集团机器智能，上海，中国[3]中国科学院自动化研究所，北京，中国xieyuankun@cuc.edu.cn

摘要

近年来，音频大语言模型（ALLM）取得了长足进展，使高质量合成音频广泛可得，也增加了语音、环境声音、歌声和音乐遭受恶意音频深度伪造的风险。因此，现实世界中的音频深度伪造检测（ADD）需要能够跨异质音频类型泛化并提供可解释判断的全类型检测器。鉴于音频大语言模型具有强大的多任务泛化能力，我们首先研究其在监督微调（SFT）和强化微调（RFT）两种设置下的全类型音频深度伪造检测性能。然而，仅使用二元真实/伪造标签进行监督微调，往往会使模型沦为黑盒分类器，牺牲可解释性。同时，在稀疏监督下，朴素强化微调容易出现奖励劫持，并生成幻觉式、缺乏依据的推理依据。为此，我们提出一种自动标注与润色流程，用于构建频率–时间结构化思维链（CoT）推理依据，并生成 ~340K条冷启动示范样本。在CoT数据的基础上，我们提出频率时间分组相对策略优化（FT-GRPO），这是一种两阶段训练范式：先通过监督微调对音频大语言模型进行冷启动，再在基于规则的频率–时间约束下应用GRPO。实验表明，FT-GRPO在全类型音频深度伪造检测上达到最先进性能，同时生成有可解释性且基于FT的推理依据。数据和代码可在线获取[1]。

1 引言

随着音频大语言模型（音频大语言模型）的最新进展，生成高质量合成音频（包括深度伪造语音、环境声音、歌声和音乐）变得越来越容易。这类深度伪造内容带来了媒体与娱乐、网络安全和政治传播领域日益加剧的风险。

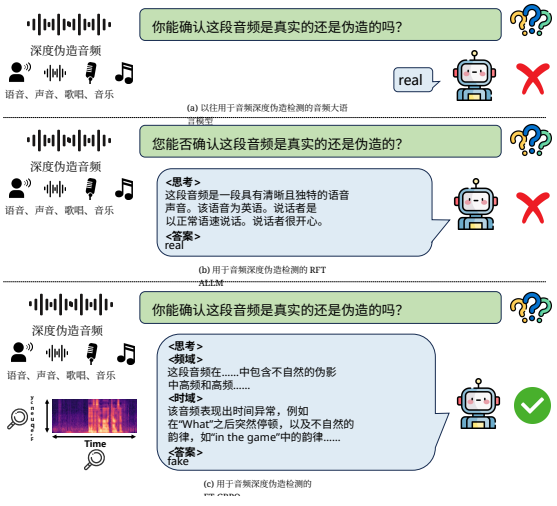

图 1：用于全类型音频深度伪造检测任务的三种基于音频大语言模型的反伪造音乐方法：（a）仅标签监督微调，（b）强化微调，以及（c）我们提出的 FT-GRPO。

近年来，音频深度伪造检测（ADD）的研究稳步发展。早期工作主要针对深度伪造语音，并在 ASVspoof 挑战赛中催生了多种对抗措施（Todisco 等，2019；Liu 等，2023；Wang 等，2024）。近年来，音频深度伪造检测的范围已从语音扩展至其他音频领域，包括歌声深度伪造检测（Zhang 等，2024d；Xie 等，2024a；Zang 等，2024）、环境声音深度伪造检测（Yin 等，2025；Xie 等，2024b）和音乐深度伪造检测（Comanducci 等，2024）。然而，现有研究大多局限于单一音频类型。在真实世界的音频深度伪造检测场景中，检测器需要能够判断任意音频类型的真伪。近期，已有一些研究探讨涉及两个或更多音频领域的跨类型设置（Gohari 等，2025；Xie 等，2025）；然而，这些工作在强大的跨类型泛化能力和可解释决策方面仍有欠缺。

*通讯作者[1]

[1] https://xieyuankun.github.io/FT-GRPO

化能力和可解释决策方面仍有欠缺。

音频大语言模型继承了大量预训练知识，并具备强大的理解能力。近期关于音频大语言模型可信度的研究（Li 等人，2025；Yang 等人，2025）开始涉及深度伪造检测，但目前仍处于初步阶段。Gu 等人（Gu 等人，2025）首次系统研究了将音频大语言模型用于音频深度伪造检测。他们发现，零样本提示往往会失败，因为音频大语言模型无法可靠地内化真实性判断目标，导致演奏不佳；采用二元真实/伪造标签进行监督微调（SFT）能够提高准确率，但通常会使模型沦为黑盒分类器，并可能抑制其固有的推理和解释能力。此外，现有基于音频大语言模型的音频深度伪造检测研究大多局限于语音领域。

这些局限凸显了实际部署面临的一项核心挑战：我们能否训练音频大语言模型来解决全类型音频深度伪造检测任务，同时实现强大的跨类型泛化能力，并给出可解释的决策依据？强化学习（RL）为实现这一目标提供了一个有前景的方向。在视觉领域，Visual-RFT（Liu 等人，2025）表明，基于强化学习的微调在更好地保留底层语言模型可解释性的同时，其表现可优于监督微调。受此证据启发，我们研究强化学习是否也能为音频大语言模型执行全类型音频深度伪造检测带来类似益处。

为此，我们从音频大语言模型和强化微调（RFT）的角度重新审视全类型音频深度伪造检测。具体而言，首先，我们在具有代表性的音频大语言模型上进行监督微调（SFT），建立一个基于音频大语言模型的全类型音频深度伪造检测设置下的反欺骗模型基准，以量化其当前性能和跨类型泛化能力。为了进一步提升可解释性和泛化能力，随后我们探索了直接强化微调。然而，如图 1 所示，单独使用强化微调通常无法产生可靠且有证据支撑的推理依据，因为其监督信号通常较为稀疏，并且往往仅限于格式和最终标签约束，对推理过程几乎无法提供指导，还会助长奖励劫持或退化式解释（Zhou 等，2025）。

受 DeepSeek-R1 采用 SFT冷启动的强化学习框架取得成功的启发（Guo等人，2025a），以及近期研究表明基于 SFT 的冷启动能够增强多模态大语言模型的强化学习效果（Chen等人，2025；Wei等人，2025a），我们首先通过自动标注与润色流程，为现有全类型音频深度伪造检测数据集构建频率—时间（FT）结构化思维链（CoT）推理依据。这为音频大语言模型提供了高质量示例，帮助其学习有原则的推理模式。尽管其他线索也可能有所帮助，我们专注于频率和时间这两个简洁且广泛适用的视角；先前研究已表明，它们有助于捕捉具有区分性的欺骗伪迹，并已应用于轻量级反欺骗模型（Tak等人，2021；Jung等人，2022）。基于这一冷启动，我们进一步通过强化学习优化模型。具体而言，我们提出频率—时间组相对策略优化（FT-GRPO），这是一种两阶段训练范式，将 SFT 初始化与基于规则的频率和时域约束下的 GRPO 相结合。我们还有效利用了经标注标记的非思考不匹配样本（其推理依据与标签相冲突），以提高检测准确率和可解释性。

我们将本工作的贡献总结如下：

- 为量化基于音频大语言模型的对抗措施在全类型音频深度伪造检测任务中的跨类型泛化能力，我们首先通过对具有代表性的音频大语言模型进行监督微调，建立了一个基准。
- 为提升音频大语言模型的可解释性，我们提出了一套自动标注与润色流程，为四个公开数据集构建思维链推理依据，生成了 ~340 K 个冷启动示例；据我们所知，这是音频深度伪造检测文献中首个带有推理依据标注的资源。
- 我们提出 FT-GRPO，这一两阶段范式将 SFT 冷启动与基于规则的微调推理约束下的 GRPO 相结合，并明确利用非思考样本来提升准确率和可解释性。
- FT-GRPO 仅使用语音数据训练的 30亿参数模型取得了最先进水平（SOTA）的结果，在 ASVspoof2019LA 评估集上达到 99.75% 的准确率，同时生成可解释的推理依据；在所有类型的数据上进行协同训练后，其在所有测试集上的平均准确率进一步达到 90.10%。

2 相关工作

2.1 音频深度伪造检测

在介绍音频深度伪造检测领域的当前工作时，我们按类型将其分为语音、声音、歌声和音乐。

语音。语音深度伪造检测已得到广泛研究，这在很大程度上得益于 ASVspoof 挑战赛。具有代表性的对抗措施包括 AASIST（Jung等人，2022）以及将 XLSR 与 AASIST 相结合的基于自监督学习的流程（Tak等人，2022）。后续工作探索了不同的自监督学习骨干网络（Phukan等人，2024年；Kheir等人，2025）、层利用（Zhang等人，2024b；Wang等人，2025c；Pan等人，2024年）以及鲁棒性（Zhang等人，2025；Kawa等人，2023）。

