LIELLA：基于光子权重流的私有边缘适配

Yuexun Huang芝加哥大学 芝加哥，美国
yesunhuang@uchicago.edu

Kfir Suli many麻省理工学院 剑桥，美国
kfir@mit.edu

Ke Lin威廉与玛丽学院 威廉斯堡，美国
leonard.keilin@gmail.com

Xiangyu Ren*芝加哥大学 f Edinb 爱丁堡，英国
xiyu.ren@outlook.com

Dirk Englund*麻省理工学院 剑桥，美国
englund@mit.edu

Liang Jiang*芝加哥大学 芝加哥，美国
liangjiang@uchicago.edu

arXiv:2611.29343v1 [cs.AR] 22 Nov 2026

摘要

尽管光子加速器已展现出实现高能效机器学习推理的潜力，但在严格的内存、精度和硬件约束下，跨多个边缘客户端对大型模型进行私有个性化仍然十分困难。我们提出 LIELLA，这是一种权重流架构，使多个边缘客户端能够通过紧凑的私有适配器对共享模型进行个性化。所提出的两种光子适配器架构同时支持反向传播和仅前向优化；与电子适配器相比，它们显著降低了建模的客户端电能消耗。在模拟硬件噪声下，共享同一流式传输骨干网络的客户端在 QNLI 和 SST-2 上实现了接近无噪声条件下的准确率，即使每个接收器的每次乘加运算仅消耗阿焦耳级光学能耗。该架构还支持在执行相同任务的客户端之间进行硬件感知适配，使其各自的私有适配器能够进一步提升准确率。对于千兆字节级的 TinyLlama 骨干网络，LIELLA 将每个客户端的推理内存占用维持在兆字节级，并将持久化适配器存储维持在千字节级，同时通过仅前向适配实现出色性能。这些发现进一步凸显了由硬件驱动的机会：开发更小巧但更高效的参数高效微调（PEFT）方法，以及更准确、更具抗噪性的仅前向优化算法。

关键词

光子架构、边缘适配、参数高效微调、仅前向优化、内存核算

1 引言

面向不同用户和任务对大型模型进行微调的需求日益增长，应用场景涵盖个人助理、推荐系统、图像生成和医疗保健 [9, 35, 46, 63] 。

在资源受限的边缘设备上部署大型个性化模型，本身就面临巨大的内存和计算挑战；而实现私有的设备端微调，则对资源提出了更为严格的要求 [15, 53]。光子计算为实现高能效推理和训练提供了一条颇具前景的路径 [7, 39, 41], , 但要将这些能力扩展到多个边缘客户端的私有微调，仍面临架构层面的挑战。参数高效微调（PEFT）减少了可训练参数的数量 [24, 26], , 但传统的本地执行仍需存储并反复访问规模大得多的冻结骨干网络 [53], , 这限制了它在解决边缘端部署所面临的更广泛内存和计算挑战方面的作用。将微调卸载到远程基础设施可以缓解这些资源限制，但传输敏感的训练数据或客户端专属的模型更新会带来显著的隐私风险 [12, 64]。

权重流计算。权重流式传输开创了分布式推理的新范式，使多个边缘客户端能够共享集中托管的模型骨干网络，同时降低客户端侧的内存和能耗需求。这一范式尤其适用于光子系统，因为高能效光电乘法和波分复用（WDM）天然支持共享权重广播 [21, 22]。Netcast 通过实验展示了在已部署的 86 公里光纤链路上进行权重流式推理，并实现了准确的图像识别；接收器侧的光能低至每次乘加运算（MAC）40 aJ [23, 41]。AIR-LLM 将权重广播原理扩展到了射频大语言模型推理 [16, 17, 49]。这些系统确立了一种前景广阔的范式，可让多个边缘客户端共享同一模型骨干网络。然而，要将这一范式从推理转变为私有个性化，仍面临尚未解决的架构挑战，因为每个客户端都必须独立微调共享模型，同时确保其训练数据和适配器状态保持私有。

私有微调中的挑战。 在共享权重流上支持独立的私有微调会带来三个架构挑战。首先，每个客户端都必须能够利用共享骨干网络独立训练自己的个性化模型，同时将其训练数据、梯度和适配器状态保留在本地。其次，该架构必须支持不同的优化方法，以评估其有效性并确定可行的微调方法——

微调策略，并刻画训练准确率、能耗和硬件需求之间的权衡。例如，反向传播（BP）需要转置骨干算子，并需要存储中间激活或进行重新计算；而零阶（ZO）优化则避免了这些依赖，但代价是需要额外的前向评估，且梯度估计可能更加嘈杂 [15, 29, 36, 60]。第三，私有适配器架构必须经过设计，以便在客户端内存和能耗成本较低的情况下实现有效微调，同时避免给共享权重流服务带来显著开销。这需要统筹考虑适配器参数化、电子与光子执行、私有状态的存储和传输、硬件噪声与精度，以及服务器端需求。要应对这些挑战，需要

进行端到端协同设计，将独立的私有适配、有效的优化方法和硬件感知适配器与共享权重流服务整合起来。

先前工作。

Netcast 展示了光学权重流，AIR-LLM 探索了射频权重广播，QAMNet 则研究了用于神经计算的高效光学编码[1, 16, 17, 23, 41, 49]。

然而，这些方法主要面向推理，并未解决设备端训练问题。包括 FICONN、L2ight、Mirage 和 Opto-ViT-v2 在内的光子训练架构，探索了原位学习、高精度训练和抗噪微调 [3, 7, 11, 20]。除光子学之外，AHWA-LoRA 将低秩适配器应用于模拟存内计算中的硬件感知微调 [34]。类似地，MobiZO 和 MeBP 通过零阶优化和内存高效反向传播，提升了数字边缘微调的效率 [15, 43]。尽管如此，这些工作主要关注在单个计算设备上进行本地模型训练，而非多个客户端共享同一骨干网络服务时的分布式私有适配。Petals 和 DLoRA 将骨干网络计算卸载到远程服务器，同时保留客户端特定适配 [4, 14]。这会将数据相关的中间表示暴露给服务器，并要求针对每个客户端的输入执行骨干网络计算，从而带来严重的隐私风险，并限制服务器端摊销。此外，这些工作都未利用与私有适配器硬件的协同设计。

LIELLA架构。为了在共享模型上实现私有微调，我们提出了轻量适配实时流式独立边缘学习（LIELLA）。这是一种分布式微调架构，允许多个边缘客户端独立地个性化共享的权重流骨干网络。在 Netcast [41],的基础上，远程服务器将冻结的骨干网络权重以光学信号的形式进行流式传输，而每个客户端则调制其本地输入，并整合接收到的信号以执行相应的乘累加。随后，LIELLA 通过将轻量级私有适配器纳入客户端侧计算，扩展了这一推理流水线。每个客户端评估自身的训练损失，并在本地更新其私有适配器，无需向服务器传输训练数据，也不会泄露其适配器参数。因此，多个客户端可以独立训练并部署自己的紧凑型适配器，将其作为针对不同任务和硬件条件的个性化模型的灵魂，同时使用相同的骨干网络服务作为躯体。总体而言，我们的主要目标是在最大限度降低边缘内存和能耗成本的同时，为多个边缘客户端实现私有个性化，并保持对硬件噪声的鲁棒性。

首先，我们提出了两种基于共享 Netcast 权重流的私有光子适配方案。Netcast 集成核心光学（NICO）利用流式光信号，在接收器处以光学方式完成所有客户端私有适配器计算。混合适配器内核集成（MAKI）以光学方式执行固定适配器运算，同时保留客户端私有适配器计算的电子实现。我们将电子本地集成（ELI）作为基线：它保留光学骨干网络流，但在客户端以电子方式执行所有适配器计算。随后，我们将 LoRA-FA [58] (FA) 和 LoRA-XS [2] (XS) 映射到这三种设计中，从而在同一权重流系统内支持不同的适配器参数化方式。

其次，我们构建了一个共享光学权重流执行框架，支持多个边缘客户端进行推理和私有微调。在低噪声服务器/简单客户端（LN/S）信令方案 [23, 41], 的基础上，我们设计了一种噪声感知缩放方法，将冻结的数字骨干网络转换为共享光学流。具体而言，我们根据任务无关的公共校准数据选择输入缩放，以尽量减小含噪光学推理与数字推理之间的差异；同时推导输出缩放，使其满足接收器在前向和转置运算中的动态范围约束。由此得到的共享服务通过前向权重流支持推理和仅前向（FO）微调，并通过额外的转置权重流扩展到反向传播（BP）。

第三，我们开发了物理感知光子模拟器和资源模型，用于评估共享光学权重流上的私有适配。我们的模拟器对不同适配器设计和执行模式下的关键硬件非理想因素进行建模。我们还基于已发表的硬件测量结果和文献中的组件参数，采用特定工作负载核算方法，对客户端内存和电能消耗进行建模。综合来看，这些模型使我们能够评估多个客户端在共享权重流上执行不同任务时的资源高效私有微调，以及在异构接收器条件下执行相同任务的客户端之间的硬件感知个性化。此外，这还支持从准确率、抗噪性和资源效率等方面，对不同 PEFT 参数化方案和优化方法进行系统比较。

最后，我们在各种任务和硬件条件下评估了我们的框架。在每次光学乘加运算噪声为 640 aJ 的模拟接收器噪声条件下，共享冻结骨干网络的客户端在 QNLI 上达到超过 85% 的准确率，在 SST-2 上达到超过 90% 的准确率。对于在不同接收器条件下执行相同任务的客户端，硬件感知私有适配器进一步将准确率最高提升了 11 个百分点。在拥有 9,856 个可训练参数（38.5 KiB）和 4.81 KiB 冻结客户端私有秩校准的情况下，XS 在无噪声光子模拟中通过仅前向优化实现了 87.7% 的 SST-2 准确率，同时建模的客户端推理内存为 7.46 MiB。在名义组件假设下，对于所评估的仅前向训练工作负载，采用 XS 的 NICO 将建模的客户端电能较 ELI 降低了 70% 以上。

我们的主要贡献如下：

- 我们实现了多个客户端在共享冻结光学流的情况下独立进行私有模型适配，且无需增加服务器端骨干网络计算或占用大量客户端存储空间。
- 我们提出了两种光子私有适配器设计，将固定计算和客户端私有计算分配到光学与电子路径中，并将其与纯电子基线进行比较。
- 我们为 ZO 和 BP 建立了仅前向和支持转置的微调服务协议，随后评估了模拟光子噪声下的学习与资源权衡，包括客户端内存和建模的电能。
- 据我们所知，我们首次系统地研究了不同的参数高效微调参数化方法和零阶方法如何影响分布式

光子权重流式传输硬件上的私有微调，并揭示了新的动机：设计更小、更有效的适配器，以及准确、查询高效且抗噪的零阶算法。

2 背景与动机

2.1 光学权重流

Netcast 通过光子矩阵-向量乘法（MVM）实现光学权重流 [23, 41]。对于权重矩阵$W \in \mathbb{R}^{M\times N}$，服务器将 w_{mn} 编码到波长 λ_m 的时间步 n上，并通过 M 个波长通道和 N 个时间步广播权重矩阵。在客户端，宽带马赫–曾德调制器（MZM）通过 x_n 对每个输入进行编码 $x_n = \cos(2\theta_n)$，同时调制所有波长通道。经过波长解复用后，平衡光电探测器在一段时间内对所得信号进行积分，以计算 MVM。

在低噪声服务器/简单客户端（LN/S）配置中， [23],服务器通过光功率重新缩放权重幅值，并在差分检测通道中表示其符号。为适应光学接收器有限的动态范围，数字权重矩阵 $W \in \mathbb{R}^{M\times N}$ 通过正对角缩放矩阵 D_{in} 和 D_{out}映射为光学编码矩阵 $\bar{W}$：

$$\bar{W} = D_{\text{out}}^{-1} W D_{\text{in}}, \qquad \bar{x} = \frac{D_{\text{in}}^{-1} x}{s}, \tag{1}$$

其中 $s = \|D_{\text{in}}^{-1}x\|_{\infty}$。缩放矩阵是公开的，并由各客户端共享，而 s则针对每个输入在本地计算。随后计算两个探测器通道的预期光电子计数，如下：

$$\mu_{mn}^{\pm} = \mathbb{E}[K_{mn}^{\pm}] = \frac{n_0|\bar{w}_{mn}|}{2}\left[1 \pm \operatorname{sgn}(\bar{w}_{mn})\bar{x}_n\right]. \tag{2}$$

