适配关键之处：面向低资源语言的高效多语言语音识别深度感知适配

Yang Xiao、Eun-Jung Holden、
Ting Dang墨尔本大学
yxiao9550@student.unimelb.edu.au, {eunjung.holden, ting.dang}@unimelb.edu.a u

arXiv:2602.01008v1 [eess.AS] 1 Feb 2026

摘要

近期的语音基础模型在高资源语言的多语言自动语音识别（ASR）方面表现出色，但由于数据稀缺和效率限制，将其适配到低资源语言仍面临挑战。全模型微调计算开销大且容易过拟合，而 LoRA 等参数高效方法会在各层均匀地应用适配，忽略内部表征，从而损害效果和效率。我们分析了多语言自动语音识别模型，发现其呈现 U 形可适配性模式：早期层和晚期层具有语言特定性，需要更多适配；中间层则保留共享语义，所需适配较少。基于这一观察，我们提出 DAMA，即深度感知模型适配框架，可根据各层的作用分配适应能力。DAMA 还引入基于奇异值分解（SVD）的初始化，以约束适配并保留 U 形模式，同时采用冻结的中间层基线以进一步提升效率。在两个基准数据集涵盖的 18 种低资源语言上进行评估后，DAMA 以减少 80% 可训练参数的代价，实现了与最先进方法相当或更高的准确率；在极端数据稀缺条件下，错误率降低了 29%，并且相较于基线方法，显著改善了内存占用、训练时间和计算效率。这些结果凸显了结构感知适配在实现高效、可扩展的多语言自动语音识别方面的优势。

1 引言

近期的语音基础模型（Pratap 等人，2024年；Cui 等人，2025年）在海量多语言语音数据上进行预训练，能够执行多种任务，例如多语言自动语音识别（自动语音识别），尤其擅长处理英语等高资源语言。然而，由于低资源语言或地方方言的（Shi 等人，2024年），数据有限导致其性能大幅下降，使代表性不足的语言之间出现显著差距。

如何利用有限的数据，高效地将这些模型适配到低资源语言，仍是一项重大挑战。在许多真实场景中，例如在新兴市场快速部署语音技术，或为地方方言开发相关技术，都迫切需要兼具准确性与高效率的适配方法。这些方法必须能够利用有限的新数据快速完成适配，在严格的内存和计算限制下运行，并在数据极其匮乏的情况下仍保持出色的性能。

传统的全参数微调在计算和内存方面开销巨大，因此在许多应用中并不实用，尤其是在边缘设备上或资源有限的情况下。低资源语言的标注数据匮乏，使适配过程更加复杂，因为全参数微调往往会导致过拟合或灾难性遗忘（Wang 等人，2024年；Chang 等人，2021年；Yang 等人，2022年；Kwok 等人，2024年）。

近期研究转向参数高效微调（PEFT）（Ding 等人，2023年），例如低秩适配（LoRA）（Hu 等人，2022年），这类方法冻结原始模型参数，仅为目标语言更新一小部分新权重，从而实现更快速、内存效率更高的适配。然而，这些方法以蛮力方式对所有层进行均匀适配，忽视了语言表示的结构，可能会限制其效果和参数效率，在极低资源设置下尤为如此。对于大型模型，即使是标准 LoRA，仍可能需要更新大量参数。

因此，要推动语音基础模型的高效多语言适配，就需要细致理解这些模型如何在内部表征和共享多语言知识。这

本研究首先逐层分析多语言语音表示如何在不同模型层中得到维持并相互作用，揭示出鲜明的可塑性U形模式：早期层和晚期层捕捉语言特定特征，因而更易适配；而中间层则保持语言无关性。这表明，适配新语言时，不同层所需的适配程度各不相同，这一发现挑战了以往研究中认为所有层都同样适合适配的主流假设（Song 等人，2024年；Kwok 等人，2025年）。

受这一 U 形模式的启发，我们提出深度感知模型适配（DAMA），通过三种机制高效且有效地将模型适配到新语言。首先，深度感知秩调度为可塑性较强的早期层和晚期层分配较高的适应能力，同时限制中间层的秩，从而兼顾参数效率与模型结构属性的保留。其次，为约束中间层的适配，我们提出基于 SVD 的初始化，使适配权重初始化在与模型主导权重正交的方向上。这有助于保留共享的语言表征并维持 U 形适配性。最后，为进一步提高效率，尤其是在低资源设置下，我们引入基底保护投影（BPP），冻结部分适配权重，从而在保留关键知识的同时减少可训练参数的数量。

我们使用 Common Voice 和 FLEURS 数据集，在 18 种低资源语言上评估了 DAMA。DAMA 的表现达到或超过最先进的基线方法，同时将参数量减少约 80%。更重要的是，在极低资源设置下（0.5 至 1 小时的数据），其词错误率（WER）相对改善最高可达 29%。效率分析显示，GPU 内存利用率提升了 24%，训练速度加快了 36%。这些结果表明，使适配与模型层属性相契合，能够构建可扩展且参数高效的多语言系统，同时不牺牲准确率。我们的贡献总结如下：

- 我们首次系统地分析了语音基础模型中逐层的多语言表征。我们揭示了语言特异性的 U 形分布，展示了这些模型如何维持并共享跨语言知识。

- 我们提出 DAMA，这是一种新颖的深度感知多语言ASR适配框架，在适配能力与参数效率之间实现了有效平衡，同时保留了必要的多语言知识。

- 我们在 18 种语言上进行的实验表明，所提出的 DAMA 实现了帕累托最优权衡。它在显著减少可训练参数、内存使用量和训练时间的同时，达到了或超过了 SOTA 性能。更重要的是，DAMA 在低资源设置下展现出更强的鲁棒性。

2 相关工作

多语言自动语音识别（MASR）。MASR（Yadav 和 Sitaram，2022年）旨在使用单一的统一基础模型，对多种语言的语音进行转录。近期的发展主要得益于训练数据和模型容量的扩展，例如 Whisper（Radford 等人，2023年）和 MMS（Pratap 等人，2024年）。尽管这些模型在识别高资源语言方面表现出色，但在应用于预训练期间未见过的低资源语言时，其性能会显著下降。由于数据可用性有限，因此需要以高效的方式适配新语言，同时避免完整重新训练所带来的高昂计算成本或灾难性遗忘的风险（Li 等人，2022年）。

高效多语言适应策略。要将语音基础模型适配到新语言，传统上默认采用全量微调。然而，在资源受限环境中，更新所有参数的计算开销高得难以承受。更重要的是，全参数微调（FFT）往往会导致灾难性过拟合：模型以牺牲可泛化的语义知识为代价，记住稀疏数据（Xiao 等人，2022年；Yang 等人，2022年；Peng 和 Xiao，2024年）。一些方法尝试通过限制重要参数的更新来保留先验知识（Xiao 等人，2025年；Xiao 和 Das，2025年）。然而，这些方法往往难以平衡可塑性-稳定性困境，限制了它们有效学习新任务的能力。为实现高效适配，参数高效微调已成为主流范式。低秩适配（LoRA）（Hu 等人，2022年）和适配器（Houlsby 等人，2019年）等方法会冻结预训练骨干网络，并引入少量可训练参数以捕获任务特定偏移。尽管这些方法