其他类型（声音、歌唱、音乐）。与语音深度伪造检测的研究相比，其他类型的研究目前仍主要处于数据集构建阶段，方法探索较少。对于声音，近期的环境声音深度伪造检测（ESDD）挑战赛（Yin等，2025）极大地推动了深度伪造声音检测研究的发展。ESDD 涵盖了广泛的基于音频大语言模型的文本转音频（TTA）和音频转音频（ATA）合成方法，目前最先进的解决方案通常依赖于声音自监督学习特征，例如 SS LAM（Guo 等，2025b）。对于歌声，SVDD（Zhang 等，2024d）推动了相关方法的发展，这些方法将 XLSR 与面向音乐的自监督学习模型相结合，例如 MERT（Li 等，2024a）和 WavLM（Chen 等，2022）（Guragain 等，2024；Zhang 等，2024c；Chen 等，2024）。对于音乐，FakeMusicCaps（Comanducci 等，2024）支持合成音乐检测，但相关方法研究仍然有限（Li 等，2024c；Wei 等，2025b）。

跨类型。少数研究探讨了不同音频类型之间的迁移，例如语音与歌声之间（Gohari等人，2025）或从语音到音乐（Li等人，2024b）。Xie等人（Xie等人，2025）进一步建立了一个基于自监督学习的全类型音频深度伪造检测基准，并提出小波提示调优以提升泛化能力。

2.2 音频大语言模型用于音频深度伪造检测

音频大语言模型拥有丰富的预训练知识和强大的理解能力，已成功应用于音频推理和音频描述（Zhifei 等，2025；Tang 等，2024；Ma 等，2025）、语音识别与生成（Fathullah 等，2024；Zhang 等，2024a；Yan 等，2025）以及质量评估（Wang 等，2025a,b）等多种任务。近期关于音频大语言模型可信度的研究（Li 等，2025；Yang 等，2025）也开始涉及深度伪造检测，但目前仍处于初步阶段。Gu 等人（Gu 等，2025）首次开展了使用音频大语言模型执行音频深度伪造检测任务的系统性研究。他们在 Qwen-Audio 上进行监督微调，以实现二元真实/伪造分类，并从提示设计到零样本、少样本和超参数设置，对音频大语言模型的性能进行了详细研究，在19LA基准上取得了 99.40% 的准确率，超过了以往的小型音频模型。然而，将输出限制为简单的真实/伪造标签，严重限制了音频大语言模型的可解释性。

3 方法

图2 概述了我们的方法。我们首先通过两步式自动生成字幕与润色流程构建FT推理依据，生成一个包含有效推理依据的思考子集和一个非思考子集。随后，我们提出FT-GRPO，这是一种两阶段训练范式：(i) 在思考子集上通过监督微调对音频大语言模型进行冷启动，以学习FT推理模式；(ii) 采用复合奖励进行GRPO训练，以提升检测准确率，同时鼓励生成完整且基于FT的推理依据。

3.1 数据构建

当代音频大语言模型的推理和解释能力仍然有限。因此，我们首先构建推理依据标注数据，以激发其推理行为并提升任务理解能力。相应地，对于音频深度伪造检测数据集 D 中的样本 $S = \{a, y\}$，其中 a 表示音频片段，$y \in \{$real，fake$\}$ 表示真实标签，我们采用两步描述生成流程。

步骤1：原始音频描述。我们首先使用最先进水平的音频描述生成模型 Qwen3-Omni-Captioner[2]，该模型具备细粒度的音频理解能力，旨在复杂多样的声学场景中生成准确、全面的描述。无需额外提示词，它便能解析并描述范围广泛的音频内容，包括语音、环境声音、音乐和电影音效，因此非常适合我们的全类型音频深度伪造检测任务。

在步骤1中，每个音频样本 Sisfed 被输入说明文字生成器。给定标签 y，我们指示模型从频域和时域两个角度进行推理，将推理过程写入 <思考>，并将标签 y 写入 <答案>，从而生成原始说明文字 R_c。然而，R_c 存在一些问题：（i）说明文字通常过于冗长——

[2]https://huggingface.co/Qwen/Qwen3-Omni-30B-A3B- 说明文字模型

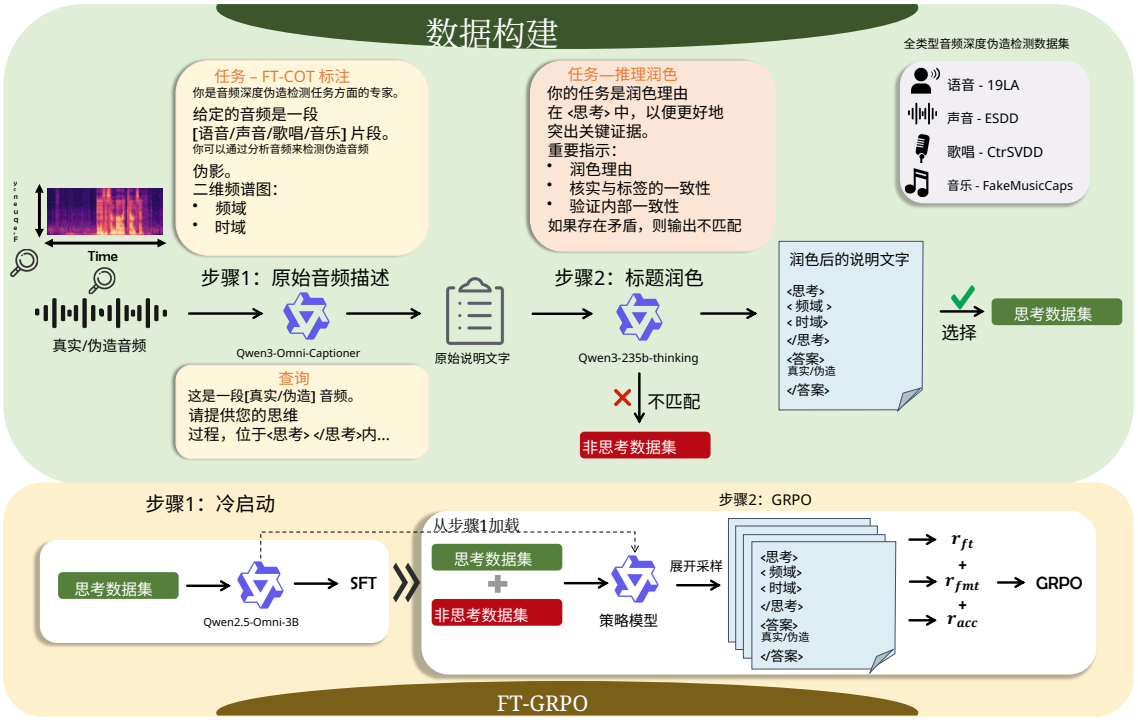

图2：流程概览。上：通过步骤1原始音频说明文字生成和步骤2说明文字润色进行数据构建。下：FT-GRPO训练，包括第一阶段监督微调冷启动和第二阶段群体相对策略优化。

冗长且包含大量与真实性判断无关的信息；（ii）有些输出内部不一致，其中 <思考> 与< 答案> 相矛盾（例如，推理表明是真实的，而答案却是伪造的）；（iii）尽管已明确提供标签y，有些输出的预测 <答案>仍与之矛盾。

步骤2：标题润色。为解决这些问题，我们使用大语言模型 Qwen3-235B[3] 来润色每条R_c。针对问题 (i)，Qwen3-235B 会重写描述，使其简洁明了，删除与真实性线索无关的内容，并确保符合所需的输出格式。针对问题 (ii) 和 (iii)，当检测到矛盾，或无法可靠地协调原始说明文字与给定标签时，Qwen3-235B 会输出一个 <不匹配> 标记。经过此步骤后，我们为大多数样本获得了润色后的说明文字 P_c。 S。

我们将被标记为 <不匹配> 的样本视为当前音频大语言模型无法稳健处理的非思考样本。对于这些非思考样本，我们仅保留其 <答案> ，并将其用作强化学习数据 D_n。对于具有有效 P_c 的样本，我们同时将其用作：（i）FT-GRPO 第一阶段监督微调的冷启动数据 D_{think}；以及（ii）第二阶段 GRPO 的训练数据。因此，第二阶段 GRPO 基于$D_{\text{r}} = D_{\text{think}} \cup D_{\text{non_think}}$ 进行训练。这一设计的动机是观察到，在缺乏可靠 <思考> 标注的数据上进行监督微调可能会对训练产生负面影响；因此，我们仅将经过正确润色且格式一致的样本用作冷启动初始化的高质量数据。

