arXiv:2602.03355v1 [cs.SD] 3 Feb 2026

PACE：预训练音频持续学习

ChangLi*，**Kanglei Zhou***，**Liyuan Wang**†清华大学心理与认知科学系

摘要

音频是分析语音、音乐和环境声音的基本模态。尽管预训练音频模型显著推动了音频理解的发展，但在数据分布随时间变化的真实场景中，这些模型仍然十分脆弱。本文首次提出针对预训练模型（PTMs）的音频持续学习（CL）系统性基准，并全面分析其独特挑战。与视觉领域中参数高效微调（PEFT）已被证明对持续学习有效不同，直接将此类策略应用于音频会导致性能不佳。这是由音频骨干网络的一项根本特性造成的：它们侧重于低层频谱细节，而非结构化语义，从而导致严重的上游—下游失配。通过广泛的实证分析，我们确定了一条有前景的技术路径，即采用首会话适应（FSA）的解析分类器，但也发现了两项主要局限：粗粒度场景中的表征饱和，以及细粒度场景中的表征偏移。为应对这些挑战，我们提出**PACE**，这是一种创新方法，通过正则化解析分类器改进 FSA，并引入基于自适应子空间正交参数高效微调的多会话适应，以实现更好的语义对齐。此外，我们设计了基于频谱图的边界感知扰动，以缓解表征重叠并提升稳定性。在六个不同的音频持续学习基准测试上的实验表明，PACE 的表现显著优于最先进的基线，为借助预训练模型实现稳健且可扩展的音频持续学习迈出了重要一步。

1 引言

音频是人类交流和环境感知的核心，支持众多应用，例如语音识别（Abdel-Hamid 等人，2014）、声学事件检测（Zhuang 等人，2010）以及声音场景理解（Nakamura 等人，2000；Zhou等人，2024c）。随着大规模监督预训练（Gong 等人，2021）和自监督预训练（Gong 等人，2022；Chen 等人，2024；2022；Li 等人，2024b）的兴起，预训练音频模型已在广泛的下游任务中取得了显著成功。然而，在音频
分布持续演变时，这些模型往往难以有效适应，同时避免灾难性遗忘，这暴露了其在音频相关应用中的一项关键局限。

持续学习（CL）旨在通过使模型在学习新任务的同时保留旧知识，来解决这一局限（Zhou等人，2025b；2024d；2025a）。尽管近年来视觉领域的进展已证明，参数高效微调（PEFT）可用于基于预训练模型（PTMs）的持续学习（Wang 等人，2022b；a），但将其扩展到音频领域仍鲜有研究，且极具挑战性。预训练视觉模型通常能够编码稳定且结构良好的语义表征（Janson等人），

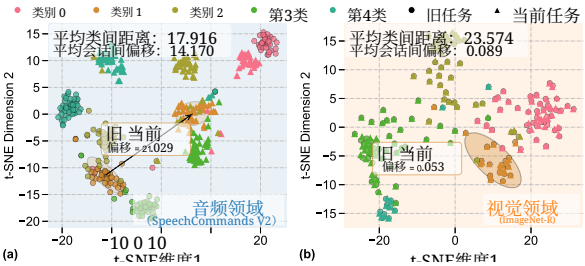


图1：与 ImageNet-R 上的视觉持续学习（b）相比，SC2 上的音频持续学习（a）在相邻会话之间遭受的**表征偏移**显然大得多。

*共同第一作者
†通讯作者：liyuanwang@tsinghua.edu.cn

因此在相邻会话之间的表征偏移相对较小（见图1(b)）。相比之下，音频识别高度依赖细粒度频谱线索（Wang 等人；Wu 等人），而持续学习则需要逐步调整高层语义目标，以获得具有判别性的表征。预训练音频模型（Chen 等人，2024；Gong 等人，2022；Alex 等人，2025）通常通过频谱图重建目标进行训练，因此更注重低层时频模式，而非结构化语义（Tabassum 等人，2024；Bao 等人；Epstein & Meir，2019）。尽管采用了这种学习目标，预训练与下游数据分布之间的差异还会进一步加剧**上下游不匹配**，如图10所示。这些不匹配迫使骨干网络在各个会话中不断重塑其内部表征，导致显著的表征偏移，其幅度往往超过类别之间细微的频谱差异，进而造成严重的灾难性遗忘（见图1(a)）。

为了系统地探究基于预训练的持续学习方法能否以及如何有效应用于音频领域，我们构建了一个全面的基准测试，并揭示了三**项关键发现**：**首先**，我们发现，具有代表性的视觉领域持续学习方法，尤其是那些依赖任务共享表征的方法，迁移到音频领域后性能显著下降。我们确定了一条简单而有效的技术路径，将首会话适应（FSA）、骨干网络冻结和二阶解析分类结合起来（McDonnell 等人，2023；Zhuang 等人，2022）。**其次**，尽管该方法在领域差距相对较小的粗粒度基准测试上表现出色，但在学习首个任务时会出现表征饱和（Li 等人，2020），从而阻碍后续适应。**第**三，当应用于要求更高的细粒度场景（例如乐器和说话人分类）时，该方法的效果会减弱，因为这些场景存在显著的上游与下游不匹配。因此，与联合训练上界相比，仍存在明显的性能差距。

为弥合这一差距，我们提出了 PACE（**预**训练**音**频**持**续学**习**），这是一种新方法，旨在充分利用预训练音频模型，同时克服粗粒度和细粒度持续学习场景中的上游与下游不匹配问题。不同于视觉持续学习（通常只需冻结预训练骨干网络即可，Zhang 等人，2023；2024），PACE 采用针对首会话适应定制的音频专用参数高效微调策略，有选择地适应骨干网络的后层，从而实现更有效的表征学习，尤其适用于粗粒度任务。为提升细粒度场景中的跨会话适应能力，PACE 进一步引入：（1）多会话适应（MSA），采用自适应子空间正交参数高效微调策略，实现渐进式适应，同时限制先前所学特征的漂移，从而以有原则的方式平衡稳定性与可塑性；（2）边界感知扰动机制，通过施加定向时频变换来近似历史决策边界，增强类内紧凑性以及学习所得表征空间中的类间可分性。

我们在三个粗粒度基准（ESC-50、US8K、SC2）和三个细粒度基准（TIMIT-2、TIMIT-3、VocalSet）上进行了广泛实验。PACE 始终优于最先进的持续学习方法，在 TIMIT-2 上至少提升 +5.3%，在 TIMIT- 3 上提升 +4.1%，在 VocalSet 上提升 +6.3%。此外，PACE 显著缩小了与联合训练上界之间的差距，在 ESC-50、US8K、SC2、TIMIT-2、TIMIT-3 和 VocalSet 上的性能分别与上界相差 0.8%、0.6%、3.5%、4.3%、1.2% 和 7.6%。为促进未来研究，我们将发布所有构建的基准测试、复现的基线以及代码库。

2 音频持续学习基准测试

为系统地研究基于预训练模型的音频持续学习，我们首先介绍问题表述，然后展示全面的基准结果，揭示这一设置所面临的独特挑战。

预训练持续学习。 基于预训练模型的持续学习假设可以访问一个预训练骨干网络 f_0，其参数为 θ_0，由源域获得，并逐步适配到一系列 T任务 $\mathcal{T}_1, \cdots, \mathcal{T}_T$，而无需从头开始重新训练。每个任务 $\mathcal{T}_t = (\mathcal{D}_t, \mathcal{Y}_t)$ 都会使用 $\mathcal{D}_t$ 将模型从 θ_{t-1} 更新为 θ_t，并在累积的标签空间 $\bigcup_{i=1}^{t} \mathcal{Y}_i$ 上进行评估。

预训练音频持续学习。 在音频持续学习中，每个输入 $x_{n,t} \in \mathcal{X}_t$ 都是原始音频信号，每个标签$y_{n,t} \in \mathcal{Y}_t$ 都属于某个任务特定类别，其中 $\mathcal{Y}_i \cap \mathcal{Y}_j = \emptyset$ 对于 $i \neq j$。目标是从序列数据集中学习 $\mathcal{D}_1, \cdots, \mathcal{D}_T$同时保持在所有先前类别上的性能。与

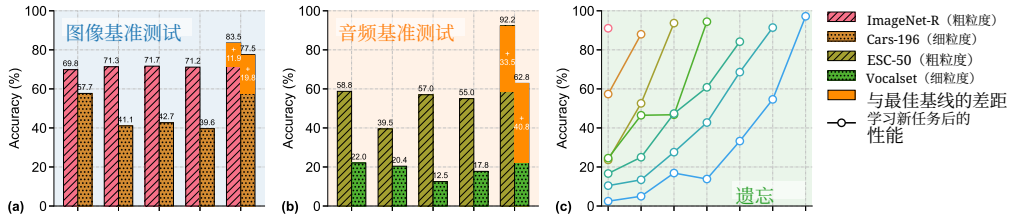


图2：视觉持续学习与音频持续学习的比较。(a) 和 (b) 展示了粗粒度和细粒度设置下，音频与图像数据集上的性能模式。(c) 表明，尽管 PEFT-FT 在首个任务上具有很强的可塑性，但较大的表征偏移会导致严重遗忘。

视觉持续学习不同，由于预训练目标与下游任务粒度之间存在根本性不匹配，预训练音频模型在这种设置下面临额外挑战。

在实证评估中，我们采用 EAT（Chen 等人，2024）作为默认的 f_0。EAT 是一种通用音频自监督模型，基于大规模音频和语音数据集（Gemmeke 等人，2017）进行预训练。我们构建了六个具有代表性的音频持续学习基准。前三个基准——ESC-50（Piczak，2015）、UrbanSound8K（Salamon 等人，2014）和 Speech Commands V2（SC2）（Warden，2018）——代表了环境声音分类和关键词检测等**粗粒度**任务。这些任务与 EAT 预训练目标（Li 和 Angelov，2024）的契合度相对较高，因此上下游不匹配程度较小。为探索涉及严重分布变化的更具挑战性的场景，我们进一步引入了三个**细粒度** 基准：用于说话人识别的 TIMIT-2&3（Garofolo 等人，1993），以及用于乐器识别的 VocalSet（Wilkins 等人，2018）。这些任务要求进行结构化语义理解，与 EAT 的预训练目标明显不匹配，因此对持续学习提出了更大的挑战。数据集的详细介绍和任务配置见第4.1节和第D节。

根据基准测试结果，我们得出三项关键的实证发现：

发现1：视觉持续学习方法在音频任务上性能下降。 如图2(a)和图2(b)所示，将持续学习方法从视觉领域直接迁移到音频理解任务会产生不同的性能模式。具体而言，基于PEFT的持续学习方法，如L2P（Wang 等人，2022c）、DualPrompt（Wang 等人，2022b）和S-Prompt++（Wang 等人，2022a），在音频领域的表现显著更差，性能下降幅度几乎是视觉领域的三倍。这些方法依赖共享表示进行提示键匹配或任务增量适应，在处理音频的细粒度频谱结构时似乎效果较差。相比之下，基于统计的方法使用一次性调优的骨干网络，并采用基于二阶

以 RanPAC（McDonnell 等人，2023）和 ACL（Zhuang 等人，2022）为代表的统计方法，在音频持续学习中始终比基于PEFT的方法取得更强、更稳定的结果。我们将这种性能差距归因于音频领域显著的表征偏移（见图 1(a) 和 1(b)）：一旦学习新的会话，就会迅速发生大幅遗忘（见图 2(c)）。**RanPAC 牺牲了模型可塑性以实现持续更新，但它缓解了更严重的音频表征偏移，因此这种权衡更适用于音频持续学习。**这些观察促使我们采用基于冻结骨干网络的二阶统计量解析分类器，作为预训练音频持续学习的**基础技术路线**。