取得了成功，但 MASR 中的标准参数高效微调方法忽视语言表征，并以均匀方式适配所有模型层，因而效率和效果均受影响（Song 等人，2024年；Li 等人，2025年；Yang 等人，2025年）。尽管 AdaLoRA（Zhang 等人，2023b）等方法引入了动态秩分配，但它们依赖于根据训练数据得出的敏感度分数。在少样本或低资源设置下，数据不足会使这些敏感度估计变得不稳定，因为其获取依赖计算开销大的搜索。这凸显了针对低资源语音数据，仍缺乏更有效且高效的适应机制。

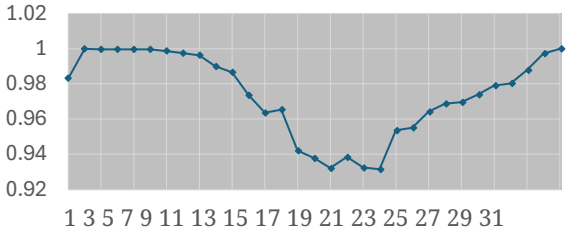
3 语言表征分析

为探究语言表征如何在潜在空间中得以维持和共享，我们对每一层进行线性探测，开展逐层分析，以执行语言识别（LID）任务。如果某一层的表示具有语言特定性，则 LID 准确率会较高；反之，较低的准确率则表明这些表示具有更强的语言无关性。通过此分析，我们能够确定各层保留语言特定信息的程度。

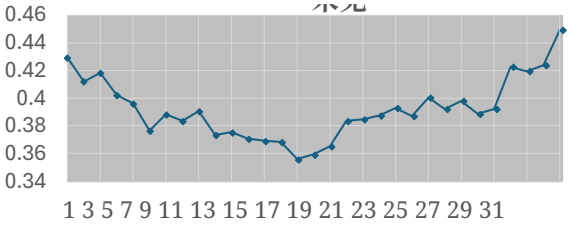
3.1 逐层线性探测

具体而言，我们在编码器-解码器框架内开展分析，该框架常用于近期的语音基础模型（Rad- ford 等人，2023年；Omnilingual 等人，2025年；Peng 等人，2023年）。给定一个语音输入序列 x_t，编码器会处理该输入并生成一系列潜在表示，记作$\mathbf{h}_t = \text{Encoder}(x_t)$。随后，这些编码器输出 $\mathbf{h}_t$ 被馈入解码器 f_θ，其参数由权重 θ确定。解码器由 L层组成，第 l层的中间激活记作 $\mathbf{z}_t^{(l)}$，其中 $l \in \{1, \ldots, L\}$。

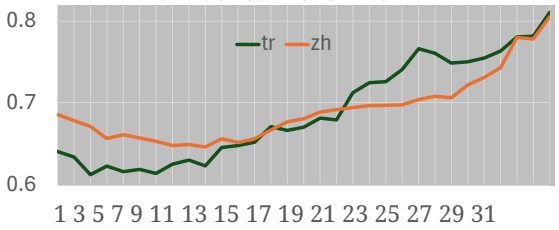
对于每个解码器层 l，我们对表示$\mathbf{z}_t^{(l)}$ 应用线性探测。具体而言，我们在激活值$\mathbf{z}_t^{(l)}$ 上训练线性分类器 $g^{(l)}(\cdot)$，以交叉熵损失执行LID任务。分类准确率定量衡量了语言特定信息在第 l层被编码的程度。这全面展示了语言表示如何在解码器网络的不同深度上被保持、减弱或抽象化。



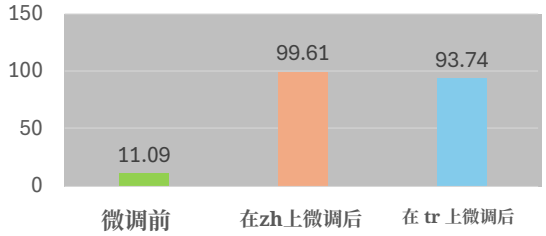
（a）五种已见语言的线性探测准确率。



（b）十种未见语言的线性探测准确率。



（c）五种已见语言在.....之后的线性探测准确率微调。



(d) 在英语上微调后的词错误率。

图 1：不同语言在微调前后的逐层探测。

数据集与语言选择：为确保分析的鲁棒性，我们从 Common Voice（Ardila等人，2020年）数据集中选取了多种语言。我们构建了两个不同的评估组，以测试模型的泛化能力。第一组包含五种在预训练期间已见的语言，这使我们能够衡量模型如何表征已知知识。第二组包含十种未见语言，使我们能够观察模型如何处理全新的语言模式。这些语言的完整列表及其具体详情见附录A。

3.2 U 形层级可塑性

已见语言分析：如图1a所示，已见语言的结果在解码器深度维度上呈现出明显的 U 形模式。早期层（第 1 至第 5 层）和晚期层（第 28 至第 32 层）的准确率接近 100%。这表明模型保留了显著的语言特定标记。相反，中间层的性能明显下降。具体而言，第 17 层附近的准确率降至约 93%。这种相对下降形成了“语义谷”，这表明中间层对语言身份不那么敏感，而更侧重于语言无关语义表示。

未见语言分析： 进一步分析预训练期间未见的语言，也会发现同样的“语义谷”，尽管准确率有所下降。这种一致的 U 形行为证实，解码器会自然地按照依赖深度的可塑性来组织信息。无论涉及何种语言，早期层和晚期层都会捕捉语言特定的语言学特征，而中间层则相对保持语言无关的表示。

值得注意的是，这种 U 形模式与近期基于文本的大语言模型研究结果一致（Kojima 等，2024年；Tang 等，2024年；Wu 等）。我们的结果表明，即使语音训练采用不同的端到端学习范式，这种层级现象在语音基础模型中同样显著。

3.3 微调的影响

我们进一步探究常用的微调如何影响 U 形可塑性。首先，我们在两种截然不同的语言——土耳其语（tr）和普通话（zh）上进行全参数微调，并重复探测分析。图 1c 所示的结果表明，鲜明的 U 形模式已彻底消失，预训练模型中观察到的“语义谷”被持续上升的趋势所取代。中间层原本捕获语言无关表征，如今却被迫编码强烈的语言特定信息。虽然这使模型能够适配新语言，但语义坍缩也伴随着对英语等已见语言的灾难性遗忘，如图 1d 所示，微调后英语的词错误率从 11% 飙升至 93% 以上。此外，这也导致显著的计算成本，包括更大的内存占用和更长的训练时间。

虽然标准 LoRA 提高了适配效率并保留了原始模型权重，但它并未考虑中间层的特殊作用。通过均匀适配所有层，包括负责语言无关处理的层，LoRA 仍可能扰乱关键语义表示，侵蚀已确立的结构，并削弱跨语言泛化能力和鲁棒性。保留 U 形可塑性，尤其是中间层的可塑性，对于保护已学习知识、提高效率至关重要；适配应主要针对早期层和晚期层。

