边缘大型 AI 模型：革新 6G 网络

Zixin Wang, Yuanming Shi, Yong Zhou, Jingyang Zhu, and Khaled. B. Letaief

摘要—大型人工智能模型（LAMs）具备解决广泛现实世界问题的人类般能力，体现了各领域和模态专家的潜力。通过利用地理位置分散的边缘设备的通信和计算能力，边缘 LAM 作为一种使能技术出现，能够推动在 6G 中提供各种实时智能服务。与主要使用小型模型支持单一任务的传统边缘人工智能（AI）不同，边缘 LAM 的特点是需要大型模型的分解和分布式部署，并且有能力支持高度通用和多样化的任务。然而，由于无线网络中的通信、计算和存储资源有限，大量的可训练神经元和显著的通信开销给边缘 LAM 的实际部署带来了巨大障碍。在本文中，我们从模型分解和资源管理的角度研究了边缘 LAM 的机会与挑战。具体来说，我们提出了协作微调和全参数训练框架，以及一个由微服务辅助推理架构支持的方法，以提高无线网络中的边缘 LAM 部署效率。此外，我们还探讨了边缘 LAM 在空中接口设计中的应用，重点在于信道预测和波束成形。这些创新的框架和应用程序为推进 6G 技术提供了宝贵的见解和解决方案。

I. 介绍

随着人工智能（AI）的显著进步，大型 AI 模型（LAMs）现在在执行现实世界中的复杂任务方面表现出色。它们是跨模态的、高度通用化的，并且擅长知识迁移，展示了人工通用智能 [1] 的巨大潜力。

通过革命性的模型架构、庞大的参数规模、丰富的数据和大量的计算资源，LAMs 从预训练中获得了前所未有的泛化能力。它们可以适应广泛的下游任务，只需少量样本甚至零样本学习就能达到与人类智力相当的性能水平。

这种能力在各个领域开启了突破性应用，承诺开启一个全能智能的新时代。例如，92%的《财富》500 强公司已将 ChatGPT 作为其运营基础设施的一部分，而配备特斯拉 FSD 的车辆已经行驶了超过 25.75 亿公里。

此外，使用高质量电信特定数据集训练 LAMs 能够解锁其在电信领域内有效处理各种任务的潜力 [2]。然而，许多新兴应用（如自动驾驶和机器人）依赖于云服务器，这不可避免地导致严重的隐私问题和高延迟决策。

Z. Wang and K. B. Letaief are with The Hong Kong University of Science and Technology, Hong Kong. Y. Shi, Y. Zhou, and J. Zhu are with ShanghaiTech University, China.

因此，在 6G 中将具备强大泛化能力的 LAMs 推向网络边缘能够为移动用户提供个性化、实时且值得信赖的智能服务。

然而，边缘 LAM 面临的挑战也是显而易见的。LAM 的主要特征可以总结为三个 “巨大” 方面：大量的参数、广泛的数据供应和巨大的计算需求。例如，GPT-3 拥有 170 亿个参数，并在数万台 Nvidia V100 GPU 上进行训练，使用了大约 3000 亿个单词和 570GB 的训练数据。传统的边缘 AI 训练依赖于联邦学习（FL），整个模型是在资源受限的边缘设备 [3] 上进行训练的。这种方法对于边缘 LAMs 的训练是不可行的，因为边缘设备的计算、存储和通信资源有限。对于边缘 LAM 推理而言，多个下游任务需要处理多模态输入数据，这些数据必须依次通过各种 LAM 模块以生成推理结果。然而，传统的边缘 AI 推理仅限于处理单任务和单一模式场景，这不足以满足多模态多任务的需求。鉴于上述边缘 LAM 部署中存在的挑战，解决这些问题需要建立分布式和可信的训练框架，确保资源高效且低延迟的推理架构，并探索其在空中接口设计中的应用，以充分利用边缘 LAM 在革新 6G 网络方面的潜力。

训练: 边缘 LAM 的训练既涉及微调，也包括从零开始学习（例如，全参数训练），使用分布在边缘设备上的敏感数据，旨在提高学习性能并增强可信度。这一过程涉及更新至少数亿个参数，这对于单个边缘设备来说是无法承受的，需要将学习模型分解到各个设备和服务端。因此，可以采用各种并行化框架，例如数据、张量和流水线并行性 [4]，以促进网络边缘实时且协作的训练。然而，现有的并行化框架主要是为云计算中心和集中式数据集设计的，在这些场景中数据隐私不是主要关注点，而在边缘 LAM 训练中则不同。因此，增强现有针对边缘 LAMs 的并行化框架至关重要，以保护本地数据隐私，并建立有效的协作机制，同时考虑各种并行化框架对计算、通信和存储资源的不同需求。

