arXiv:2603.02914v1 [eess.AS] 3 Mar 2026

强化学习微调能否提升二元语音深度伪造检测的泛化能力?

Xin Wang[1], *Wanying Ge*[1] *Junichi Yamagishi*[1]

[1] 日本国立信息学研究所

wangxin@nii.ac.jp

摘要

构建对未见攻击具有泛化能力的语音深度伪造检测模型仍然是一个具有挑战性的问题。尽管该领域已转向采用语音基础模型的预训练与微调范式，但大多数方法仅依赖监督微调（SFT）。受大型语言模型领域使用强化学习（RL）进行模型微调的启发，我们研究了强化学习的影响，具体而言是组相对策略优化（GRPO）。使用多个检测器和测试集开展的实验结果表明，基于纯 GRPO 的微调能够提升域外测试集上的性能，同时保持目标域测试数据上的性能。该方法优于仅使用 SFT 的方案以及混合方案。我们的消融研究进一步表明，GRPO 中的负奖励可能是这一改进的关键因素。

索引词：语音反欺骗、深度伪造检测、强化学习

1. 引言

语音深度伪造检测是一种重要的对策，甚至可以说是应对语音深度伪造的终极对策 [1]。按照惯例，它被定义为一个二分类任务——给定一条输入语句，语音深度伪造检测模型需要判断该语句是否由真实人类说话者发出 [2]。其目标是开发能够泛化到未见深度伪造类型、未见说话者的数据等情形的检测模型。

过去十年来，使用包含真实与伪造数据的精选数据集训练检测模型一直是常见做法 [3], [4]。然而，该领域日益兴起的趋势是采用预训练加微调的框架 [5], [6], [7]。预训练阶段 [8], 通常由计算能力和数据资源充足的团队完成，他们使用海量且多样化的真实数据以及自监督学习（SSL）准则来构建语音基础模型（例如 wav2vec 2.0 [9]）。随后，使用少量数据，通过监督微调（SFT）对基础模型进行微调，以用于语音深度伪造检测。采用这种两阶段方法构建的模型，在近期的基准测试中展现了顶尖性能 [10], [11]。预训练与微调框架最近
扩展后在中间加入了后训练阶段 [12],，如图1所示。后训练利用来自不同领域的大量真实、伪造和模拟伪造数据（例如声码合成数据 [12]），对通用基础模型进行适配；其前提是适配后的模型能够生成更适合深度伪造检测、来源追踪 [13], 及其他相关下游任务的表征。

显然，上述多阶段训练流程与大型语言模型（大语言模型） [14] 的流程相似。一个大语言模型

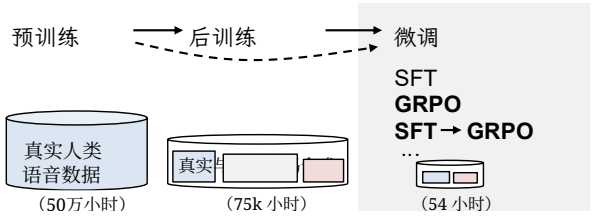


图1： 语音深度伪造检测模型多阶段训练流程示意图。本文研究了阴影区域中的不同微调算法。

也会经历预训练以学习通用语言模式，并可选择进行中期训练 [15], , 随后通过微调使大语言模型与人类偏好或特定任务对齐。然而，两个领域共同存在的一个问题是，针对特定任务或数据领域进行微调的模型容易在其他任务或数据领域上出现性能下降，这也被称为灾难性遗忘 [16]。大语言模型领域探索了多种微调设置，例如将监督微调与强化学习（RL）相结合 [14], , 或仅通过组相对策略优化（GRPO）进行强化学习微调 [14]。无论采用何种流程，强化学习都被发现是提升大语言模型跨任务和跨领域性能 [14] 并更好地保留预训练期间所获先验知识的关键步骤 [17]。

由于大多数语音深度伪造检测研究仅采用监督微调（即使在后训练阶段也是如此 [12]），我们受到大语言模型领域进展的启发，研究了基于强化学习的微调能否提升语音深度伪造检测器的泛化性能。因此，我们采用强化学习，尤其是 GRPO，对语音深度伪造检测模型进行微调，如图1所示。我们的研究发现和贡献总结如下。

- 我们将 GRPO 应用于多个经过后训练的基于 SSL 的语音深度伪造检测器，发现基于纯 GRPO 的微调提升了所有域外测试集上的检测性能，优于仅使用 SFT 或 SFT 加 GRPO 等其他微调方案。
- 我们进行了消融研究，发现 GRPO 中的负奖励可能是这一改进的关键组成部分。GRPO 中的正则化项影响较小。
- 我们使用一款近期发表的工具 [18] 测量了测试数据分布的漂移，并找到了支持上述发现的证据。

近期有一项研究将 GRPO 应用于基于大语言模型的语音深度伪造检测器 [19], , 但据我们所知，我们上述发现对于采用基于自监督学习的前端的主流方法而言是新的。论文评审结束后，我们将发布代码。