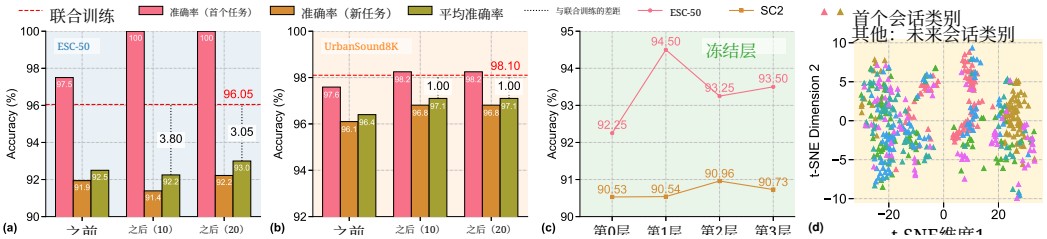


图 3：持续学习中的表征调优分析。(a) 和 (b) 展示了 RanPAC 在 FSA训练轮次中首个会话、后续会话及平均性能相对于联合训练的表现。(c) 展示了仅冻结浅层带来的提升。(d) 为 FSA 后 VocalSet 的 t-SNE可视化。

发现2：粗粒度数据集上的表征饱和。如图 3(a) 和 3(b) 所示，即使不使用 FSA，RanPAC 在粗粒度数据集上也能取得较高的首会话准确率，这表明在这些场景中，预训练与下游任务之间的领域差距相对较小。尽管 FSA 能进一步提高首会话准确率，但即使延长训练时间，也无法有效提升未来任务准确率，与联合训练上界之间仍存在明显差距。这表明预训练骨干网络在第一会话中已经捕获了大部分相关信息，限制了其提取额外判别性特征以使后续任务受益的能力；我们将这一现象称为**表征饱和**。此外，图 3(c) 显示，在 FSA 期间冻结浅层能够提升性能，而全层微调往往会使性能下降，甚至低于无FSA基线。这凸显了盲目微调所有层的风险，因为这可能会损害预训练获得的、具有泛化能力的低层表征。

发现3：细粒度基准测试上的性能差距更大。将已确定的技术路线（FSA 和解析分类）应用于细粒度场景时，其性能会迅速下降，暴露出上下游不匹配造成的显著局限。如图3(d)所示，尽管 FSA 提升了首会话类别的聚类效果，却无法为未来会话类别生成连贯的分布。此外，表1显示，在第一会话上延长训练会导致后续任务的性能下降，这表明模型存在明显的过拟合倾向。这种影响这一现象在语义复杂度较高的数据集中尤为明显。例如，在 VocalSet 上，相较于联合训练，性能差距达到 13.8%；而在 ESC-50 和 UrbanSound8K 上，这一差距分别仅为 3% 和 1%。这些结果表明，仅依靠首个会话的数据不足以弥合预训练目标与细粒度音频任务下游目标之间的语义差距。综合来看，这些发现凸显了逐步在**多个会话**中使预训练表征与下游任务对齐的必要性，同时还要避免早期适应阶段出现过拟合。

方法	TIMIT-2	VocalSet
不使用 FSA	75.87	61.51
朴素 FSA	89.92	62.85
扩展FSA	83.25	61.18
联合训练	95.22	76.65

表1：细粒度基准测试的初步结果。

3 我们的 PACE 方法：**预**训练**音**频**持**续**学**习

3.1 符号与设计概述

符号。令 $f(\cdot)$ 表示具有 L 层的预训练骨干网络，$g(\cdot)$ 表示分类头。给定一个输入音频信号，我们首先计算其时频图 $x \in \mathbb{R}^{T_x \times F_x}$（例如，先进行短时傅里叶变换，再进行梅尔滤波），其中 T_x 和 F_x 分别为帧数和梅尔频带数。骨干网络生成表征 $z = f(x) \in \mathbb{R}^D$，其中 D 为特征维度。随后将 z 输入分类头，以预测 $\mathcal{Y}$ 中的类别概率 $\hat{y} = g(z)$。

设计概述。受第2节实证发现的启发，我们提出**PACE**（图4），一个统一的分阶段框架，旨在使预训练音频表示与持续学习（CL）目标重新对齐。我们将问题分解为两个组成部分：**预训练骨干网络**和**输出头**，并针对各自设计相应策略。

改进的首会话适应（FSA）。**发现1** 的实证证据以及图1中的嵌入空间分析表明，朴素的骨干网络适配会导致**严重的跨会话表征偏移**。因此，我们采用解析分类器作为默认技术路径。如**发现2**所示，预训练音频模型已经编码了强大的粗粒度语义，因此容易出现**首会话饱和**。朴素地微调输出头会扭曲这些表征，限制前向学习。为解决这一问题，我们提出**改进的首会话适应（FSA）**，冻结输出头，仅使用LoRA模块 $\{A_1^l B_1^l \mid L_{\text{tune}} < l \leq L\}$适配更深层。一旦完成适配，便用解析分类器 $\phi^1(\cdot)$替换参数化头 $h^1(\cdot)$，以保留预训练语义，同时避免不必要的参数化。

采用子空间投影的多会话适应（MSA）。尽管 FSA 在粗粒度任务上表现良好，但在上下游失配更严重的细粒度场景中仍显不足。因此，我们引入**多会话适应（MSA）**，利用多个会话的分布逐步对齐表征。然而，正如**发现 1 和 3**所揭示的，朴素的 MSA 会加剧**表征偏移**，导致灾难性遗忘。为实现无回放解决方案，我们仅在需要时进行适应，并通过

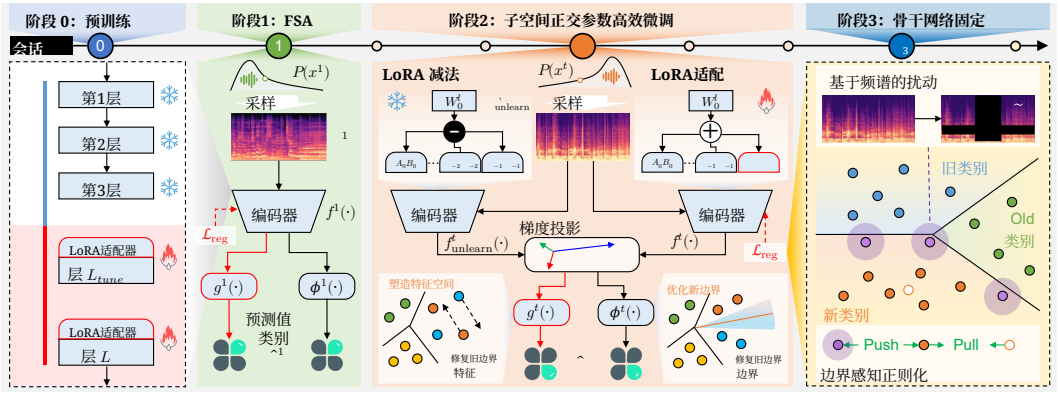


图 4：提出的 PACE 框架。阶段1使用 LoRA 进行首会话适应，随后进行解析推理。阶段2通过 LoRA 减法和梯度投影引入子空间正交参数高效微调。边界感知正则化涉及前两个阶段中的适应。阶段3冻结骨干网络，以实现稳定适应。: 冻结；: 调优；⟶: 适应路径。

子空间正交多会话适应加以约束。在会话 $2 \le t \le T_3$ 中，梯度被投影到一个无干扰子空间 $\mathcal{U}_t^l$（通过 LoRA 减法实现（Liu 与 Chang，2025年）），从而实现受控更新，同时避免扭曲先前的知识。推理仍通过 $\phi_t(\cdot)$ 以解析方式进行。

边界感知正则化。 **发现3** 指出的一个尚待解决的挑战是**类别边界重叠**。随着表征趋于稳定，新类别可能被迫进入次优的重叠区域。这会削弱适应能力和泛化能力。为缓解这一问题，我们引入**边界感知正则化**。对于每个样本 $x_{i,\ t}$，我们生成扰动变体 $\tilde{x}_{i,\ t}$，以近似类别边界区域 $\mathcal{B}t$。训练期间，$x_{i,\ t}$ 会被拉近至其类别原型μ（x_t），并被推远离 $\mathcal{B}_t$，从而增大类间间隔，促进形成紧凑且可分离的表征。这能减少边界碰撞，并提升 MSA 期间的稳定性。

PACE 的各个组成部分将在以下章节中详细介绍，包括改进FSA（第3.2节），以及结合边界感知正则化的子空间正交多会话适应（第3.3节）。

3.2 改进的首会话适应 （FSA）

实证分析表明，在粗粒度音频持续学习中，预训练模型已能提供强大的语义先验。然而，在第一会话中对整个模型进行直接微调，往往会破坏这些对齐良好的表征，导致模型过早饱和。这表明，第一会话应侧重于有针对性的细化，而非全面进行骨干网络适配。为此，我们改进的 FSA 融入了三个关键思路：（1）将更新限制在输出头，使梯度主要流入骨干网络；（2）仅适配更深层的语义相关层；（3）适配完成后，以解析分类器替换可训练头，以确保后续会话中的稳定性。

受限分类头学习。 现有的 FSA 方法将线性输出头与基本冻结的骨干网络联合训练，这往往导致输出头过拟合，同时使骨干网络未能得到充分适配，难以实现有意义的改进。为解决这一问题，我们引入两项改进：（1）通过将输出头的学习率 η_{head}设置为显著低于骨干网络 η_{bb}，强制实施不平衡优化；（2）采用分阶段策略，先在骨干网络冻结的情况下训练输出头 E_{head} 个训练轮次，随后在输出头固定的情况下进行骨干网络微调E_0 个训练轮次（$E_{\text{head}} \ll E_0$）。这种非对称训练方案促使骨干网络吸收大部分梯度信号，在数据有限的情况下逐步提升表征质量，并使性能接近上界。值得注意的是，我们的策略与 LAE（Gao 等人，2023年）和 SLCA（Zhang 等人，2023年）相反，在这些方法中，骨干网络

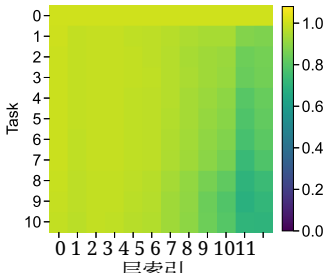


图 5：CKA（Kornblith 等人，2019年）可视化展示了持续学习中首次会话类别的表征在各层间的变化。

更新受到抑制以减轻遗忘。这一对比凸显了音频骨干网络的独特特性：有选择地鼓励骨干网络适配，对于在不引发灾难性遗忘的情况下实现有效迁移至关重要。

后层LoRA。 图 3(c) 的实证分析表明，当适应被限制在更深层时，即冻结早期层而调优后续层，首会话性能可能会得到提升。这与音频模型的层次结构相一致：浅层通常编码与领域无关的时频和声学模式（Niizumi等人，2021年；Cramer等人，2019年），而更深层则捕捉更高层次、更加特定于任务的语义抽象（Liu等人，2020；Baevski等人，2020）。图 5 中的中心核对齐（CKA）可视化进一步支持了这一点，该可视化显示模型在早期层中倾向于保持相对稳定。因此，我们冻结编码器层 1 至 $L_{\text{tune}}-1$，仅适应层 $l \geq L_{\text{tune}}$。边界层L_{tune} 通过全量微调期间的表示偏移分析确定：我们选择 CKA 相对于预训练模型的偏差超过阈值 ρ_{layer} 的最浅层。实现细节见算法 1，并在第 B 节中进一步阐述。

结合上述策略，我们将 LoRA 应用于可调层 $l \in [L_{\text{tune}}, L]$，其中 L 表示总层数。对于每个此类层，适配后的权重定义为:

$$W_1^l = W_0^l + A_1^l B_1^l, \quad \text{for } L_{\text{tune}} \leq l \leq L, \tag{1}$$

其中 $W_0^l \in \mathbb{R}^{d\times d}$ 是预训练权重，$A_1^l \in \mathbb{R}^{d\times r}$和 $B_1^l \in \mathbb{R}^{r\times d}$ 是秩为 $r \ll d$的可训练低秩矩阵。该设计使我们能够高效地细化与任务相关的语义特征，同时保留通用音频表示，从而无需微调整个模型即可实现稳健的 FSA。