其中， $K_{mn}^{\pm}$ 分别表示 $\pm$ 差分通道检测到的光电子计数。因此，积分电荷为$Q_m^{\pm} = e\sum_{n=1}^{N} K_{mn}^{\pm}$，由此得到平衡光电探测计数

$$\frac{\mathbb{E}[Q_m^{+} - Q_m^{-}]}{e n_0} = \frac{1}{n_0}\sum_{n=1}^{N}(\mu_{mn}^{+} - \mu_{mn}^{-}) = \sum_{n=1}^{N}\bar{w}_{mn}\bar{x}_n = (\bar{W}\bar{x})_m. \tag{3}$$

这里， n_0 是单位幅值权重对应的校准电子计数， e为元电荷。因此，原始数字结果可还原为 $Wx = sD_{\text{out}}(\bar{W}\bar{x})$。在 LN/S 信号传输下，每个符号预期检测到的光电子总数为 $\mu_{mn}^{+} + \mu_{mn}^{-} = n_0|\bar{w}_{mn}|$。因此，缩放矩阵的选择会直接影响散粒噪声和接收器动态范围，使光学缩放成为一项关键的设计考量，而非单纯的数值归一化。

重要的是，LIELLA 旨在提供一种共享光学服务，使多个客户端能够在使用同一串流式骨干网络的同时，独立微调各自不同的私有适配器。服务器不得执行特定于客户端的编码或计算，也不得为单个适配器引入额外的乘积累加，否则将削弱光学广播的计算复用效果。所有特定于客户端的输入、适配器计算和信号累加都应保留在接收器端。因此，尽管各客户端可能面临不同的任务、输入分布和硬件噪声条件，仍需在客户端之间采用通用的光学权重表示和共享的缩放矩阵。这些约束促成了一种与任务和客户端无关的编码与计算策略，在不增加服务器端负担的情况下，平衡光学噪声与动态范围要求。

2.2 参数高效微调

参数高效微调（PEFT）通过优化一组紧凑的参数，同时冻结骨干网络，实现对大型预训练模型的适配。与全量微调相比，它大幅减少了可训练状态和优化开销。在低秩适配（LoRA）中，每个适配的权重矩阵 W都被替换为 $W + \Delta W$，其中 ΔW是一个低秩扰动 [26]。在本文中，我们考虑两种具有不同可训练参数结构的低秩 PEFT 方法，

$$\begin{aligned} &\text{LoRA-FA:} \quad \Delta W = BA, \\ &\text{LoRA-XS:} \quad \Delta W = S_{\text{out}} C S_{\text{in}}, \end{aligned} \tag{4}$$

其中，LoRA-FA 冻结 A 并训练 B [58]，而 LoRA-XS 根据预训练权重矩阵W的秩-r截断奇异值分解构造固定因子 $S_{\text{out}} = U_r\Sigma_r$和 $S_{\text{in}} = V_r^{\top}$，并仅训练微小核心 C [2]。对于 L 个适配的线性层，相应的可训练私有参数数量为

$$\begin{aligned} P_{\text{FA}} &= r_{\text{FA}}\sum_{\ell=1}^{L} d_{\text{out},\ell}, \\ P_{\text{XS}} &= L r_{\text{XS}}^2, \end{aligned} \tag{5}$$

其中， r_{FA} 和 r_{XS} 分别表示相应的秩，$d_{\text{out},\ell}$ 是模块 ℓ的输出维度。对于我们评估的骨干网络——采用 1.1 B TinyLlama 配置、包含 22 个 Transformer 块且每个块包含七个适配线性层（$L = 154$）——秩为 16 的 LoRA-FA 需要 6,307,840 个私有参数（FP32 下为 24.06 MiB），而秩为 8 的 LoRA-XS 仅需 9,856 个（38.5 KiB），与 LoRA-FA 相比，将持久化私有参数存储减少了 640×。

然而，仅减少可训练参数并不能真正带来显著的内存优势，因为仍有必要存储冻结骨干网络。如图1（a）所示，对于 FP16 格式的 1.1 B 模型，仅骨干权重就已占用约 2 GiB，远大于私有适配器的大小。因此，只要骨干网络仍驻留在本地，即使采用日益紧凑的 PEFT 方法，端侧模型存储的改善也十分有限。光学权重流消除了本地骨干存储的需求，但每个客户端仍必须执行其私有适配器。如图1（b）所示，LoRA-FA 和 LoRA-XS 中彼此独立的固定因子与可训练因子，为通过光电硬件协同设计降低内存访问和能耗成本提供了机会，从而实现在边缘端进行超紧凑且高能效的私有适配。固定变换可以按照与流式传输权重相同的流程通过光学方式执行，而客户端私有操作则可以在接收器处通过光学方式执行，也可以通过电子方式执行；这两种方式对内存访问、调制和能耗的影响各不相同。

2.3 光子优化

微调私有适配器需要对客户端侧的私有参数 θ进行反复优化。仅前向（FO）零阶优化仅使用损失评估来估计梯度，而无需

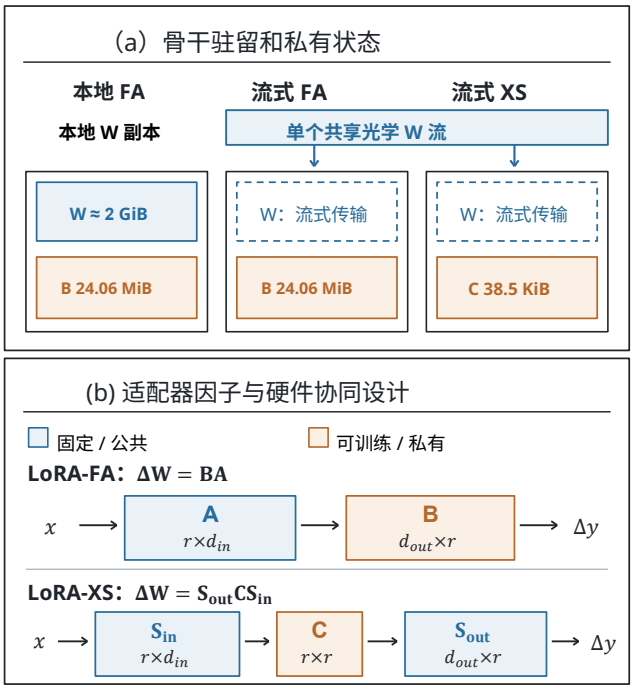


图 1：权重流式传输与适配器–硬件协同设计。(a) 流式传输无需客户端驻留的骨干网络副本（FP16 下为 ≈2GiB）。私有 FP32 可训练权重中，FA 占用 24.06 MiB，XS 占用 38.5 KiB。(b) 两者各自不同的固定因子和可训练因子，使其能够采用不同的适配器设计。

通过模型进行反向传播 [15, 36]。一种典型的两点梯度估计器是

$$\hat{g}_{\mathrm{FO}} = \frac{\mathcal{L}(\theta + \epsilon u) - \mathcal{L}(\theta - \epsilon u)}{2\epsilon} u, \tag{6}$$

其中， u 是随机扰动方向， ϵ 是扰动幅度。因此，FO 可以重复利用与推理兼容的光学权重流，而无需反向传播或相关的激活存储。然而，估算梯度需要进行多次前向评估，从而增加光学计算请求的数量。包括 SubZero 和 SubZero+ 在内的近期方法采用低维扰动来提高优化效率 [56, 57, 60]。

相比之下，常见的反向传播（BP）通过链式法则计算网络中的梯度。对于线性层 $y = Wx$，反向操作可计算为

$$\frac{\partial \mathcal{L}}{\partial x} = W^{\top} \frac{\partial \mathcal{L}}{\partial y}, \tag{7}$$

这需要转置权重算子。因此，光学反向传播需要同时使用前向权重流和面向转置的权重流，还需要保留激活值或进行重新计算 [29, 43]。与 FO 相比，BP 无需反复进行零阶梯度查询，但会带来额外的反向传播服务和内存需求。

这些差异在光子硬件噪声下尤为重要。有限的光能会引入散粒噪声，而电子读出和探测器动态范围限制会进一步降低计算精度 [21, 39, 41]。对于 FO，重复损失评估中的噪声会传递到有限差分中，增大梯度估计方差，尤其是在 ϵ 较小时。对于反向传播，光学误差会同时影响前向激活和反向梯度传播。因此，优化方法的鲁棒性和有效性不仅取决于其无噪声收敛能力，还取决于其光学查询需求、接收器噪声敏感性和客户端内存占用。这些挑战促使我们系统地研究共享光子硬件约束下的优化方法，尤其是查询高效且抗噪的 FO 方法。

3 LIELLA架构

3.1 共享光学服务

LIELLA将公共且不可变的骨干服务与由各客户端独立拥有的适配相分离。对于客户端 i 和经适配的线性层 ℓ，前向计算如下：

$$y_{i,\ell} = W_{\ell} x_{i,\ell} + \Delta W_{i,\ell} x_{i,\ell}, \tag{8}$$

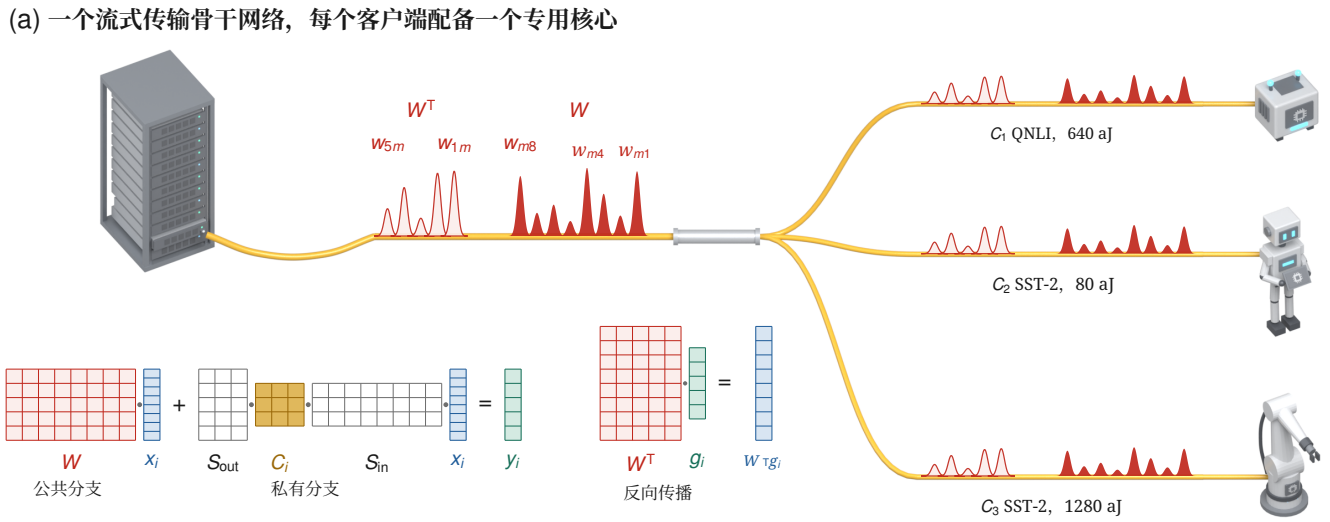
其中 W_{ℓ} 是所有客户端共享的公共冻结骨干权重，而输入 $x_{i,\ell}$ 和适配器扰动 $\Delta W_{i,\ell}$ 则由各客户端私有持有。服务器向所有客户端广播相同的光学编码权重流 W_{ℓ}。每个客户端使用本地输入 $x_{i,\ell}$ 调制接收到的光信号，并执行平衡光电探测和时间积分，以获得对 $W_{\ell} x_{i,\ell}$ 的估计。每个接收器将骨干网络计算结果与自身的适配器校正相结合，并在本地执行其余数据相关操作。

LIELLA的整体架构如图 2 (a)所示。该架构通过接收端光电集成，将共享光学骨干服务与客户端专属的私有适配解耦。单一光学权重流服务于多个独立进行适配的客户端。在此示例中，一个客户端正在针对 QNLI 进行微调，另外两个客户端则在不同的接收器条件下针对 SST-2 进行适配。它们使用相同的公共权重流，但各自独立执行输入调制、光电探测和私有适配器更新。每个客户端均保留自己的输入、标签、损失、私有适配器参数和优化器状态，并在本地进行适配。这样便能实现客户端专属的个性化，同时无需增加服务器端骨干网络计算，也不会泄露客户端的私有数据。