2. 方法

2.1. 本文的训练范式与研究范围

给定输入波形 $\boldsymbol{x} \in \mathbb{R}^T$，我们的检测模型预期会输出标签 $y \in \mathcal{Y} = \{\text{真实}, \text{伪造}\}$，用于标示输入是否真实。在模型训练过程中，我们通过训练流程调整模型 $p_\Theta(\mathbf{y}|\mathbf{x})$ 的参数集 Θ。使用训练好的模型进行决策时，我们会计算输入样本 $\boldsymbol{x}^{\text{test}}$ 的概率 $p_\Theta(\mathbf{y} = \text{真实}|\mathbf{x} = \boldsymbol{x}^{\text{test}})$（或其对数几率）。如果概率高于预先设定的阈值，我们就判定为 $\widehat{y} = \text{真实}$；否则判定为 $\widehat{y} = \text{伪造}$。

本文研究多阶段训练流程（图1），以及采用基于自监督学习的前端和二分类后端的检测模型。我们假设参数集 Θ（包括前端和后端）已经过后训练，并且仅关注使用来自特定领域的小规模训练集 $\mathcal{D} = \{\boldsymbol{x}^{(k)}, y^{(k)}\}_k$ 对 Θ 进行微调。[1] 请注意，对于仅包含预训练和微调的传统设置，Θ 由预训练前端的参数和随机初始化后端的参数组成。

2.2. 监督微调

最常见的微调方式是采用监督微调。该方法通过最小化训练数据 $\mathcal{D}$ 的负对数似然 $\mathcal{L}_{\text{ne}}$ 来进行训练，可写为：

$$\mathcal{L}_{\text{ne}}(\Theta) = -\mathbb{E}_{\boldsymbol{x},y\sim\mathcal{D}}\Big[\log p_\Theta(y|\boldsymbol{x})\Big] \tag{1}$$

$$\approx -\sum_{i=1}^{|\mathcal{D}|}\sum_{\tilde{y}\in\mathcal{Y}} \delta(\tilde{y} = y^{(i)}) \log p_\Theta(\tilde{y}|\boldsymbol{x}^{(i)}), \tag{2}$$

其中，$\delta(\cdot)$ 是指示函数。

2.3. 强化学习

强化学习是监督微调的一种替代方法。在众多基于强化学习的算法中，我们考虑一种基于策略的算法，其损失函数定义为

$$\mathcal{L}_{\text{RL}}(\Theta) = -\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D},\tilde{y}\sim p_{\widehat{\Theta}}(\mathbf{y}|\mathbf{x})}\Big[\mathcal{A}(\tilde{y}, y)\frac{p_\Theta(\tilde{y}|\boldsymbol{x})}{p_{\widehat{\Theta}}(\tilde{y}|\boldsymbol{x})} - \beta\mathcal{L}_{\text{p}}(\dot{\Theta}, \Theta)\Big], \tag{3}$$

其中，$\tilde{y}$ 是从模型中采样得到的标签，采样时使用参数集 $\widehat{\Theta}$ 的“旧”快照（例如，前一个训练轮次使用的参数集）。优势函数 $\mathcal{A}(\tilde{y}, y)$：$\mathcal{Y}\times\mathcal{Y} \to \mathbb{R}$ 用于衡量采样标签 $\tilde{y}$ 与真实标签 y 的匹配程度。正则化项 $\mathcal{L}_{\text{p}}$ 用于惩罚微调后的模型与参考模型之间的差异（例如，具有冻结参数集 $\dot{\Theta}$ 的预训练/后训练模型），尽管有观点认为正则化并非必需 [17]。它可以通过近似的库尔贝克-莱布勒散度 [20, § 2.3]来实现。

GRPO是微调大语言模型时常用的一种强化学习算法 [17]。它通过对每个训练输入 $\boldsymbol{x}$ 采样多个实例$\{\tilde{y}_j\}_{j=1}^G$，并使用以下公式计算组归一化优势 $\mathcal{A}(\tilde{y}_i, \{\tilde{y}_j\}_{j=1}^G, y)$，从而改进公式（3）

$$\mathcal{A}(\tilde{y}_i, \{\tilde{y}_j\}_{j=1}^G, y) = \frac{r(\tilde{y}_i, y) - \bar{r}(\{\tilde{y}\}_{j=1}^G, y)}{\sigma(\{\tilde{y}\}_{j=1}^G, y)}, \tag{4}$$