[3]https://huggingface.co/Qwen/Qwen3-235B-A22B

3.2 FT-GRPO

在完成数据构建后，我们引入**FT-GRPO**，这是我们提出的训练范式，由两个步骤组成：（i）步骤1：SFT冷启动；以及（ii）步骤2：GRPO。

步骤1：SFT冷启动。在此步骤中，我们对 D_{think}（所有带有经过润色且合理的 <思考> 标注的样本）进行SFT。这使音频大语言模型能够从频域和时域两个角度学习如何进行音频深度伪造检测任务推理，并按照所需的模式生成输出。

设 π_θ 是一个参数为 θ 的音频大语言模型。对于每个SFT示例 $(a_i, P_{c,i}) \in D_{\text{think}}$，目标$P_{c,i}$ 是一个包含 <思考>和 <答案>的结构化序列。我们通过最大化在给定音频输入a的条件下生成 P_c 的条件似然来训练 π_θ：

$$\mathcal{L}_{\text{SFT}}(\theta) = -\mathbb{E}_{(a,P_c)\sim D_{\text{think}}} \left[\sum_{t=1}^{|P_c|} \log \pi_\theta(x_t \mid x_{<t}, a) \right]. \quad (1)$$

其中 $P_c = (x_1, \ldots, x_{|P_c|})$，$x_{<t}$ 表示

前缀标记（$x_1, \ldots, x_{t-1}$）。

步骤2：GRPO 优化。在强化学习过程中，我们进一步使用 GRPO 优化第 1 步 SFT 模型，将其作为策略模型，并采用复合奖励，同时鼓励预测正确、确保输出格式良好，并促使模型完成频率–时间（FT）推理。给定音频片段 a 及其真实标签$y \in$ {真实，伪造}，策略模型会生成输出 $o = \{\tau, y_{\text{pred}}\}$，其中 τ 表示 <思考> 中的推理路径，y_{pred} 表示 <答案>中的预测标签。

准确率奖励。我们定义一种稀疏的准确率奖励，以鼓励正确分类：

$$r_{\text{acc}} = \begin{cases} 1, & y_{\text{pred}} = y, \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

格式奖励。为稳定训练，我们强制采用固定的输出模式，并定义二元格式奖励：

$$r_{\text{fmt}} = \mathbb{I}(\mathcal{F}(o)), \tag{3}$$

其中，如果推理内容 τ 由<思考> 包裹，且预测标签y_{pred} 由 <答案>包裹，则 $\mathcal{F}(o)$ 为真。

FT 推理路径奖励。由于 r_{acc} 和r_{fmt} 都较为稀疏，且无法直接塑造 τ，我们引入 FT 推理奖励 r_{ft}，鼓励模型提供来自频域和时域的证据。令$\mathcal{A} =$ {FD，TD} 表示两个元步骤，其声明标签分别为 $a_{\text{FD}} =$<频域> 和$a_{\text{TD}} =$<时域>。我们定义：

$$r_{\text{ft}} = \frac{1}{2} \sum_{i \in \mathcal{A}} \mathbb{I}(a_i \in \tau \ \wedge \ g(a_i) = \texttt{True}), \tag{4}$$

其中 $g(a_i)$ 是一种轻量级、基于规则的完整性检查（标签存在，且后面至少跟有一个完整句子）。因此， $r_{\text{ft}} \in \{0, 0.5, 1.0\}$ 分别表示缺失、不完整和完整的 FT 推理。

总奖励。 输出 o的最终奖励计算如下

$$r = r_{\text{acc}} + \alpha\, r_{\text{fmt}} + \beta\, r_{\text{ft}}. \tag{5}$$

Type	来源	训练集	Dev	Eval
语音	19LA	25,380	24,844	71,237
声音	ESDD	139 055 39	710	4 000
歌唱	CtrSVDD	84,404	43,625	92,769
音乐	FakeMusicCaps	20,861	6,058	6,122
All	合并来源	199,023	84,438	189,951

表 1：全类型音频深度伪造检测数据集中的原始训练集、开发集和评估集。

GRPO 目标。 我们采用 GRPO 来优化策略模型。对于每个训练实例（音频查询） a，当前策略 π_θ 会采样一组 G 响应 $\{o_1, \ldots, o_G\}$。我们使用公式（5）计算奖励 $\{r_1, \ldots, r_G\}$，并估算第 i个响应的组相对优势，计算如下：

$$A_i = \frac{r_i - \text{mean}(\{r_1, \ldots, r_G\})}{\text{std}(\{r_1, \ldots, r_G\}) + \epsilon}, \tag{6}$$

其中，mean(·) 和 std(·) 分别表示组内的均值和标准差， ϵ 是一个用于数值稳定性的小常数。随后，策略模型会进行更新，以提高具有更高 A_i 的响应的似然，同时通过相对于参考策略 π_{ref}（监督微调快照）施加 KL惩罚来约束更新，过程与标准 GRPO 相同。

4 实验

4.1 数据集

要评估 CM 在全类型音频深度伪造检测中的能力，选择合适的数据集至关重要。我们优先选择相对干净的数据集，并排除部分伪造的案例，使评估聚焦于深度伪造痕迹，而非背景噪声或真实与伪造片段混合等混杂因素。这种受控设置有助于更清晰地分析跨类型音频深度伪造检测的泛化能力。原始数据集的详细信息汇总于表1。关于 FT-GRPO 训练与验证数据，以及思考与非思考样本的划分详情，请参阅附录 A。

Speech-19LA。这是一个广泛采用的基准，包含 12,456 个真实样本和 108,978 个由 11 种文本转语音系统及 8 种语音转换系统（A01–A19）生成的伪造样本。按照标准协议，A01–A06 用于训练，A07–A19 用于评估，以确保评估集中的伪造系统在训练期间未曾出现。

Sound-ESDD。ESDD（Yin 等，2025）是首个专门针对环境声音深度伪造检测的挑战赛。它涵盖了七种基于音频语言模型的

Type	ALLM	语音、声音、歌唱			音乐	平均值
语音	Qwen2-Audio-Chat-7B	89.99	67.05	**84.85**	82.33	81.06
语音	Qwen2.5-Omni-3B	**99.04**	**73.31**	66.29	**91.16**	**82.45**
语音	Qwen2.5-Omni-7B	96.32	61.95	80.30	90.80	82.34
声音	Qwen2-Audio-Chat-7B	32.24	62.60	85.40	14.88	48.78
声音	Qwen2.5-Omni-3B	**86.70**	**67.55**	77.99	**93.58**	**81.46**
声音	Qwen2.5-Omni-7B	77.23	64.32	**85.68**	90.36	79.40
歌唱	Qwen2-Audio-Chat-7B	49.01	52.22	79.53	65.39	61.54
歌唱	Qwen2.5-Omni-3B	39.78	**67.13**	**83.97**	**73.39**	66.07
歌唱	Qwen2.5-Omni-7B	**73.29**	47.55	81.50	70.81	**68.29**
音乐	Qwen2-Audio-Chat-7B	54.49	**74.42**	37.10	**88.96**	63.74
音乐	Qwen2.5-Omni-3B	83.72	73.40	23.06	77.34	64.38
音乐	Qwen2.5-Omni-7B	**89.28**	70.57	**83.41**	88.47	**82.93**

表2：基于音频大语言模型的反伪造音乐在全类型音频深度伪造检测数据集上训练与测试的准确率结果（%）。

合成方法，包括文本转音频（TTA）和音频转音频（ATA）生成。训练集和验证集划分包含方法 G01–G04，而测试集划分包含此前未见过的方法 G05–G07。

歌声-CtrSVDD。SVDD（Zhang等人，2024d）是首个专门面向歌声深度伪造检测的挑战赛。我们使用其 CtrSVDD 子集，该子集基于普通话和日语歌唱语料库构建，并由 14 个歌声合成/歌声转换系统合成。我们遵循官方划分：A01–A08 用于训练，A09–A14 用于测试。

音乐-FakeMusicCaps。FakeMusicCaps（Comanducci等人，2024年）是一个用于音乐深度伪造检测的数据集。其真实域音频来自 MusicCaps（Agostinelli等人，2023），其中包含来自 AudioSet（Gemmeke等人，2017年）的 5.5k 个10 秒音乐片段，并附有专业音乐家标注。FakeMusicCaps 使用六种合成方法（TTM01–TTM05 以及一种未知方法），根据 SunoCaps 风格说明文字生成伪造音乐。真实片段按 7:1:2 的比例划分为训练集/验证集/测试集，而伪造片段则使用 TTM01–TTM03 进行训练，TTM04 用于验证，TTM05 及未知方法用于测试。

4.2 基线

我们在全类型训练与评估协议下评估音频大语言模型，包括Qwen2-Audio- Chat-7B[4]，Qwen2.5-Omni-3B[5],和 Qwen2.5- Omni-7B[6]。在 FT-GRPO 阶段，我们选择 Qwen2.5-Omni-3B 进行实验，以兼顾效果与效率。在最先进水平对比中，我们纳入了一种此前的小模型全类型音频深度伪造检测方法（Xie等人，2025）以及一种近期的大模型方法（Gu 等人，2025）。