3.2 光子适配器设计

图2(b)展示了两种光子适配器架构——网络广播集成核心光学（NICO）和混合适配器内核集成（MAKI），并将它们与电子本地集成（ELI）这一基线进行比较。它们共享公共骨干流，但适配器计算的执行位置不同。对于因子适配（FA）， A_{ℓ} 是公共且冻结的，而 $B_{i,\ell}$ 是私有且可训练的。对于交叉适配（XS）， $S_{\mathrm{in},\ell}$ 和 $S_{\mathrm{out},\ell}$ 都是公共的冻结因子，只有 $C_{i,\ell}$ 是可训练的。NICO和MAKI将这些公共因子复用为与客户端无关的光学流，因此接收器端执行时无需在每个客户端存储其权重矩阵。

NICO。 NICO在接收器端以光学方式实现私有可训练适配器变换，同时不向



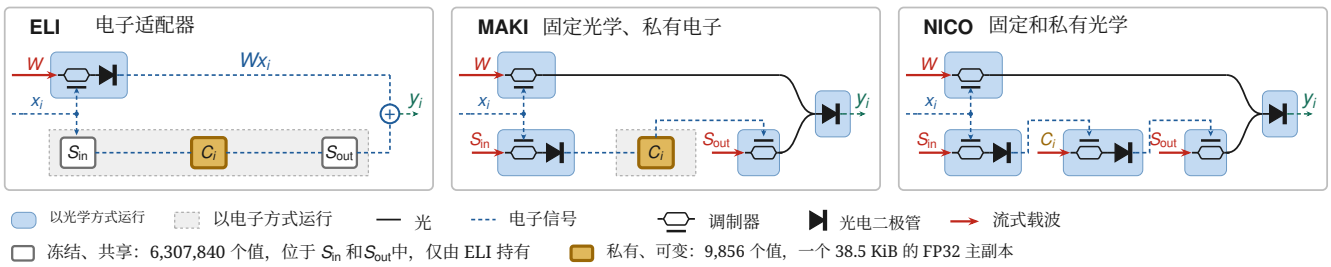


图2：LIELLA 服务架构与私有适配器设计示意图。(a) 相同的公共权重流支持 QNLI [51] 和 SST-2 [42] 客户端，即使接收器条件不同，也无需服务器端个性化计算。(b) 对于图示的 LoRA-XS 适配器，ELI 以电子方式执行所有适配器因子，MAKI 将固定因子映射到光路上，而 NICO 还会在接收端光路上执行专用核心。

服务器泄露其私有权重。源端对共享主干网络和公共固定适配器因子进行编码。对于交叉适配，无噪声接收器端将其分解为三层光学计算，

$$\begin{aligned} h^{(1)}_{i,\ell} &= S_{\text{in},\ell} x_{i,\ell}, \\ h^{(2)}_{i,\ell} &= C_{i,\ell} h^{(1)}_{i,\ell}, \\ \Delta y_{i,\ell} &= S_{\text{out},\ell} h^{(2)}_{i,\ell}. \end{aligned} \tag{9}$$

此处，第一和第三阶段使用公开的光学权重流。为实现私有中间阶段，服务器提供未加权光载波，客户端将其 $r \times r$ 核心 $C_{i,\ell}$ 通过本地电光权重调制器编码到载波上。在数字状态实现中，从客户端内存读取私有系数，并通过数模转换器和驱动器进行转换；系数的幅值控制载波衰减，其符号则决定平衡探测通道。第一阶段的秩宽输出 $h^{(1)}_{i\ell}$，经探测后用于驱动本地加权载波上的输入调制器。与式 2中的 LN/S 信号一样，预期的输出通道 j 中的差分电荷满足

$$\mathbb{E}[Q^+_{C,j} - Q^-_{C,j}] \propto \sum_{k=1}^{r} C_{i,\ell,jk} h^{(1)}_{i,\ell,k}, \tag{10}$$

最多进行客户端本地光学归一化和增益恢复。所得的秩宽 $h^{(2)}_{i\ell}$，会重新调制到公共 $S_{\text{out},\ell}$ 流上，以生成最终输出。因子适配同样采用类似方式，使用公共 A_ℓ 和本地编码的私有 $B_{i,\ell}$ 光学阶段。该设计以客户端侧调制和秩宽转换换取光学私有核心乘加运算，而空白载波传输会带来额外的服务器端光学供给开销。不过，需要注意的是，无需向服务器传输任何私有适配器系数、激活值或梯度。

MAKI。 MAKI 将可变的适配器计算保留在本地电子电路中，同时以光学方式执行固定适配器因子。在 FA 中，固定的 A_ℓ 变换以光学流的方式传输，而私有的 $B_{i,\ell}$ 变换则以电子方式执行。在 XS 中，两个固定变换 $S_{\text{in},\ell}$ 和 $S_{\text{out},\ell}$ 均作为公共光学流提供，并在二者之间以电子方式进行 $C_{i,\ell}$ 计算。中间秩信号被检测后传送至

电子核心，并重新调制到公共输出因子流上。

XS 秩域校准与缩放。对于每个 XS 层，令$Q_{1,i,\ell}$和 $Q_{2,\ell}$ 为正的对角秩-r 缩放因子。客户端根据 $h_i^{(1)}$,,,的本地训练时秩激活统计，对$Q_{1,i\ell}$ $S_{\text{in},\ell}x_{i\ell}$ r ℓ =,进行校准，随后冻结并持久化存储其对角元素。在当前 LIELLA 默认校准中，服务器将$g_{\ell,j}$ 计算为$D_{\text{out},\ell}^{-1}S_{\text{out},\ell}$的列范数的倒数，而客户端将$v_{i,\ell}$ 计算为 $h_{i\ell}^{(1)}$,的归一化逐坐标均根。在共享且固定的 λ_q条件下， $Q_{2,\ell}$和$Q_{1,\ell}$ 可表示为

$$Q_{2,\ell} = \lambda_q \operatorname{diag}(g_\ell), \tag{11}$$

$$Q_{1,i,\ell} = \lambda_q(\min_j g_{\ell,j}) \operatorname{diag}(v_{i,\ell}). \tag{12}$$

因此，$Q_{2,\ell}$ 仅取决于公开的固定因子和共享缩放设置，而$Q_{1,i,\ell}$ 仍为本地私有信息。我们定义

$$\begin{aligned} \bar{C}_{i,\ell} &= Q_{2,\ell}^{-1} C_{i,\ell} Q_{1,i,\ell}, \\ \bar{S}_{\text{out},\ell} &= D_{\text{out},\ell}^{-1} S_{\text{out},\ell} Q_{2,\ell}, \\ u_{i,\ell} &= \bar{C}_{i,\ell} Q_{1,i,\ell}^{-1} h_{i,\ell}^{(1)} = Q_{2,\ell}^{-1} C_{i,\ell} h_{i,\ell}^{(1)}. \end{aligned} \tag{13}$$

它们的乘积恰好为 $\bar{S}_{\text{out},\ell} u_{i,\ell} = D_{\text{out},\ell}^{-1} S_{\text{out},\ell} C_{i,\ell} h_{i\ell}^{(1)}$,。NICO 在本地为其光学私有核心构造 $\bar{C}_{i,\ell}$，而 MAKI 则以电子方式计算 $C_{i,\ell} h_{i\ell}^{(1)}$,。两者都使用相同的公开 $\bar{S}_{\text{out},\ell}$ 光学流，且都不会将$Q_{1,i,\ell}$ 发送给服务器。

ELI. ELI 仅以光学方式执行共享骨干乘法运算，并将所有适配器变换保留在客户端侧电子设备中。这与 Opto-ViT-v 的做法类似2 [7]，并作为参考设计。该设计无需额外的公共分布式权重流供各个独立客户端共享。不过，这需要保留所有本地固定因子存储和电子适配器乘加运算，因而有所代价。

除了骨干服务外，光学部署还需要公共固定因子流；对于反向传播（BP），还需要这些流的转置方向。接收端转换、光学噪声和资源成本将在第 4 节讨论。

3.3 训练执行

图 2(a) 展示了推理和仅前向（FO）训练所使用的共享前向W流，以及反向传播（BP）所需的额外W_{T}流。对于仅前向（*FO*）训练，每个客户端都会扰动其私有适配器，并仅利用前向服务计算损失。这种方式复用了供 NICO 和 MAKI 使用的公共骨干网络和固定因子光学流，无需请求其转置。因此，这也提供了额外的隐私保护，因为执行训练的客户端与执行推理的客户端大体上表现相同。

对于反向传播（*BP*），共享服务还会为 W_ℓ 以及任何通过光学方式执行的公共适配器因子广播转置方向流。每个客户端通过其私有适配器传播导数，并在本地计算参数梯度。对于 XS，适配器的反向路径依次应用 $S_{\text{out},\ell}^{\mathrm{T}}$, $C_{i,\ell}^{\mathrm{T}}$, 和 $S_{\text{in},\ell}^{\mathrm{T}}$。BP 还需要保留前向激活，或通过额外的前向流传递重新计算。

在这两种模式下，客户端的数据和权重同样保持私密，因为共享源不需要任何客户端专属信息。

由公共流调度服务的客户端可以同时使用相同的广播，而反向请求可能需要额外的服务时隙。因此，LIELLA 在 FO 的查询量和准确率与 BP 的转置服务需求及激活内存之间进行权衡。评估所用的优化器方案及其服务核算详见第 5 节。

4 硬件与资源模型

本节将建立一个统一的操作级模型，用于刻画光接收器噪声、客户端电能和内存需求。通过将光子计算、电子接口和训练服务请求纳入考量，该模型使我们能够在各种训练方法下，对 NICO、MAKI 和 ELI 的资源需求进行一致比较。

4.1 光学噪声模型

接收器光子预算。根据第2节中的 LN/S 信号，探测电子数与编码权重绝对值成正比。令 $\mu_M = \text{mean}_{jk}|\bar{M}_{jk}|$ 表示归一化光学矩阵 $\bar{M}$ 的平均绝对权重，并令 $E_{\text{rx},M}$ 表示每次有效乘加运算的平均入射接收器光能量（已计入任何局部幅值衰减）。相应的满量程（$|\bar{M}_{jk}| = 1$）探测电子数为

$$n_{0,M} = \frac{\eta E_{\text{rx},M}}{(hc/\lambda)\mu_M}, \tag{14}$$

其中， η 为探测器量子效率， hc/λ 为光子能量。

散粒噪声和读出噪声。我们将式 2 中的差分光电子计数建模为相互独立的泊松变量。由于对于每个有效乘加运算，其均值之和为$n_{0,M}|\bar{M}_{jk}|$，因此归一化输出噪声方差为

$$\begin{aligned} \operatorname{Var}(\xi_{\text{shot},j}) &= \frac{\sum_k(\mu_{jk}^+ + \mu_{jk}^-)}{n_{0,M}^2} = \frac{\|\bar{M}_{j,:}\|_1}{n_{0,M}}, \\ \operatorname{Var}(\xi_{\text{read},j}) &= \left(\frac{\sigma_e}{n_{0,M}}\right)^2. \end{aligned} \tag{15}$$

其中， σ_e 为每次接收器积分读出的均方根电子噪声。对于一次有效读出，散粒噪声方差取决于编码权重的幅值，与归一化输入值无关。我们使用从具有上述方差、相互独立且均值为零的高斯分布中抽取的样本来模拟散粒噪声和读出噪声。接收器模型采用 $\lambda = 1550$ nm、 $\eta = 0.8$ 和 $\sigma_e = 127.03$ 个电子的均方根噪声。此外，在恢复 LN/S 缩放之前，会将含噪输出限制在接收器范围内，并省略精确为零的输入。

适配器阶段与公共读出。在 NICO 和 XS/MAKI 中，最终的光学适配器分支（用于 FA 的私有 B_i，或用于 XS 的公共 S_{out}）与骨干网络W共用一个差分接收器。由于两个分支使用独立归一化的输入，因此必须在公共读出时进行增益匹配。令$s_{x,i}$表示骨干网络输入归一化， u_i表示最终光学适配器因子的局部秩坐标输入（XS 见式 13）。对于每个活动输入向量，客户端设置 $t_i = \max(s_{x,i}, \|u_i\|_\infty)$ 和 $\rho_i = t_i/s_{x,i} \geq 1$。这种增益匹配通过在接收器侧对具有固定余量的公共适配器流进行额外衰减 $\eta_i = \rho_i/R \leq 1$ 来实现，其中 R 限定了所有获准的 ρ_i。骨干网络光子预算和