解析分类器。为了最大限度地利用先前表征的稳定性，并防止可训练头 $h^t(\cdot)$ 产生偏差累积，我们采用无样例递归解析分类器（McDonnell 等人，2023）进行最终预测。给定一个用于增强特征可区分性的随机投影器 $W_{\text{proj}} \in \mathbb{R}^{D\times D_{\text{proj}}}$（Tran等人，2025），我们可以得到投影后的特征矩阵 $\hat{} Z_t = W_{\text{proj}} Z_t = [\hat{}z_{i,\,t}, \cdots, \hat{}z_{N_t,\,t}]^\top \in \mathbb{R}^{N_t\times D_{\text{proj}}}$ 以及对应的独热标签矩阵 $Y_t \in \mathbb{R}^{N_t\times|\mathcal{Y}_t|}$。自相关矩阵 R^t 以正则化项 $\gamma > 0$初始化，即$R_t = (\hat{}Z_t^\top \hat{}Z_t + \gamma I)^{-1}$。随后使用Woodbury恒等式递归更新:

$$R_t = R_{t-1} - R_{t-1}\hat{Z}_t^\top (I + \hat{Z}_t R_{t-1} \hat{Z}_t^\top)^{-1} \hat{Z}_t R_{t-1}. \tag{2}$$

$\hat{}$ 分类器权重 W_t 随后通过闭式更新规则进行更新:

$$\hat{W}_t = \hat{W}_{t-1} - R_t \hat{Z}_t^\top \hat{Z}_t \hat{W}_{t-1} + R_t \hat{Z}_t^\top Y_t. \tag{3}$$

在推理阶段，新特征 $z_{i,\,t}$ 通过以下方式进行分类：$\hat{}y_{i,\,t} = \phi_t(W_{\text{proj}} z_{i,\,t}) = \hat{}z_{i,\,t}\hat{}W_t$。这种设计能够持续更新决策边界，无需使用样本，同时保留与稳定后的表征空间之间的对齐，从而确保学习过程稳健且不会造成破坏。

3.3 自适应多会话子空间正交参数高效微调

FSA（Zhuang 等人，2022年；McDonnell 等人，2023年）通过在首个任务期间优化预训练表征，缓解音频领域中的灾难性表征偏移。对于领域差距很小的粗粒度数据集，这通常已经足够；但它会严重限制骨干网络的可塑性，从而限制其在细粒度场景中的适用性。在此类情形下，要弥合预训练与下游任务之间的语义错配，就需要学习超越第一会话所建立表征的新表征。为应对这一挑战，我们提出**自适应多会话子空间正交参数高效微调**，使模型能够在多个会话中充分适应，同时避免与此前获得的表征发生破坏性干扰。我们的核心思路是利用多个任务的数据重塑表征空间，使其在领域差距较大时与下游语义对齐，同时保留先前表征空间的几何结构，以维持与解析分类器的兼容性。当进一步适配骨干网络的收益逐渐减小时，我们便冻结骨干网络以实现长期稳定性，并进入阶段3（$t > T_3$）。

多会话适应（MSA）。我们通过引入会话专属 LoRA（Hu 等人，2021年），将 FSA 扩展到多会话设置，使每个会话都能适应骨干网络，同时保留先前会话的参数。具体而言，对于每个会话 $t \in (1, T_3]$，我们在基础权重 W_0 上增加新的低秩更新 $A_t B_t$，形成会话专属参数

$W_t = W_0 + \sum_{\tau=0}^{t-1} B_\tau A_\tau + B_t A_t$，其中 $\{A_\tau, B_\tau\}_{\tau=0}^{t-1}$ 被冻结以防止追溯性干扰。然而，更新仍可能与早期表征不一致，导致决策边界发生灾难性遗忘。为此，我们旨在确保骨干网络更新 g_{update} 不会显著改变过去任务的表征。形式上，对于任意旧样本 $x_{i,\tau} \in \mathcal{X}_\tau$，若满足 $\tau < t$，表征变化应满足

$$\Delta f_t(x_{i,\tau}) = -\eta_{\text{bb}}\, g_{\text{update}}^\top x_{i,\tau} \approx 0, \tag{4}$$

其中，η_{bb} 是骨干网络的学习率。为实现这一点，我们约束 g_{update} 位于先前表征张成的子空间的零空间中。具体而言，令 g_{original} 表示根据当前模型预测 g_t（f_t（$\mathcal{X}_t$））与其标签 $\mathcal{Y}_t$ 之间的交叉熵损失 $\mathcal{L}_{\text{ce}}$ 计算得到的梯度。随后，我们将投影后的更新定义为

$$g_{\text{update}} = P_{\mathcal{U}_t}\, g_{\text{original}} = P_{\mathcal{U}_t} \nabla_\theta \mathcal{L}_{\text{ce}}(g_t(f_t(\mathcal{X}_t)), \mathcal{Y}_t), \tag{5}$$

其中，$P_{\mathcal{U}_t} = U_t U_t^\top$，而 U_t 是张成所有先前获取的表征的零子空间 $\mathcal{U}_t$ 的标准正交基。该投影可确保适应更新对旧样本的影响最小，从而避免已学习表征发生扭曲，并维持持续学习的稳定性。

为计算零空间 $\mathcal{U}_t$，一种简单的方法是存储来自 $\mathcal{X}_1$、.....、$\mathcal{X}_{t-1}$ 的所有历史特征（Wang 等人，2021年；2024b），但这会带来巨大的存储开销。受 LoRA 减法（Liu 与 Chang，2025；Ilharco 等人，2023）的启发，我们改为通过减去所有先前的 LoRA 参数 $W_{\text{unlearn}_t} = W_0 - \sum_{\tau=0}^{t-1} A_\tau B_\tau$ 来构建一个未学习模型。

为提高计算效率，我们随后通过 $X_t^{\text{ucov}} = f_{\text{取消学习}t}(\mathcal{X}_t)^\top f_{\text{取消学习}t}(\mathcal{X}_t) \in \mathbb{R}^{D\times D}$ 计算当前会话特征的未中心化协方差矩阵，该矩阵与 $f_t^{\text{unlearn}}(\mathcal{X}_t)^\top \in \mathbb{R}^{N_t \times D}$ 共享相同的主子空间，其中 $f_{\text{取消学习}t}(\cdot)$ 表示使用 $W_{\text{取消学习}t}$ 的冻结特征提取器，D 为特征维度。对 X_t^{ucov} 执行奇异值分解（SVD）后，我们得到其特征分解：$X_t^{\text{ucov}} = U_t^l \Lambda_t^l (U_t^l)^\top$，并将逐层投影算子定义为 $P_{\mathcal{U}_t^l} = U^{l,\,t\,(1:m)} (U^{l,\,(1:\,m)\,t})^\top$，其中

$$m = \arg\max_m \frac{\sum_{i=1}^m \lambda_i}{\sum_{i=1}^D \lambda_i} > \rho_{\text{svd}},\ \lambda_i : i\text{-th max diagonal entry of } \Lambda_t^l. \tag{6}$$

尽管采用了正交投影，持续适应仍可能累积偏移，从而细微地降低早期会话兼容性，尤其是在类别规模较小时。我们通过实验证实，控制已见样本的累计数量能够在稳定性–可塑性权衡之间取得良好平衡。具体而言，我们将阈值定义为 $T_3 = \text{argmax}_{T_3} \sum_{i=0}^{T_3} N_t > N_{\text{stop}}$，其中 N_{stop} 是用于控制稳定性–可塑性权衡的阈值。骨干网络稳定后，我们通过冻结骨干网络以保留已学习的知识，并在此后仅更新解析分类器，从而转入阶段3。

边界感知正则化。虽然 MSA 保留了先前会话的表征，但可能会降低学习新会话所需的可塑性（Liu 与 Chang，2025），尤其是在新类别与旧类别的表征变得纠缠时。这种纠缠可能导致更新后的决策边界混淆过去的表征。为增强区分度，我们引入边界感知正则化，以促进类内紧凑性和类间间隔扩大。

具体而言，对于会话 t 中的每个输入 $x_{i,\,t}$，我们使用时频掩蔽（Park 等人，2019）生成扰动样本 $\tilde{x}_{i,\,t}^k = \mathcal{Q}(x_{i,\,t},\ r_T,\ r_F)$，其中 $\mathcal{Q}$ 以比率 $r_T \le R_T$ 和 $r_F \le R_F$ 随机掩蔽时间和频率维度，每个输入生成 N_p 个扰动样本。为检测易处于边界的样本，我们检查这些扰动样本是否会被临时模型 θ_{temp} 持续误分类。该模型冻结了前一个模型 θ_{t-1} 的骨干网络，仅调整分类头以适应会话 t。我们将此类样本纳入边界集 $\mathcal{B}_t$：

$$\mathcal{B}_t = \Big\{ \bar{f}_{t-1}(x_{i,t}) \;\Big|\; \frac{1}{N_p} \sum_{k=1}^{N_p} \mathbf{1}\big[\hat{y}_{\theta_{\text{temp}}}(\tilde{x}_{i,t}^k) \neq y_{i,t}\big] > \rho_p \Big\}. \tag{7}$$

其中 $\mathbf{1}(\cdot)$ 是指示函数，ρ_p 是误分类阈值。

To r对表征进行正则化，我们为每个干净输入及其扰动定义如下损失 :

$$\mathcal{L}_{\text{reg}}(i) = \max\left(0,\quad \delta + \frac{1}{|\mathcal{S}_i|} \sum_{u\in\mathcal{S}_i} \big\|\bar{f}_t(u) - \bar{\mu}(x_c)\big\|_2^2 - \min_{b\in\mathcal{B}_t} \big\|\bar{f}_t(x_{i,t}) - b\big\|_2^2\right), \tag{8}$$

其中，$\mathcal{S}_i = \{x_{i,t}, \tilde{x}_{i,t}^1, \ldots, \tilde{x}_{i,t}^{N_p}\}$ 和 $\bar{\mu}(x_c)$ 是类别 c 的质心。这会将特征拉近其类中心，并使其远离附近的边界点，从而减轻未来的混淆。

4 实验

4.1 实验设置

我们采用 EAT（Chen 等人，2024）作为默认实现。EAT 是一种通用自监督音频骨干网络，在 AudioSet-2M（∼ 5000 小时）（Gemmeke 等人，2017）上进行了预训练。EAT 是一种基于频谱图的掩码预测模型，采用视觉 Transformer 架构，包含 12 个 Transformer（Vaswani 等，2017）块。我们在 3 个粗粒度数据集和 3 个细粒度数据集上评测性能，涵盖环境音、语音和音乐。所有数据集均随机按 8:2 的比率划分为训练集和测试集。更多详情见第 D 节。

粗粒度音频持续学习基准。**ESC-50**（Piczak，2015）由来自 50 个类别的 2000 个样本组成（每个样本 5 秒），涵盖了吠叫、雨声、门铃和锯木等多种环境声音。**UrbanSound8K（US8K）**（Salamon 等人，2014）包含来自 10 个城市声音类别的 8732 个样本（每个样本最长 4 秒），代表了空调声和汽车喇叭声等典型城市声音。**Speech Command v2（SC2）**（Warden，2018）包含 10.5 万段一秒录音，涵盖 35 个语音类别，用于关键词识别。对于持续学习，ESC-50 被划分为 10 个会话，每个会话包含 5 个类别；US8K 被划分为 5 个会话，每个会话包含 2 个类别；SC2 被划分为 7 个会话，每个会话包含 5 个类别。这些数据集要么与预训练模型的领域相匹配，要么已在第一会话中得到良好适配。

细粒度音频持续学习基准。**TIMIT**（Garofolo 等人，1993）是一个经典的语音语料库，包含来自 8 种美国方言的 630名说话人，每人朗读 10 个语音学内容丰富的句子，总计约 6300 条语音。我们将 TIMIT 重新构建为持续学习说话人识别基准，设有 2 种配置：**TIMIT-2**（315 个任务，每个任务包含 2 名说话人）和 **TIMIT-3**（210 个任务，每个任务包含 3 名说话人）。**VocalSet**（Wilkins 等人，2018）是一个专为演唱技巧识别和歌手识别整理的数据集。其中包含来自 20 位歌手的约 3560 段音频片段（≈10 小时），涵盖颤音和强声唱法等 16 种演唱技巧。为平衡数据集，我们针对每种技巧随机抽取 79 段音频（64 段用于训练，15 段用于测试），并进一步将其划分为 8 个会话，每个会话包含 2 个类别。