其中，$r(\tilde{y}_i, y)$ 是一个奖励函数，用于衡量 $\tilde{y}_i$ 与 y 或预定义规则 [21], 进行比较时的优劣

[1]为避免符号过于繁杂，除非必要，我们省略样本索引（k）。我们不区分随机变量及其取值（例如，$\mathbf{y}$ 与 y），除非这种区分是必要的。

其中，$\bar{r}$ 和 σ 分别是 r 在样本输出组 $\{\tilde{y}_j\}_{j=1}^G$中的均值和标准差。

GRPO 还会裁剪比率 $\frac{p_\Theta(\tilde{y}|\boldsymbol{x})}{p_{\widehat{\Theta}}(\tilde{y}|\boldsymbol{x})}$，但为避免符号过于繁杂，此处略去。GRPO 还需要三个参数集：每个训练步都要更新的参数集（Θ）、更新频率较低的“旧”参数集（$\widehat{\Theta}$），以及冻结参考（$\dot{\Theta}$）。为简化起见，我们可以设定 $\widehat{\Theta} = \Theta$，并在式（3）的 $\frac{p_\Theta(\tilde{y}|\boldsymbol{x})}{[p_\Theta(\tilde{y}|\boldsymbol{x})]_{\text{lg}}}$ 中加入停止梯度算子。我们还可以移除裁剪函数。这种简化的GRPO（GRPO_{s}）会在每批采样数据之后更新模型，根据定义，这与同策略强化学习类似 [22]。

2.4. 将GRPO应用于深度伪造检测器

使用GRPO进行微调可以直接应用于二分类深度伪造检测器。与大语言模型的情况相比，最本质的区别在于，深度伪造检测器采样得到的输出 $\tilde{y}$ 要么是伪造，要么是真实，而不是一系列词元。奖励可以简单地定义为指示函数 δ $r = 1$ 0 $(\tilde{y}_i,$ $(\tilde{y}_i = y)$，使得对于 $\tilde{y}_i = y$；否则，。参考参数 $\dot{\Theta}$ 可以采用后训练模型的参数。

除了直接应用GRPO之外，我们还旨在理解GRPO与SFT之间可能影响微调检测模型泛化能力的差异。主要差异如下所列。

a) **正则化项** $\mathcal{L}_{\text{p}}$：相对于一个“老而可靠”模型进行正则化，被认为可以缓解微调期间的灾难性遗忘问题 [23]。尽管少数研究认为，对于大语言模型而言，$\mathcal{L}_{\text{p}}$ 并非必不可少 [24], ，但它可能会影响二分类检测器的性能，因为该检测器的输出是二值的。

b) **优势函数** $\mathcal{A}(\tilde{y}_i, \{\tilde{y}_j\}_{j=1}^G, y)$：GRPO 中的优势函数会累积来自匹配（$\tilde{y}_i = y$）和未匹配（$\tilde{y}_i \neq y$）样本的正奖励与负奖励，即使我们定义的是 1-0 奖励 $r(\tilde{y}_i, y) = \delta(\tilde{y}_i = y)$。[2]这与监督微调不同，后者仅使用真实标签计算损失，即在式（2）中，$\widehat{y}$ 始终等于 y。

为分析这两个因素的影响，我们引入了几种变体，并通过实验比较其性能：

- GRPO（$\beta = 0$）：该变体将 $\mathcal{L}_{\text{p}}$ 的权重设为 0，以检验 GRPO 是否能从正则化中获益；
- 使用较大 β 的 GRPO（GRPO（$\beta = 1$））[3]：该变体用于考察过度正则化是否会降低性能。
- 不使用任何负样本的 GRPO（GRPO w/o neg.）：它定义了 $\mathcal{A}(\tilde{y}_i, y) = r(\tilde{y}_i, y) = \delta(\tilde{y}_i = y)$，因此移除了负例 $\widehat{y}_i \neq y$。该变体类似于监督微调加上正则化项（参见补充材料）。

上述变体可应用于简化GRPO（GRPO_{s}），但本文不采用这些配置。

3. 实验

3.1. 模型配置

我们遵循图1中的训练流程，并比较监督微调、GRPO 及其变体在微调中的表现（§ 2）。

对于后训练模型的微调，我们使用 AntiDeepfake 项目发布的后训练模型 [12]，具体来说，是采用多语言 SSL 前端的模型：

[2]具体而言，我们针对匹配（$r(\tilde{y}_i, y) = 1$）和不匹配（$r(\tilde{y}_i, y) = 0$）样本，获取 $\mathcal{A}(\tilde{y}_i, \{\tilde{y}_j\}_{j=1}^G, y) = (1-\bar{r})/\sigma$ 和 $-\bar{r}/\sigma$。[3]$\beta = 1$ 是默认值 $\beta = 0.04$ [21] 的 25 倍。

表1：测试集等错误率（%）。每个等错误率均为经过三 个训练评估轮次后的平均值。*"ave."*列列出域内或域外的平均等错误率。*GRPO*默认使用 $\beta = 0.04$ （*§3.1*）。每列中，等错误率越高，颜色越深。