4.3 实现细节

所有实验均在 8 张 NVIDIA A100 GPU 上进行，我们的微调流程使用 ms-swift（Zhao 等人，2025）实现。对于监督微调和群体相对策略优化，我们均采用 LoRA（Hu 等人，2022）进行参数高效微调，LoRA 秩为$r = 64$，LoRA alpha（缩放因子）为 $= 16$，LoRA 丢弃率为 $= 0.05$，遵循先前工作中探索的设置（Gu 等人，2025）。训练期间，大语言模型和音频编码器均设置为可训练。对于监督微调基线，每种音频类型均训练 5 轮。对于 FT-GRPO，第一阶段监督微调冷启动在语音和音乐上训练 3 轮，在声音和歌唱上训练 2 轮；第二阶段群体相对策略优化则对所有类型训练 2 轮。对于 FT-GRPO 的两个阶段，我们均采用 1×10^{-5} 的学习率和每设备批量大小 16。在第二阶段群体相对策略优化中，我们将 G 设为 8，采样温度设为 0.9。在总奖励公式中，α 和 β 均设为 0.1。评估方面，所有实验均报告准确率（ACC），其计算方式是直接比对模型响应中的预测标签与真实标签。与先前音频深度伪造检测基准中常用的等错误率相比，准确率更能反映现实世界中的决策情况。

5 结果与分析

5.1 音频大语言模型在全类型音频深度伪造检测上的演奏

音频大语言模型通常使用异构音频数据和多项任务进行训练；然而，它们在音频深度伪造检测任务上的有效性仍不明确。我们首先

[4]https://huggingface.co/Qwen/Qwen2-Audio-7B-Instruct [5] https://huggingface.co/Qwen/Qwen2.5-Omni-3B [6] https://huggingface.co/Qwen/Qwen2.5-Omni-7B

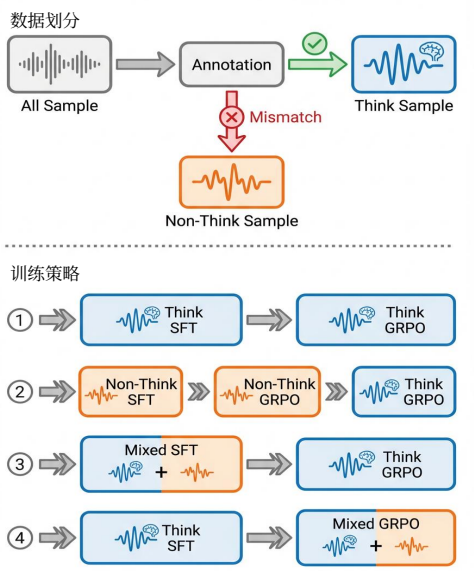

图 3：FT-GRPO 中的四种不同训练策略。

策略	语音、	声音、	歌唱	音乐	AVG
SFT	99.04	73.31	66.29	**91.16**	82.45
GRPO	96.32	68.14	78.21	88.75	82.86
S①	99.30	71.75	83.24	89.76	86.01
S②	98.36	66.94	**85.43**	89.67	85.10
S③	99.46	71.68	83.65	90.01	86.20
S④	**99.75**	**75.05**	84.26	90.44	**87.38**

表3：语音训练的 Qwen2.5-Omni-3B 在不同训练策略下的消融结果（%）。

按照第 4.1 节所述的协议评估具有代表性的音频大语言模型，并在表 2 中报告结果。

在语音数据上训练时，Qwen2.5-Omni-3B 的域内（ID）准确率最高，达到 99.04%，平均演奏也最高，为 82.45%，表明其对其他音频类型具有相对较强的泛化能力。相比之下，对于其他三种训练类型（声音、歌唱和音乐），所有接受评估的音频大语言模型均未能持续取得较高的域内演奏，这表明仅靠通用的全用途音频指令微调，无法稳定解决音频深度伪造检测问题。值得注意的是，声音训练的反伪造音乐退化幅度最大：所有模型的域内准确率（声音列）仅维持在 ~60% 左右（62.60–67.55%），凸显了环境声音深度伪造检测的难度，以及当前音频大语言模型在检测声音类型方面有限的泛化能力。

5.2 FT-GRPO 消融研究

鉴于仅监督微调训练的局限性——包括演奏效果欠佳和可解释性有限，我们进一步探索将强化学习用于音频深度伪造检测。鉴于该策略已被证明在视觉领域优于常规监督微调（Liu 等人，2025），我们首先直接对经过语音训练的 Qwen2.5-Omni-3B 应用 GRPO，以检验其有效性。结果见表3。与仅监督微调训练相比，GRPO使平均准确率小幅提升0.48%；然而，其域内语音准确率降至 96.32%。此外，对生成输出的检查表明，尽管设置了微调奖励并规定了思维模式，推理仍缺乏有效约束，且常常退化为无关或杂乱的陈述，这可能是奖励劫持或捷径行为所致（Baker 等人，2025）。这是意料之中的，因为我们的奖励主要针对预测正确性和格式合规性，而没有明确监督推理轨迹。受 DeepSeek-R1 中冷启动强化学习范式的启发，我们因此利用冷启动示例引导模型进行微调结构化思维链推理，为后续的 GRPO 提供更稳定的初始化，从而提升可读性并减少退化推理。 +

这引出了另一个问题：在数据标注阶段，我们应如何处理不匹配的样本？具体而言，没有思维过程的样本是否会影响演奏？为回答这一问题，我们设计了四种消融训练策略，以研究应将这些非思考样本置于何处，如图4所示。具体来说，S1 会丢弃非思考样本；S2 包含三个阶段：首先在非思考样本上进行监督微调，然后在带思考标注的样本上继续进行监督微调，最后应用 GRPO；S3 将非思考样本混入监督微调阶段；而 S4 仅使用带思考标注的样本进行监督微调，同时在强化学习阶段使用混合数据进行 GRPO。结果如表3所示。我们发现，S4取得了最佳演奏，域内语音准确率最高，达到 99.75%，平均准确率也达到最佳的 87.38 %。此外，我们观察到，在冷启动阶段注入没有思考标注的样本，会显著降低监督微调期间的词元级准确率，表明模型对期望的结构化推理格式遵循得更差，这可能会阻碍后续的推理式学习。

5.3 与最先进模型的比较

完成消融研究后，我们将经过语音训练的 FT-GRPO-Qwen2.5-Omni-3B 与现有的最先进方法进行比较。如表 4 所示，我们的 FT-

模型	语音、声音、歌唱			音乐可解释性		AVG
W2V2-AASIST (Tak等人, 2022)	99.58	37.17	59.08	58.23	×	63.52
WPT-W2V2-AASIST (Xie等人,2025)	99.06	46.37	38.56	92.87	×	69.22
ALLM4ADD (Gu 等人, 2025)	99.43	72.00	80.85	**91.09**	×	85.84
FT-GRPO	**99.75**	**75.05**	**84.26**	90.44	✓	**87.38**

表4：与最先进方法的演奏性能比较。

训练类型	方法	语音	声音	歌唱	音乐	AVG
语音	SFT	99.04	73.31	66.29	91.16	82.45
语音	FT-GRPO	**99.75**(+0.99)	**75.05**(+1.74)	**84.26**(+17.97)	90.44(-0.72)	**87.38**(+5.00)
声音	SFT	86.70	67.55	77.99	93.58	81.46
声音	FT-GRPO	**89.68**(+2.98)	**75.00**(+7.45)	**85.34**(+7.35)	91.21(-2.37)	**85.31**(+3.85)
歌唱	SFT	39.78	67.13	83.97	73.39	66.07
歌唱	FT-GRPO	**57.98**(+18.20)	**75.02**(+7.89)	**86.58**(+2.61)	**90.58**(+17.19)	**77.54**(+11.47)
音乐	SFT	**83.72**	**73.40**	23.06	77.34	64.38
音乐	FT-GRPO	60.54(-23.18)	59.11(-14.29)	**60.49**(+37.43)	**83.57**(+6.23)	**65.93**(+1.55)
协同训练	SFT	98.55	68.17	79.73	93.34	84.95
协同训练	FT-GRPO	**98.74**(+0.19)	**77.15**(+8.98)	**86.16**(+6.43)	**98.33**(+4.99)	**90.10**(+5.15)

表5：SFT 与 FT-GRPO 在所有音频类型上的对比（%）。

GRPO 始终优于以往的小模型基线（例如 W2V2-AASIST 和 WPT-W2V2-AASIST）以及此前的大模型方法。除了可解释性有所提升外，FT-GRPO 的语音准确率达到 99.75%，平均准确率也最高，达到 87.38%。值得注意的是，小模型报告的准确率是在固定决策阈值 0.5 下得到的，而 ALLM4ADD 则通过将模型响应与真实标签直接匹配来进行评估。这也表明小模型在实践中存在局限：选择合适的阈值——尤其是在域外条件下——可能并非易事。