基线方法。我们采用 EAT（Chen 等人，2024）骨干网络，该网络使用 ∼5,000 小时音频在 AudioSet-2M（Gemmeke 等人，2017）上进行预训练。EAT 遵循视觉 Transformer（ViT）的设计，包含 12 个 Transformer（Vaswani 等人，2017）模块。为验证预训练音频持续学习模型的效果，我们考虑了为视觉领域开发的最先进水平（SoTA）基线，包括：（1）基于PEFT的方法，如 L2P（Wang 等人，2022c）、DualPrompt（Wang 等人，2022b）、S-Prompt++（Wang 等人，2022a）、LoRASub（Liu 与 Chang，2025）和 HiDe-PET（Wang 等，2023；2025）；（2）基于统计的方法，如最近原型分类（NPC）（Rebuffi 等，2017b）、RanPAC（McDonnell 等人，2023）和 ACL（Zhuang 等人，2022）。

4.2 实验结果

总体性能。我们使用自监督预训练音频模型 EAT（Chen 等人，2024），在六个音频持续学习基准上将 PACE 与最先进方法进行比较。如表 2 所示，PACE 始终优于所有基线，在粗粒度和细粒度基准测试中均取得最佳性能。此外，我们还针对非人类音频和跨域评估开展了额外的基准测试，并使用另一种自监督预训练音频模型 SSLAM（Alex 等人，2025）进一步验证了 PACE 的有效性，相关结果分别见 E.3 节和 E.5 节。

在**粗粒度基准**（ESC-50、US8K、SC2）上，LoRA 联合训练通过充分利用任务特定监督取得了出色的结果，凸显了骨干网络在非持续学习条件下的潜在可塑性。然而，L2P、DualPrompt 和 S-Prompt++ 等基于提示的方法在持续学习中表现不佳，因为它们依赖共享提示键，而这些键极易受到遗忘的影响。混合使用任务共享组件和任务特定组件往往会导致表征偏移，有时表现甚至不如朴素的参数高效微调。HiDe-PET 通过特征回放在一定程度上缓解了分类器遗忘，但其有效性有限，因为存储的特征本身也会持续发生表征偏移。LoRASub 在一定程度上减轻了漂移，但仍无法避免持续的分类器退化，并且在较长的任务序列（如 TIMIT）中需要进行参数扩展。相比之下，RanPAC 和 ACL 等基于统计的方法会冻结骨干网络，并依靠 FSA 与解析分类器，通过避免漂移实现更稳健的性能。尽管如此，它们仍受表征饱和所限，尤其是在适应早期。PACE 通过层感知调优和自适应子空间正交正则化克服了这一局限，

表 2：不同方法在六个音频持续学习基准测试上的平均 top-1 准确率（%）示例。

方法	粗粒度			细粒度		
	ESC-50	US8K	SC2	TIMIT-2	TIMIT-3	VocalSet
朴素方法						
EAT（LoRA）+联合训练	96.50	98.07	95.91	95.22	95.22	76.65
EAT（冻结）+线性探针	40.50	49.99	32.84	2.30	4.92	13.82
EAT（提示）+线性探针（Lester 等，2021年）	58.75	44.98	41.54	1.67	1.43	22.04
EAT（适配器）+线性探针（Houlsby 等人，2019年）	59.25	38.76	55.33	6.22	14.12	16.45
EAT（LoRA）+线性探针（Hu 等人，2021年）	64.00	49.68	30.56	0.00	2.62	15.79
基于*PEFT*的持续学习方法						
L2P（Wang 等人，2022年c）	39.50	38.75	14.70	1.50	2.53	20.39
DualPrompt（Wang 等人，2022年b）	57.00	42.40	21.92	5.87	10.00	12.50
S-Prompt++（Wang 等人，2022年a）	55.00	42.57	27.23	6.43	8.25	17.76
HiDe-Prompt（Wang 等人，2023年）	83.75	79.89	40.10	47.78	49.60	48.36
HiDe-LoRA（Wang 等人，2025年）	88.75	76.48	33.66	47.30	49.60	46.05
HiDe-Adapter（Wang 等人，2025年）	82.75	78.03	33.71	7.14	12.22	49.67
LoRASub（Liu 与 Chang，2025年）	57.50	57.81	34.24	0.00	0.00	24.01
基于统计的持续学习方法						
最近类均值（NCM）（Rebuffi 等，2017b）	33.25	36.09	9.30	6.90	6.83	32.89
使用 FSA 的最近类均值（NCM）	49.00	42.44	57.60	23.97	34.68	34.53
ACL（高阶）（Zhuang 等人，2022年）	90.00	95.98	80.29	75.56	75.56	62.50
不使用 FSA 的 RanPAC（高阶）	92.50	96.49	81.22	75.87	75.87	61.51
RanPAC（高阶）（McDonnell 等人，2023年）	92.25	97.08	90.53	85.63	89.92	62.82
PACE（我们的方法）	**95.75**	**97.49**	**91.87**	**90.95**	**94.05**	**69.08**

表3：我们提出的改进FSA在粗粒度数据集上的消融结果。

方法	ESC-50	US8K	SC2
不使用 FSA	92.50	96.49	81.22
朴素 FSA	$92.25^{-0.27\%}$	$97.08^{+0.61\%}$	$90.53^{+11.46\%}$
低学习率	$93.75^{+1.35\%}$	$97.35^{+0.89\%}$	$90.95^{+11.98\%}$
学习与冻结	$94.50^{+2.16\%}$	$97.38^{+0.92\%}$	$91.30^{+12.41\%}$
我们的 FSA	$95.75^{+3.51\%}$	$97.49^{+1.04\%}$	$91.87^{+13.11\%}$

表 4：我们在细粒度数据集上对关键组件进行的消融结果。

方法	TIMIT-2	TIMIT-3	VocalSet
Ours	90.95	94.05	69.08
我们的（无 FSA）	$75.87^{-16.57\%}$	$75.87^{-19.35\%}$	$61.51^{-10.97\%}$
我们的（无 MSA）	$85.63^{-5.86\%}$	$89.92^{-4.40\%}$	$62.82^{-9.06\%}$
我们的（无 $\mathcal{L}_{\text{reg}}$	$89.21^{-1.91\%}$	$93.73^{-0.34\%}$	$66.78^{-3.33\%}$
无GP的Ours	$88.01^{-3.23\%}$	$89.05^{-5.31\%}$	$58.55^{-15.26\%}$

在保持稳定性的同时增强表征可塑性。因此，它始终能缩小与联合训练上界之间的差距（例如，在 ESC-50 上为 0.75%，在 US8K 上为 0.58%）。

在细粒度基准（TIMIT-2&3、VocalSet）上，我们观察到所有基线（包括联合训练）的性能都显著下降，这凸显了这些任务中固有的领域差距。TIMIT尤其容易受到不稳定性的影响，因为其类别数量众多（最多达630个）。基于提示的方法在这些设置下始终无法奏效。基于统计的方法虽然更加稳定，但与联合训练相比仍存在较大差距。相比之下，PACE显著缩小了这一差距，因为仅依靠基于FSA的统计方法不足以有效地使表征与下游领域对齐。通过将MSA融入FSA，同时缓解表征遗忘，并结合音频专用PEFT和定制的扰动损失，PACE获得了更具判别性和稳定性的表征。因此，在细粒度场景下，PACE展现出更强的预训练与下游任务之间的对齐能力，其与联合训练之间的性能差距显著缩小（在TIMIT-2和TIMIT-3上分别为4.27%和1.17%，在VocalSet上为7.57%）。

消融实验。由于粗粒度任务仅靠FSA即可得到良好处理，而细粒度任务需要通过MSA进一步适应，我们在粗粒度数据集上评估FSA的设计选择（表3），并在细粒度数据集上对所有核心组件进行消融（表4）。

表3表明，我们改进的FSA（见第3.2节）优于朴素 FSA，这得益于结合了两项关键设计选择：受限分类头学习（通过低学习率和提前冻结分类头实现）以及对后续 Transformer 层的受控适应。为进一步考察层冻结阈值的敏感性，我们通过网格搜索在细粒度数据集上选取 $\rho_{\text{layer}} = 0.94$。该阈值决定了适应过程中哪些层保持冻结（见图6(a)）。如图6(b)所示，恰好冻结这些层可获得最佳性能，这证实了后层自适应的重要性和有效性。除FSA外，表4还表明，完整的PACE模型得益于其核心组件的互补作用，包括MSA

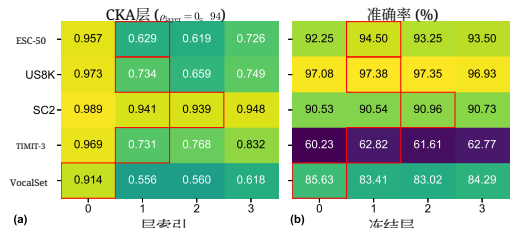


图6：我们的FSA策略中冻结层数量（L_{tune}）的敏感性。

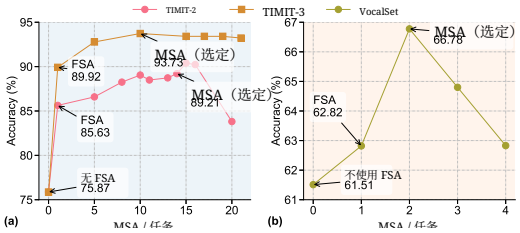


图7：我们的MSA策略中适应会话数量（L_3）的敏感性。

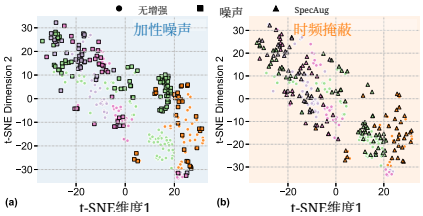


图8：比较不同扰动效果的t-SNE可视化。

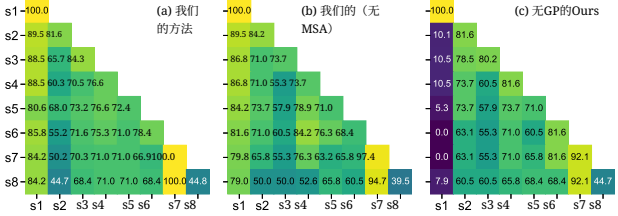


图9：跨会话准确率的热力图可视化：(a) Ours，(b) 我们的（无 MSA），以及 (c) 无GP的Ours。

（见第 3.3 节）、梯度投影和边界感知正则化（$\mathcal{L}_{\text{reg}}$）。每个组件都带来独特的优势，移除其中任何一项都会导致细粒度任务的性能明显下降。我们还考察了MSA策略中适应会话数量的敏感性。如图7所示，在较大范围的适应会话设置下，性能均保持稳定，表明PACE对这一超参数具有鲁棒性，且我们的早停机制能够有效确定合适的适应时长，更多细节见第E.1节。我们还在第E.6节考察了其他超参数的敏感性。

可视化。为了更好地理解我们的方法，我们提供了额外的可视化分析。在图8中，我们使用t-SNE比较扰动效果：加性噪声（见图8(a)）显著扭曲了数据流形，而时频掩蔽（见图8(b)）保留了局部邻域结构和类别一致性，因此适合用于边界感知正则化。图9可视化了跨会话的准确率，突出了MSA和梯度投影的稳定作用。在没有梯度投影的情况下（见图9(c)），模型遭受严重的灾难性遗忘；例如，在完成会话8之后，会话1类别上的准确率从其初始的100%下降到7.9%，表明存在显著的表征漂移和边界坍塌。相比之下，完整的PACE（见图9(a)）在整个过程中保持了较高的准确率，表明梯度投影和边界感知正则化有效地约束了跨会话干扰并保留了早期知识。