	Pre-训练	后训练	微调 tune	DFE24评估集（域内）						其他评估集（域外）					Row ID
				4 s	10 s	13 s	30 s	50 s	ave.	ADD23	FoR	DV	ItW	ave.	
XLS-R-2B	✓	✓	×	27.73	23.45	21.74	20.62	19.64	22.64	4.67	2.61	2.23	1.24	2.69	1
	✓	✓	SFT	12.17	10.28	10.11	9.50	9.23	10.26	6.09	3.92	8.75	6.35	6.28	2
			$GRPO_s$	12.45	11.07	11.06	10.44	10.98	11.20	3.95	0.65	6.72	2.44	3.44	3
			GRPO	11.06	9.99	9.90	9.45	9.27	9.93	5.34	0.47	2.76	2.19	2.69	4
	✓	✓	监督微调 → $GRPO_s$	12.17	10.20	10.07	9.36	9.43	10.24	6.42	3.11	8.20	6.24	5.99	5
			监督微调 → GRPO	11.53	9.94	9.88	8.93	8.76	9.81	5.78	2.75	7.04	5.89	5.37	6
			GRPO（$\beta = 0$）	11.58	10.32	10.00	9.64	9.51	10.21	4.43	0.47	6.50	2.71	3.53	7
			GRPO（$\beta = 1$）	19.37	15.61	15.27	14.63	14.24	15.83	4.02	1.34	1.66	1.13	2.04	8
			不使用负样本的 GRPO。	13.30	11.60	11.23	10.64	10.31	11.41	6.69	0.83	3.09	3.14	3.44	9
	✓	×	SFT	12.24	10.34	10.20	9.63	9.83	10.45	17.81	29.83	16.21	13.49	19.34	10
			$GRPO_s$	14.33	12.65	12.02	11.27	10.87	12.23	22.88	29.71	23.67	23.03	24.82	11
			GRPO	12.93	11.18	11.06	10.24	10.35	11.15	17.57	25.68	25.19	19.00	21.86	12
MMS-1B	✓	✓	×	27.68	23.49	22.94	22.64	22.59	23.87	9.06	1.23	2.47	1.82	3.64	13
	✓	✓	SFT	11.86	9.87	9.51	9.07	9.08	9.88	4.14	1.35	5.35	6.59	4.36	14
			$GRPO_s$	12.47	10.90	10.22	9.45	9.47	10.50	3.73	0.73	2.68	4.52	2.92	15
			GRPO	11.87	10.41	9.87	9.21	9.47	10.17	4.25	1.78	2.84	2.98	2.96	16
MMS-300M	✓	✓	×	32.87	30.08	29.38	26.33	27.51	29.23	7.96	1.51	2.35	2.91	3.68	17
	✓	✓	SFT	14.45	11.83	11.55	11.90	11.41	12.23	8.81	3.61	4.37	14.17	7.74	18
			$GRPO_s$	13.46	11.60	11.16	11.27	10.74	11.65	5.70	1.40	2.91	8.31	4.58	19
			GRPO	14.05	11.73	11.42	11.12	11.37	11.94	4.99	1.57	2.64	5.91	3.78	20

XLS-R-2B [25], 、MMS-1B [26], 或 MMS-300M。[4] 所有模型均采用相似的架构：基于自监督学习的前端处理输入波形，并通过全局平均池化将其最后一层的特征合并为固定维度的向量。随后，使用线性层和 Softmax 将该向量转换为伪造和真实的概率。在推理过程中，使用真实类别的对数几率进行评分。

作为参考，我们还纳入了两种设置：1）预训练加后训练，不进行微调；2）预训练加微调，不进行后训练。在后一种设置中，前端使用由 Fairseq 预训练的对应 SSL 模型进行初始化 [27],，后端则随机初始化。

所有 模型的微调方案与 An-tiDeepfake [12] 相同，但根据计算预算进行了调整：最多训练 10 个 epoch，并且每 20k 个训练步进行一次提前停止验证。使用验证集上等错误率最低的检查点进行评估。对于使用 GRPO 及其变体的模型，我们遵循 DeepSeekMath 方案 [21] ，并分别将 rollout 数量和惩罚权重设置为 $G = 64$ 和 $\beta = 0.04$。对于使用 SFT 和 GRPO（SFT→ GRPO）进行微调的情况，每个阶段最多运行 10 个 epoch。

微调与评估循环共进行三轮，并报告平均等错误率（EER）。所有实验均使用单张 H100 GPU 卡完成。

3.2. 数据与协议

我们采用 AntiDeepfake [12] 微调实验中的数据协议。微调数据（以及目标域）为 Deepfake-Eval-2024 数据集的训练划分（DFE24） [28]。它包含约 50 小时来自 2024 年社交网络的真实与伪造语音数据，其中背景噪声和声学条件具有多样性。为适应 GPU 显存，微调集中的语音片段被切分为最长 15 秒。

域内评估数据来自 DFE24 的评估分区。具体而言，我们使用 [12], 提供的五个分段版本，其最长语音时长分别为 4、10、13、30 和 50 秒。域外评估数据集包括音频深度伪造检测挑战赛 Track 1.2 评估集（ADD23） [4], Fake-OrReal 测试集（FoR） [29], 分段式DEEP-VOICE数据集（DV） [30], 以及 In-the-Wild数据集（ItW） [31]。这些数据集涵盖了来自 DFE24 目标领域的、具有不同声学条件的未见数据，因此可用于衡量微调模型的泛化性能。[5]