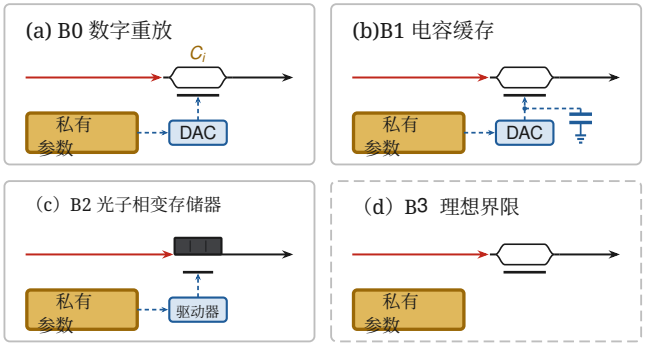


图3：私有状态驻留场景示意图。B0 通过数字方式重新加载私有权重。B1 缓存模拟偏置并进行刷新。B2 使用非易失性光子存储，但需承担编程成本。B3 假设传输成本为零。

表1：标称客户端电能系数（pJ/事件）。B1/B2 的编程成本为基于器件估算值的条件性数据；B2 的数值为下界。

事件（单位）	标称值
数模转换器 + 负载调制器（标量输入） [5, 37]	2.1
平衡接收器 +模数转换器（标量输出） [41, 65]	2.5
复合秩光电光桥（标量转换） [5, 37, 65]	5.0
集成电子（乘加运算） [48, 66]	2.9
本地 SRAM 访问（字节传输） [25, 31]	2.5
B1 模拟状态写入（偏置更新） [33]	56
B2 光子状态编程（单元更新） [13]	≥ 4.8

散粒噪声保持不变，而适配器接收的光子计数和散粒噪声方差取决于 ρ_i。对于 XS 公共接收器，输出通道 j 处的条件归一化方差为

$$\mathrm{Var}(\xi_{\mathrm{W+FT},j} \mid \rho_i) = \frac{\|\bar{W}_{j,:}\|_1 + \rho_i \|\bar{S}_{\mathrm{out},j,:}\|_1}{n_{0,W}} + \frac{\sigma_e^2}{n_{0,W}^2}. \tag{16}$$

4.2 电能模型

为量化共享光子权重流式传输和私有适配在客户端侧产生的电能成本，我们对流式传输与适配（SA）子系统进行建模。其建模的操作集合 $\mathrm{K_{SA}}$ 涵盖共享W流（以及用于 BP 的 W^{T}）所需的客户端 DAC／调制器和接收器／ADC 接口、光学与电子适配器操作、私有状态传送以及本地优化。相应的电能总和为

$$E_{\mathrm{SA,client}} = \sum_{k \in \mathcal{K}_{\mathrm{SA}}} n_k e_k, \tag{17}$$

其中 n_k 和 e_k 分别表示类别 k的执行次数和每事件电能。然后，我们将其按已服务的骨干网 MAC 运算进行归一化，得到$e_W = 10^{15} E_{\mathrm{SA,client}}/N_W$，单位为每次已服务骨干网 MAC 运算的 fJ。入射接收器的光能量使用第 4.1 节中的参考光学 MAC 能耗单独计入。

能耗核算反映了每种架构的实际执行情况。共享的光学 W+适配器输出只产生一次公共接收器/模数转换器读取，而中间秩转换则分别计入。复合操作成本仅计入一次。FO 和 BP 成本依据各自的模型调用次数计算，其中包括额外的前向评估和反向传播轮次。我们还计入运行时归一化、私有因子生成及相关内存访问。源等效模型计入 FP32 除法和逐向量 SRAM 读取。此外，我们将针对固定系数的预计算倒数以及每次调用的系数复用视为条件式算术与流量优化，同时计入其额外的设置成本和缓存成本。一次性校准设置成本不计入，但运行时使用校准系数的成本仍计入。

私有状态驻留。为考察重复私有权重传送与编程的能耗开销，我们进一步考虑以下四种私有状态驻留情景，如图 3所示：

（1）B0 以数字方式重放私有权重。（2）B1 使用易失性电容保持模拟缓存。（3）B2 假设采用非易失性光子存储。（4）B3 将私有权重传送能耗理想化为零。

状态编程能耗的假设列于表 1中。B1 假设每隔 $\tau/100$ 刷新一次，其中 τ为器件保持时间。B2 的推理设置能耗按 100 次请求分摊。

能耗系数。表1 列出了图中能耗结果所采用的名义组件成本。6。

4.3 客户端内存和存储模型

客户端峰值内存。我们通过统计各执行阶段中同时存活的张量来估算，包括持久状态和临时工作区。共享分配只计算一次，各阶段中的最大内存占用决定了实际峰值内存。为简化起见，我们假设输入已预先嵌入为 FP32，这对应于 $X \in \mathbb{R}^{B\times S\times 2048}$（在 $B = 1$，$S = 72$ 时为 B，其中 B 为批大小）。因此，不计入嵌入查找，因为其公开权重也可以流式传输以供客户端本地计算，从而避免持久存储 250 MiB 的表，同时不泄露私有令牌 ID。客户端保留 45 个冻结的 FP32 RMSNorm 向量（0.3515625 MiB）、私有状态和临时注意力／层工作区。自回归 KV 缓存已禁用，且工作区会在各层间重复使用。除私有适配器存储外，我们还计入注意力、中间激活和损失计算。尤其是，全词表交叉熵可能占用大量训练内存，而选择受监督的位置或使用融合内核可以减少其工作区 [52]。在比较推理、FO 和反向传播时，我们将这些潜在优化与已实现的模拟区分开来。

持久状态。对于因子适配，私有 $B_{i,\ell}$因子需要 24.06 MiB，此外客户端私有秩校准还需要 9.63 KiB $Q_{1,i,\ell}$。对于交叉适配，私有核心 $C_{i,\ell}$和校准 $Q_{1,i,\ell}$分别需要 38.5 KiB 和 4.81 KiB，总计 43.31 KiB。这两项估算均假设有 154 个适配层，且采用 FP32 精度。

物理容量假设。表2 体现了器件密度与保持能力之间的权衡。B1 使用相对较大的易失性电容器

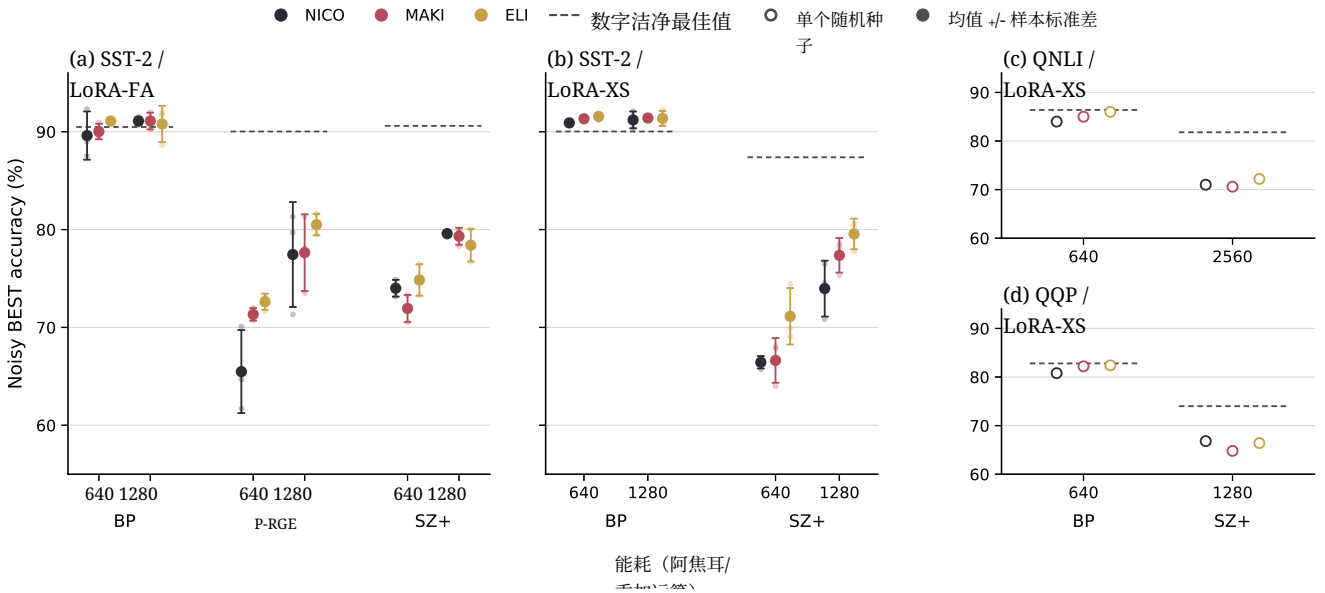


图 4：光子接收器噪声下的私有适配。(a,b) 在每次乘加运算耗能 640 和 1280 aJ 时，SST-2 上 LoRA-FA 和 LoRA-XS 在反向传播和仅前向优化下的准确率。(c) QNLI 上的 LoRA-XS 使用反向传播时为 640 aJ，使用 SubZero+ 时为 2560 aJ。(d) QQP 上的 LoRA-XS 使用反向传播时为 640 aJ，使用 SubZero+ 时为 1280 aJ。SST-2 使用 872 个验证样例；QNLI 和 QQP 则各自使用独立且固定的 500 个样例验证面板。颜色分别表示 NICO、MAKI 和 ELI。误差条表示通过筛选的硬件噪声随机种子所得的样本标准差；空心标记表示单个随机种子。灰色虚线表示与任务和方法对应、分别经数字洁净训练得到的最佳参考值。

表 2：B0-B2 模型中的私有状态容量。

资源	FA16	FA8	XS8
私有参数	6,307,840	3,153,920	9,856
B1：电容器组	90,112	45,056	64
B2：光子相变存储器单元	6,307,840	3,153,920	9,856

来保持模拟调制器偏置。因此，我们复用一个最大的模块组，并计入重新编程和刷新的开销。B2 使用密度高得多的非易失性光子相变存储器单元，因此无需在模块间重新加载即可实现全模型驻留，但需计入初始编程和训练更新的开销。根据参考器件 [13, 33], , B1 参考器件提供 5.33–5.97 位有效精度，而 B2 报告了 209 个电平（约 7.71 位）以及单个存储单元 6,000 次循环的耐久性指标。我们的容量和能耗模型假设，B1 的每个系数对应一个由电容器保持的模拟状态，或 B2 的每个系数对应一个非易失性光子单元；这两者都不意味着具备 FP32 光学精度。FP32 主副本和校准仍以数字方式实现。在固定检查点推理中，对应电平的简单量化（XS，针对 FA 重新缩放范围）在所测试的能耗下，造成的噪声准确率变化小于 1 个百分点。

5 实验设置

5.1 模型与工作负载

我们使用 FP16 1.1B TinyLlama 对 LIELLA 进行评估。该模型是一个拥有 22 层的 Transformer，隐藏维度为 2048，词汇表包含 32,000 个词元 [59]。本研究评估了三个 GLUE 任务（SST-2、QNLI、QQP [42, 51]）。对于 SST-2，我们使用 1,000 个训练样本和完整的 872 个样本验证集。QNLI 和 QQP 的评估子集各包含 500 个样本。鉴于在模拟硬件噪声下反复训练的计算成本，我们使用通用训练随机种子（42），以便在不同硬件配置之间进行受控比较。

5.2 适配与优化配置

我们对 FA 和 XS 均评估反向传播（BP）。对于仅前向（FO）适配，FA 使用 MobiZO 的并行随机梯度估计（P-RGE） [15] 和 SubZero+ [56]；在主要的噪声训练比较中，XS 使用 SubZero+。P-RGE 使用四对反向方向（每次更新进行八次损失评估）和 SGD。SubZero+ 使用 99 个单侧方向以及一个共享基线、投影子空间 Adam，并每 50 次更新进行一次基刷新。我们采用参考方法中具有代表性的批次设置：P-RGE 每次评估使用四个样本，Sub- Zero+的批大小为 16。对于 FA，SubZero+通过秩八优化子空间更新完整的私有因子，同时不改变适配器参数化方式。对于每种优化方法，我们都会进行有限的

初步超参数搜索，以选择其配置。通常会持续训练，直至验证性能趋于稳定；预先设定更新预算的实验除外。

5.3 评估协议与指标

验证准确率报告为最佳检查点（最佳，若结果并列则取最早的一次更新）。噪声关闭（关闭）是在禁用接收器噪声的情况下评估同一检查点，不进行重新训练。硬件噪声随机种子仅在配置匹配的情况下汇总。准确率通过纳入光学散粒噪声和读出噪声的模拟进行评估。电能和客户端内存根据第4节中的模型估算。资源估算采用的参考序列长度为 $S = 72$，而训练仍使用实际填充后的序列长度。使用全位置词汇表头时，每个序列涉及 $N_W \sim 74$B 次骨干网 MAC 运算。我们报告每次骨干网 MAC 运算的推理电能，以及每次更新的训练能耗，并将各方法的实际批大小和查询次数纳入计算。