5 结论

尽管采用参数高效微调（PEFT）的预训练模型（PTMs）已推动视觉持续学习取得显著进展，但由于上下游不匹配严重，将其直接应用于音频领域仍面临根本性挑战。通过系统的基准测试，我们发现了音频持续学习中的独特难题，例如早期适应阶段的表征饱和，以及细粒度场景中的表征偏移。为解决这些问题，我们提出了 PACE，这是一个统一框架，结合了选择性首会话适应、自适应子空间正交参数高效微调和边界感知扰动，以增强表征对齐并维持适当的可塑性–稳定性权衡。PACE 在六个不同的音频持续学习基准测试中取得了最先进的性能。尽管其适应时间略长于 RanPAC，但总体训练成本仍显著低于以往基于参数高效微调的基线方法，展现了其在粗粒度和细粒度场景中的有效性。除技术贡献外，我们的工作还为预训练音频模型的稳健持续适应奠定了基础，对语音识别、音频描述、智能家居、环境声音理解等现实应用具有广泛意义。

致谢。本研究得到国家自然科学基金项目编号62406160和北京市自然科学基金 L247011 资助。

伦理声明。我确认，我和本研究的所有共同作者均已阅读并承诺遵守 ICLR 道德准则。

可复现性声明。我们已提供源代码及清晰的说明，并将在论文被接收后发布这些内容。

参考文献

Ossama Abdel-Hamid、Abdel-rahman Mohamed、Hui Jiang、Li Deng、Gerald Penn 和 Dong Yu。用于语音识别的卷积神经网络。*IEEE/ACM* 音频、语音与语言处理汇刊，22(10)：1533–1545，2014。

Tony Alex、Sara Atito、Armin Mustafa、Muhammad Awais 和 Philip JB Jackson。SSLAM：利用音频混合增强自监督模型，以处理多声部声景。载于国际学习表征会议，2025年。

Alexei Baevski、Yuhao Zhou、Abdelrahman Mohamed 和 Michael Auli。wav2vec 2.0：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。

Ye Bai、Jingping Chen、Jitong Chen、Wei Chen、Zhuo Chen、Chuang Ding、Linhao Dong、Qianqian Dong、Yujiao Du、Kepan Gao 等。Seed-ASR：通过基于大语言模型的语音识别理解多样化的语音与语境。*arXiv* 预印本 *arXiv:2407.04675*，2024。

Hangbo Bao、Li Dong、Songhao Piao 和 Furu Wei。BEiT：图像变换器的 BERT 预训练。载于国际学习表征会议。

Ruchi Bhatt、Pratibha Kumari、Dwarikanath Mahapatra、Abdulmotaleb El Saddik 和 Mukesh Saini。音频分析持续学习场景与策略的特征分析。*arXiv* 预印本 *arXiv:2407.00465*，2024。

Umberto Cappellazzo、Daniele Falavigna 和 Alessio Brutti。通过适配器软混合高效微调音频频谱 Transformer。*arXiv* 预印本 *arXiv:2402.00828*，2024。

Heng-Jui Chang、Saurabhchand Bhati、James Glass 和 Alexander H Liu。Usad：通过蒸馏实现通用语音与音频表征。*arXiv* 预印本 *arXiv:2506.18843*，2025年。

Sanyuan Chen、Chengyi Wang、Zhengyang Chen、Yu Wu、Shujie Liu、Zhuo Chen、Jinyu Li、Naoyuki Kanda、Takuya Yoshioka、Xiong Xiao 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年。

Wenxi Chen、Yuzhe Liang、Ziyang Ma、Zhisheng Zheng 和 Xie Chen。Eat：采用高效音频 Transformer 的自监督预训练。发表于国际人工智能联合会议，2024。

Aurora Linh Cramer、Ho-Hsiang Wu、Justin Salamon 和 Juan Pablo Bello。看、听并了解更多：深度音频嵌入的设计选择。发表于*IEEE*国际声学、语音与信号处理会议，第 3852–3856 页。IEEE，2019年。

Arthur Douillard、Alexandre Ram´e、Guillaume Couairon 和 Matthieu Cord。Dytox：采用动态令牌扩展的持续学习 Transformer。载于*IEEE/CVF* 计算机视觉与模式识别会议论文集，第 9285–9295 页，2022年。

Baruch Epstein 和Ron Meir。自编码器无监督与半监督学习的泛化界。*arXiv* 预印本 *arXiv:1902.01449*，2019年。

Qiankun Gao、Chen Zhao、Yifan Sun、Teng Xi、Gang Zhang、Bernard Ghanem 和 Jian Zhang。采用通用参数高效微调的统一持续学习框架。载于*IEEE/CVF*国际计算机视觉会议论文集，第 11483–11493 页，2023年。

John S Garofolo、Lori F Lamel、William M Fisher、David S Pallett、Nancy L Dahlgren、Victor Zue 和 Jonathan G Fiscus。TIMIT 语音声学-语音学连续语料库。（无标题），1993年。Jort F Gemmeke、Daniel PW Ellis、Dylan Freedman、Aren Jansen、Wade Lawrence、R Channing Moore、Manoj Plakal 和 Marvin Ritter。Audio Set：音频事件本体与人工标注数据集。载于*IEEE*国际声学、语音与信号处理会议，第 776–780 页。IEEE，2017年。Yuan Gong、Yu-An Chung 和 James Glass。AST：音频频谱图变换器。载于*INTERSPEECH*，第 571–575 页，2021年。Yuan Gong、Cheng-I Lai、Yu-An Chung 和 James Glass。SSAST：自监督音频频谱图变换器。载于*AAAI*人工智能会议论文集，第 36 卷，第 10699–10709 页，2022年。Neil Houlsby、Andrei Giurgiu、Stanislaw Jastrzebski、Bruna Morrone、Quentin De Laroussilhe、An-drea Gesmundo、Mona Attariyan 和 Sylvain Gelly。面向自然语言处理的参数高效迁移学习。载于国际机器学习会议，第 2790–2799 页。PMLR，2019年。Edward J Hu、Yelong Shen、Phillip Wallis、Zeyuan Allen-Zhu、Yuanzhi Li、Shean Wang、Lu Wang 和 Weizhu Chen。LoRA：大语言模型的低秩适配。*arXiv* 预印本 *arXiv:2106.09685*，2021年。Gabriel Ilharco、Marco Tulio Ribeiro、Mitchell Wortsman、Ludwig Schmidt、Hannaneh Hajishirzi 和 Ali Farhadi。使用任务算术编辑模型。载于国际学习表征会议，2023年。Paul Janson、Wenxuan Zhang、Rahaf Aljundi 和 Mohamed Elhoseiny。一个简单的基线方法：质疑持续学习中使用预训练模型的必要性。载于*NeurIPS*分布偏移研讨会：连接方法与应用。Yuxuan Jiang、Zehua Chen、Zeqian Ju、Chang Li、Weibei Dou 和 Jun Zhu。FreeAudio：无需训练的时间规划，实现可控长篇文本到音频生成。载于第*33*届 *ACM*多媒体国际会议论文集，第 9871–9880 页，2025年。Menelaos Kanakis、David Bruggemann、Suman Saha、Stamatios Georgoulis、Anton Obukhov 和 Luc Van Gool。通过卷积重参数化实现无任务干扰的增量式多任务学习。载于欧洲计算机视觉会议，第 689–707 页。Springer，2020年。James Kirkpatrick、Razvan Pascanu、Neil Rabinowitz、Joel Veness、Guillaume Desjardins、Andrei A Rusu、Kieran Milan、John Quan、Tiago Ramalho、Agnieszka Grabska-Barwinska 等。克服神经网络中的灾难性遗忘。美国国家科学院院刊，2017年。Simon Kornblith、Mohammad Norouzi、Honglak Lee 和 Geoffrey Hinton。重新审视神经网络表征的相似性。载于国际机器学习会议，第 3519–3529 页。PMLR，2019年。Minh Le、An Nguyen、Huy Nguyen、Trang Nguyen、Trang Pham、Linh Van Ngo 和 Nhat Ho。专家混合与基于提示的持续学习相结合。神经信息处理系统进展，第 37 卷：119025–119062，2024年。Brian Lester、Rami Al-Rfou 和 Noah Constant。规模对参数高效提示调优的力量。*arXiv* 预印本 *arXiv:2104.08691*，2021年。Chang Li、Ruoyu Wang、Lijuan Liu、Jun Du、Yixuan Sun、Zilu Guo、Zhenrong Zhang、Yuan Jiang、Jianqing Gao 和 Feng Ma。用于增强音乐生成的质量感知掩码扩散变换器。*arXiv* 预印本 *arXiv:2405.15863*，2024a。Chang Li、Zehua Chen、Liyuan Wang 和 Jun Zhu。基于潜在桥接模型的音频超分辨率。*arXiv* 预印本 *arXiv:2509.17609*，2025a。Tingle Li、Baihe Huang、Xiaobin Zhuang、Dongya Jia、Jiawei Chen、Yuping Wang、Zhuo Chen、Gopala Anumanchipalli 和 Yuxuan Wang。让物体发声：交互式、物体感知的图像到音频生成。*arXiv* 预印本 *arXiv:2506.04214*，2025b。

Xiang Lisa Li 和 Percy Liang。前缀调优：优化用于生成的连续提示。*arXiv* 预印本 *arXiv:2101.00190*，2021年。Xingjian Li、Haoyi Xiong、Haozhe An、Cheng-Zhong Xu 和 Dejing Dou。Rifle：通过重新初始化全连接层，实现深度迁移学习中的深层反向传播。载于国际机器学习会议，第 6010–6019 页。PMLR，2020。Yi Li 和 Plamen P Angelov。通过原型对比掩码自编码器实现可解释音视频表征学习。载于*NeurIPS* 自监督学习理论与实践研讨会，2024年。

Yizhi Li、Ruibin Yuan、Ge Zhang、Yinghao Ma、Xingran Chen、Hanzhi Yin、Chenghao Xiao、Chenghua Lin、Anton Ragni、Emmanouil Benetos 等。MERT：采用大规模自监督训练的声学音乐理解模型。载于国际学习表征会议，2024b。

Yan-Shuo Liang 和 Wu-Jun Li。Inflora：用于持续学习的无干扰低秩自适应。在 *IEEE/CVF*计算机视觉与模式识别会议论文集，第 23638–23647 页，2024。

Alexander H Liu、Sang-gil Lee、Chao-Han Huck Yang、Yuan Gong、Yu-Chiang Frank Wang、James R Glass、Rafael Valle 和 Bryan Catanzaro。UniWav：迈向语音表征学习与生成的统一预训练。*arXiv* 预印本 *arXiv:2503.00733*，2025年。

Andy T Liu、Shu-wen Yang、Po-Han Chi、Po-chun Hsu 和 Hung-yi Lee。Mockingjay：利用深度双向Transformer编码器进行无监督语音表征学习。在*IEEE*国际声学、语音与信号处理会议，第 6419–6423 页。IEEE，2020a。

Xialei Liu、Chenshen Wu、Mikel Menta、Luis Herranz、Bogdan Raducanu、Andrew D Bagdanov、Shangling Jui 和 Joost van de Weijer。面向类增量学习的生成式特征回放。载于*IEEE/CVF*计算机视觉与模式识别会议研讨会论文集，第 226–227 页，2020b。

Xuan Liu 和 Xiaobin Chang。用于无示例持续学习抗漂移空间的 LoRA 减法。载于计算机视觉与模式识别会议论文集，第 15308–15318 页，2025年。

Mark D McDonnell、Dong Gong、Amin Parvaneh、Ehsan Abbasnejad 和 Anton Van den Hengel。Ranpac：用于持续学习的随机投影与预训练模型。神经信息处理系统进展，36:12022–12053，2023。Shentong Mo、Weiguo Pian 和 Yapeng Tian。持续音视频学习的类增量分组网络。载于*IEEE/CVF*国际计算机视觉会议论文集，第7788–7798页，2023。Manjunath Mulimani 和 Annamaria Mesaros。使用预训练音频模型进行音频分类的在线增量学习。*arXiv*预印本 *arXiv:2508.20732*，2025。Satoshi Nakamura、Kazuo Hiyane、Futoshi Asano、Takanobu Nishiura 和 Takeshi Yamada。用于声音场景理解和免提语音识别的真实环境声学声音数据库。2000。Daisuke Niizumi、Daiki Takeuchi、Yasunori Ohishi、Noboru Harada 和 Kunio Kashino。用于音频的 BYOL：通用音频表征的自监督学习。载于国际神经网络联合会议，第1–8页。IEEE，2021。Daniel S Park、William Chan、Yu Zhang、Chung-Cheng Chiu、Barret Zoph、Ekin D Cubuk 和 Quoc V Le。Specaugment：一种用于自动语音识别的简单数据增强方法。*arXiv*预印本 *arXiv:1904.08779*，2019。Weiguo Pian、Shentong Mo、Yunhui Guo 和 Yapeng Tian。音视频类增量学习。载于*IEEE/CVF*国际计算机视觉会议论文集，第7799–7811页，2023。