3.3. GRPO 与监督微调的结果比较

表1列出了三轮实验的平均等错误率。首先可以观察到，**通过监督微调进行微调的模型在域外测试集上的性能退化程度各不相同**。与未经过微调的后训练模型（第1、13和17行）相比，对应的监督微调模型（分别为第2、14和18行）在域内评估集（DFE24）上的等错误率更低，但在大多数域外测试集上的等错误率更高。例如，使用 XLSR-R-2B 的后训练模型在 ItW 上的等错误率为 1.24%（第1行），而经过监督微调后升至 6.35%（第2行）。

[4]对应的检查点文件为 xlsr 2b.ckpt、mms 1b.ckpt 和 mms 300m.ckpt（DOI:10.5281/zenodo.15580543）。

[5]由于后训练检查点是在多个数据库及其评估分区上训练的，因此一些常用的测试数据（例如 ASVspoof [3]）无法用于评估。

与监督微调相比，采用纯 GRPO（或 GRPO$_s$）进行微调，在保持域内测试数据表现的同时，缓解了域外测试数据上的性能退化。[6] 使用 XLS-R-2B 时，对比监督微调（第2行）和 GRPO（第4行）可见，后者在全部四个域外测试集上的等错误率都更低。例如，ItW 上的等错误率从 6.35%（第2行）降至 2.19%（第4行）。同时，采用 GRPO 后，域内测试集上的等错误率为 10%，与采用监督微调的结果相当。当然，域内测试集上的等错误率远低于未进行微调时的结果（第1行）。

使用GRPO$_s$ 进行微调，在所有三类检测模型（含后训练）上均取得了与使用GRPO相近的结果。这表明实现上的差异，例如奖励裁剪以及保留单独的"旧"参数集$\hat{\Theta}$，至少在我们的实验中并不关键。

从实验的其他设置来看，我们还有两项补充观察。首先，**监督微调→GRPO并不优于仅GRPO微调**。尽管域内测试集上的等错误率相当，但前者（第6行）在域外测试集上的等错误率高于后者（第4行），尤其是在DV（7.04%对2.76%）和 ItW（5.89%对2.19%）上。通过比较监督微调→GRPO$_s$（第5行）与纯 GRPO$_s$（第3行），也观察到了相同趋势。其次，**GRPO应用于后训练模型时有效，但应用于预训练模型时无效**。若比较第4行与第12行在域外测试数据上的结果，这一点显而易见。[7]

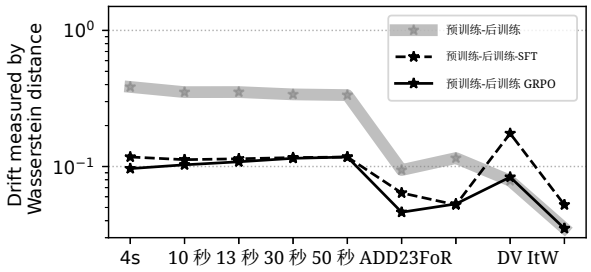
总而言之，结果表明，在后训练模型上应用监督微调能够提升域内数据集上的性能，但可能会遗忘后训练阶段获得的部分知识，从而降低域外检测性能。相比之下，经过 GRPO 调优的模型倾向于保留后训练阶段获得的知识，同时从新的训练数据中学习。

3.4. 漂移分析结果

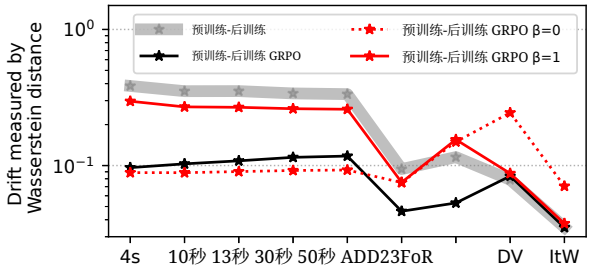
我们进一步使用 [18] 中的工具分析了数据分布的漂移。总体流程是：为测试集和参考集中的每条伪造 语音提取固定维度的嵌入向量（经过平均池化后）。随后，测量测试集中嵌入向量分布与参考集中嵌入向量分布之间的 Wasserstein 距离，即漂移。我们遵循了 [18] 的方法，并使用 ASVspoof 2019 开发集作为参考集。

我们使用后训练的 XLS-R-2B 测量了漂移模型、经过监督微调的版本，以及经过 GRPO 微调的版本。结果绘于图 2 (a)。对于训练后的模型（灰色粗线），在五个 DFE24 测试集上测得的漂移高于其他四个测试集（ADD23、FoR、DV、ItW）。这并不令人意外，因为 DFE24 测试集相对较新，且声学条件更多样。使用领域匹配数据进行监督微调降低了 DFE24 测试集上的漂移值（黑色虚线），但 DV 和 ItW 上的漂移却有所升高。相比之下，GRPO 微调降低了 DFE24 测试集上的漂移，同时没有导致域外数据发生漂移。这一结果进一步支持了以下发现：使用 GRPO 进行微调可以缓解域外数据造成的性能退化。