4 提出的 DAMA

为在适配过程中保持 U 形结构，我们提出了高效的深度感知模型适配（DAMA）。首先，我们引入深度感知秩调度方案，为早期和晚期层分配更高的适配容量，同时限制中间层的适配程度。其次，我们提出一种用于中间层 LoRA 的基于 SVD 的初始化，以显式保留语言无关表示，从而实现高效的低秩适配。第三，我们设计了基底保护投影（BPP），保护并冻结中间层的参数，以进一步提高适配效率和稳定性，尤其是在低资源设置下。

4.1 深度感知秩分配

标准 LoRA（Hu 等人，2022年）通过微调每个权重矩阵的低秩更新来适配预训练模型。给定一个权重矩阵为 $W_0 \in \mathbb{R}^{d\times k}$的预训练模型，LoRA 将更新 ΔW 限制为两个低秩矩阵的乘积：$B \in \mathbb{R}^{d\times r}$ 和 $A \in \mathbb{R}^{r\times k}$，其中 $r \ll d,\ k$。前向传播如下：

$$h = (W_0 + \Delta W)x = W_0x + BAx, \quad (1)$$

在训练过程中，W_0 被冻结。A 进行随机初始化，B 被设为零，因此在初始阶段，更新不会产生任何影响。标准 LoRA 为所有层分配相同的秩 r，从而为早期、中期和晚期层分配均匀的适应能力。

相反，DAMA采用遵循U形分布的深度感知秩调度，并通过依赖于层的秩函数 $r(l)$灵活调整适应能力，其依据是

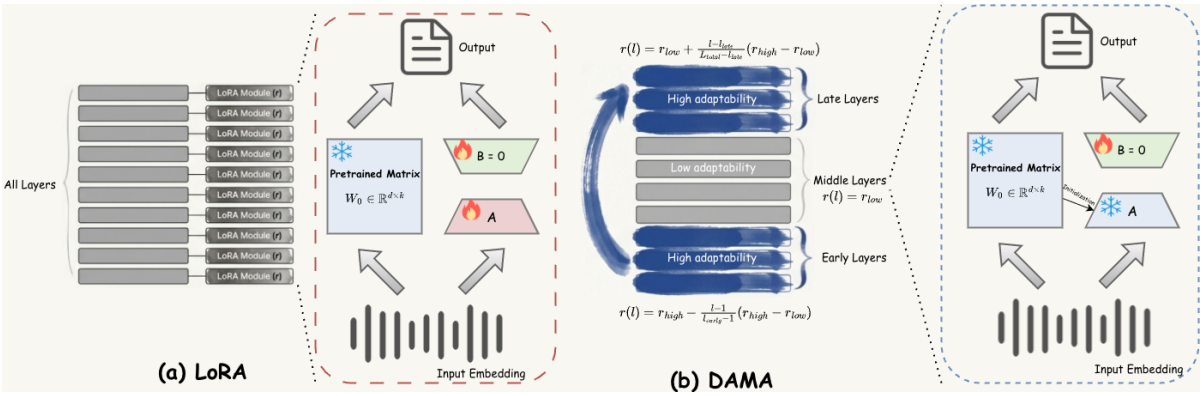


图2：DAMA框架与 LoRA 的对比概览。（a）采用统一秩的标准 LoRA。（b）DAMA框架。具体而言，深度感知秩调度为早期层和晚期层分配较高的可塑性，而基底保护投影则从物理上锁定中期层，以保护“语义谷”。所有层均指解码器中的所有层。解码器处理来自编码器的声学嵌入，并将其转录为输出词元。

每个解码器层的位置 l。具体而言，我们将 L 个解码器层划分为三个区段：早期（l_{early}： $1 \le l \le l_{\text{early}}$）、中期（$l_{\text{mid}}$： $l_{\text{early}} < l < l_{\text{late}}$）和晚期（$l_{\text{late}}$： $l_{\text{late}} \le l \le L_{\text{total}}$），其中 $l_{\text{early}} = \lfloor \theta_1 L_{\text{total}} \rfloor$和 $l_{\text{late}} = \lfloor \theta_2 L_{\text{total}} \rfloor$针对 $0 < \theta_1 < \theta_2 < 1$。我们为每一层分配一个LoRA秩 $r(l)$，并采用以最大秩r_{high} 和最小秩 r_{low}为界的U形调度：

$$r(l_{\text{early}}) = r_{\text{high}} - \frac{l-1}{l_{\text{early}}-1}(r_{\text{high}} - r_{\text{low}}), \quad (2)$$

$$r(l_{\text{mid}}) = r_{\text{low}}, \quad (3)$$

$$r(l_{\text{late}}) = r_{\text{low}} + \frac{l-l_{\text{late}}}{L_{\text{total}}-l_{\text{late}}}(r_{\text{high}} - r_{\text{low}}), \quad (4)$$

如图2所示，该调度为早期层分配更多适配能力，因为模型需要在这些层中适应新目标语言的特定特征；同时也为晚期层分配更多适配能力，因为模型会在这些层中从处理语义表示过渡到生成目标语言中的特定词汇输出；而在中期层中尽量减少适配，以保留语言无关空间。这提高了所有投影模块的效率和稳定性。

4.2 基于 SVD 的初始化

尽管中间层中的低秩 r 有助于保留特征性的 U 形表征结构，但我们进一步采用基于 SVD 的初始化进行适配，以强化这一点。该方法将 LoRA 更新限制在对语言无关语义子空间影响最小的方向上，从而防止这些关键层发生过度适配和语义漂移。不同于标准 LoRA 所采用的随机高斯初始化（可能引入“几何噪声”），基于 SVD 的初始化有助于维持 U 形表征的完整性。

给定预训练语音基础模型中期层的权重矩阵 $W \in \mathbb{R}^{m \times n}$，我们执行 SVD：

$$W = U \Sigma V^T, \quad (5)$$

其中，$\Sigma = \text{diag}(\sigma_1, \ldots, \sigma_{\min(m,n)})$ 按降序排列包含奇异值。前r 个主成分对应于最大的奇异值 $\sigma_1, \ldots, \sigma_r$，捕获了在适配过程中应予保留的语言无关语义知识。

为将适配限制在与该子空间正交的方向上，我们通过选取末尾的奇异向量来选择残差成分 $V_{\text{tail}} = [v_{r+1}, \ldots, v_{\min(m,n)}]$。然后，我们将 LoRA 适配矩阵 A（参见式（1））初始化为：

$$A = V_{\text{tail}}^T, \quad (6)$$

这种初始化将 LoRA 更新限制在与语言无关语义子空间重叠最小的方向上。因此，预训练模型的核心语义结构得以保留，从而降低了适配过程中发生语义漂移的风险。

4.3 基底保护投影（BPP）

为进一步稳定适配并提升计算效率，我们引入 BPP 模块，在中间层冻结 LoRA 矩阵 A，仅更新 B。通过冻结经奇异值分解初始化的 A，我们确保训练更新不会无意间将适配重新引导回

受保护的语义子空间，从而严格保留核心的语言无关表示，使语义漂移几乎不可能发生。与同时更新 A 和B的标准 LoRA 不同，我们的方法仅通过更新 B，将适配限制在秩显著更低的空间中。在极低资源设置下，这一策略尤其有利，因为它大幅减少了可训练参数的数量。因此，过拟合风险得以降低，泛化能力也可能得到提升，尤其是在下游任务数据集规模有限时。