5.4 全类型音频深度伪造检测任务的 FT-GRPO

在本节中，我们将 FT-GRPO 从语音训练扩展到全类型音频深度伪造检测任务。结果见表 5。在所有音频类型上，FT-GRPO 的表现始终优于仅监督微调训练。具体而言，在语音上训练时，FT-GRPO 将平均准确率提高了 +5.00%，其中歌唱上的提升尤为显著（+17.97%）。在声音上训练时，FT-GRPO 将平均准确率提高了 +3.85%。在歌唱上训练时，FT-GRPO 的整体提升最大，平均准确率提高了 +11.47%，并分别将语音和音乐的准确率提高了 +18.20% 和 +17.19%。相比之下，在音乐上训练带来的平均准确率提升较小（+1.55%。这主要是因为歌唱（+37.43%）和音乐（+6.23%）上的提升，部分被语音（−23.18%）和声音（−14.29%）上的表现下降所抵消，这表明仅在音乐上训练时跨类型泛化能力较弱。最后，在协同训练设置下，协同训练的 FT-GRPO 表现最佳，平均准确率达到 90.10%（比 SFT 高 +5.15% ）。这一提升主要得益于声音（+8.98%）和音乐（+4.99%）上的大幅进步，同时歌唱（+6.43%）和语音（+0.19%）上的表现也稳步提升。

6 结论

我们从音频大语言模型的视角研究全类型音频深度伪造检测任务，旨在兼顾强大的跨类型性能与可解释决策。我们首先建立了一个全类型ADD基准，并对具有代表性的音频大语言模型开展系统评估。为提升音频大语言模型的可解释性，我们提出了一种自动标注与清洗流程，可大规模生成微调思维链推理依据，最终得到34万条高质量实例。基于这些数据，我们提出FT-GRPO，这是一种两阶段强化微调框架：先进行SFT冷启动，再在复合奖励机制下进行GRPO，以奖励正确性、格式合规性和完整的FT推理，同时纳入音频大语言模型往往表现不佳的非思考样本。实验表明，FT-GRPO使用一个3B音频大语言模型便达到了最先进准确率，并通过协同训练进一步提升了全类型性能。

局限性

尽管本研究表明，音频大语言模型能够在全类型音频深度伪造检测任务上达到最先进水平，但仍有若干局限有待解决。首先，我们的实验仅涵盖有限数量的音频大语言模型，且主要来自Qwen-Audio/Omni系列，因此研究结论未必能充分推广至其他架构。其次，尽管我们的推理依据通过包含一致性检查与清洗的两阶段自动标注与润色流程生成，但受当前音频大语言模型音频理解能力有限的影响，部分标注可能仍不够理想。第三，尽管我们在四个广泛使用、覆盖主要音频类型的公开数据集上进行了评估，测试条件仍不全面，例如声学环境、合成方法、部分伪造音频或混合音频类型场景。未来工作将重点解决这些局限。

参考文献

Andrea Agostinelli、Timo I Denk、Zalán Borsos、Jesse Engel、Mauro Verzetti、Antoine Caillon、Qingqing Huang、Aren Jansen、Adam Roberts、Marco Tagliasacchi 等 2 人。2023。《Musi-clm：根据文本生成音乐》。《*arXiv* 预印本 *arXiv:2301.11325*》。

Bowen Baker、Joost Huizinga、Leo Gao、Zehao Dou、Melody Y Guan、Aleksander Madry、Wojciech Zaremba、Jakub Pachocki 和 David Farhi。2025。《监控推理模型的不当行为及推动混淆所带来的风险》。《*arXiv*预印本 *arXiv:2503.11926*》。

Sanyuan Chen、Chengyi Wang、Zhengyang Chen、Yu Wu、Shujie Liu、Zhuo Chen、Jinyu Li、Naoyuki Kanda、Takuya Yoshioka、Xiong Xiao 等 2 人。2022。Wavlm：面向全栈语音处理的大规模自监督预训练。《*IEEE*信号处理选题期刊》，16(6):1505–1518。

Shuang Chen、Yue Guo、Zhaochen Su、Yafu Li、Yu- lun Wu、Jiacheng Chen、Jiayu Chen、Weijie Wang、Xiaoye Qu 和 Yu Cheng。2025。推进多模态推理：从优化的冷启动到分阶段强化学习。*arXiv*预印本 *arXiv:2506.04207*。

Xuanjun Chen、Haibin Wu、Roger Jang 和 Hung-yi Lee。2024。用于 SingFake 检测的歌声图建模。发表于*Interspeech 2024* 论文集，第 4843–4847 页。

Luca Comanducci、Paolo Bestagini 和 Stefano Tubaro。2024年。FakeMusicCaps：用于检测和归因通过文本到音乐模型生成的合成音乐的数据集。*arXiv* 预印本 *arXiv:2409.10684*。

Yassir Fathullah、Chunyang Wu、Egor Lakomkin、Jun- teng Jia、Yuan Shangguan、Ke Li、Jinxi Guo、Wenhan Xiong、Jay Mahadeokar、Ozlem Kalinli 等。2024年。利用语音识别能力提示大语言模型。载于*ICASSP 2024-2024 IEEE*声学、语音与信号处理国际会议*(ICASSP)*，第13351–13355页。IEEE。

Jort F Gemmeke、Daniel PW Ellis、Dylan Freedman、Aren Jansen、Wade Lawrence、R Channing Moore、Manoj Plakal 和 Marvin Ritter。2017年。AudioSet：一种用于音频事件的本体和人工标注数据集。发表于*2017*年 *IEEE* 声学、语音与信号处理国际会议（*ICASSP*），第 776–780 页。IEEE。

Mahyar Gohari、Davide Salvi、Paolo Bestagini 和 Nicola Adami。2025。用于歌声深度伪造检测的音频特征研究。载于*ICASSP 2025- 2025 IEEE*声学、语音与信号处理国际会议 *(ICASSP)*，第 1–5 页。IEEE。

Hao Gu、Jiangyan Yi、Chenglong Wang、Jianhua Tao、Zheng Lian、Jiayi He、Yong Ren、Yujie Chen 和 Zhengqi Wen。2025。ALLM4ADD：释放音频大语言模型在音频深度伪造检测中的潜力。载于第 *33* 届 *ACM*多媒体国际会议论文集，第 11736–11745 页。

Daya Guo、Dejian Yang、Haowei Zhang、J unxiao Song、Ruoyu Zhang、Runxin Xu、Qihao Zhu、Shirong Ma、Peiyi Wang、Xiao Bi 等。2025a。DeepSeek-R1：通过强化学习激励大语言模型的推理能力。*arXiv* 预印本 *arXiv:2501.12948*。

Xiaoxuan Guo、Hengyan Huang、Jiayi Zhou、Renhe Sun、Jian Liu、Haonan Cheng、Long Ye 和 Qin Zhang。2025b。Envsslam-ffn：面向 ESDD 2026挑战赛的轻量级层融合系统。*arXiv* 预印本 *arXiv:2512.20369*。

Anmol Guragain、Tianchi Liu、Zihan Pan、Hardik B. Sailor 和 Qiongqiong Wang。2024年。用于受控歌声深度伪造检测（ctrsvdd）2024挑战赛的语音基础模型集成。载于*2024 IEEE* 语音语言技术研讨会，第 774–781 页。

Edward JHu、yelong shen、Phillip Wallis、ZeyuanAllen-Zhu、Yuanzhi Li、Shean Wang、Lu Wang 和 Weizhu Chen。2022。LoRA：大语言模型的低秩适配。载于国际学习表征会议。

Jee-weon Jung、Hee-Soo Heo、Hemlata Tak、Hye-jin Shim、Joon Son Chung、Bong-Jin Lee、Ha-Jin Yu 和 Nicholas Evans。2022。AASIST：采用集成时频图注意力网络的音频反欺骗。载于 *ICASSP* 会议论文集，第 6367–6371 页。

Piotr Kawa、Marcin Plata 和 Piotr Syga。2023。针对音频深度伪造检测的对抗攻击防御。载于 *Interspeech 2023* 会议论文集，第 5276–5280 页。

Yassine El Kheir、Youness Samih、Suraj Maharjan、Tim Polzehl 和 Sebastian Möller。2025年。《用于音频深度伪造检测的自监督学习模型全面逐层分析》。*arXiv*预印本 *arXiv:2502.03559*。

Kai Li、Can Shen、Yile Liu、Jirui Han、Kelong Zheng、Xuechao Zou、Zhe Wang、Shun Zhang、Xingjian Du、Hanjun Luo 等。2025。《Au-diotrust：音频大语言模型多维可信度基准测试》。*arXiv*预印本 *arXiv:2505.16211*。