Karol J Piczak。ESC：用于环境声音分类的数据集。载于*ACM*多媒体国际会议论文集，第1015–1018页，2015。Sylvestre-Alvise Rebuffi、Hakan Bilen 和 Andrea Vedaldi。使用残差适配器学习多种视觉域。神经信息处理系统进展，30，2017a。Sylvestre-Alvise Rebuffi、Alexander Kolesnikov、Georg Sperl 和 Christoph H Lampert。iCaRL：增量分类器与表征学习。载于*IEEE*计算机视觉与模式识别会议论文集，第2001–2010页，2017b。David Rolnick、Arun Ahuja、Jonathan Schwarz、Timothy Lillicrap 和 Gregory Wayne。持续学习的经验回放。神经信息处理系统进展，32，2019。Justin Salamon、Christopher Jacoby 和 Juan Pablo Bello。用于城市声音研究的数据集与分类体系。载于*ACM*多媒体国际会议论文集，第1041–1044页，2014。Riyansha Singh、Parinita Nema 和 Vinod K Kurmi。利用对比表征实现音频分类中稳健的少样本类增量学习。*arXiv*预印本 *arXiv:2407.19265*，2024。Bob L Sturm。GTZAN音乐流派数据集分析。载于第二届国际 *ACM* 音乐信息检索与以用户为中心的多模态策略研讨会论文集，第7–12页，2012。Bob L Sturm。GTZAN数据集：其内容、缺陷、对评估的影响及未来用途。*arXiv*预印本 *arXiv:1306.1461*，2013。Afrina Tabassum、Dung Tran、Trung Dang、Ismini Lourentzou 和 Kazuhito Koishida。UAMix-MAE：利用无监督音频混合高效微调预训练音频变换器。载于*IEEE*国际声学、语音与信号处理会议，第5435–5439页。IEEE，2024。Quyen Tran、Tung Lam Tran、Khanh Doan、Toan Tran、Dinh Phung、Khoat Than 和 Trung Le。为基于预训练的持续学习增强多视图。载于国际学习表征会议，2025。Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszkoreit、Llion Jones、Aidan N Gomez、Łukasz Kaiser 和 Illia Polosukhin。注意力机制就是你所需要的一切。神经信息处理系统进展，30，2017。Helin Wang、Yuexian Zou、Dading Chong 和 Wenwu Wang。采用并行时频注意力的环境声音分类。*INTERSPEECH 2020*论文集。Liyuan Wang、Jingyi Xie、Xingxing Zhang、Mingyi Huang、Hang Su 和 Jun Zhu。基于提示的持续学习的层次分解：重新思考被遮蔽的次优性。神经信息处理系统进展，36:69054–69076，2023。Liyuan Wang、Xingxing Zhang、Hang Su 和 Jun Zhu。持续学习综合综述：理论、方法与应用。*IEEE*模式分析与机器智能汇刊，46(8):5362–5383，2024a。Liyuan Wang、Jingyi Xie、Xingxing Zhang、Hang Su 和 Jun Zhu。HiDe-PET：通过参数高效微调的层次分解实现持续学习。*IEEE*模式分析与机器智能汇刊，2025。Shipeng Wang、Xiaorong Li、Jian Sun 和 Zongben Xu。在特征协方差的零空间中训练持续学习网络。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第184–193页，2021。Shipeng Wang、Xiaorong Li、Jian Sun 和 Zongben Xu。利用自监督在特征协方差的零空间中训练网络以实现增量学习。*IEEE* 模式分析与机器智能汇刊，2024b。

Yabin Wang、Zhiwu Huang 和 XiaopengHong。使用预训练变换器进行 S-prompts 学习：一种适用于领域增量学习的奥卡姆剃刀。神经信息处理系统进展，35：5682–5695，2022a。

Zifeng Wang、Zizhao Zhang、Sayna Ebrahimi、Ruoxi Sun、Han Zhang、Chen-Yu Lee、Xiaoqi Ren、Guolong Su、Vincent Perot、Jennifer Dy 等。DualPrompt：用于无复习持续学习的互补提示。载于欧洲计算机视觉会议，第 631–648 页。Springer，2022b。

Zifeng Wang、Zizhao Zhang、Chen-Yu Lee、Han Zhang、Ruoxi Sun、Xiaoqi Ren、Guolong Su、Vin- cent Perot、Jennifer Dy 和 Tomas Pfister。学习如何为持续学习设计提示。载于*IEEE/CVF* 计算机视觉与模式识别会议论文集，第 139–149 页，2022c。

Pete Warden。《Speech Commands：用于有限词汇语音识别的数据集》。*arXiv* 预印本 *arXiv:1804.03209*，2018。

Julia Wilkins、Prem Seetharaman、Alison Wahl 和 Bryan Pardo。VocalSet：一个歌唱声音数据集 *ISMIR*，第 468–474 页，2018。
集。发表于

Yusong Wu、Christos Tsirigotis、Ke Chen、Cheng-Zhi Anna Huang、Aaron Courville、Oriol Nieto、Prem Seetharaman 和 Justin Salamon。FLAM：逐帧语言-音频建模。发表于机器学习国际会议。

Yi Yuan、Xubo Liu、Haohe Liu、Mark D Plumbley 和 Wenwu Wang。Flowsep：基于整流流匹配的语言查询式声音分离。载于*IEEE*国际声学、语音与信号处理会议，第 1–5 页。IEEE，2025年。

Friedemann Zenke、Ben Poole 和 SuryaGanguli。通过突触智能实现持续学习。载于国际机器学习会议，第 3987–3995 页。PMLR，2017年。

Gengwei Zhang、Liyuan Wang、Guoliang Kang、Ling Chen 和 Yunchao Wei。SLCA：用于预训练模型持续学习的慢速学习器与分类器对齐。载于*IEEE/CVF*国际计算机视觉会议论文集，第 19148–19158 页，2023年。

Gengwei Zhang、Liyuan Wang、Guoliang Kang、Ling Chen 和 Yunchao Wei。SLCA++：释放预训练下顺序微调在持续学习中的潜力。*arXiv*预印本 *arXiv:2408.08295*，2024。

Linglan Zhao、Xuerui Zhang、Ke Yan、Shouhong Ding和Weiran Huang。Safe：面向预训练模型持续学习的慢速与快速参数高效微调。神经信息处理系统进展，37：113772–113796，2024。

Da-Wei Zhou、Hai-Long Sun、Jingyi Ning、Han-Jia Ye和De-Chuan Zhan。基于预训练模型的持续学习：综述。载于人工智能国际联合会议，2024a。

Da-Wei Zhou、Hai-Long Sun、Han-Jia Ye和De-Chuan Zhan。面向基于预训练模型的类增量学习的可扩展子空间集成。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第23554–23564页，2024b。

Kanglei Zhou、Ruizhi Cai、Liyuan Wang、Hubert PH Shum和Xiaohui Liang。动作质量评估综合综述：方法与基准。*arXiv*预印本 *arXiv:2412.11149*，2024c。

Kanglei Zhou、Liyuan Wang、Xingxing Zhang、Hubert PH Shum、Frederick WB Li、Jianguo Li 和 Xiaohui Liang。Magr：用于持续动作质量评估的流形对齐图正则化。发表于欧洲计算机视觉会议，第 375–392 页。Springer，2024d。

Kanglei Zhou、Zikai Hao、Liyuan Wang 和 Xiaohui Liang。面向虚拟现实中360度视频持续感知质量评估的自适应分数对齐学习。*IEEE*可视化与计算机图形学汇刊，31(5):2880–2890，2025a。

Kanglei Zhou、Qingyi Pan、Xingxing Zhang、Hubert PH Shum、Frederick WB Li、Xiaohui Liang 和 Liyuan Wang。基于自适应流形对齐图正则化的持续动作质量评估。*arXiv*预印本 *arXiv:2510.06842*，2025b。

Huiping Zhuang、Zhenyu Weng、Hongxin Wei、Renchunzi Xie、Kar-Ann Toh 和 Zhiping Lin。ACIL：通过绝对记忆和隐私保护实现解析类增量学习。神经信息处理系统进展，35：11602–11614，2022年。

XiaodanZhuang、Xi Zhou、Mark AHasegawa-Johnson 和 Thomas S Huang。真实世界声学事件检测。《模式识别快报》，31(12):1543–1551，2010。

声明

致谢。 本工作得到国家自然科学基金项目编号62406160和北京自然科学基金L247011的资助。

道德声明。 我确认，我和本工作的所有合著者均已阅读并承诺遵守 ICLR 道德准则。

可复现性声明。 我们已附上源代码及清晰的使用说明，并将在论文被接收后发布代码。

大语言模型辅助。 我们使用大语言模型润色了稿件。作者已全面审阅和编辑所有内容，并对已发表的作品承担全部责任。

B 改进的首会话适应伪代码

算法 1： 改进的首会话适应

需要： 预训练骨干网络 $f_0, E_0, E_{\text{head}}$，CKA 阈值 ρ_{layer}，学习率 $\eta_{\text{bb}} > \eta_{\text{head}}$

阶段 A（检测层边界）：
$\theta_1^{\text{pre}} \leftarrow \theta_0$
对 训练轮次 $= 1, \cdots, E_0$ **执行**
  $\theta_1^{\text{pre}} \leftarrow \theta_1^{\text{pre}} - \eta_{\text{bb}} \nabla_\theta \mathcal{L}_{\text{ce}}(h_{\text{ce}}(f_1^{\text{pre}}(\mathcal{X}_1)),\ \mathcal{Y}_1)$ {对所有层进行参数高效微调}
结束循环
循环 $k = 1, \cdots, L$ **do**
  $s^k \leftarrow \text{CKA}(f_1^{\text{pre}}(\mathcal{X}_1)^k, f_0(\mathcal{X}_1)^k)$
结束循环
$k^\star \leftarrow$ 参数最大值$_k[\, s^k < \rho_{\text{layer}}]$
阶段 B（头部预热，η_{head}）：
遍历训练轮次 $= 1, \cdots, E_{\text{head}}$ **执行**
  $h_1 \leftarrow h_1 - \eta_{\text{head}} \nabla_h \mathcal{L}_{\text{ce}}(h_1(f_0(\mathcal{X}_1)),\ \mathcal{Y}_1)$
结束循环
阶段 C（骨干网络适配，η_{bb}）：
冻结 h_{ce}，θ_0；$\theta_1 \leftarrow \theta_0 + \theta_1^{\text{LoRA}}$
遍历训练轮次 $= 1, \cdots, E_0$ **执行**
  $\theta_1 \leftarrow \theta_1 - \eta_{\text{bb}} \nabla_\theta \mathcal{L}_{\text{ce}}(h_1(f_1(\mathcal{X}_1)),\ \mathcal{Y}_1)$ {仅层 $k^\star{+}1(L_{\text{tune}}){:}L$ 在 f_{θ_1}可训练}
结束循环
阶段 D（解析阶段）：
冻结 θ_1；丢弃 h_1；在 $\phi(\cdot) = f_1(\cdot)$ 上初始化统计量；继续参见第 3.3 节。

算法 1 概述了改进的首会话适应流程，旨在为音频持续学习中的 FSA 建立有效且稳定的表征迁移。阶段 A 对所有层进行轻量级参数高效微调更新，以探测敏感性，随后使用 CKA 衡量 f_1^{pre} 与预训练骨干网络 f_0 在每一层的相似度。边界层 $L_{\text{tune}} = k^\star + 1$ 被确定为表征相似度开始低于阈值 ρ_{layer}的位置，这表明适当的适应应从该处开始。随后，阶段 B 以较小的学习率η_{head} 和较少的训练轮数 E_{head}训练受限的分类器头 h_1。阶段 C 仅对 $k^\star$ 之后的更深层采用 LoRA 风格的更新进行适应，并使用较大的学习率 η_{bb}，同时冻结浅层和分类器头，以保留预训练知识并增强骨干网络适应。最后，阶段 D 冻结适配后的骨干网络，丢弃临时头，并为解析阶段初始化统计量（参见第 3.3 节）。