5 实验设置

5.1 数据集

我们使用两个多语言数据集来评估我们的方法，以确保广泛适用性：Common Voice（Ardila等人，2020年）和 FLEURS 数据集（Conneau等人，2023年）。Common Voice 是一个众包语料集合，为多种语言提供经过验证的转写文本。遵循既有方案（Della Libera等人，2024年；Kwok 等人，2025年），我们选取了十种在训练期间未见过的语言，包括卢旺达语、世界语、卡拜尔语、卢干达语、草原马里语、中部库尔德语、阿布哈兹语、库尔曼吉库尔德语、弗里西语和国际语。对于这些语言，我们采用标准的数据划分：十小时用于训练，一小时用于验证，一小时用于测试。

FLEURS 数据集（Conneau等人，2023年）源自 FLoRes 机器翻译基准，包含平行语音和文本。我们从 FLEURS 构建了两个不同的评估组。第一组由“已见薄弱”语言组成，例如印地语和威尔士语，模型曾接触过这些语言，但仍难以有效处理。第二组由“未见”语言组成，包括干达语和索拉尼库尔德语，这些语言对模型而言完全陌生。

为确保评估公平且严谨，我们依据三项严格标准选择了这些语言。首先，我们纳入来自非洲、欧洲和亚洲的语言，以最大限度地提高地理多样性，从而避免结果出现地域偏差。其次，我们优先选择资源有限的语言。这些语言往往缺乏充足的训练数据，其挑战性也高于英语等广泛使用的语言。最后，我们根据任务难度选择语言。值得注意的是，基线 Whisper 模型完全无法转写“未见”组的内容，也无法在“已见但较弱”组中取得良好表现。有关该数据集的详细信息，请参阅附录 B。

5.2 实现细节

我们采用 Whisper large v2 模型（Radford 等人，2023年）作为骨干模型，该模型采用标准的编码器-解码器架构，是一种常用的多语言语音基础模型。遵循先前的基准研究（Della Libera 等人，2024年），我们为未见语言初始化并训练新的词元嵌入，以确保模型能够正确识别这些语言。为评估性能，我们将本方法与微调及最先进的参数高效微调方法进行比较，包括标准 LoRA（Hu 等人，2022年）及其变体：DoRA（Liu 等人，2024年）、LoRA-FA（Zhang 等人，2023a）、LoRA-XS（Bałazy 等人，2024年）、VB-LoRA（Li 等人，2024年）和 AdaLoRA（Zhang 等人，2023b）。我们选择这些方法，以涵盖多种效率策略，例如动态秩分配和权重分解。

对于我们的方法，我们将其应用于所有线性投影矩阵，包括查询、键、值、输出和前馈网络层。我们将r_{high}和 alpha 优化为 32，同时将 r_{low} 设为 8。我们将 θ_1 和 θ_2 分别设为 0.3 和 0.7。训练共进行两个训练轮次，批量大小为 6。我们采用 AdamW 优化器并结合动态学习率调度器。最后，所有推理任务均采用贪心解码策略。我们使用词错误率（WER）准确率指标和效率指标评估系统，效率指标包括参数数量、MACs（乘加运算）和 GPU 显存。MACs 专门衡量可训练适配器模块引入的额外浮点运算，不包括冻结的骨干网络。完整的配置详情和基线设置请参阅附录 B。

6 结果与讨论

6.1 性能比较

在 Common Voice 和 FLEURS 数据集上的实验（表 1）表明，DAMA 在准确率与效率之间实现了更优的权衡。DAMA 的平均词错误率为 39.73%，与表现最强的基线 LoRA（39.71%）基本持平。值得注意的是，DAMA 在不同语言上的鲁棒性更强：它成功避免了 VB-LoRA 等压缩方法在“未见”语言上出现的显著性能下降（43.20% 对 59.81%），并且

表 1：Common Voice 和 FLEURS 上平均词错误率与参数效率的性能比较。最右侧一列单独报告了各方法额外的乘加运算次数。

方法	参数量（M）	未见语言 CommonVoice	未见语言 FLEURS	已见-弱 FLEURS	已见-弱 FLEURS	平均额外乘加运算量（G）
微调	906.5	43.87	50.05	27.55	41.34	-
LoRA (Hu 等人，2022年)	68.2	43.25	47.13	**25.19**	39.71	102.2
DoRA (Liu 等人，2024年)	68.7	43.61	**46.46**	25.24	39.73	1415.4
LoRA-FA (Zhang 等人，2023a)	34.1	46.00	48.56	25.65	41.55	51.1
LoRA-XS (Bałazy 等人，2024年)	**1.3**	57.36	55.23	30.29	50.06	104.2
VB-LoRA（Li 等人，2024年）	4.9	59.81	50.75	28.00	49.59	758.3
AdaLoRA（Zhang 等人，2023b）	51.1	51.66	52.36	28.41	46.02	76.7
DAMA（本文方法）	14.9	**43.20**	47.25	25.26	39.73	**22.3**

表 2：Common Voice（10 种语言）在 2 小时、1 小时和 0.5 小时低资源设置下的平均词错误率。

方法	平均词错误率（0.5 小时）	平均词错误率（1小时）	平均词错误率（2小时）
微调	68.11	61.79	54.60
LoRA (Hu 等人，2022年)	64.84	59.96	54.28
DoRA(Liu 等人，2024年)	64.73	59.93	54.24
LoRA-FA (Zhang 等人，2023a)	68.63	63.23	57.48
LoRA-XS（Bałazy 等人，2024年）	80.57	73.99	67.58
VB-LoRA（Li 等人，2024年）	73.05	76.64	77.81
AdaLoRA（Zhang 等人，2023b）	90.16	72.73	65.94
DAMA（我们的方法）	**64.11**	**58.80**	**54.20**

在“已见但薄弱”语言上的表现优于全量微调（47.25% 对 50.05%）。然而，DAMA 仅使用 1,490 万个参数，约为标准 LoRA 所需参数量（6,820 万个）的五分之一。相比之下，全量微调虽然使用了超过 9 亿个参数，是我们方法的 60 多倍，但词错误率仍显著更高，达到 41.34%。这表明，更新所有参数的效率低下，而且在数据有限时可能会损害源模型。

6.2 低资源情境下的数据效率

我们进一步考察了 DAMA 在数据稀缺场景下的平均词错误率：如表 2 所示，对于每种未见语言，使用 Common Voice 中 0.5 至 2 小时的数据。与其他方法相比，DAMA 保持了出色的泛化能力，而其他方法则明显退化。具体而言，在 1 小时设置下，DAMA 的平均词错误率为 58.80%，优于微调基线的 61.79%。这表明，在低数据量情境下更新所有参数会导致严重过拟合。将 DAMA 与 AdaLoRA 和 LoRA-XS 等方法进行比较，可以看出其鲁棒性。这些基线方法未能有效适配，错误率飙升至 72% 以上。尽管 AdaLoRA 采用动态秩 al-