推理：边缘 LAM 推理指的是为用户请求的不同模态的各种应用程序提供实时智能服务，从而提高整体推

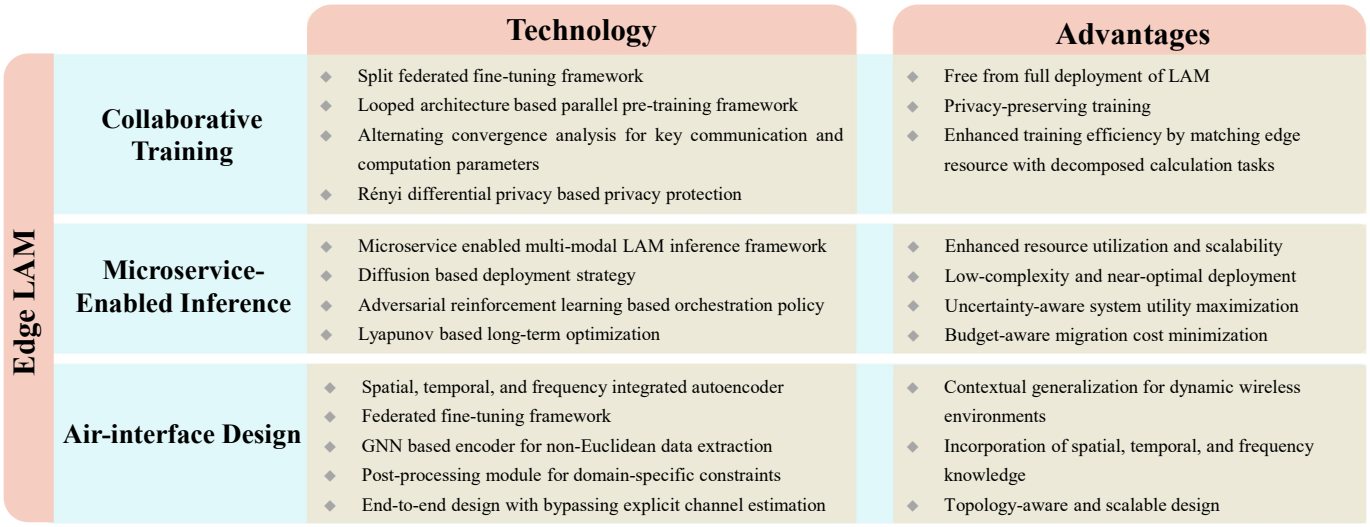

图 1. 技术和优势的概述，在提出的边缘 LAM 框架中。

理的及时性和性能。这些多模态应用，如自动驾驶和具身 AI，涉及通过一系列训练良好的神经网络顺序处理用户请求。然而，与传统边缘 AI 中的单一模态推理任务不同，这些过程面临着由于在网络边缘重复部署模块而导致的冗余计算和资源利用率低等重大挑战。这种冗余在调用期间占据了有限的边缘资源，并增加了整体推理延迟。

应用：人工智能是设计和优化空中接口技术的关键工具，促进了智能和高效的无线系统的发展 [5]。尽管近期取得了进展，现有的基于 AI 的方法通常被设计为解决特定场景中的具体问题。特别是在模型尺寸受限且训练数据有限的情况下，现有的基于 AI 的方法往往难以实现良好的泛化能力，这限制了它们在各种场景下的多任务处理能力。这一问题在高度动态的无线网络中进一步加剧，在这种网络中需要频繁重新训练，从而在网络通信开销、计算成本和可实现性能方面带来了显著挑战。参数达数十亿的 LAMs 展现了前所未有的通用智能水平，并且可以被微调以同时支持多样化的网络管理和优化任务 [6], [7]。尽管取得了显著进展，将 LAM 应用于空中接口技术的设计与优化仍然具有挑战性，因为需要考虑无线信道的复杂性、复杂的网络架构以及低延迟决策的需求。

考虑到边缘 LAM 在未来无线通信系统中的潜在应用，本文旨在从模型分解和资源分配的角度提供一个全面的框架，用于部署边缘 LAM 及其在空中接口设计中的应用。

- 在第 II 节中，我们介绍了一种专为边缘网络设计的协作训练框架，该框架涵盖了联邦微调（FedFT）和全参数并行训练，并结合了新颖的模型分解、资源分配和同步方案，以支持在网络边缘高效地进行设备上的训练。[8]
- 在第 III 节中，我们提出了一种基于微服务的边缘 LAM 推理框架，该框架将 LAM 分解为部署在边缘设备上的功能模块，并重新组织多模态下游推理服务作为动态微服务流。我们进一步研究了边缘 LAM 推理中的微服务部署、健壮编排和在线迁移，利用实时学习和优化技术来提升性能。
- 在第 IV 节中，我们介绍了用于空口设计的边缘 LAM，特别关注信道预测和波束成形设计。详细阐述了使边缘 LAM 能够有效捕捉无线数据的基本特征、适应动态信道条件和网络拓扑变化以及在各种任务中做出低延迟决策的设计原则。此外，在第 V 节中提供了一个简要案例研究，以展示所提出的框架将联邦 LAM 应用于信道预测的有效性。