C 相关工作

持续学习（CL）使模型能够持续学习，同时减轻对先前已学习知识的灾难性遗忘（Wang 等人，2024a）。持续学习领域的大多数研究都聚焦于视觉领域，代表性方法包括基于正则化的方法（Kirkpatrick 等人，2017；Zenke 等人，2017）、基于回放的方法（Rolnick 等人，2019；Liu 等人，2020b）以及基于架构的方法（Douillard

等人，2022；Kanakis 等人，2020）。近来，基于预训练模型（PTMs）的持续学习受到越来越多的关注，因为预训练模型能够提供强大的通用初始化（Zhou 等人，2024a），无需从头开始训练。这种范式通常采用参数高效微调（PEFT）策略（Rebuffi 等人，2017a；Li & Liang，2021；Hu 等人，2021），在冻结的骨干网络上添加轻量级模块（Wang 等人，2022c;b；Tran 等人，2025；Le 等人，2024；Gao 等人，2023；Zhao 等人，2024），或将其与正则化原则相结合（Liu & Chang，2025；Liang & Li，2024；Zhou 等人，2024b）。在视觉领域，预训练模型编码的表征相对稳定，且语义结构良好（Janson等人），因此不同学习会话之间的表征偏移较小（见图 1(b)），使基于参数高效微调的持续学习十分有效。然而，这一范式在音频领域的有效性仍缺乏充分研究，亟待系统性探究。

音频识别涵盖语音识别（Abdel-Hamid 等人，2014；Bai 等人，2024）、声学事件检测（Yuan 等人，2025年；Li 等人，2025b）和音乐理解（Li 等人，2024b）等任务。近期的预训练模型（Gong 等人，2022年；Chen 等人，2024；Alex 等人，2025年；Liu 等人，2025年；Chang 等人，2025年）在这些领域均取得了出色表现，并展现出为音频生成任务提供条件信号或正则化信号的潜力（Li 等人，2025年a；2024a；Jiang 等人，2025年）。然而，这些进展通常假设能够获取完整数据进行离线训练，忽视了现实场景中不断演变的非平稳分布（Bhatt 等人，2024）。关于音频持续学习的新兴研究（Mulimani & Mesaros，2025；Cappellazzo 等人，2024；Singh 等人，2024；Mo 等人，2023；Pian 等人，2023）往往借鉴视觉领域的方法或采用简化的设置，却未能应对音频特有的挑战。与图像不同，音频依赖细粒度的频谱线索，使其表征对分布偏移高度敏感，也容易发生灾难性遗忘（见图 1(a)）。因此，需要制定适用于音频的持续学习策略，使其兼容预训练骨干网络，并符合音频的频谱特性。

D 额外实验细节

所有实验均在 NVIDIA A800 GPU 上进行。输入大小为 512×128。每段音频片段均截取前 5.12 秒，批量大小为 24。对于所有任务，我们设置 $\eta_{\text{bb}} = 0.05$， $\eta_{\text{head}} = 0.01$，并通过网格搜索确定训练轮数。对于每项任务，我们报告所有任务的平均准确率，即 $\text{Acc} = \frac{1}{T}\sum_{t=1}^{T}\text{Acc}_t$，作为持续学习性能的主要衡量指标。对于超参数，为了改进首次会话适配，我们设置 $E_{\text{head}} = 1$， $\rho_{\text{layer}} = 0.94$；对于多会话适配，设置 $\rho_{\text{svd}} = 0.99$， $N_{\text{stop}} = 220$；对于边界感知扰动，设置 $N_p = 20$， $\rho_p = 0.3$， $\delta = 0.25$；解析分类器则设置为 $D_{\text{proj}} = 8192$。我们还在表5中提供了各数据集的详细统计量及其对应的训练轮数 E_0。

表5：我们实验所用数据集的统计量。

数据集	训练轮次	类别	会话	总计训练样本	总计测试样本
ESC-50（Piczak，2015）	10	50	10	1600	400
US8K（Salamon 等人，2014）	15	10	5	8000	2000
SC2（Warden，2018）	1	35	7	84651	21178
TIMIT-2（Garofolo 等人，1993）	30	630	315	5040	1260
TIMIT-3（Garofolo 等人，1993）	30	630	210	5040	1260
VocalSet（Wilkins 等人，2018）	6	16	8	1216	304

E 补充实验结果

E.1 MSA 停止会话

当我们设置 N_{stop} 来估计截断点时，会为每个任务得到一个特定的截断点，如表 6 所示。为了进一步评估这一任务级停止机制的有效性，我们针对所选 $T_3 = t^*$ 附近的相邻会话进行消融实验（即 t^*-1、 t^*-2、 t^*+1、 t^*+2）。我们观察到，在大多数数据集（US8K、SC2、TIMIT-3、VocalSet）上， t^* 取得了局部最优结果，

数据集	已选择 t^*	不含 FSA	FSA	t^*-2	t^*-1	t^*	t^*+1	t^*+2
ESC-50	2	92.50	92.25	N/A	**92.25**	91.75	91.75	92.00
US8K	1	96.49	97.08	N/A	N/A	**97.08**	95.15	81.44
SC2	1	81.22	90.53	N/A	N/A	**90.53**	89.46	81.14
TIMIT-2	14	75.87	85.63	88.49	88.73	89.21	**90.40**	90.23
TIMIT-3	10	75.87	89.92	92.78	**93.73**	**93.73**	93.41	93.41
VocalSet	2	61.51	62.82	N/A	61.51	**66.78**	64.80	62.83

表 6：不同适应会话下，FSA 和 MSA（无 $\mathcal{L}_{\text{reg}}$ 无改进FSA）在不同数据集上的结果。

显著优于不使用 FSA 和原始FSA，同时在现实的持续学习设置中（TIMIT-2/3）较长会话序列上保持了更稳定的性能。尽管在 ESC-50 和 TIMIT-2 上，t^* 并非总是全局最优，但以朴素方式应用 MSA，与最优结果相比并未造成明显性能下降，这验证了我们停止策略的必要性。此外，值得注意的是，随着适应所涉及的会话数量增加，性能通常会有所下降；当会话相对较小时，这种现象更加明显。我们将此归因于骨干网络已充分适应下游领域，因此，额外适应带来的边际收益不足以抵消累积的表征漂移，而这些漂移会导致先前任务上的遗忘，这凸显了停止准则的重要性。

E.2 学习与遗忘

为阐明持续更新期间稳定性与可塑性如何相互作用，我们在 **VocalSet**基准上评估了有梯度投影（GP）和无梯度投影的多会话适应（MSA）。我们报告五项关键指标：(1) 会话 $t \le T_3$ 中的遗忘，(2) 通过平均最大准确率衡量的可塑性，(3) T_3之前的平均准确率，(4) T_3之后的平均准确率，以及 (5) 后向迁移（*BWT*）。结果如表 7 所示。

表 7：VocalSet 上有无梯度投影的 MSA 对比。

设置	遗忘	可塑性	平均准确率（$t \le T_3$）	平均准确率（$t > T_3$）	BWT
FSA	27.63	**92.10**	**64.48**	60.52	–10.90
不使用梯度投影的 MSA	57.90	**92.10**	34.21	66.67	–9.02
MSA（使用 GP）	**23.69**	88.16	64.47	**70.62**	**–7.14**

结果表明，梯度投影在保持可塑性的同时大幅减少了遗忘。尽管两个 MSA 变体都达到了相近的最高准确率，但移除 GP 会导致显著的表征漂移，进而造成严重遗忘和后向迁移性能下降。相比之下，梯度投影稳定的 MSA 能够保持早期会话中的稳定性能，并在后续会话中取得更高的准确率，这表明投影能有效限制破坏性的特征偏移，同时保留适应能力。

E.3 非人类音频的补充基准测试与跨域评估

为更全面地评估持续学习在多种音频条件下的性能，我们引入了两个补充基准测试：（1）一个细粒度非人类音频数据集；（2）一个跨领域声音—语音数据集。这些基准测试针对类内变化显著且存在分布不匹配的场景，而这正是本研究所探讨挑战的核心。

GTZAN：细粒度非人类音频。许多与乐器相关的数据集（例如 GTMUSIC（Sturm，2012））的语义粒度较粗。初步分析表明，非音乐预训练骨干网络（如 EAT）结合单会话适应步骤，准确率最高可达**99.8%**，这说明其分布复杂度极低，因此不太适合用于评估持续学习

。相比之下，**GTZAN** 数据集（Sturm，2013）包含 10 种音乐流派，类内多样性更丰富，呈现出更贴近现实的细粒度分布偏移，能够更好地反映非人类音频持续学习中不断演变的语义结构。

ESC–Speech：跨域声音–语音基准。为评估异质域组成下的鲁棒性，我们通过组合 ESC-50（环境声音）和 SpeechCommands V2（口语词汇），构建了一个名为 **ESC–Speech** 的合成跨域数据集。这种混合在非语言声学事件与人类语音之间造成了显著的域不匹配，为预训练音频模型构成了一个具有挑战性的场景。

实验协议。两个基准测试均采用针对各自特点设计的持续学习设置进行评估：GTZAN 使用 5 阶段划分，每类包含 20 个样本，以捕捉其细粒度音乐变化；而 ESC–Speech 采用 10 阶段划分，每类包含 50 个样本，以体现其跨域（声音–语音）偏移。结果见表8。

表8：更多样化基准测试上的持续学习结果。

数据集	属性	L2P	HiDe-Prompt	RanPAC	PACE
GTZAN	非人类声音（乐器）	10.00	51.00	73.00	**78.00**
ESC–Speech	跨域（ESC + SC2）	21.50	52.58	57.00	**72.17**

观察。 在这两个基准测试中，PACE 均取得了最高性能，展现出对类内变化（GTZAN）和跨域分布偏移（ESC–Speech）的强大适应能力。音乐域评估进一步揭示，预训练音频模型存在显著的表征偏移，这加大了持续适应的难度。最初为基于视觉的持续学习设计的方法（例如 L2P）在此类场景中的效果有限。相比之下，PACE 的多会话适应策略能够逐步对齐不断演变的语义空间，在域内和域间分布变化下均带来稳定提升。

E.4 计算成本与训练开销

PACE 的计算开销主要来自早期阶段的骨干网络更新，这些更新是获得语义对齐表征所必需的。在最初几个适应会话之后，该方法转为解析分类器，并结合轻量级子空间正交更新，从而使后续所有会话的训练都非常高效。

在粗粒度基准上，改进FSA 提供了充分的对齐，无需额外的骨干网络更新。在这些设置下，PACE 以与标准微调基本相同的计算成本，达到接近联合训练性能。

在细粒度基准上，PACE 相对于 RanPAC 引入了适度的开销，但仍比 HiDe-Prompt 等基于提示的方法高效得多；后者会在所有会话中反复优化提示或低秩适配器。结果报告于表 9。

表 9：细粒度数据集上的训练时间与训练时间比率。训练时间在单块 GPU 上计算，比率相对于 RanPAC 计算。

数据集	训练时间（秒/样本）	PACE / RanPAC	HiDe-Prompt / RanPAC
VocalSet	0.22	1.22	5.44
TIMIT-3	0.12	2.96	146.98
TIMIT-2	0.31	3.13	124.19

这些结果表明，PACE 在显著提升准确率的同时，只带来了适度的计算开销。相比之下，HiDe-Prompt 的成本高出 $5\times$–$40\times$，性能却更差。尽管我们的方法在细粒度数据集上的训练时间长于 RanPAC，但在单块 GPU 上仍能达到约 0.2 秒/样本的平均速度，实际应用中效率很高。总体而言，PACE 实现了良好的效率与效果平衡，尤其适用于细粒度持续学习场景。