位置上，它依赖数据驱动的敏感度评分来剪枝参数。在极低资源环境下，由于数据稀疏性，这些评分会变得不可靠，进而导致架构决策不佳——

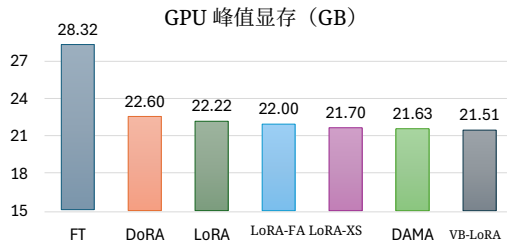


图 3：Common Voice 上所有 10 种语言采用不同适配方法时的平均峰值 GPU 显存使用量。

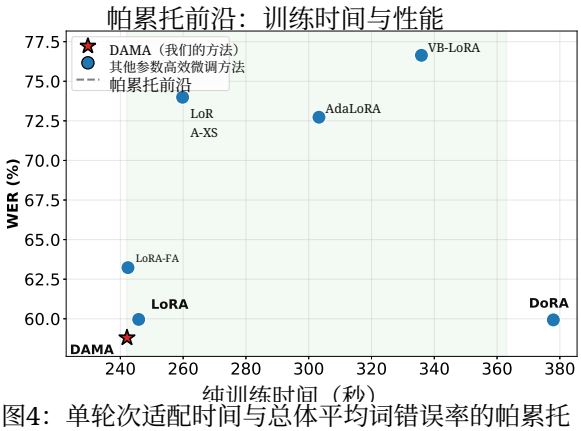


图4：单轮次适配时间与总体平均词错误率的帕累托前沿分析。

策。相比之下，DAMA 利用基于语音层级特性的结构化秩调度。这一固定先验引导适应过程，使其不会对有限数据过拟合。因此，即使在具有挑战性的训练条件下，DAMA 也能持续取得与 LoRA 和 DoRA 等稳健方法相当的高准确率，证明了其可靠性。

6.3 计算效率

乘加运算。 我们进一步使用额外乘加运算评估计算开销。如表 1 所示，DAMA 在所有方法中实现了最低的计算开销，仅需 22.3G 次乘加运算。相比之下，标准 LoRA 需要 102.2G 次乘加运算，几乎高出五倍。此外，DoRA 和 VB-LoRA 等先进方法会产生巨大的计算成本，分别达到 1415.4G 和 758.3G 次乘加运算。而 LoRA-FA 和 AdaLoRA 的乘加运算量低于标准基线方法，但仍高于我们的方法。此外，这两种方法的词错误率也更高，尤其是在未见语言上。这证实了我们的稀疏保护基设计最适合实时应用。

进方法如 DoRA 和 VB-LoRA 会产生巨大的计算成本，分别达到 1415.4G 和 758.3G 次乘加运算。虽然 LoRA-FA 和 AdaLoRA 的乘加运算量低于标准基线方法，但其乘加运算量仍高于我们的方法。此外，这两种方法的词错误率也更高，尤其是在未见语言上。这证实了我们的稀疏保护基设计最适合实时应用。

GPU 峰值显存。 图 3 展示了不同适配策略下的 GPU 峰值显存使用量。与资源消耗较大的微调相比，我们的方法显著降低了 GPU 峰值显存使用量，从 28.32 GB 降至 21.63 GB，降幅约为 24%，这使我们的方法更适用于资源受限环境。

与其他参数高效方法相比，DAMA 依然展现出更高的效率。它所需的内存少于标准 LoRA 和 DoRA。这种效率源于我们的自适应秩策略，它避免了不必要的梯度更新。VB LoRA 的内存使用量虽与之相当，为 21.51 GB，但准确率却因此大幅下降，词错误率也显著更高。因此，在所有评估方法中，DAMA 在内存开销与性能之间实现了最理想的平衡。

训练时间效率分析。 我们进一步比较了单个训练轮次的适配时间，如图 4 所示。帕累托图以秒为单位，在横轴上显示训练时间，在纵轴上显示词错误率。DAMA 位于左下角的最优位置，表明它是所有基线方法中效率最高的方法。

具体而言，DAMA 以 242.11 秒实现了最快的训练时间，同时保持了最低的词错误率 58.80%。相比之下，尽管 LoRA FA 以 242.43 秒达到了与我们相当的速度，但其词错误率却显著更高，为 63.23%。这表明 LoRA FA 所做的简化损害了模型质量。相反，DoRA 取得了具有竞争力的 59.93% 错误率，但所需训练时间长得多，达 377.93 秒。训练时间增加 56%，使 DoRA 不太适合快速适配。最后，AdaLoRA 和 VB-LoRA 等方法远远落后于前沿，因为它们既训练速度更慢，错误率也更高。因此，DAMA 为实时应用提供了速度与准确率的最佳平衡。

表 3：使用 Common Voice 数据集对 DAMA 的三个关键组件进行消融研究。

深度感知（U 形）	基保护（BPP）	基于 SVD 的初始化	平均词错误率
均匀	✓	✓	43.67
✓	完全适配	✓	**42.98**
✓	✓	随机	43.95

6.4 消融研究

我们在表 3 中考察了 DAMA 三项设计的贡献：深度感知秩调度、基于 SVD 的初始化以及 BSP。首先，我们将深度感知调度替换为均匀秩分布，即便使用更复杂的适配器，词错误率仍升至 43.67%。这一结果证实了“U 形”可塑性假说：模型在早期层和晚期层需要较高的适配能力，但在中间层必须严格限制干预，以保持性能。

其次，我们将基于 SVD 的初始化方法与标准随机初始化进行比较。随机初始化方法的错误率更高，达到 43.95%，这证明 SVD 将 LoRA 更新限制在对语言无关语义子空间影响最小的方向上，并保持 U 形结构，从而提升性能。最后，我们用完整适配替换所提出的 BSP，准确率仅略微提升了 0.22%（43.20% 对比 42.98%）。然而，这一微小提升是以牺牲结构安全性和增加中间层参数为代价的。

7 结论

本研究首次系统地探究了 MASR 中的逐层可塑性，并揭示了一种独特的 U 形模式。基于这一发现，我们提出了 DAMA，这是一种新颖的适配框架，利用这种 U 形结构，更高效、更有效地适配新语言。实验结果表明，DAMA 在各种语言上均取得了出色的性能，尤其是在低资源设置下，同时显著减少了可训练参数数量，并提高了内存使用、训练时间和浮点运算方面的效率。这些结果凸显了根据基础模型内部组织特点定制适配策略的重要性。我们的发现为利用模型内部表征开发可扩展且鲁棒的多语言与低资源适配方法开辟了新途径，也为更易于普及、更高效的语音技术铺平了道路。

局限性

尽管 DAMA 展现出很强的鲁棒性，未来仍有若干改进方向。首先，尽管我们的方法已在18种语言多样性丰富的语言上得到验证，但将评估扩展到更多语言，包括极其稀有的方言，将进一步检验“U 形”先验的泛化能力。其次，我们的方法专门针对低资源适配进行了优化，通过确保结构完整性来实现这一目标；在高资源场景中，放宽对语义谷的限制或许能带来更大的性能提升。最后，将本研究拓展到探索语义谷现象是否适用于自动语音识别之外的其他下游任务，是一个令人振奋的未来研究方向。