Yizhi Li、Ruibin Yuan、Ge Zhang、Yinghao Ma、Xingran Chen、Hanzhi Yin、Chenghao Xiao、Chenghua Lin、Anton Ragni、Emmanouil Benetos 等。2024a。《MERT：采用大规模自监督训练的声学音乐理解模型》。载于 *ICLR*。

Yupei Li、Manuel Milling、Lucia Specia 和 Björn W Schuller。2024b。《从音频深度伪造检测到 AI 生成音乐检测：路径与概述》。*arXiv* 预印本 *arXiv:2412.00571*。

Yupei Li、Qiyang Sun、Hanqian Li、Lucia Specia 和 Björn W Schuller。2024c。《借助可解释性检测机器生成音乐：挑战与早期基准测试》。*arXiv* 预印本 *arXiv:2412.13421*。

Xuechen Liu、Xin Wang、Md Sahidullah、Jose Patino、Héctor Delgado、Tomi Kinnunen、Massimiliano Todisco、Junichi Yamagishi、Nicholas Evans、An- dreas Nautsch 等 2 人。2023。《ASVspoof 2021：迈向真实环境中的欺骗性语音与深度伪造语音检测》。*IEEE/ACM* 音频、语音与语言处理汇刊。

ZiyuLiu、Zeyi Sun、Yuhang Zang、Xiaoyi Dong、Yuhang Cao、Haodong Duan、Dahua Lin 和 Jiaqi Wang。2025年。Visual-RFT：视觉强化微调。*arXiv* 预印本 *arXiv:2503.01785*。

Ziyang Ma、Yinghao Ma、Yanqiao Zhu、Chen Yang、Yi-Wen Chao、Ruiyang Xu、Wenxi Chen、Yuanzhe Chen、Zhuo Chen、Jian Cong 等人。2025。MMAR：一个用于语音、音频、音乐及其混合内容深度推理的高难度基准。*arXiv* 预印本 *arXiv:2505.13032*。

Zihan Pan、Tianchi Liu、Hardik B Sailor 和 Qiongqiong Wang。2024。对预训练语音模型的隐藏嵌入进行注意力融合，以实现反欺骗检测。发表于*Interspeech 2024* 论文集，第 2090–2094 页。

Orchid Chetia Phukan、Gautam Siddharth Kashyap、Arun Balaji Buduru 和 Rajesh Sharma。2024年。同质性还是异质性：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*。

Hemlata Tak、Massimiliano Todisco、Xin Wang、Jee-weon Jung、Junichi Yamagishi 和 Nicholas Evans。2022年。使用 wav2vec 2.0 和数据增强进行自动说话人验证欺骗与深度伪造检测。发表于说话人与语言识别研讨会（*Odyssey 2022*）。ISCA。Jee-weon Tak、H.、J. Patino、M. Kamble、M. Todisco 和 N. Evans。2021。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。*arXiv*预印本 *arXiv:2107.12710*。

Changli Tang、Wenyi Yu、Guangzhi Sun、Xianzhao Chen、Tian Tan、Wei Li、Lu Lu、Zejun Ma 和 Chao Zhang。2024年。扩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*），第 11236–11240 页。IEEE。

Massimiliano Todisco、Xin Wang、Ville Vestman、Md. Sahidullah、Héctor Delgado、Andreas Nautsch、Junichi Yamagishi、Nicholas Evans、Tomi H. Kinnunen 和 Kong Aik Lee。2019 年。ASVspoof 2019：欺骗与伪造音频检测的未来前景。载于*Interspeech 2019* 论文集，第 1008–1012 页。

Hui Wang、Jinghua Zhao、Yifan Yang、Shujie Liu、Junyang Chen、Yanzhe Zhang、Shiwan Zhao、Jinyu Li、Jiaming Zhou、Haoqin Sun 等 1 人。2025a。SpeechLLM-as-Judges：迈向通用且可解释的语音质量评估。*arXiv* 预印本 *arXiv:2510.14664*。

Siyin Wang、Wenyi Yu、Xianzhao Chen、Xiaohai Tian、Jun Zhang、Lu Lu、Yu Tsao、Junichi Yamagishi、Yux- uan Wang 和 Chao Zhang。2025b。QualiSpeech：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*。

Xin Wang、Héctor Delgado、Hemlata Tak、Jee-weon Jung、Hye-jin Shim、Massimiliano Todisco、Ivan Kukanov、Xuechen Liu、Md Sahidullah、Tomi Kin- nunen、Nicholas Evans、Kong Aik Lee 和 Ju- nichi Yamagishi。2024年。ASVspoof 5：大规模众包语音数据、深度伪造和对抗攻击。载于*ASVspoof 2024*研讨会（已录用）。

Zhiyong Wang、Ruibo Fu、Zhengqi Wen、Jianhua Tao、Xiaopeng Wang、Yuankun Xie、Xin Qi、Shuchen Shi、Yi Lu、Yukun Liu 等 1 人。2025c。使用冻结的 wav2vec 2.0 进行伪造音频检测的专家混合融合。载于*ICASSP 2025-2025 IEEE* 国际声学、语音与信号处理会议（*ICASSP*），第 1–5 页。IEEE。

Lai Wei、Yuting Li、Kaipeng Zheng、Chen Wang、Yue Wang、Linghe Kong、LichaoSun 和 Weiran Huang。2025a。通过冷启动强化学习推进多模态推理。*arXiv* 预印本 *arXiv:2505.22334*。

Zhaolin Wei、Dengpan Ye、Jiacheng Deng 和 Yuhan Lin。2025b。从人声到节拍：增强音乐

通过识别背景中的伪造内容进行深度伪造检测。在*ICASSP 2025-2025 IEEE* 国际声学、语音与信号处理会议（*ICASSP*），第 1–5 页。IEEE。

Yuankun Xie、Ruibo Fu、Zhiyong Wang、Xiaopeng Wang、Songjun Cao、Long Ma、Haonan Cheng 和 Long Ye。2025年。检测所有类型的深度伪造音频：通过小波提示调优增强听觉感知。*arXiv*预印本 *arXiv:2504.06753*。

Yuankun Xie、Jingjing Zhou、Xiaolin Lu、Zhenghao Jiang、Yuxin Yang、Haonan Cheng 和 Long Ye。2024a。FSD：用于歌曲伪造检测的首个中文数据集。载于 *ICASSP 2024-2024*年 *IEEE* 国际声学、语音与信号处理会议 *(ICASSP)*，第 4605–4609 页。IEEE。

Zeyu Xie、Baihan Li、Xuenan Xu、Zheng Liang、Kai Yu 和 Mengyue Wu。2024b。FakeSound：通用音频深度伪造检测。载于*Interspeech 2024* 论文集，第 112–116 页。

Canxiang Yan、Chunxiang Jin、Dawei Huang、Haib- i ng Yu、Han Peng、Hui Zhan、Jie Gao、Jing Peng、Jingdong Chen、Jun Zhou及其他1位作者。2025年。Ming- uniaudio：用于联合理解、生成和编辑的统一表征语音大语言模型。*arXiv* 预印本 *arXiv:2511.05516*。

Chih-Kai Yang、Neo S Ho和Hung-yi Lee。2025年。迈向大型音频语言模型的全面评估：综合综述。*arXiv* 预印本 *arXiv:2505.15957*。

Han Yin、Yang Xiao、Rohan Kumar Das、Jisheng Bai、Haohe Liu、Wenwu Wang和Mark D Plumb- ley。2025年。EnvSDD：环境声音深度伪造检测基准。载于*Interspeech 2025*，第201–205页。

Yongyi Zang、Jiatong Shi、You Zhang、Ryuichi Yamamoto、Jionghao Han、Yuxun Tang、Shengyuan Xu、Wenxiao Zhao、Jing Guo、Tomoki Toda及其他1位作者。2024年。Ctrsvdd：受控歌声深度伪造检测的基准数据集与基线分析。*arXiv* 预印本 *arXiv:2406.02438*。

Dong Zhang、Xin Zhang、Jun Zhan、Shimin Li、Yaqian Zhou 和 Xipeng Qiu。2024a。Speechgpt-gen：扩展信息链语音生成。*arXiv* 预印本 *arXiv:2401.13527*。

Qishan Zhang、Shuangbing Wen 和 Tao Hu。2024b。基于自监督 xls-r 和 sls 分类器的音频深度伪造检测。发表于 *ACMMultimedia2024*。

Qishan Zhang、Shuangbing Wen、Fangke Yan、Tao Hu 和 Jun Li。2024c。Xwsb：一种用于 svdd 2024 挑战赛的融合系统，结合了 xls-r 和 WavLM，并采用 sls 分类器检测系统。载于*2024 IEEE* 语音语言技术研讨会 *(SLT)*，第 788–794 页。IEEE。