[6]这一设置在概念上类似于 DeepSeek-R0 [14] 无需监督微调（SFT）即可进行强化学习（RL）的思路。相比之下，通过 SFT→GRPO 进行微调是 DeepSeek-R1 流程的一部分。[7]事实上，在域外测试集上，对后训练模型进行 SFT（第 10 行）的表现也优于对预训练模型进行 SFT（第 2 行），这支持了后训练具有预期优势的假设 [12]。



(a) 比较采用不同微调算法的模型



(b) 不同 *GRPO* 变体之间的比较

图 2：在不同测试集上测得的漂移值

3.5. GRPO 消融研究

接下来，我们考察了 GRPO 中正则化项 $\mathcal{L}_p$ 的影响。与默认设置（第 4 行）相比，不使用正则化的 GRPO（$\beta = 0$，第 7 行）在域内测试集上取得了相近的等错误率，但在域外 DV 上的等错误率升至 6.50%。然而，当正则化更强时（$\beta = 1$，第 8 行），模型在域内测试集上的等错误率更高，似乎对目标域拟合不足。因此，与关于大语言模型的研究不同 [17], [24], ，我们无法得出正则化并非必要的结论。图 2 (b) 中的漂移分析也支持正则化项的影响。通过适当的正则化，监督微调也可能得到改进，但这并非我们当前工作的重点。

就 GRPO 中的负奖励而言，不使用负奖励（第 9 行）时，所有测试集上的等错误率都高于默认的 GRPO 配置（第 4 行）。负奖励可能是监督微调与 GRPO 之间的一个关键差异。尽管还需要进一步探索以解释 GRPO 带来的改进，但这留待未来工作。

4. 结论

本文研究了类似 GRPO 的强化学习是否适用于微调基于自监督学习的语音深度伪造检测模型。我们使用多个采用不同基于自监督学习前端的后训练检测模型开展实验，发现与传统的监督微调（SFT）相比，采用 GRPO 进行微调能够减轻域外数据上的性能退化，同时保持域内测试集上的性能。

我们还确定了 GRPO 的优选方案。具体而言，仅GRPO微调优于先SFT后GRPO；与采用预训练SSL前端和随机初始化后端的模型相比，将 GRPO 应用于后训练模型时效果更好。研究还发现，GRPO 中的负奖励至关重要。正则化项会产生影响，但过强的正则化可能导致欠拟合。我们希望这些发现能为更多相关研究带来启发。

5. 致谢

本研究部分得到日本科学技术振兴机构 PRESTO 计划（JPMJPR23P9）以及由新能源与产业技术综合开发机构（NEDO）委托的项目（JPNP22007）的支持。本研究使用 TSUBAME4.0 完成。

6. 生成式人工智能使用披露

生成式人工智能仅用于修正草稿中的语法错误。随后，论文通过付费服务进行了人工校对。

7. 参考文献

[1] 欧盟执法合作署，直面现实？：执法与深度伪造的挑战：来自欧洲刑警组织创新实验室的观察报告。英文。卢森堡：出版局，2024.[2] Z. Wu等人，“ASVspoof：自动说话人验证欺骗攻击与对策挑战赛”，*IEEE*信号处理精选主题期刊，第11卷，第4期，第588–604页，2017.[3] M. Todisco等人，“ASVspoof 2019：欺骗与伪造音频检测的未来展望”，载于*Interspeech*会议论文集，2019，第1008–1012页。[4] J. Yi等人，“ADD 2023：第二届音频深度伪造检测挑战赛”，载于*IJCAI*深度伪造音频检测与分析研讨会论文集，2023.[5] J. M. Martin-Donas与A. Alvarez，“面向2022年ADD挑战赛的基于Wav2vec2的Vicomtech音频深度伪造检测系统”，载于*ICASSP*会议论文集，2022，第9241–9245页。[6] H. Tak等人，“使用wav2vec 2.0与数据增强的自动说话人验证欺骗攻击及深度伪造检测”，载于*Odyssey*会议论文集，2022，第112–119页。[7] Q. Zhang、S. Wen与T. Hu，“基于自监督XLS-R与SLS分类器的音频深度伪造检测”，载于*ACM MM*会议论文集，ACM，2024，第6765–6773页。[8] A. Mohamed等人，“自监督语音表征学习：综述”，*IEEE*信号处理精选主题期刊，第16卷，第6期，第1179–1210页，2022.[9] A. Baevski、Y. Zhou、A. Mohamed与M. Auli，“Wav2vec 2.0：语音表征自监督学习框架”，载于*NIPS*会议论文集，第33卷，2020，第12 449–12 460页。[10] X. Wang等人，“ASVspoof 5：大规模众包语音数据、深度伪造与对抗攻击”，载于*2024*年*ASVspoof*研讨会，2024，第1–8页。[11] S. Dowerah等人，“Speech DF Arena：语音深度伪造检测模型排行榜”，*IEEE*信号处理开放期刊，第7卷，第73–81页，2026.[12] W. Ge、X. Wang、X. Liu与J. Yamagishi，“面向深度伪造语音检测的后训练”，载于*ASRU*会议论文集，2025，（已录用）。[13] N. Klein、T. Chen、H. Tak、R. Casal与E. Khoury，“音频深度伪造系统的来源追踪”，载于*Interspeech*会议论文集，ISCA，2024，第1100–1104页。[14] D. Guo等人，“DeepSeek-R1通过强化学习激励大语言模型进行推理”，英文，《自然》，第645卷，第8081期，第633–638页，2025.