伦理声明

本文使用的所有数据均可公开获取，并依据以下许可协议使用：知识共享署名-非商业性使用-禁止演绎 4.0 国际许可协议和知识共享署名 4.0 国际许可协议、TED 使用条款、YouTube 服务条款以及 BBC 使用条款。

参考文献

Rosana Ardila、Megan Branson、Kelly Davis、Michael Kohler、Josh Meyer、Michael Henretty、Reuben Morais、Lindsay Saunders、Francis Tyers 和 Gregor Weber。2020年。Common Voice：一个大型多语言语音语料库。载于第十二届语言资源与评估会议论文集，第 4218–4222 页。

Klaudia Bałazy、Mohammadreza Banaei、Karl Aberer 和 Jacek Tabor。2024年。LoRA-XS：使用极少参数数量的低秩适配。*arXiv* 预印本 *arXiv:2405.17604*。

Heng-Jui Chang、Hung-yi Lee 和 Lin-shan Lee。2021年。迈向端到端自动语音识别的终身学习。载于 *Interspeech 2021* 论文集，第 2551–2555 页。

Alexis Conneau、Min Ma、Simran Khanuja、Yu Zhang、Vera Axelrod、Siddharth Dalmia、Jason Riesa、Clara Rivera 和 Ankur Bapna。2023年。Fleurs：通用语音表征的小样本学习评估。载于*2022*年 *IEEE* 口语技术研讨会（*SLT*），第 798–805 页。IEEE。

Wenqian Cui、Dianzhi Yu、Xiaoqi Jiao、Ziqiao Meng、Guangyan Zhang、Qichao Wang、Steven Y Guo 和 Irwin King。2025年。《语音语言模型最新进展：一项综述》。载于第*63*届计算语言学协会年会论文集（第*1*卷：长篇论文），第13943–13970页。

Luca Della Libera、Pooneh Mousavi、Salah Zaiem、Cem Subakan 和 Mirco Ravanelli。2024年。Cl-masr：一个多语言自动语音识别持续学习基准。*IEEE/ACM* 音频、语音与语言处理汇刊。

Ning Ding、Yujia Qin、Guang Yang、Fuchao Wei、ZonghanYang、YushengSu、Shengding Hu、Yulin Chen、Chi-Min Chan、Weize Chen 等。2023年。大规模预训练语言模型的参数高效微调。《自然·机器智能》，5(3)：220–235。

Neil Houlsby、Andrei Giurgiu、Stanislaw Jastrzebski、Bruna Morrone、Quentin De Laroussilhe、Andrea Gesmundo、Mona Attariyan 和 Sylvain Gelly。2019年。面向自然语言处理的参数高效迁移学习。载于机器学习国际会议，第2790–2799页。PMLR。

Edward J Hu、Yelong Shen、Phillip Wallis、Zeyuan Allen-Zhu、Yuanzhi Li、Shean Wang、Lu Wang、Weizhu Chen 等。2022年。LoRA：大语言模型的低秩适配。国际学习表征会议，1(2):3。

Takeshi Kojima、Itsuki Okimura、Yusuke Iwasawa、Hitomi Yanaka 和 Yutaka Matsuo。2024年。论基于解码器的预训练语言模型的多语言能力：发现并控制语言特定神经元。载于计算语言学协会北美分会*2024*年会议论文集：人类语言技术（第*1*卷：长篇论文），第6919–6971页。

Chin Yuen Kwok、Hexin Liu、Jia Qi Yip、Sheng Li 与 Eng Siong Chng。2025年。一种用于扩展多语言自动语音识别模型语言能力的两阶段 LoRA 策略。《*IEEE* 音频、语音与语言处理汇刊》。

Chin Yuen Kwok、Jia Qi Yip 与 Eng Siong Chng。2024年。面向多语言自动语音识别系统自回归解码器的持续学习优化。收录于《*Interspeech 2024*年会议论文集》，第1225–1229页。

Bo Li、Ruoming Pang、Yu Zhang、Tara N Sainath、Trevor Strohman、Parisa Haghani、Yun Zhu、Brian Farris、Neeraj Gaur 与 Manasa Prasad。2022年。大规模多语言自动语音识别：一种终身学习解决方案。收录于《*2022*年 *IEEE* 国际声学、语音与信号处理会议（*ICASSP*）》，第6397–6401页。IEEE。

Jiahong Li、Yiwen Shao、Jianheng Zhuo、Chenda Li、Liliang Tang、Dong Yu 和 Yanmin Qian。2025年。通过 LoRA 语言专家实现高效的多语言自动语音识别微调。载于 *Interspeech*，第 1138–1142 页。

Yang Li、Shaobo Han 和 Shihao Ji。2024年。VB-LoRA：利用向量库进行极致参数高效微调。神经信息处理系统进展，37：16724–16751。

Shih-Yang Liu、Chien-Yi Wang、Hongxu Yin、Pavlo Molchanov、Yu-ChiangFrank Wang、Kwang-Ting

Cheng 和 Min-Hung Chen。2024年。DoRA：权重分解低秩适配。发表于第四十一届国际机器学习会议。

ASR Omnilingual、Gil Keren、Artyom Kozhevnikov、Yen Meng、Christophe Ropers、Matthew Setzler、Skyler Wang、Ife Adebara、Michael Auli、Can Balioglu 等。2025年。全语种自动语音识别：面向 1600+ 语言的开源多语言语音识别。*arXiv*预印本 *arXiv:2511.09690*。

Tianyi Peng和Yang Xiao。2024年。面向增量关键词检测的暗经验。*arXiv*预印本：*2409.08153*。

Yifan Peng、Jinchuan Tian、Brian Yan、Dan Berrebbi、Xuankai Chang、Xinjian Li、Jiatong Shi、Siddhant Arora、William Chen、Roshan Sharma等。2023年。使用开源工具包和公开可用数据复现Whisper风格训练。收录于*2023*年*IEEE*自动语音识别与理解研讨会（*ASRU*），第1–8页。IEEE。

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, AliElkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi 等。2024年。将语音技术扩展至 1,000+ 种语言。机器学习研究期刊，25(97):1–52。

Alec Radford、Jong Wook Kim、Tao Xu、Greg Brock- man、Christine McLeavey 和 Ilya Sutskever。2023年。通过大规模弱监督实现鲁棒语音识别。收录于国际机器学习会议，第 28492–28518 页。PMLR。

Jiatong Shi、Shih-Heng Wang、William Chen、Martijn Bartelds、Vanya Bannihatti Kumar、Jinchuan Tian、Xuankai Chang、Dan Jurafsky、Karen Livescu、Hung-yi Lee 等。2024年。ML-SUPERB 2.0：在建模约束、语言和数据集方面对多语言语音模型进行基准测试。收录于*Proc. Inter-speech2024*，第 1230–1234 页。

Zheshu Song、Jianheng Zhuo、Yifan Yang、Ziyang Ma、Shixiong Zhang和Xie Chen。2024年。LoRA-Whisper：参数高效且可扩展多语言自动语音识别。载于*Interspeech 2024*会议论文集，第3934–3938页。