6 结果

6.1 光子噪声下的学习

在接收器光子预算较低时，私有适配仍然可行。图4 显示，在接收器噪声下，BP 和 FO 均可在同一个冻结的光子骨干网络上进行独立的私有微调；在 NICO、MAKI 和 ELI 上，两者的 SST-2 准确率总体相近。在每次有效光学乘加运算能耗为 640 aJ 时，BP 使用 FA 和 XS 时在 SST-2 上的准确率最高分别达到 91.1% 和 91.6%；使用 XS 时，在 QNLI 和 QQP 上分别达到 84.0–86.0% 和 80.8–82.4%，后者接近其 82.8% 的数字无噪声参考。SubZero+ 在相同配置下，也支持在 2560 aJ 时通过全部三种适配器进行带噪声的 QNLI 适配，以及在 1280 aJ 时进行 QQP 适配。

SubZero+ 在消耗更多前向查询的同时，提高了低能耗下的 FO 准确率。在 640 aJ 下，FA–SubZero+ 在全部三种架构中的平均准确率均高于 P-RGE，其中在 NICO 中的提升最大（74.0% 对比 65.5%）。每次 SubZero+ 梯度估计使用 100 次前向评估，相当于 12.5 次 P-RGE 更新。在 640 aJ 下每种方法的九条轨迹中，所选最佳检查点平均每次运行分别需要约 44k 和 13k 次估计器前向评估（3.3×）。

BP 在光子噪声下具有显著更强的鲁棒性。在 640 aJ 下，FA–BP 和 XS–BP 的准确率均已接近各自的数字无噪声参考值。相比之下，即使在 1280 aJ 下，FA–P-RGE 和 XS–SubZero+ 在 ELI 中仍分别比各自的无噪声参考值低约 9.5 和 7.8 个百分点。这些结果推动了更具抗噪性的 FO 优化方法的开发。对于 XS，我们聚焦于 SubZero+，因为无噪声初步实验表明，P-RGE 需要显著更多的前向评估才能达到相当的准确率，但仍低于 BP。

6.2 硬件感知适配

低接收器能耗下，噪声数据训练的反向传播优于干净数据训练的反向传播。在图5 (a)中，我们将独立经过噪声数据训练的 XS–BP/ELI 最佳检查点与固定的、经数字方式干净数据训练的检查点进行比较，并在接收器噪声条件下对其进行评估。在 80 和 160 阿焦耳/乘加运算下，噪声数据训练的反向传播平均准确率分别达到 62.9% 和 86.6%，而干净数据训练的基准模型分别为 50.9% 和 58.4%。在 640 和 1280 阿焦耳/乘加运算下，高能耗条件对应的最佳检查点准确率分别达到 91.6% 和 91.4%，而同一干净数据训练的基准模型分别为 83.3% 和 87.0%。这些提升表明，通过私有参数高效微调进行接收器专用硬件适配具有潜力，这与先前关于低秩硬件误差补偿和抗噪微调的光子学研究结果一致 [7, 27, 47]。

客户端特定适配可适应异构光学接收器。 图 5(a) 进一步展示了在三种不同适配器下，于 80 和 1280 阿焦耳/乘加运算条件下训练的冻结 XS–BP；检查点从共同的 2,000 次更新 最佳 预算内选取，并在八种接收器能耗下进行评估。在 80 阿焦耳/乘加运算下，低能耗训练的 ELI 适配器达到 65.2%，而高能耗训练的适配器为 54.2%；在 1280 阿焦耳/乘加运算下，高能耗训练的适配器达到 91.1%，而低能耗训练的适配器为 81.9%。在这两种情况下，与不匹配的对应方案相比，接收器匹配训练带来了超过 11% 的相对准确率提升；MAKI 和 NICO 也得出了类似结论。在共享光学广播系统中，传播损耗和耦合效率会使不同客户端获得不同的接收器光子预算。这些结果表明，应为每个客户端单独进行接收器感知的私有微调，而非依赖单一共享适配器。

在接收器能耗较高时，FO 的准确率接近无噪声水平。图5(b) 对比了高能耗下含噪 FO 的最佳准确率与同一检查点在噪声关闭时的读出结果，以及分别在无噪声条件下训练的 ELI 参考结果。在 5120 阿焦耳/乘加运算下，ELI 中的 FA–P-RGE 和 XS–SubZero 含噪准确率分别为 86.6% 和 82.8%；在这些相同检查点上关闭推理噪声后，准确率分别为 87.9% 和 82.6%。在 10240 阿焦耳/乘加运算下，含噪准确率达到 86.8% 和 84.4%，同一检查点在噪声关闭时的读出结果则为 88.3% 和 84.2%。两项含噪结果与分别在无噪声条件下训练的 88.4% 和 86.2% 参考结果相差均不超过 2 个百分点。XS 在噪声关闭时的准确率偶尔略有下降，这可能是统计波动所致。

反向传播的准确率优势伴随着额外的硬件和信任要求。图 5(c) 显示，零向量门控使 FA–BP 在 1280 阿焦耳/乘加运算下的准确率从 55.7% 提升至 91.9%。在一项单独的固定状态诊断中，近 65 % 的转置输入恰为零，因此被省略；这导致其物理读取引入的接收器噪声传播至上游误差信号。此外，反向传播需要额外的W_T数据流；不可信的服务器可能扣留该数据流以暂停训练，或篡改它以破坏客户端侧梯度。拆分学习中相关的服务器端操纵风险也说明了这一点 [55]。相比之下，无门控的 P-RGE FO 仅使用前向数据流便达到 84.9%。这些硬件和信任方面的依赖进一步支持了我们优先选择准确率更高、查询高效的 FO 方法。

6.3 客户端侧电能

MAKI 减少了电子计算。 通过将固定的适配器变换卸载到光学流中，MAKI 消除了相应的本地电子乘加运算和权重访问，同时保留电子方式的私有计算。图 6 显示，即使在 B0 条件下，与 ELI 相比，基线模型中的 FA 推理能耗降低了 7.4%，FA 反向传播能耗降低了 10.7%。对于 XS 推理，由于两个固定因子均被卸载，能耗节省进一步达到 28.0%。

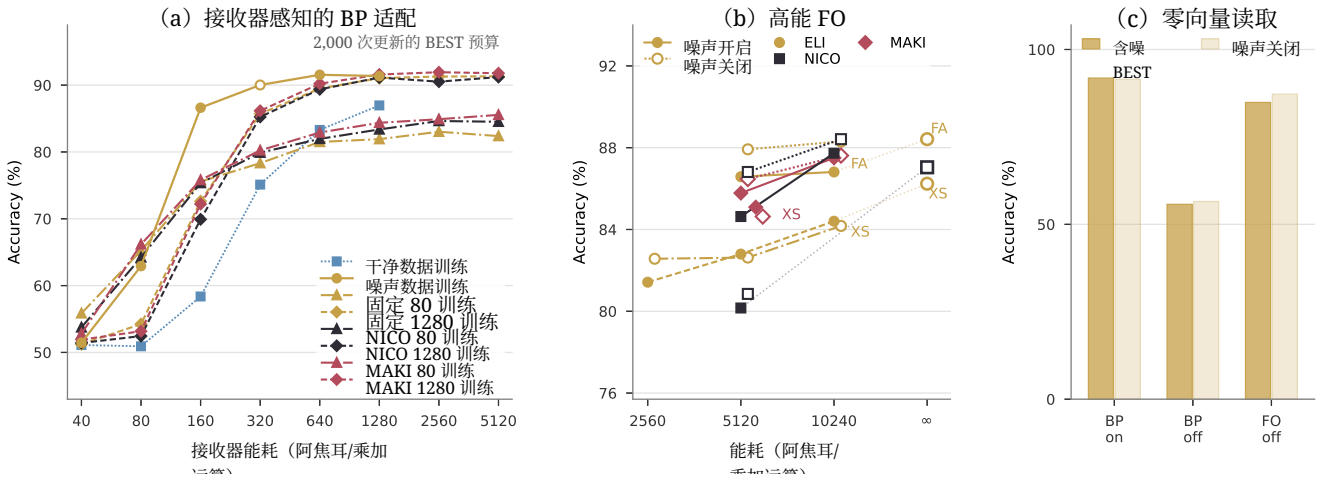


图 5：低能耗 BP 适配、高能耗 FO 恢复和零向量门敏感性。(a) 在各种能耗下，2,000 次更新预算内的 XS–BP 噪声推理曲线。ELI 噪声数据训练曲线取三个独立训练的检查点的平均值，而每条固定能耗曲线取所选单个冻结检查点的三次噪声推理结果的平均值。(b) 高能 FO 准确率。噪声开启/关闭读出使用不同的线型，其中 ELI 使用圆形标记、NICO 使用方形标记、MAKI 使用菱形标记。5120 阿焦耳/乘加运算处的 XS/NICO 噪声关闭点为 80.8%。FA/ELI、XS/ELI 和 XS/NICO 的独立无噪声数据训练参考结果显示在概念性 ∞ 处，仅通过视觉参考线连接，并非匹配的检查点。(c) FA/ELI 零向量门控：在最佳检查点下，BP 开启/关闭以及未门控 P-RGE FO 的噪声和噪声关闭准确率。

私有权重驻留对 NICO 的益处远大于对 MAKI 的益处。 尽管 FA-NICO 以光学方式执行私有-B 乘加运算，B0 仍会反复通过静态随机存取存储器和数模转换器/调制器提取并驱动 630 万个私有系数，使推理能耗升至 8.51 mJ/请求，比 ELI 高出 48.1%。易失性模拟缓存（B1）将其降至 3.45 mJ，而非易失性驻留（B2）则进一步降至 3.02 mJ，相较 ELI 分别节省 40.0% 和 47.5%。相比之下，FA-MAKI 从驻留中获益甚微，B1 的能耗甚至比 B0 高出 6.5%。由于 MAKI 保留了电子方式的私有-B 乘加运算，模拟驻留会带来额外的读出、转换和暂存开销，却无法消除这些计算。

算术优化带来了更显著的光学节能效果。进一步的能耗分解表明，重复进行归一化和重新缩放会大幅抵消光学计算的优势，在 MAKI 和 NICO 中尤为明显。因此，我们考虑通过折叠固定缩放因子，并在归一化操作间重复使用缓存的倒数来进行优化，同时通过数值检查和条件回退避免出现极小数值。这些优化对 MAKI 和 NICO 的益处大于对 ELI 的益处，因为 ELI 的电子私有核心乘加成本基本不受影响。在采用相同的受保护算术策略时，FA-MAKI B0 相比 ELI 节能 25.6%，而 FA-NICO B1/B2 的节能幅度分别达到 64.9%/73.9%。FA BP 也呈现出相同趋势：MAKI B0 节能 21.8%，NICO B1/B2 分别节能 48.7%/52.5%。对于 XS 推理，即使没有光学私有权重驻留，MAKI B0 也能节能 61.3%。

紧凑的XS私有核心可减轻权重驻留开销。 XS仅包含9,856个私有 C 系数，而FA16包含630万个，因而大幅减少了私有权重传输。在受保护的B0下，XS-NICO的能耗仅为1.25 mJ/请求，比XS-ELI低61.1%。其能耗在B0–B3各阶段的变化不到0.7%，而XS-MAKI无需光学私有核心，便能实现几乎相同的能耗。为单独考察适配器秩的影响，图 6(g)–(h)进一步比较了FA8和XS。在受保护的NICO B0下，XS可将推理能耗降低68.6%；不过，在B1下这一优势缩小至 14.7%，在B2下几乎消失，而在B2 BP下，XS仍可节省16.5%的能耗。这些发现为开发更高效的PEFT方法提供了新的硬件驱动的动机：无需依赖代价高昂的私有权重驻留，也能实现光子能耗节省。

6.4 客户端内存和私有存储

XS 可在私有存储之外进一步减少推理内存。 在 B0 下，XS-NICO 仅需 7.46 MiB 的客户端推理内存，而 FA16-NICO 需要 32.87 MiB，减少了 4.4× 。采用 MLP 分块后 [30], , 这一数值进一步降至 5.35 MiB。值得注意的是，这两种方案的内存占用都小于仅存储 FA16 私有权重所需的 24.06 MiB。然而，ELI 并不具备这一优势，因为它需要存储固定适配器因子，而 XS-MAKI 达到了与 XS-NICO 相同的 7.46 MiB。尽管共享执行缓冲区会削弱 640× 私有参数缩减的效果，XS 与光子流式传输相结合，仍能实现极低的运行时内存预算。