[15] K. Mo 等，《大型语言模型的中期训练：综述》，arXiv: 2510 06826 [ ], 。2025 无同行评审版本。[16] L. Wang、X. Zhang、H. Su 和 J. Zhu，“持续学习综述：理论、方法与应用”，《IEEE 模式分析与机器智能汇刊》，第 46 卷，第 8 期，第 5362–5383 页，2024 年。[17] I. Shenfeld、J. Pari 和 P. Agrawal，“强化学习的剃刀：为何在线强化学习遗忘较少”，载于 NeurIPS 2025 持续与兼容基础模型研讨会论文集，2025 年。[18] X. Wang、W. Ge 和 J. Yamagishi，“MLOps 背景下语音深度伪造检测的数据漂移监测”，载于 ICASSP 论文集，2026 年，（已录用）。[19] Y. Xie 等，“通过频时强化学习实现可解释的全类型音频深度伪造检测与音频大语言模型”，2026 年，arXiv:2402。03300 [无同行评审版本]。[20] A. Amini、T. Vieira 和 R. Cotterell，“更准确地估计语言模型之间的库尔贝克-莱布勒散度”，载于 NeurIPS 论文集，2025 年。[21] Z. Shao 等，《DeepSeekMath：突破开放语言模型型数学推理的极限》，arXiv:2402。03300 [无同行评审版本], 2024。[22] N. Lambert，《基于人类反馈的强化学习》。在线，2024 年，https://rlhfbook.com/。[23] L. Wang、X. Zhang、H. Su 和 J. Zhu，“持续学习综述：理论、方法与应用”，《IEEE 模式分析与机器智能汇刊》，第 46 卷，第 8 期，第 5362–5383 页，2024 年。[24] Q. Yu 等，“DAPO：大规模开源大语言模型强化学习系统”，载于 NeurIPS 论文集，2025 年。[25] A. Babu 等，“XLS-R：大规模自监督跨语言语音表示学习”，载于 Interspeech 论文集，2022 年，第 2278–2282 页。[26] V. Pratap 等，“将语音技术扩展至 1,000+种语言”，《机器学习研究期刊》，第 25 卷，第 1 期，2024 年。[27] M. Ott 等，“Fairseq：快速且可扩展的序列建模工具包”，载于 NAACL-HLT 论文集，2019 年。[28] N. A. Chandra 等，《Deepfake-Eval-2024：对流传于 *2024*, 2503 02857 [ ], 的深度2025 伪造内容进行的多模态真实场景基准测试》，arXiv:。无同行评审版本。[29] R. Reimao 和 V. Tzerpos，“FoR：合成语音检测数据集”，载于 2019 年国际语音技术与人机对话会议，2019 年，第 1–10 页。[30] J. J. Bird 和 A. Lotfi，《实时检测用于深度伪造语音转换的 AI 生成语音》，arXiv:2308。12734 [无同行评审版本], 2023。[31] N. M¨uller、P. Czempin、F. Diekmann、A. Froghyar 和 K. B¨ottinger，“音频深度伪造检测能否泛化？”，载于 Interspeech 论文集，2022 年，第 2783–2787 页。

A. 附录

A.1. 关于所应用的GRPO损失的说明

A.1.1. 不含负奖励的 *GRPO*

在 § 2.4 中，我们提到，不含任何负样本的 GRPO（GRPO w/o neg.）与监督微调加上正则化项“等价”。这种等价性体现在根据 GRPO 和监督微调损失函数计算出的梯度上。为说明这一点，我们对式 (3) 两边关于 Θ 求导，得到

$$\nabla_\Theta \mathcal{L}_{\text{RL}}(\Theta) = -\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D},\tilde{y}\sim p_{\widehat{\Theta}}(\mathbf{y}|\mathbf{x})}\Big[\mathcal{A}(\tilde{y},y)\nabla_\Theta \frac{p_\Theta(\tilde{y}|\boldsymbol{x})}{p_{\widehat{\Theta}}(\tilde{y}|\boldsymbol{x})} - \nabla_\Theta \beta \mathcal{L}_{\text{p}}(\dot{\Theta},\Theta)\Big] \tag{5}$$

$$= -\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\Bigg\{\sum_{\tilde{y}}\Big[\mathcal{A}(\tilde{y},y)\nabla_\Theta \frac{p_\Theta(\tilde{y}|\boldsymbol{x})}{p_{\widehat{\Theta}}(\tilde{y}|\boldsymbol{x})} - \nabla_\Theta \beta \mathcal{L}_{\text{p}}(\dot{\Theta},\Theta)\Big] p_{\widehat{\Theta}}(\mathbf{y}|\mathbf{x})\Bigg\} \tag{6}$$