E.5 其他预训练骨干网络

为评估 PACE 在单一预训练检查点之外的通用性，我们进一步使用**SSLAM** （Alex 等人，2025）对该框架进行了基准测试。SSLAM 是一种近期提出的源感知预训练音频模型，基于复调混合音频训练，并与 EAT （Chen 等人，2024）采用相同的视觉 Transformer 骨干网络。实验在两个粗粒度数据集和两个细粒度数据集上进行，并将 PACE 与 L2P、HiDe-Prompt 和 RanPAC 进行了比较。结果见表10。

表10：使用 SSLAM 骨干网络在粗粒度和细粒度数据集上的持续学习性能。

数据集	粒度	骨干网络	L2P	HiDe-Prompt	RanPAC	PACE
ESC-50	粗粒度	SSLAM	40.50	82.00	95.75	**96.25**
SC2	粗粒度	SSLAM	15.24	37.85	88.59	**90.39**
VocalSet	Fine	SSLAM	17.76	47.22	63.83	**68.42**
TIMIT-2	Fine	SSLAM	0.32	46.24	90.08	**93.81**

结果表明，PACE 在两种预训练骨干网络及各个粒度级别上均持续优于所有基线。这表明，预训练音频模型中发现的挑战（如表征饱和和语义错位）并非 EAT 所特有，在 SSLAM 等更新的模型中同样会出现。此外，PACE 在所有设置下的强劲表现凸显了其作为音频持续学习框架的鲁棒性与通用适用性。

E.6 超参 数选择与敏感性分析 析

本研究在 PACE 的设计中引入了多个连续变量。为明确这些选择流程，我们提供了全面的敏感性分析。

SVD 阈值 ρ_{svd}。 遵循先前基于参数高效微调的持续学习研究（Liu 与 Chang，2025；Wang 等人，2021），我们直接采用相同的 SVD 能量阈值。在所有数据集上，将 ρ_{svd} 调整至 $[0.90, 0.99]$ 范围内，结果几乎完全相同，这证实该超参数主要影响数值秩的选择，对学习动态的影响极小。

停止阈值 N_{stop} **和冻结阈值** ρ_{layer}。 这两个超参数控制 MSA，并在细粒度的**VocalSet** 和 **TIMIT-3** 数据集上联合调优。我们仅使用 FSA，在 $\{0.90, 0.92, 0.94, 0.96, 0.98, 1.00\}$ 范围内搜索，并选取最能平衡适应性与稳定性的值。固定ρ_{layer} 后，我们使用不带正则化的 MSA 调优 N_{stop}。如图 7(a) 和 7(b) 所示，模型性能在 200 至 250 个梯度步数之间趋于平稳。尽管在 VocalSet 上，$N_{\text{stop}} = 250$ 略高于最优 T_3，但其性能仍比 FSA 高约 2%，说明该方法能够适应小幅偏差。

验证 ρ_{layer}。 如图 6 所示，所选值 $\rho_{\text{layer}} = 0.94$ 成功避免了可适应的更深层过早冻结，同时保持了稳定的低级特征提取能力，尤其适用于 TIMIT-3 等细粒度数据集。

敏感性分析。 我们还进一步考察了其余所有连续超参数的敏感性。为提高可读性，我们仅呈现平均性能，并突出显示所选配置。结果汇总于表 11 至表 14 中。

总体而言，PACE 在广泛的超参数配置下均表现出较强的鲁棒性。所选值可作为稳定的默认设置，仅需进行最少调参，凸显了该框架在多样化音频场景中用于持续学习的实用性。

E.7 边界感知正则化中时频掩蔽的理论依据

选择时频掩蔽，是基于音频信号的频谱结构以及边界感知正则化的要求。我们采用 SpecAugment 式掩蔽（Park 等人，2019年）作为一种直接作用于时频域的保标签扰动。通过有选择地去除狭窄的时间或频谱带，该操作会对频谱图造成局部失真，同时保留信号的全局语义身份。这类扰动

表 11：MSA 停止阈值 N_{stop} 在三个细粒度数据集上的敏感性 s.

方法	TIMIT-2	TIMIT-3	VocalSet	Avg.
FSA	85.63	89.92	62.82	**79.46**
MSA（150）	88.25	92.78	62.82	81.28
MSA（含 175）	89.05	93.24	66.78	83.02
MSA（含 200）	88.49	93.33	66.78	82.87
MSA（含 210）	89.21	93.33	66.78	83.11
MSA（含 220）	89.21	93.73	66.78	**83.24**
MSA（含 230）	90.39	93.73	66.78	83.63
MSA（含 240）	90.23	93.73	64.80	82.92
MSA w/ 250	90.23	93.41	64.80	82.81

表 12：层冻结阈值 ρ_{layer} 在不同数据集上的敏感性。

ρ_{layer}	TIMIT-3	SC2	VocalSet	ESC-50	US8K	Avg.
0.90	83.41	90.96	62.82	94.50	97.38	85.01
0.91	83.41	90.96	62.82	94.50	97.38	85.01
0.92	85.63	90.96	62.82	94.50	97.38	86.26
0.93	85.63	90.96	62.82	94.50	97.38	86.26
0.94	85.63	90.96	62.82	94.50	97.38	**86.26**
0.95	85.63	90.54	62.82	94.50	97.38	86.17
0.96	85.63	90.54	62.82	92.25	97.38	85.72
0.97	85.63	90.54	60.23	92.25	97.38	85.21
0.98	85.63	90.54	60.23	92.25	97.08	85.15
0.99	85.63	90.53	60.23	92.25	97.08	85.14
1.00	85.63	90.53	60.23	92.25	97.08	85.14
w/o	85.63	90.53	60.23	92.25	97.08	85.14

表 13：边界比率 ρ_{ratio} 在 VocalSet 基准测试上的敏感性。

比率	0.20	0.25	0.30	**0.35**	0.40	0.45	0.50
Acc	65.79	67.43	67.43	**69.08**	68.75	64.47	64.47

表14：增强强度对 VocalSet 基准测试的敏感性。

强度	0.10	0.20	0.30	0.40	**0.50**	0.60	0.70	0.80	0.90
Acc	65.79	66.12	64.80	63.49	**69.08**	68.10	65.79	64.47	61.51

能够对每个样本周围的邻域进行受控探索，从而构造易处于边界的样本，同时不破坏类别一致性。

为评估所提出的正则化项 $\mathcal{L}_{\text{reg}}$ 是否影响鲁棒性，我们进一步在测试时应用和不应用时频掩蔽的情况下，对 **VocalSet** 上的模型进行了评估。结果汇总于表15。使用 $\mathcal{L}_{\text{reg}}$ 训练的模型在掩蔽后的测试输入下性能下降幅度明显更小（1.83% 对比 6.26%），这表明边界感知正则化增强了鲁棒性，而非损害鲁棒性。

E.8 粗粒度与细粒度音频分类案例研究

为进一步阐明细粒度与粗粒度数据集之间的差异，我们在图10中提供了案例研究。在这两类数据集上使用PEFT-FT，我们跟踪了第一会话中三个类别在随后三个会话中的预测动态，同时报告预训练模型中的最大类间距离（在第一会话中测量）以及平均特征

表15：在VocalSet数据集上对时频掩蔽（TFM）鲁棒性的评估。

评估	测试（无TFM）	测试（有TFM）
训练（无TFM）	66.78	60.52
训练（有TFM）	69.08	67.25

会话间偏移。从图10(a)可以明显看出，粗粒度数据集呈现出相对显著的频谱模式（例如*Cat* 中的谐波结构，或*Doorwood knock*中的周期性能量突发）。如此显著的差异带来了较大的类间分离度，使得模型更容易在第一会话中捕捉这些差异；这也反映在表 2 中，我们的改进FSA取得了接近联合训练的性能。因此，应用朴素PEFT-FT时，模型能够在表征未发生显著语义变化的情况下完成适应；与类别距离相比，这对应着较小的跨会话特征偏移。值得注意的是，具有高度鲜明结构的类别（例如*Door wood knock*）在这种朴素适应下更能抵抗遗忘。

对于细粒度场景，如图10(b)所示，我们观察到不同类别具有高度相似的频谱模式。在这种情况下，模型需要区分630名说话人，仅凭时频细节不足以提供具有判别性的信息。相反，模型必须在每个会话中仅利用有限的数据调整其表征，以形成更具判别性的深层特征。这会导致类间分离度相对较小（几乎只有粗粒度数据集的一半），但需要进行更大幅度的表征变化。最终影响是灾难性的：尽管模型最初具有较高的可塑性，并在当前会话中取得了良好的准确率，但朴素PEFT-FT缺乏保留先验知识的记忆容量，导致模型在后续会话中出现严重的错误分类。这凸显了PACE方法的必要性：该方法能够有效利用多个会话中的数据，持续调整网络，同时约束先前表征保持稳定。

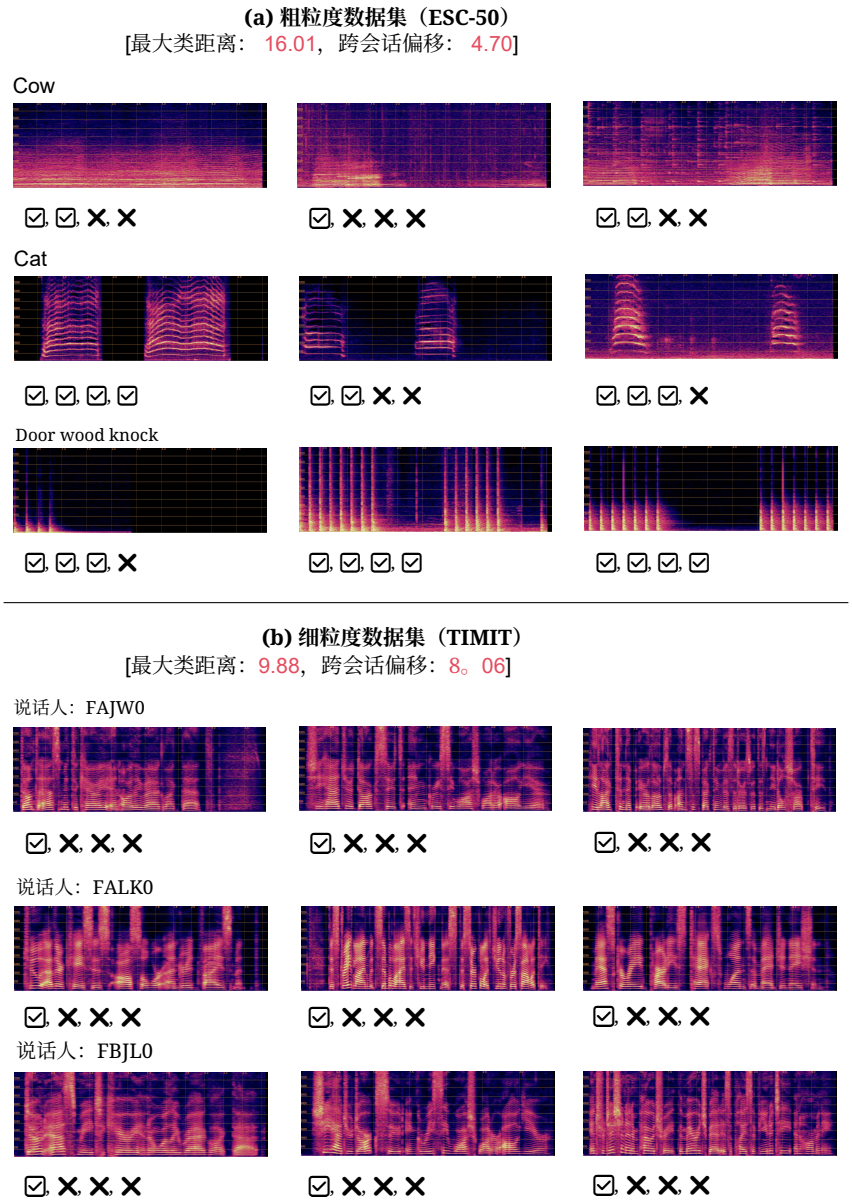


图10：粗粒度数据集（ESC-50）和细粒度数据集（TIMIT-3）的案例研究：第一会话中三个类别的线性幅度谱图，以及它们在朴素PEFT-FT下的性能轨迹。