You Zhang、Yongyi Zang、Jiatong Shi、Ryuichi Yamamoto、Tomoki Toda 和 Zhiyao Duan。2024d。Svdd 2024：首届歌声深度伪造检测挑战赛。载于*2024 IEEE* 语音语言技术研讨会 *(SLT)*，第 782–787 页。IEEE。

Zirui Zhang、Wei Hao、Aroon Sankoh、William Lin、Emanuel Mendiola-Ortiz、Junfeng Yang 和 Chengzhi Mao。2025。我能听见你：用于深度伪造音频检测的选择性鲁棒训练。发表于第十三届国际学习表征会议。

Yuze Zhao、Jintao Huang、Jinghan Hu、Xingjun Wang、YunlinMao、Daoze Zhang、ZeyinziJiang、ZhikaiWu、Baole Ai、Ang Wang 等 1 人。2025。Swift：用于微调的可扩展轻量级基础设施。发表于*AAAI* 人工智能会议论文集，第 39 卷，第 29733–29735 页。

Xie Zhifei、Mingbao Lin、Zihang Liu、Pengcheng Wu、Shuicheng Yan 和 Chunyan Miao。2025 年。Audio- reasoner：提升大型音频语言模型的推理能力。载于*2025* 年自然语言处理实证方法会议论文集，第 23840–23862 页，中国苏州。计算语言学协会。

Hengguang Zhou、Xirui Li、Ruochen Wang、Minhao Cheng、Tianyi Zhou 和 Cho-Jui Hsieh。2025。R1-zero 在一个 2B 非监督微调模型上的视觉推理中迎来了“顿悟时刻”。*arXiv* 预印本 *arXiv:2503.05132*。

A 数据构建详情

如图 6 所示，我们展示了数据构建中使用的提示词。具体而言，我们将所提出的标注流程应用于 19LA、ESDD、CtrSVDD 和 FakeMusicCaps 的训练集和开发集划分。标注完成后，每种音频类型都被划分为思考子集和非思考子集，汇总见表 6。

由此得到的非思考率也可作为衡量 Qwen-Omni-Captioner-30B 能否在不同音频类型中可靠地产生连贯 FT推理依据的替代指标：对于以人声内容为主的语音和歌声，该比率仍然较低（约为 8%）；但对于环境声音和音乐，该比率显著上升，分别达到 14.44% 和 37.47%。更广泛地说，这凸显了当前音频大语言模型存在的能力不均衡：它们在理解以人声为中心的内容方面似乎更强，但在捕捉对真实性判断至关重要的细粒度声学线索和多样化非人声声音模式方面，可靠性却较低。

B 训练超参数详情

本节总结了我们在实验中使用的训练超参数。我们首先研究学习率和 LoRA 配置，然后比较冻结与微调不同模块的效果。最后，我们报告两阶段 FT-GRPO 流程的训练提示词和完整超参数设置，包括跨类型训练和协同训练的训练细节。

B.1 学习率和 LoRA 配置

我们首先在 Qwen2.5-Omni-3B 上确定优化超参数，采用完全微调设置，即音频编码器、对齐器和大语言模型均可训练。表7 报告了常用学习率以及两种具有代表性的 LoRA 配置下的 ID 准确率：$(r{=}8, \alpha{=}32)$，这是 `ms-swift` 和许多大语言模型微调框架的默认配置；以及 $(r{=}64, \alpha{=}16)$，Gu 等人报告其表现出色（Gu 等人，2025）。总体而言，较大的 LoRA 秩往往能为音频深度伪造检测带来更高的准确率，这表明更大的适配器容量有助于建模音频特定伪影。同时，学习率仍然至关重要：学习率过大会降低演奏，而适中的数值（1×10^{-5}）在我们的研究中取得了最佳结果。除非另有说明，我们均使用 LoRA $(r{=}64, \alpha{=}16)$，学习率为 1×10^{-5}，用于以下消融实验。

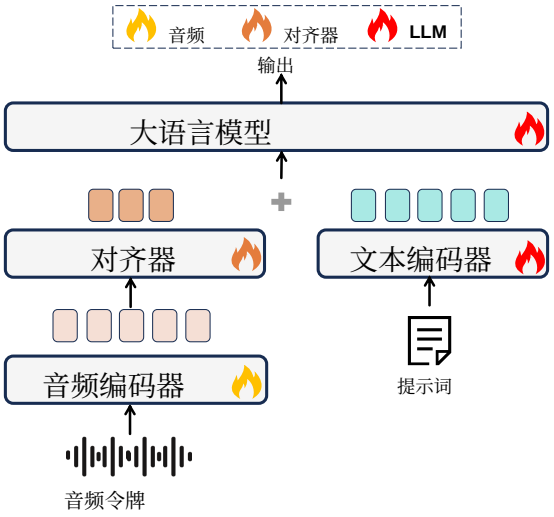

图 4：音频大语言模型中的三个可微调模块。

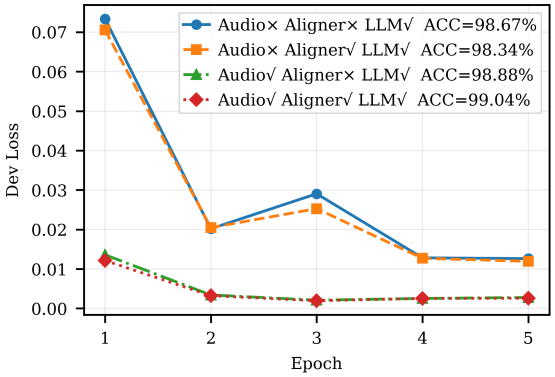

图 5：语音训练的 Qwen2.5-Omni-3B 的四种训练模式。✓ 表示可训练，× 表示冻结。准确率在 19LA 评估集上测得。

B.2 冻结与微调

如正文所述，我们采用 LoRA 和全层可训练设置。实际上，若干模块级开关仍可配置。图 4 展示了典型的音频大语言模型架构。我们使用 `ms-swift` 作为微调框架，并以 Qwen2.5-Omni 为例。该框架提供三个冻结开关：冻结_视觉编码器、冻结_对齐器和冻结_大语言模型。具体而言，冻结_视觉编码器 用于控制是否更新音频/视觉编码器；冻结_对齐器用于控制是否更新模态到词元投影（即将最后一层编码器特征映射到输入词元的对齐模块）；而冻结_大语言模型 用于控制是否更新大语言模型主干网络。

Type	训练集		Dev		非思考率（%）
	思考	非思考	思考	非思考	
语音	23,426	1,954	23,049	1,795	8.06
声音	121,544	17,511	34,817	4,893	14.44
歌唱	77,247	7,157	40,256	3,369	8.06
音乐	15,958	4,903	4,636	1,422	37.47

表6：训练集和开发集划分中思考与非思考样本的统计信息。

lr	r	α	准确率（%）
5×10^{-5}	8	32	91.70
1×10^{-5}	8	32	97.04
5×10^{-6}	8	32	98.70
5×10^{-5}	64	16	93.48
1×10^{-5}	64	16	**99.04**
5×10^{-6}	64	16	98.71

表7：学习率（lr）、LoRA 秩 r、LoRA alpha α 的超参数研究。在此表中，音频编码器、对齐器和大语言模型均经过微调。

阶段1（SFT冷启动）	设置
基础模型	Qwen2.5-Omni-3B
精度	bfloat16
PEFT	LoRA
LoRA 秩 / alpha	64 / 16
最大序列长度	768
训练轮数	3
学习率	1×10^{-5}
权重衰减	0.1
预热比例	0.01
每设备训练批次大小	16
梯度累积	1
数据加载器工作线程	8
分布式训练	DeepSpeed ZeRO-2

表8：FT-GRPO 步骤1的超参数。

使用选定的超参数（LoRA(r=64, α=16) 和学习率 1×10^{-5}），我们在 19LA 训练/测试协议下比较了四种训练变体，并监测开发集损失，如图 5所示。根据音频编码器是否冻结，我们观察到两种明显不同的损失区间。开发集损失和评估准确率均表明，微调音频编码器的效果显著优于冻结音频编码器。此外，与保持对齐器冻结相比，启用对齐器能进一步提升演奏表现。三个组件全部经过微调时，取得了最佳结果。

第二阶段（GRPO）	设置
强化学习方法	GRPO（rlhf_type=grpo）
初始检查点	监督微调检查点
精度	bfloat16
PEFT	LoRA
LoRA 秩 / alpha	64 / 16
最大提示长度	768
最大补全长度	768
训练轮数	2
学习率	1×10^{-5}
权重衰减	0.1
预热比例	0.05
每设备训练批次大小	32
梯度累积	1
组大小 G	8（生成数量_）
奖励	准确率 +格式 +微调
奖励权重	[1, 0.1, 0.1]
数据加载器工作线程	8
分布式训练	DeepSpeed ZeRO-2