$$= -\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\Bigg\{\sum_{\tilde{y}}\Big[\mathcal{A}(\tilde{y},y)\nabla_\Theta p_\Theta(\tilde{y}|\boldsymbol{x}) - \nabla_\Theta \beta \mathcal{L}_{\text{p}}(\dot{\Theta},\Theta) p_{\widehat{\Theta}}(\mathbf{y}|\mathbf{x})\Big]\Bigg\} \tag{7}$$

$$= -\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\Bigg\{\sum_{\tilde{y}}\Big[\mathcal{A}(\tilde{y},y)\nabla_\Theta p_\Theta(\tilde{y}|\boldsymbol{x})\Big] - \nabla_\Theta \beta \mathcal{L}_{\text{p}}(\dot{\Theta},\Theta)\Bigg\} \tag{8}$$

$$= -\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\Bigg\{\sum_{\tilde{y}}\Big[\mathcal{A}(\tilde{y},y) p_\Theta(\tilde{y}|\boldsymbol{x})\nabla_\Theta \log p_\Theta(\tilde{y}|\boldsymbol{x})\Big] - \nabla_\Theta \beta \mathcal{L}_{\text{p}}(\dot{\Theta},\Theta)\Bigg\}. \tag{9}$$

请注意，最后一行基于对数导数技巧 $\nabla_\theta f(x;\theta) = f(x;\theta)\nabla_\theta \log f(x;\theta)$。如果令 $A(\tilde{y},y) = \delta(\tilde{y}=y)$，即仅当采样标签与真实标签匹配时奖励为 1，则有

$$\nabla_\Theta \mathcal{L}_{\text{RL}}(\Theta) = -\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\Bigg\{\sum_{\tilde{y}}\Big[\mathcal{A}(\tilde{y},y) p_\Theta(\tilde{y}|\boldsymbol{x})\nabla_\Theta \log p_\Theta(\tilde{y}|\boldsymbol{x})\Big] - \nabla_\Theta \beta \mathcal{L}_{\text{p}}(\dot{\Theta},\Theta)\Bigg\} \tag{10}$$

$$= -\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\Bigg\{p_\Theta(y|\boldsymbol{x})\nabla_\Theta \log p_\Theta(y|\boldsymbol{x}) - \nabla_\Theta \beta \mathcal{L}_{\text{p}}(\dot{\Theta},\Theta)\Bigg\}. \tag{11}$$

与监督微调（即 $\nabla\mathcal{L}_{\text{ne}} = -\mathbb{E}_{\boldsymbol{x},\boldsymbol{y}\sim\mathcal{D}}\Big[\nabla \log\ p_\Theta(y|\boldsymbol{x})\Big]$）相比，梯度类似，只是多了对梯度 $\nabla_\Theta \log\ p_\Theta(y|\boldsymbol{x})$ 进行加权的项 $p_\Theta(y|\boldsymbol{x})$，以及正则化项。

这表明，不使用负样本的 GRPO 会优先考虑模型现有的高概率模式（通过权重 $p_\Theta(y|\boldsymbol{x})$），其作用类似于一种针对异常数据的隐式正则化。相比之下，标准监督微调会平等对待所有（正）样本。该权重项源自比率 $^{p_\Theta(\tilde{y}|\boldsymbol{x})}/_{p_{\widehat{\Theta}}(\tilde{y}|\boldsymbol{x})}$，也就是许多强化学习算法中使用的重要性采样比。

A.2. GRPO实现的详细信息

在实现中，我们遵循 GRPO 论文，并加入了裁剪函数：

$$\mathcal{L}_{\text{RL}}(\Theta) = \mathbb{E}_{x\sim\mathcal{D},\tilde{y}\sim p_{\widehat{\Theta}}(\mathbf{y}|\mathbf{x})}\left[\left(\min\left(\frac{p_\Theta(\tilde{y}|\boldsymbol{x})}{p_{\widehat{\Theta}}(\tilde{y}|\boldsymbol{x})}\mathcal{A}(\tilde{y},y), \text{clip}\left(\frac{p_\Theta(\tilde{y}|\boldsymbol{x})}{p_{\widehat{\Theta}}(\tilde{y}|\boldsymbol{x})}, 1-\epsilon, 1+\epsilon\right)\mathcal{A}(\tilde{y},y)\right) - \beta\mathcal{L}_{\text{p}}(\dot{\Theta},\Theta)\right)\right], \tag{12}$$

其中，优势函数的分母设有一个下限值，以防标准差为 0。

$$\mathcal{A}(\tilde{y}_i, \{\tilde{y}_j\}_{j=1}^G, y) = \frac{r(\tilde{y}_i,y) - \bar{r}(\{\tilde{y}\}_{j=1}^G, y)}{\sigma(\{\tilde{y}\}_{j=1}^G, y) + \rho}, \tag{13}$$

超参数设置如下：

表 2： 所实现 *GRPO* 的超参数

超参数	值
ϵ	0.2
ρ	1e-5
更新 $\widehat{\Theta}$ 前的训练步数	1k
每 $\boldsymbol{x}$ 采样的 $\tilde{y}$ 数量	64