训练内存可以单独配置。训练工作区在很大程度上抵消了适配器尺寸带来的节省：基线 XS-NICO FO 需要 436.79 MiB，而全缓存 BP 需要 3486.82 MiB，约为 FP16 骨干权重的 1.7× 。不过，额外内存只需在适配期间配置，可供多个设备的训练重复使用，并在部署时释放。此外，我们还考察了现有的节省内存策略，包括 MLP 分块 [30], 选择性监督令牌头部计算

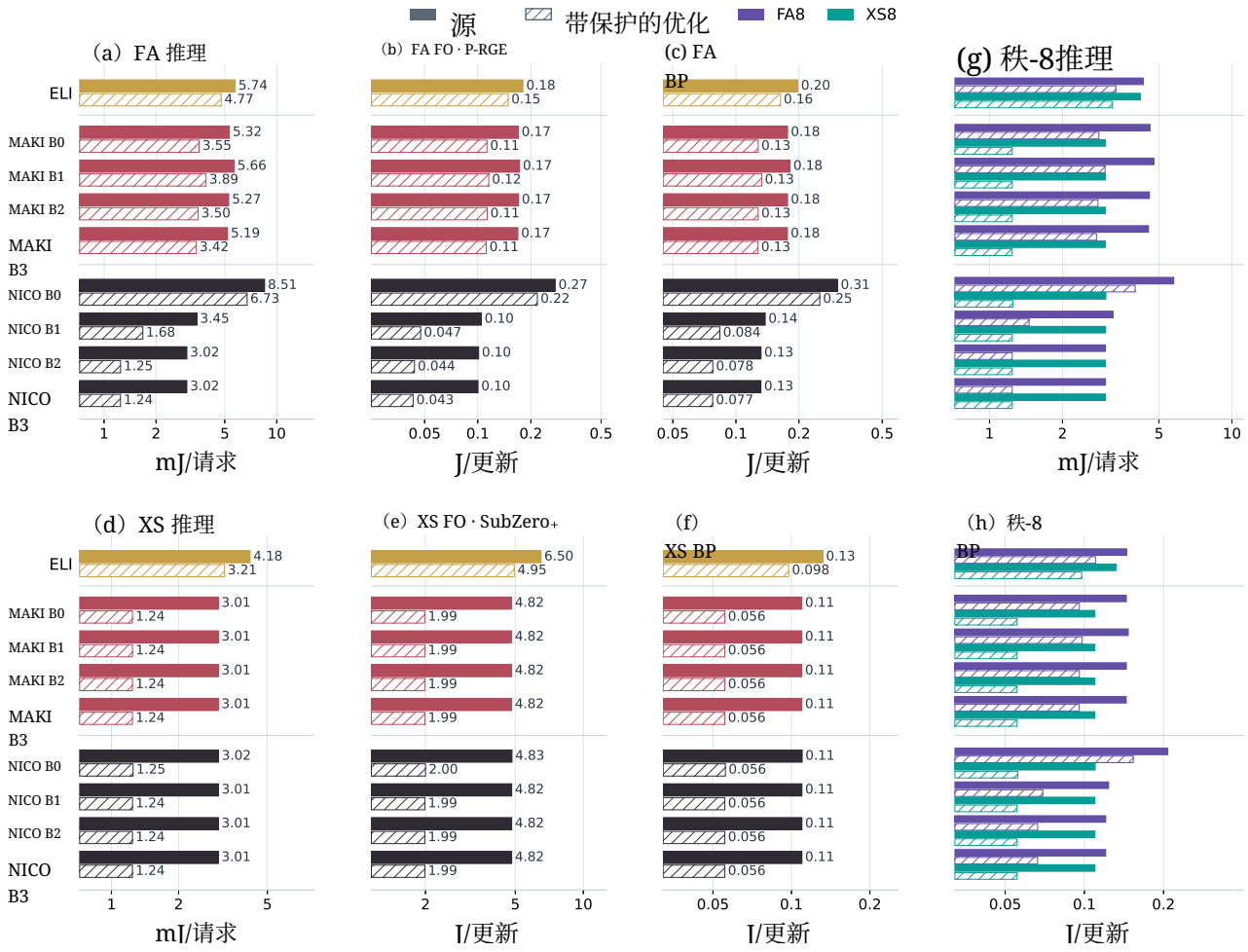


图 6：客户端 SA 电能消耗。（a）–（f）分别为 FA（上）和 XS（下）的推理、FO 及 BP；（g、h）为秩匹配的 FA8/XS8 推理和 BP。每个面板均比较 ELI 与 MAKI/NICO B0–B3。实心/阴影柱分别表示模拟源/带保护的优化算术。

与高效内存的全词汇表损失 [52], 和激活重计算 [6]。对于采用保护性算术的 XS-NICO，这些优化将 FO 内存降至名义值 77.85 MiB（有界静态随机存取存储器为 79.09 MiB），约 27× 小于 FP16 TinyL- lama 权重。FA-FO 的受益较少，因为估计器返回状态占据了其大部分内存。

反向传播以主干网络重新流式传输来换取更低的激活内存。反向传播必须保留中间激活以进行反向计算，因此激活存储成为主要的内存瓶颈。为降低这项开销，我们考虑采用激活检查点 [6]，即丢弃中间激活，并通过重新流式传输主干网络来重新计算这些激活。对于 XS-NICO，每次更新额外进行一次完整主体W遍，可将内存占用从 3486.82 MiB 降至 643.93 MiB。将重新流式传输次数增至 11.5 次，可在进一步优化头部和多层感知机之前，将内存占用进一步降至约 450 MiB。采用全部三项建模优化后，带保护机制的反向传播达到 170.19 MiB（有界静态随机存取存储器为 180.02 MiB），并需要 2.28× 这一有界内存占用与优化后的 XS-FO 相当，同时还需进行转置权重流式传输并承担额外的重计算成本。这种组合除了 g22 额外增加的 253 次主体块前向计算外，还增加了 88 次全词元门控/升维矩阵前向计算；图 (g) 单独展示了检查点的效果。这些要求进一步凸显了开发更准确、查询高效且抗噪的 FO 方法的必要性。

7 讨论与相关工作

以往研究探索了广播推理（Netcast 和 AIR-LLM） [16, 23, 41], 光子与模拟加速或学习（Albireo、FICONN、L2ight、Mirage、Opto-ViT-v2、ALoRA 和 AHWA-LoRA） [3, 7, 11, 20, 34, 40, 47], 以及分布式或边缘微调（Petals、DLoRA、MobiZO 和 MeBP） [4, 14, 15, 43]。LIELLA 在这些方向的基础上，将共享的、与客户端无关的权重传输与本地可变的私有适配分离开来，无需进行逐客户端服务器端骨干网络计算。它还可以被视为一种通用架构，并不局限于 Netcast 风格

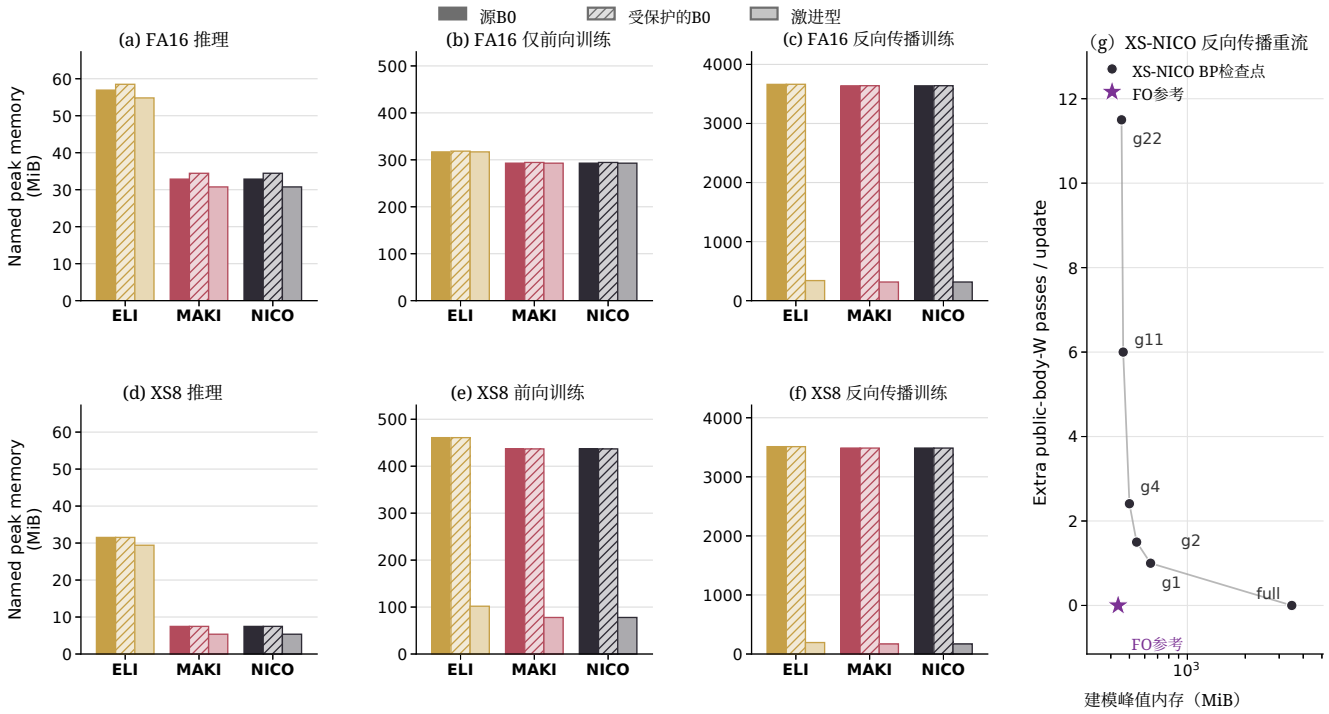


图 7：客户端内存。(a)–(f) ELI、MAKI 和 NICO 下 FA 与 XS 的峰值命名缓冲区内存，对比源实现、保护性算术和仅内存激进优化投影。MAKI 和 NICO 使用 B0 数字重放。所有方法均使用 $S = 72$；推理使用 (Batch)$B = 1$，FA 使用 P-RGE FO $B = 4$，XS 使用 SubZero+ FO 和 BP $B = 16$。(g) 仅检查点 XS-NICO BP 内存与额外主干网络重新流式传输的对比；紫色星号标示未优化的 XS FO 参照。“g”表示组大小，即每个检查点分段中的 Transformer 块数量；“full”表示完整激活缓存。

服务，还可以扩展到 AIR-LLM 风格的无线射频权重广播，或 QAMNet 的相干同相/正交光学计算 [1], , 后者借助不同的接收器、校准和噪声模型，可以自然地保持公有/私有分离。相关工作进一步通过量子安全多方学习展示了光学计算实现强安全性保障的潜力 [44]。

评估的 LoRA、LoRA-FA 和 LoRA-XS 变体 [2, 26, 58] 实例化了 LIELLA 的适配器接口，而非定义该接口。FourierFT 使用稀疏可训练频谱系数 [18], , 这表明光学傅里叶处理可用于客户端特定的适配器。Quantum-PEFT 使用紧凑的基于泡利矩阵的酉参数化 [32], , 这表明可编程光子量子电路是另一种潜在的私有计算途径。这些方法均兼容基于光的计算，并推动进一步开发有利于光学计算的 PEFT。

仅前向和无梯度方法涵盖 MeZO、SubZero 和 SubZero + [36, 56, 57]；方差缩减或自适应零阶优化（ZO）方法，如 MeZO-SVRG、ZO-AdaMU、HiZOO、PaZO 和 DiZO [19, 28, 45, 61, 62]；以及 LOZO、AdaZeta、EGGROLL 和 FZOO 中的结构化或批量搜索 [8, 10, 38, 54]。与反向传播相比，这些零阶优化（ZO）方法可以进一步提升安全性并带来内存优势，但可能需要大量含噪前向评估，且抗噪性较弱。更根本的是，Transformer [50]并不特别友好，因为注意力和训练时激活缓冲区仍然会带来较高的客户端内存需求，这可能意味着需要更优的、面向大语言模型的原生光学模型架构。

我们共同展望 LIELLA 的硬件实现，将共享光学权重流式传输与客户端私有适配器相结合。此类实现将支持对能耗、延迟和可扩展性进行端到端表征，同时为原生光子模型架构、硬件高效的 PEFT 以及抗噪 FO 优化开辟新的机遇。