Tianyi Tang、Wenyang Luo、Haoyang Huang、Dongdong Zhang、Xiaolei Wang、Wayne Xin Zhao、Furu Wei和Ji-Rong Wen。2024年。语言特定神经元：大语言模型多语言能力的关键。载于计算语言学协会第*62*届年会论文集（第*1*卷：长文），第5701–5715页。

Liyuan Wang、Xingxing Zhang、Hang Su和Jun Zhu。2024年。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。

Zhaofeng Wu、Dani Yogatama、Jiasen Lu、Yoon Kim 等。《语义枢纽假说：语言模型跨语言共享语义表示》和模态。《载于第十三届国际学习表征会议。

Yang Xiao与Rohan Kumar Das。2025年。倾听、分析并适应以学习新型攻击：一种用于音频深度伪造来源追踪的无样本类增量学习方法。载于*Interspeech 2025*，第1563–1567页。

Yang Xiao、Nana Hou与Eng Siong Chng。2022年。Rainbow Keywords：面向在线口语关键词检出的高效增量学习。载于*Interspeech*会议录，第3764–3768页。

Yang Xiao、Peng Tianyi、Rohan Kumar Das、Yuchen Hu与Huiping Zhuang。2025年。Analytickws：面向小型化关键词检出的无样本解析类增量学习。载于计算语言学协会论文集发现：*ACL 2025*，第14147–14158页。

Hemant Yadav与Sunayana Sitaram。2022年。面向自动语音识别的多语言模型综述。载于第十三届语言资源与评测会议论文集，第5071–5079页。

HongliYang、ShengLi、Hao Huang、AyiduosiTuohan 和 Yizhou Peng。2025年。面向多语言自动语音识别的参数高效无缝语言扩展语言感知提示调优。发表于 *Interspeech*，第 1133–1137 页。

Muqiao Yang、Ian Lane 和 Shinji Watanabe。2022年。端到端语音识别模型的在线持续学习。发表于 *2022* 年 *Interspeech* 会议论文集，第 2668–2672 页。

Longteng Zhang、Lin Zhang、Shaohuai Shi、Xiaowen Chu 和 Bo Li。2023a。Lora-fa：面向大语言模型微调的内存高效低秩适配。*arXiv* 预印本 *arXiv:2308.03303*。

Qingru Zhang、Minshuo Chen、Alexander Bukharin、Pengcheng He、Yu Cheng、Weizhu Chen 和 Tuo Zhao。2023b。参数高效微调的自适应预算分配。载于第十一届国际学习表征会议。

A 深度感知分析详述

A.1 探测方法

为确保我们的分析反映的是模型的内在表征，而非探针的能力，我们采用了轻量级线性探测协议。在探测过程中，我们将 Whisper 主干网络完全冻结，仅在其隐藏状态之上训练一个轻量级线性分类器。对于解码器中选定的第 l 层，我们提取每个话语的完整隐藏状态序列，然后采用时间均值池化策略，为每个样本获得一个单维表征。随后，将这些池化后的表征输入单个线性层，以预测语言标识；各区域设置通过标准化的区域设置词汇表映射为整数标签。我们仅使用 AdamW 优化该线性探针，学习率为 1×10^{-3}，采用交叉熵损失进行语言识别，批量大小设为 256，并在实验中训练 5 轮；根据验证损失选出最佳模型，然后报告测试集上的最终准确率。

A.2 分析的语言

为验证 U 形可塑性曲线的普适性，我们从 Common Voice 数据集中选取了一组多样化的语言。

- **已见语言：** 英语、土耳其语、俄语、德语和中文。
- **未见语言：** 卢旺达语、世界语、卡拜尔语、卢干达语、草原马里语、中库尔德语、阿布哈兹语、库尔曼吉库尔德语、弗里西语和国际语。

A.3 语言选择的语言学依据

为避免选择偏差，我们选择“已见”语言，以最大限度地提高类型学多样性。具体而言，我们选择了在三个结构维度上与英语不同的语言，以确保观察到的“语义谷”是一种普遍现象。

- 句法语序：我们纳入了主宾谓语言（例如土耳其语），以对比英语的主谓宾语序，并测试模型面对显著语序变化时的鲁棒性。
- 形态学：我们选取了屈折语（例如俄语、德语），通过复杂的屈折变化和词元稀疏性来挑战具有可塑性的早期层。
- 信息密度：我们纳入了中文，以验证语义谷的稳定性，即使处理与字母子词截然不同的高密度汉字时也是如此。

因此，对于“未见”的目标语言，我们遵循了既定的 CL-MASR 基准（Della Libera等人，2024年）（例如卢旺达语、卢干达语），以确保与基线方法进行公平且可复现的比较。

B 实验

B.1 数据集与语言统计

在本小节中，我们介绍实验数据集的设计方式。对于 Common Voice 中的“未见”语言，我们采用了标准的 CL-MASR 基准（Della Libera 等人，2024年），其中包括卢旺达语和卢干达语等语言，以确保基线比较公平且可复现。

对于 FLEURS 数据集，我们通过筛选 Whisper-tiny 与更大的检查点之间词错误率明显下降、但仍有可见改进空间的语言，选取了 Whisper 已见但表现较弱的语言组（印地语、威尔士语、白俄罗斯语、波斯语、斯瓦希里语）。这种模式表明，扩大模型规模有所帮助，因此这些语言适合进行参数高效微调，因为即使在更大的模型上，它们的错误率也没有“卡”在极高水平。我们还依据 FLEURS 分组考虑了地理多样性，以及书写系统和语言结构的多样性，从而降低观察到的提升仅适用于某一地区或某种文字的可能性。

优先考虑覆盖 FLEURS 的多个地区，并纳入近邻迁移案例（阿斯图里亚斯语）以及结构多样的语言：撒哈拉以南非洲的班图语卢干达语、中东具有丰富形态结构并使用波斯-阿拉伯文字的中库尔德语、南亚使用奥里亚文字的奥里亚语，以及东南亚的南岛语宿务语，以便评估不同语言和地理条件下的改进效果。

B.2 基线方法

为全面验证所提出的 DAMA 框架，我们将其与三类适配方法进行了比较。这些方法涵盖了从标准统一方法到先进动态分配策略的各种方案。

语言	ISO 639-1	时长（分钟）		
		训练	验证	Test
Common Voice				
卢旺达语	rw	600	60	60
世界语	eo	600	60	60
卡拜尔语	kab	600	60	60
卢干达语	lg	600	60	60
草原马里语	mhr	600	60	60
中库尔德语	ckb	484	60	60
阿布哈兹语	ab	600	60	60
库尔曼吉库尔德语	kmr	296	60	60
弗里西语	荷兰弗里斯兰语	330	60	60
国际语	ia	313	60	60
FLEURS				
印地语	hi	399	60	60
威尔士语	cy	600	60	60
白俄罗斯语	be	571	60	60
波斯语	fa	600	60	60
斯瓦希里语	sw	600	60	60
卢干达语	lg	600	60	60
中库尔德语	ckb	584	60	60
阿斯图里亚斯语	ast	452	60	60
宿务语	ceb	600	60	60
奥里亚语	or	206	60	60