表9：FT-GRPO 步骤2的超参数。

B.3 训练提示

与数据构建阶段类似，在所有实验中，我们都使用由 FT 引导的提示词，引导音频大语言模型同时依据频率域和时间域线索进行推理。对于真实/伪造 SFT 实验，我们对提示词进行了略微简化，因为不要求严格的输出格式，如图 6所示。对于 FT-GRPO，SFT 阶段和 GRPO 阶段均使用与数据构建阶段相同的系统提示词；唯一的变化是将用户提示替换为：“您能否验证这段音频是真实的还是伪造的？”

B.4 FT-GRPO超参数

我们在表中列出了语音训练的FT-GRPO在步骤1（SFT冷启动）和步骤2（GRPO）中的超参数 8 和表9。对于其他音频类型，除训练轮数外，所有超参数均保持不变：步骤1中，语音和音乐使用3轮训练，声音和歌唱使用2轮训练；步骤2中，所有四种类型均使用2轮GRPO训练。对于协同训练设置，由于数据规模大得多，冷启动使用2轮训练，GRPO使用1轮训练。

P数据构建步骤1：原始音频的提示词详情 说明文字

系统提示词

你是音频深度伪造检测任务方面的专家。给定音频是一段 [语音/环境音/歌声/音乐] 片段。伪造音频完全由深度伪造合成方法生成，而非通过手动拼接或数字编辑制作。你可以通过从频域和时域两个维度分析音频频谱图，检测伪造音频中的伪影。

在推理时，请指出伪影出现的位置，以帮助区分真实录音与深度伪造。请简要概述思维过程，只总结最重要的要点。在 <思考> 中，请按以下结构组织推理过程：

- 首先， <频域> 描述有效频率范围内的频谱特征和任何不自然的伪影，并重点关注高频。
- 接着， <时域>描述时间连贯性和任何不自然的伪影，尤其是出现在特定时间区间内的伪影，并特别关注静音或低能量片段。

推理过程和答案分别包含在 <think> </think>和 <answer> </answer> 标签中，即 <think> 此处填写推理过程 </think><answer> 此处填写答案 </answer>。

格式示例（必须严格遵循此结构）：<think><频域>[在此处填写分析]<时域>[在此处填写分析]</think><answer>[真实/伪造]</answer>

用户提示

<音频>这是一个 [真实/伪造]音频。请在 <think> </think> 标签内提供你的思维过程，并在最终答案 nal r（真实/伪造）位于 <答案> </答案> 标签内。

数据构建步骤2的提示词详情：标题润色

系统提示词

Y 我们的任 kis poli 在 <思考> 中补充原因内容，以便更好地突出关键证据，并判断 <思考> 部分中的真实/伪造描述是否与 <答案> 部分中的真实/伪造标签一致。

重要说明：。

1 润色原因：仅保留与两个维度的真实性判断相关的描述。由于没有涉及任何编辑，因此删除所有与编辑相关的思考描述；2. 验证与标签的一致性：如果用户消息指出“这是一个真实/伪造音频”，但描述或最终答案与之矛盾，则输出“ 不匹配”。3. 验证内部一致性：如果‘<思考>’中的描述所提供的证据与最终答案矛盾，则输出“不匹配”。

格式示例（必须严格遵循此结构）：<思考><频域>[此处填写你的分析]<时域>[此处填写你的分析]</思考><答案>[真实/伪造]</答案>

用户提示

T这是对一个音频样本的真实性描述。请严格遵循重要说明进行润色 it.

SFT 步骤中的训练提示词

系统提示词

作为音频深度伪造检测领域的人工智能专家，你可以识别真实音频和深度伪造音频。通过沿频率和时间两个关键维度分析频谱图，你可以检测伪影，从而有效 y 区分真实音频与深度伪造音频。

用户提示

<音频>你能验证这段音频是真是假吗?

图6：数据构建和 SFT 步骤的提示词详情。

真实示例（语音）

文件： LA_T_1078395.flac
<思考>
[频域] 该音频呈现出自然、完整的频谱特征，在低频、中频和高频范围内均清晰可见，符合真实录音的特征。未观察到通常与深度伪造合成相关的高频伪影或失真。

[时域]整个音频中的人声节奏、语调和特征始终稳定且自然。未检测到突发异常、不自然的停顿，或表明存在深度伪造生成的语音模式不一致现象。

</思考>
<答案>真实</答案>

伪造示例（语音）

文件： LA_T_7794385.flac
<思考>
[频域] 这段音频在中高频和高频部分包含不自然的高频伪影，与自然的人类语音不符。这些伪影呈现出合成频谱包络，缺乏真实语音录音中常见的平滑度和自然变化，表明其可能由深度伪造合成而成。

[时域] 这段音频存在时间异常，例如“What”之后突然停顿，以及“in the game”等短语中不自然的韵律。这些异常反映的是算法生成模式，而非自然的语音节奏，进一步表明其来源为合成。

</think><答案>伪造</answer>

图7：语音标注示例。

真实示例（声音）

文件： EnvSDD/dev/real_audio/TUTSED2017Dev/a131_14.wav
<思考>
<频域> 该音频呈现出持续稳定的低频嗡鸣，这是室内环境录音中附近有电子设备时的典型特征。其频谱特征随时间保持稳定，频带能量没有突然变化。未观察到合成伪影（例如频谱间隙、异常谐波或数字削波），表明其频率响应平滑且自然。
<时域> 该信号保持着稳定的背景嗡鸣，没有突然的振幅变化或不自然的停顿。未检测到瞬态数字伪影（咔哒声／爆裂声），且整个片段的时间连续性良好。没有编辑痕迹，加上噪声自然平稳，说明该录音真实可信。</think>
<答案>真实</答案>

伪造示例（声音）

文件： EnvSDD/dev/fake_audio/TTA/audioldm1/TUTASC2019Dev/airport-prague-1173-45640-a_0.wav<思考>
<频域> 该音频表现出异常的高频能量和金属音色，频谱纹理过于均匀，缺乏真实录音中通常具有的自然频谱复杂度。这种频域规律性与音频合成模型引入的伪影相一致。

<时域>该信号呈现出突兀的起始/结束，没有自然的淡入/淡出或混响衰减。音高和振幅保持异常稳定，背景噪声也缺乏真实的波动，表明环境连续性较弱。这些时域特征是人工智能生成音频的典型表现。</思考>

<答案>伪造</答案>

图8：声音标注示例。

真实示例（歌唱）

文件： CtrSVDD_0010_T_0008078.flac
<思考>
<频域> 音频呈现出自然的声乐谐波和平滑的频谱分布，没有突兀或合成的高频截止。能量均匀分布在整个频率范围内，不存在通常与深度伪造生成相关的人工噪声模式。

<时域> 在整个短语“tenden mushimushi katatsumuri”中，时间连续性得以保持，音素之间过渡流畅，呼吸控制稳定。韵律和时间安排符合人类语音生成的特点，没有算法操纵的迹象。</thi
nk>
<答案>真实</答案>

伪造示例（歌唱）

文件： CtrSVDD_0055_T_0019326.flac
<思考>
<频域> 该音频在咝音辅音和元音过渡处呈现出不自然的高频伪影，频谱包络过于平滑且受到压缩，与自然发声不符。持续的电子嗡鸣也进一步表明这是合成生成的音频。

<时域> 音节过渡的时序不自然，尤其是被拉长的末尾音节 ỳuan，¨缺乏人类的呼吸控制和颤音调制。音节之间的人工停顿和微弱的静电噪声，都符合深度伪造合成的特征。

</思考>
<答案>假</答案>

图9：歌声标注示例。

真实示例（音乐）

文件： lePP2BaZhJU.wav
<思考>
<频域> 音频呈现出音乐录音典型的完整频率范围，低频、中频和高频清晰且分明。没有混叠、量化噪声或频谱失真等数字伪影，表明其保真度很高，且不存在不自然的频率失衡。

<时域> 音频始终保持稳定的节奏和速度，各音乐片段的能量一致，没有突变、信号中断或不自然的静音。音乐元素之间的过渡平滑自然，体现出真实的演奏，而非合成生成。

</think>
<答案>真实</答案>

伪造示例（音乐）

文件： TTM02_7jWRIjFaoeU.wav
<思考>
<频域> 这段音频呈现出不自然的高频伪影，其特点是刺耳的削波质感，且缺乏自然的谐波泛音。这些特征表明它是深度伪造合成音频，而非真实的声学录音。失真的贝斯和鼓声主导着低频频谱，而钹声和吉他失真则呈现出与自然放大不符的合成高频分量。

<时域>节奏精确性和动态压缩缺乏变化，表明其由算法生成。音乐段落之间的过渡时机机械刻板，缺少人类演奏所特有的微时值细微差异。此外，音乐以突兀的截断结束，缺少自然衰减，这是深度伪造音频合成中常见的伪影。

</思考>
<答案>伪造</答案>

图10：音乐标注示例。