8 结论

我们介绍了 LIELLA，使共享冻结的光子权重流的各客户端能够进行独立私有适配。通过光子适配器协同设计与光电布局，LIELLA 显著降低了客户端存储和电力成本。我们进一步刻画了光子噪声下的仅前向服务和反向传播服务，揭示了它们在准确率、能耗和内存方面的权衡。除架构本身外，我们的研究结果还为开发更紧凑、更有效的 PEFT 方法，以及更准确、查询高效且抗噪的 ZO 优化算法提供了新的硬件驱动动机。

9 致谢

感谢 Allen Zang 和 Junming Gong 富有帮助的讨论。我们感谢美国陆军研究办公室（ARO）（W911NF-23-1-0077）提供的支持，

ARO MURI（W911NF-21-1-0325）、AFOSR MURI（FA9550-21-1-0209、FA9550-23-1-0338）、ONR MURI（N000142612102）以及美国国家科学基金会（NSF）（OSI-2326767、OSI-2426975、EEC-2550064、OSI-2553578、OSI-2553619）提供的支持。本研究还借助了芝加哥大学研究计算中心提供的资源完成。我们感谢 NVIDIA 通过 NVIDIA 学术资助计划为本研究所使用的图形处理器提供支持。我们还感谢本研究使用了 Google 云平台（GCP）的计算资源。K.S. 感谢欧盟地平线欧洲研究与创新计划根据玛丽·斯克沃多夫斯卡-居里资助协议（编号：101202109）提供的支持。

10 生成式人工智能披露。

研究构想、系统设计、实验方法和科学解释均由人类作者提出并主导。生成式人工智能工具，包括 OpenAI ChatGPT 和 Codex，在作者监督下协助完成代码实现、数据组织和稿件编辑。作者审阅并验证了所有技术内容、实验结果和结论，并对本研究承担全部责任。

参考文献

[1] Marc Gong Bacvanski、Sri Krishna Vadlamani、Kfir Sulimany 和 Dirk Robert Englund。2024年。QAMNet：快速高效的光学 QAM 神经网络。arXiv: 2409 12305 2409 12305 。https://arxiv.org/abs/ 。Klaudia [2] Bałazy、Mohammadreza Banaei、Karl Aberer 和 Jacek Tabor。2024年。LoRA-XS：以极少参数实现低秩适配。arXiv:2405.17604 https://arxiv.org/abs/2405.17604[3] Saumil Bandyopadhyay、Alexander Sludds、Stefan Krastanov、Ryan Hamerly、Nicholas C. Harris、Darius Bunandar、Matthew Streshinsky、Michael Hochberg 和 Dirk Englund。2024年。采用仅前向训练的单芯片光子深度神经网络。《自然·光子学》18卷，第12期（2024年），1335–1343。doi:10.1038/ s41566-024-01567-z[4] Alexander Borzunov、Dmitry Baranchuk、Tim Dettmers、Maksim Riabinin、Younes Belkada、Artem Chumachenko、Pavel Samygin 和 Colin Raffel。2023年。Petals：大型模型的协作推理与微调。载于《计算语言学协会第61届年会论文集》（第3卷：系统演示）。计算语言学协会，558–568。doi:10.18653/v1/2023.acl-demo.54[5] R. L. Chen、H. W. Ting 和 S. J. Chang。2012年。用于超宽带收发器的六位 2.7-GS/s、5.4-mW 奈奎斯特互补金属氧化物半导体数模转换器。《IET 电路、器件与系统》6卷，第2期（2012年），95–102。doi:10.1049/iet-cds.2011.0192[6] Tianqi Chen、Bing Xu、Chiyuan Zhang 和 Carlos Guestrin。2016年。以次线性内存开销训练深度网络。arXiv:1604.06174 https://arxiv.org/abs/ 1604.06174[7] Xuming Chen、Deniz Najafi、Mehrdad Morsali、Chengwei Zhou、Zahra Ghanaa- tianjobzari、Mahdi Nikdast、Shaahin Angizi 和 Gourav Datta。2026年。Opto-ViT-v2：用于光子近传感器视觉变换器加速器的抗噪片上微调。arXiv: 2607 19421 2607 19421 。[8] https://arxiv.org/abs/ 。Yiming Chen、Yuan Zhang、Liyuan Cao、Kun Yuan 和 Zaiwen Wen。2025年。利用低秩结构提升语言模型的零阶微调。在国际学习表征会议论文集。https://proceedings.iclr.cc/paper_files/paper/2025/hash/9ccc9d814d3dee4750debaf23061e733-Abstract-Conference.html[9] Meihua Dang、Anikait Singh、Linqi Zhou、Stefano Ermon 和 Jiaming Song。2025年。扩散模型的个性化偏好微调。载于 IEEE/CVF 计算机视觉与模式识别会议论文集。8020–8030。https://openaccess.thecvf.com/content/CVPR2025/html/Dang_Personalized_Preference_Fine-tuning_of_Diffusion_Models_CVPR_2025_paper.html[10] Sizhe Dang、Yangyang Guo、Yanjun Zhao、Xiaodong Zheng、Guang Dai、Ivor Tsang 和 Haishan Ye。2026年。FZOO：面向大语言模型微调的快速零阶优化器，达到 Adam 级速度。载于国际学习表征会议论文集。https://proceedings.iclr.cc/paper_files/paper/2026/hash/b8d4c82b365d9361863000f8d1e86f1b-Abstract-Conference.html[11] Cansu Demirkiran、Guowei Yang、Darius Bunandar 和 Ajay Joshi。2024年。Mirage：用于深度神经网络训练的基于 RNS 的光子加速器。载于第51届 IEEE 国际计算机体系结构研讨会论文集。doi:10.1109/ ISCA59077.2024.00016

[12] AbdAllah M. A. Elsheikh、Jon Rokne 和 Reda Alhajj。2026年。《基础模型中的隐私保护：技术、威胁与权衡的系统综述》。《人工智能评论》59，文章编号 145（2026年）。doi:10.1007/s10462-026- 11535-4[13] Chen Gao、Zhou Han、Gaofei Wang 和 Wentao Huang。2025年。《结构优化的光子相变存储器：实现高存储密度与耐久性，面向可重构电信系统》。《光子学》 12, 11 2025 1130 10 3390 12111130 ( ), 。doi: 。/photonics Chao [14] Gao 和 Sai Qian Zhang。。DLoRA：面向大语言模型的分布式参数高效微调解决方案。载于计算语言学协会自然语言处理实证方法会议 2024 年论文集。计算语言学协会，2024 13703 13714 10 18653 1 2024 802 –。doi:。/v / . [15] findings-emnlp。Lei Gao、Amir Ziashahabi、Yue Niu、Salman Avestimehr 和 Murali Annavaram。2025年。MobiZO：借助推理引擎实现边缘端高效大语言模型微调。载于《2025年自然语言处理实证方法会议论文集》。计算语言学协会，20206–20223。doi: 10 18653 1 2025 1022 。/v / .emnlp-main。Zhihui [16] Gao、Tingjun Chen 和 Dirk Englund。2026年。AIR-LLM：通过无线电广播人工智能权重，借助射频计算实现无存储器边缘大语言模型推理。arXiv:2610.00465 https://arxiv.org/abs/2610.00465[17] Zhihui Gao、Sri Krishna Vadlamani、Kfir Sulimany、Dirk Englund 和 Tingjun Chen。2026年。通过射频物理内计算实现解耦式机器学习。《科学进展》12，2（2026年），eadz0817。doi:10.1126/sciadv.adz0817[18] Ziqi Gao、Qichao Wang、Aochuan Chen、Zijing Liu、Bingzhe Wu、Liang Chen 和 Jia Li。2024年。基于离散傅里叶变换的参数高效微调。载于第41届国际机器学习会议论文集（《机器学习研究论文集》，第235卷）。PMLR，14884–14901。https://proceedings. mlr.press/v235/gao24o.html[19] Tanmay Gautam、Youngsuk Park、Hao Zhou、Parameswaran Raman 和 Wooseok Ha。2024年。用于微调语言模型的方差缩减零阶方法。载于第41届国际机器学习会议论文集（《机器学习研究论文集》，第235卷）。PMLR，15180–15208。https://proceedings.mlr.press/v235/gautam24a.html [20] Jiaqi Gu、Hanqing Zhu、Chenghao Feng、Zixuan Jiang、Ray T. Chen 和 David Z. Pan。2021年。L2ight：通过高效的原位子空间优化实现光学神经网络的芯片上学习。载于《神经信息处理系统进展》。arXiv:2110.14807 https://arxiv.org/abs/2110.14807[21] Ryan Hamerly、Liane Bernstein、Alexander Sludds、Marin Soljačić 和 Dirk Englund。2019年。基于光电乘法的大规模光学神经网络。《物理评论 X》9，2（2019年），021032。doi:10.1103/PhysRevX.9. 021032[22] Ryan Hamerly、Alexander Sludds、Saumil Bandyopadhyay、Liane Bernstein、Zai-jun Chen、Manya Ghobadi 和 Dirk R. Englund。2021年。通过波分复用权重广播实现光学神经网络边缘计算。载于《人工智能新兴主题》（ETAI）2021（《SPIE论文集》，第11804卷）。SPIE，118041R。doi:10.1117/12.2594886[23] Ryan Hamerly、Alexander Sludds、Saumil Bandyopadhyay、Zaijun Chen、Zhizhen Zhong、Liane Bernstein 和 Dirk Englund。2022年。Netcast：基于波分复用定义的光学神经网络的低功耗边缘计算。arXiv:2207.01777 https://arxiv.org/abs/ 2207 01777 。Zeyu [24] Han、Chao Gao、Jinyang Liu、Jeff Zhang 和 Sai Qian Zhang。2024年。大模型参数高效微调：全面综述。《机器学习研究汇刊》（2024年）。https://openreview.net/forum?id=lIsCS8b6zj [25] Mark Horowitz。2014年。计算能耗问题（以及我们能做些什么）。载于IEEE国际固态电路会议（ISSCC）。doi:10.1109/ISSCC.2014.6757323[26] Edward J. Hu、Yelong Shen、Phillip Wallis、Zeyuan Allen-Zhu、Yuanzhi Li、Shean Wang、Lu Wang 和 Weizhu Chen。2022年。LoRA：大语言模型的低秩适配。载于国际学习表征会议。https://arxiv.org/abs/2106.09685[27] Yuyao Huang、Tingzhao Fu、Honghao Huang、Run Sun、Wencan Liu、Sigang Yang 和 Hongwei Chen。2024年。通过低秩适配校准芯片上衍射光学神经网络的误差。载于 CLEO 2024，《技术文摘系列》。Optica出版集团。doi:10.1364/CLEO_SI.2024.SM3M.1 论文 SM3M.1。[28] Shuoran Jiang、Qingcai Chen、Youcheng Pan、Yang Xiang、Yukang Lin、Xiangping Wu、Chuanyi Liu 和 Xiaobao Song。2024年。ZO-AdaMU优化器：在零阶优化中利用动量和不确定性调整扰动。载于《AAAI人工智能会议论文集》，第38卷。18363–18371。doi:10.1609/aaai.v38i16.29796[29] Yeachan Kim、Mingyu Lee 和 SangKeun Lee。2026年。大语言模型内存高效微调综述。《计算语言学协会汇刊》 14 2026 960 982 10 1162 692 ( ), –。doi: 。/ [30] tacl.a。Nikita Kitaev、Łukasz Kaiser 和 Anselm Levskaya。。Reformer：高效的 2020 Transformer。载于国际学习表征会议。https: //arxiv.org/abs/2001.04451