表4：各语言训练集、验证集和测试集的数据时长（分钟）。

1. 统一适配基线：这些方法在所有层中采用固定秩，将模型视为同质结构。

- **LoRA（Hu 等人，2022年）：**这是最广泛使用的参数高效微调方法。它将可训练的低秩矩阵（A 和 B）以统一秩注入每一层。

- **DoRA（Liu 等人，2024年）：**LoRA 的一种稳健变体，将权重分解为幅值和方向。它在准确率方面是一个强有力的基线，但训练时计算成本较高。

2. 数据驱动动态基线方法：这些方法尝试根据数据敏感度分配参数，这与我们基于结构的方法形成对比。

- **AdaLoRA（Zhang 等人，2023b）：**它根据梯度得出的重要性分数，在各层之间动态分配秩预算。如第 6.2 节所述，在低资源场景中，梯度信号存在噪声，因此该方法往往表现不佳。

3. 注重效率的变体：这些方法优先考虑参数或内存效率。

- **LoRA-FA（Zhang 等人，2023a）：**冻结下投影矩阵 A（随机初始化）并且仅训练 B。虽然内存效率较高，但它缺乏我们基底保护投影所具备的结构化初始化。

- **LoRA-XS（Bałazy 等人，2024年）：**利用奇异值分解（SVD）对权重更新进行静态压缩。

- **VB-LoRA（Li 等人，2024年）：**采用共享的“向量库”来构造低秩矩阵。这是一种极端压缩方法，但往往会损害模型在未见语言上的性能。

将 DAMA 与这些不同的基线进行比较，可以验证我们的策略优于统一方法和数据驱动方法（后者在低资源场景中容易过拟合）。

B.3 基线方法超参数设置

通用训练设置：所有方法的训练批次大小均为 6，验证批次大小为 16，采样率为 16000 Hz。目标序列长度上限为 448。所有方法均使用 AdamW 优化器训练 2 轮，最大梯度范数为 5.0，采用 FP16精度，并通过 NewBobScheduler 进行学习率调度（改进阈值为 0.0025，退火因子为 0.8）。对于 Common Voice 数据集，会过滤掉时长超过 10 秒的语句，生成词元上限为 80。对于 FLEURS 数据集，会过滤掉时长超过 30 秒的语句，生成词元上限为 120。

统一低秩方法 对于统一低秩方法，我们在所有层中保持一致的结构约束。具体而言，LoRA 和 DoRA 均采用固定秩 $r = 64$ 和 alpha $\alpha = 64$。二者的主要区别在于优化方式：DoRA 采用不含丢弃的权重分解，以分离幅值更新；而 LoRA 使用 0.1 的丢弃率。

高级与高效变体关于高级高效变体，我们采用了针对其动态架构量身定制的特定配置。AdaLoRA采用自适应预算分配策略，初始秩为 48，并根据经指数移动平均因子 0.85 平滑处理的敏感度分数，逐步剪枝至目标秩 32。相比之下，VB-LoRA采用极致压缩方案，使用包含 90 个向量（维度为 1280）的向量库，并进行Top-2选择；值得注意的是，为确保

表5：Common Voice 数据集按语言划分的不同适配方法结果。

Lang	FT	LoRA	DoRA	LoRA-FA	LoRA-XS	VB-LoRA	AdaLoRA	DAMA
ab	60.12	61.59	61.79	64.58	81.09	73.91	73.35	62.01
ckb	51.42	50.64	51.04	51.04	63.81	58.77	59.03	50.97
eo	19.64	15.36	15.51	16.11	20.92	22.36	18.61	15.36
荷兰弗里斯兰语	28.92	26.45	25.63	29.87	40.94	36.17	34.27	26.41
ia	16.49	9.04	9.31	10.25	13.60	12.74	11.52	9.33
kab	63.73	68.29	68.75	72.91	85.34	79.29	79.71	67.14
kmr	39.97	39.40	39.52	42.19	55.13	45.36	48.97	38.61
lg	58.37	59.42	60.51	63.81	77.84	68.67	69.59	60.07
mhr	32.26	33.06	33.19	36.56	50.32	44.67	43.33	32.93
rw	67.81	69.24	70.89	72.66	84.65	77.48	78.19	69.13

表6：FLEURS 数据集按语言划分的不同适配方法结果。“*”表示已见语言。

Lang	FT	LoRA	DoRA	LoRA-FA	LoRA-XS	VB-LoRA	AdaLoRA	DAMA
ast	22.69	17.28	17.32	19.17	24.56	19.93	21.06	17.64
be*	23.23	19.43	19.53	20.03	27.91	22.88	22.78	18.69
ceb	21.45	18.91	17.80	19.52	23.37	20.67	20.16	18.70
ckb	57.12	58.29	58.18	60.47	71.79	66.43	66.76	58.83
cy*	28.98	27.73	27.52	27.77	28.96	27.90	29.47	27.77
fa*	23.16	19.95	19.91	20.05	22.87	22.30	21.53	20.15
hi*	25.64	24.78	24.78	24.90	27.38	26.15	26.90	24.82
lg	58.81	62.95	60.72	63.85	72.21	64.92	68.97	61.71
or	90.18	78.22	78.28	79.77	84.22	81.78	84.89	79.37
sw*	36.76	34.08	34.46	35.50	44.37	40.79	41.35	34.88

在受限参数空间内收敛，需要采用更高的学习率 A。最后，LoRA-XS通过对预训练LoRA模块（$r = 64$）进行奇异值分解（10 次迭代）构建；我们冻结得到的低秩矩阵，仅训练中间的 $r \times r$ 潜在映射矩阵。我们还评估了 LoRA-FA，它随机初始化下投影矩阵 A 并将其冻结，仅更新上投影矩阵 B。该变体同样采用秩 $r = 64$，作为静态子空间约束的基线。

C 其他结果

C.1 不同语言的结果。

Common Voice 和 FLEURS 数据集上的详细性能见表5 和表6。该评估涵盖多种低资源语言，分为“未见”目标语言和“已见-弱”语言（表中以“*”标注）。具体而言，这些结果凸显了 DAMA 相较于 LoRA 等均匀基线方法以及 AdaLoRA 等数据驱动变体所具有的更优泛化能力。尽管标准方法因过拟合而在未见语言上出现性能下降，DAMA 在两类语言上均持续保持较高准确率。

表7：适配卢旺达语（未见）后英语（已见）上的灾难性遗忘比较。

Lang	Base	微调	LoRA	DoRA	DAMA
英语	11.09	105.50	12.93	13.09	12.56
卢旺达语	–	67.81	69.24	70.89	69.13

C.2 遗忘估计

为评估适应过程是否会损害模型的既有知识，我们分析了在目标低资源语言（卢旺达语）上训练后，模型在源语言（英语）上的性能。结果见表7。具体而言，数据揭示了与无约束更新相关的严重灾难性遗忘风险。如表所示，微调（FT）使英语上的词错误率从基线 11.09% 急剧飙升至 105.5%。这表明，尽管微调能够适配新语言，但它完全覆盖了模型预训练的语义核心。

相比之下，DAMA 展现出更优异的稳定性。其英语词错误率为 12.56%，明显更接近原始基线，并优于 LoRA（12.93%）和 DoRA（13.09%）。因此，这证实了我们的基底保护投影（BPP）成功锁定了“语义谷”，确保模型在学习新语言的同时，不会牺牲其内在能力。