[31] Adrian Kneip 和 David Bol。2023年。一款采用 7T-NDR 双电源、28 纳米 FD-SOI 工艺的超低功耗 SRAM：在 64 MHz 下实现 0.23-nW/kB 睡眠保持功耗和 0.8-pJ/32b 访问，并采用正向背栅偏置。《IEEE 电路与系统汇刊 I：常规论文》 70, 3 2023 1311 1323 10 1109 2022 3230984 ( )，–。doi: 。/TCSI. 。[32] Toshiaki Koike-Akino、Francesco Tonin、Yongtao Wu、Frank Zhengqing Wu、Leyla Naz Candogan 和 Volkan Cevher。2025年。Quantum-PEFT：超高效参数微调。载于国际学习表征会议。https://proceedings.iclr.cc/paper_files/paper/2025/hash/4e52bbb99690d1e05c7ef7b4c8b3569a-Abstract-Conference.html[33] Sean Lam、Ahmed Khaled、Simon Bilodeau、Bicky A. Marquez、Paul R. Prucnal、Lukas Chrostowski、Bhavin J. Shastri 和 Sudip Shekhar。2026年。采用电光模拟存储器的神经形态光子计算。《自然·通讯》17（2026年），2472。doi:10.1038/s41467-026-69084-x[34] Chen Li、Elena Ferro、Corey Lammie、Manuel Le Gallo、Irem Boybat 和 Bipin Rajendran。2026年。通过低秩适配器实现面向模拟存内计算的高效 Transformer 适配。《神经形态计算与工程》6，1（2026年），014011。doi:10.1088/2634-4386/ae405e[35] Xiaopeng Li、Yuanjin Zheng、Wanyu Wang、Wenlin Zhang、Pengyue Jia、Yingyi Zhang、Haiying He、Mengyang Ma、Yiqi Wang、Maolin Wang、Xuetao Wei 和 Xiangyu Zhao。2026年。MTA：面向个性化大语言模型的先合并后适配框架。载于计算语言学协会第64届年会论文集（第1卷：长篇论文）。计算语言学协会，44078–44091。doi:10.18653/v1/2026.acl-long.2040[36] Sadhika Malladi、Tianyu Gao、Eshaan Nichani、Alex Damian、Jason Lee、Danqi Chen 和 Sanjeev Arora。2023年。仅通过前向传播微调语言模型。载于《神经信息处理系统进展》，第36卷。https://proceedings.neurips.cc/paper_files/paper/2023/hash/a627810151be4d13f907ac898ff7e948-Abstract-Conference.html[37] Shaoyuan Ou、Kaiwen Xue、Lian Zhou、Chun ho Lee、Alexander Sludds、Ryan Hamerly、Ke Zhang、Hanke Feng、Yue Yu、Reshma Kopparapu、Eric Zhong、Cheng Wang、Dirk Englund、Mengjie Yu 和 Zaijun Chen。2025年。基于超多路复用集成光子学的光学张量处理器。《科学进展》11，23（ 2025 0228 10 1126 0228），eadu。doi: 。/sciadv.adu Bidipta Sarkar、Mattie Fellows、Juan Agustin Duque、Alistair Letcher、Antonio León Villares、Anya Sims、Clarisse Wibault、Dmitry Samsonov、Dylan Cope、Jarek Luca Liesen、Kang Li、Lukas Seier、Theo Wolf、Uljad Berdica、Valentin Mohl、Alexander David Goldie、Aaron Courville、Karin Sevegnani、Shimon Whiteson 和 Jakob Nicolaus Foerster。2026年。超大规模进化策略。载于第43届国际机器学习会议论文集（《机器学习研究论文集》，第306卷）。PMLR，107670–107741。https://proceedings.mlr.press/v306/sarkar26a.html [39] Bhavin J. Shastri、Alexander N. Tait、Thomas Ferreira de Lima、Wolfram H. P. Pernice、Harish Bhaskaran、C. David Wright 和 Paul R. Prucnal。2021年。用于人工智能和神经形态计算的光子学。《自然·光子学》15，2（2021年），102–114。doi:10.1038/s41566-020-00754-y [40] Kyle Shiflett、Avinash Karanth、Razvan Bunescu 和 Ahmed Louri。2021年。Albireo：通过硅光子学实现卷积神经网络的能效加速。载于第48届年度国际计算机体系结构研讨会论文集。860–873。doi:10.1109/ISCA52012.2021.00072[41] Alexander Sludds、Saumil Bandyopadhyay、Zaijun Chen、Zhizhen Zhong、Jared Cochrane、Liane Bernstein、Darius Bunandar、P. Ben Dixon、Scott A. Hamilton、Matthew Streshinsky、Ari Novack、Tom Baehr-Jones、Michael Hochberg、Manya Ghobadi、Ryan Hamerly 和 Dirk Englund。2022年。互联网边缘的去中心化光子深度学习。《科学》（ 2022 10 1126 8271）。doi: 。/science.abq[42] Richard Socher、Alex Perelygin、Jean Wu、Jason Chuang、Christopher D. Manning、Andrew Ng 和 Christopher Potts。2013年。基于情感树库的语义组合性递归深度模型。载于2013年自然语言处理实证方法会议论文集。计算语言学协会，1631–1642。https://aclanthology.org/D13-1170/ [43] Congzheng Song 和 Xinyu Tang。2025年。面向资源受限移动设备上大语言模型微调的内存高效反向传播。载于《2025年自然语言处理实证方法会议论文集：产业分会》。计算语言学协会，766–774。doi:10.18653/v1/2025.emnlp-industry.52[44] Kfir Sulimany、Sri Krishna Vadlamani、Ryan Hamerly、Prahlad Iyengar 和 Dirk Englund。2025年。量子安全多方深度学习。《物理评论 X》15，4（2025年），041056。doi:10.1103/k8wg-qmbh[45] Qitao Tan、Jun Liu、Zheng Zhan、Caiwen Ding、Yanzhi Wang、Xiaolong Ma、Jaewoo Lee、Jin Lu 和 Geng Yuan。2025年。分歧中的协调：迈向快速、精准且内存高效的大语言模型零阶微调。载于《神经信息处理系统进展》，第38卷。doi:10.52202/085713-5819[46] Zhaoxuan Tan、Qingkai Zeng、Yijun Tian、Zheyuan Liu、Bing Yin 和 Meng Jiang。2024年。通过个性化参数高效微调普及大语言模型。载于《2024年自然语言处理实证方法会议论文集》。计算语言学协会，6476–6491。doi:10.18653/v1/2024.emnlp-main.372

[47] Taichi Taniguchi、Mitsumasa Nakajima、Kohei Ikeda、Toshikazu Hashimoto 和 Satoshi Kawakami。2025年。ALoRA：面向光子大型神经网络的硬件感知微调。载于《2025年第三十届光电子与通信会议（OECC）暨光子交换与计算国际会议（PSC）》。IEEE。doi: 10 23919 62146 2025 11109873 。/OECC/PSC。.[48] Yvan Tortorella、Luca Bertaccini、Davide Rossi、Luca Benini 和 Francesco Conti。2022年。RedMulE：面向基于RISC-V的超低功耗SoC中自适应深度学习的紧凑型FP16矩阵乘法加速器。arXiv:2204.11192 https://arxiv.org/abs/ 2204 11192 。Sri Krishna Vadlamani、Kfir Sulimany、Zhihui Gao、[49] Tingjun Chen 和 Dirk Englund。2025年。无线边缘网络中的机器智能。arXiv:2506.12210 https://arxiv.org/abs/2506.12210[50] Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszkoreit、Llion Jones、Aidan N. Gomez、Łukasz Kaiser 和 Illia Polosukhin。2017年。注意力就是一切。载于《神经信息处理系统进展》第30卷。https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html[51] Alex Wang、Amanpreet Singh、Julian Michael、Felix Hill、Omer Levy 和 Samuel R. Bowman。2018年。GLUE：自然语言理解的多任务基准测试与分析平台。载于《2018年EMNLP研讨会：BlackboxNLP——分析和解读用于NLP的神经网络论文集》。计算语言学协会，353–355。doi:10.18653/v1/W18-5446[52] Erik Wijmans、Brody Huval、Alexander Hertzberg、Vladlen Koltun 和 Philipp Krähenbühl。2025年。降低大词汇量语言模型的损失。载于《国际学习表征会议》。https://proceedings.iclr.cc/paper_files/paper/2025/hash/aa963ac256590bb7ad5fc26c68229a3a-Abstract-Conference.html[53] Mengwei Xu、Dongqi Cai、Wangsong Yin、Shangguang Wang、Xin Jin 和 Xuanzhe Liu。2025年。基础模型的资源高效算法与系统：综述。《计算调查》57卷5期，文章110（2025年），39页。doi:10.1145/3706418[54] Yifan Yang、Kai Zhen、Ershad Banijamali、Athanasios Mouchtaris 和 Zheng Zhang。2024年。AdaZeta：面向内存高效大语言模型微调的自适应零阶张量列适配。载于《2024年自然语言处理实证方法会议论文集》。计算语言学协会， 977 995 10 18653 1 2024 56 – 。doi: 。/v / .[55] emnlp-main。Fangchao Yu、Bo Zeng、Kai Zhao、Zhi Pang 和 Lina Wang。2024年。慢性投毒：针对拆分学习的后门攻击。载于《AAAI人工智能会议论文集》第38卷。16531–16538。doi:10.1609/aaai.v38i15. 29591[56] Ziming Yu、Shuyao Xiao、Xingyu Zhao、Sike Wang、Pan Zhou、Peiyu Zang、Xiangda Yan、Yongjie Yang 和 Jia Li。2026年。SubZero+：在随机子空间中进行内存高效的自适应大语言模型零阶微调。arXiv:2608.15665 https://arxiv.org/abs/ 2608 15665 2 。v Ziming Yu、Pan Zhou、Sike Wang、Jia Li、Mi Tian[57] 和 Hua Huang。2025年。在随机子空间中对大语言模型进行零阶微调。arXiv:2410.08989 https://arxiv.org/abs/2410.08989。ICCV 2025年最终定稿版本。[58] Longteng Zhang、Lin Zhang、Shaohuai Shi、Xiaowen Chu 和 Bo Li。2023年。LoRA-FA：高效且有效的低秩表示微调。arXiv:2308.03303 https://arxiv.org/abs/2308.03303。版本3，2026年。[59] Peiyuan Zhang、Guangtao Zeng、Tianduo Wang 和 Wei Lu。2024年。TinyLlama：开源小型语言模型。arXiv:2401.02385 https://arxiv.org/abs/ 2401.02385[60] Yihua Zhang、Pingzhi Li、Junyuan Hong、Jiaxiang Li、Yimeng Zhang、Wenqing Zheng、Pin-Yu Chen、Jason D. Lee、Wotao Yin、Mingyi Hong、Zhangyang Wang、Sijia Liu 和 Tianlong Chen。2024年。重新审视内存高效大语言模型微调的零阶优化：一项基准测试。载于《第41届国际机器学习会议论文集》（《机器学习研究论文集》第235卷）。PMLR，59173–59190。https://proceedings.mlr.press/v235/ zhang24ad.html [61] Hanzhen Zhao、Shihong Ding、Cong Fang 和 Zhouchen Lin。2025年。PaZO：用于大语言模型微调的预条件加速零阶优化。载于《神经信息处理系统进展》第38卷。doi:10.52202/085713- 3711[62] Yanjun Zhao、Sizhe Dang、Haishan Ye、Guang Dai、Yi Qian 和 Ivor Tsang。2025年。轻松实现大语言模型二阶微调：一种海森矩阵引导的零阶优化器。载于《国际学习表征会议》。https://proceedings.iclr.cc/paper_files/paper/2025/hash/6bf82cc56a5fa0287c438baa8be65a70-Abstract-Conference.html[63] He Zhu、Ren Togo、Takahiro Ogawa 和 Miki Haseyama。2026年。通过高效微调与不确定性感知解耦实现医学视觉语言模型的个性化联邦学习。《生物医学信息学杂志》178卷（ 2026 105014 10 1016 2026 105014），。doi: 。/j.jbi. 。Ligeng Zhu、Zhijian Liu 和 Song Han。2019年。梯度深度泄漏。载于《神经信息处理系统进展》第32卷。[64] https://proceedings.neurips.cc/paper/2019/hash/60a6c4002cc7b29142def8871531281a-

摘要 .html

[65] Yuanming Zhu、Shengchang Cai、Shiva Kiran、Yang-Hang Fan、Po-Hsuan Chang、Sebastian Hoyos 和 Samuel Palermo。2020年。一款采用输出电平移位稳定技术的14纳米CMOS 1.5-GS/s 8位流水线逐次逼近型模数转换器。载于*IEEE*定制集成电路会议（*CICC*）。doi:10.1109/CICC48029.2020.9075942[66] Brian Zimmer、Rangharajan Venkatesan、Yakun Sophia Shao、Jason Clemons、Matthew Fojtik、Nan Jiang、Ben Keller、Alicia Klinefelter、Nathaniel Pinckney,

Priyanka Raina、Stephen G. Tell、Yanqing Zhang、William J. Dally、Joel S. Emer、C. Thomas Gray、Stephen W. Keckler 和 Brucek Khailany。2019年。一种基于地参考信号传输、采用16纳米工艺的0.11 pJ/Op、0.32–128 TOPS、可扩展多芯片模块深度神经网络加速器。载于 超大规模集成电路研讨会。 doi:10.23919/VLSIC.2019.8778056