II. 边缘 LAMs 的协作训练

训练大型模型以发现潜在模式对于实现各种智能应用至关重要，然而这需要大量的资源和数据消耗。为了解决这个问题，可以利用分布式资源和数据进行并行训练以加速训练过程。不过，现有的分布式训练框架严重依赖云计算，并且需要收集大规模的数据来进行训练。这导致了巨大的通信开销并引发了严重的隐私问题。这些问题可以通过设备上的训练来缓解，该方法利用本地计算资源进行训练，同时确保原始数据保持原地 [9]。这种方法最小化了向云端服务器传输数据的量，并

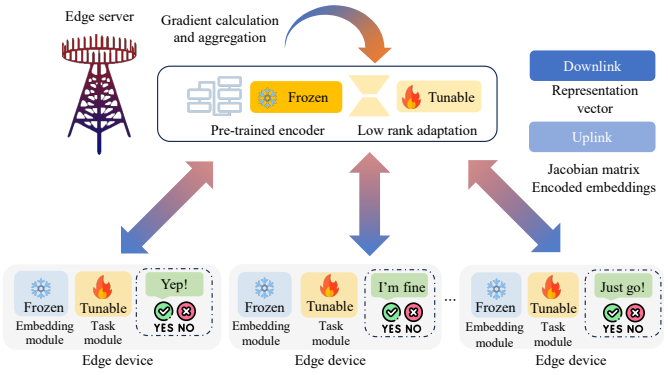

图 2. 无线网络上的联邦微调。

减少了通信开销。然而，边缘设备有限的数据、计算和存储资源限制了直接应用设备上训练到大型模型的训练中。为了充分利用大型模型的表现力并最大限度地减少通信成本同时确保数据隐私，我们提出了一种协作训练框架，该框架有助于预训练大型模型的微调以及轻量级大型模型的全参数训练。

A. 边缘 LAMs 的联合微调

支持个性化 LAM 服务需要对预训练的 LAM 进行特定任务数据集的微调，然而这些数据集通常分布在不同的边缘设备上并且高度敏感。为了解决这些问题，我们提出了一种 FedFT 框架，如图 2 所示，该框架通过协调分布式的边缘设备 [8] 以隐私保护的方式协作调整全局基础模型。然而，在每个存储和计算能力有限的边缘设备上部署完整的预训练 LAM 是不切实际的。这个问题可以通过构建分割 FedFT 框架来解决，该框架将 LAM 分解为嵌入模块、解码器模块和任务模块，其中轻量级的嵌入和任务模块被部署在边缘设备上，而计算密集型的解码器模块则保留在边缘服务器上。为了减少由于顺序通信造成的训练延迟，我们提出了一种基于低秩适应 (LoRA) 的 FedFT 框架，该框架将解码器的可训练参数分解为一系列低秩矩阵，并冻结原始参数。因此，边缘服务器并行处理接收到的嵌入向量、添加的低秩矩阵和原始编码器以形成表示向量。同时，在反向传播过程中，使用从边缘设备聚合的雅可比向量来更新这些低秩矩阵。与传统的联邦学习框架相比，这些雅可比向量导致了显著更低的传输开销。所提出的 FedFL 框架在两个关键方面与传统的 FL 框架不同。首先，在所提框架中边缘服务器使用单播传输向边缘设备发送表示向量，而不是像传统 FL 框架中的广播传输。其次，在所提框架中边缘服务器对低秩矩阵的梯度进行联邦平均，而非像传统 FL 框架中那样对完整模型进行联邦平均。通过将从理论分析中获得的见解融入到传输方案设计中，我们开发了一种在线资源分配算法以增强学习性能 [8]。

尽管 FedFT 通过在边缘服务器上聚合损失相对于表示向量的梯度来保护原始数据隐私，这些梯度仍然容易受到推理攻击，例如不受信任服务器的梯度逆向。为了解决这一问题，我们通过集成差分隐私（DP）增强 FedFT 系统，在该系统中边缘服务器聚合带噪梯度以确保样本级别的隐私保护。由于在 FedFT 上下文中对 DP 的分析特性是隐含的，我们引入了一个基于 Rényi DP (RDP) 的 FedFT 框架，该框架利用传输梯度的概率密度函数之间的分歧（在当地边缘设备上计算）来确定最佳噪声水平。与采用差分隐私的常规联邦学习系统不同，后者是服务器聚合整个模型的噪声梯度，FedFT 系统通过链式法则引入级联噪声，因为损失函数的梯度会乘以低秩矩阵的梯度。这种差异可能会阻碍所考虑的 FedFT 系统的收敛。为了解决这个问题，需要进行理论分析来评估添加的噪声对收敛行为和隐私保证的影响。然后，可以通过根据信道条件自适应地调整噪声注入来平衡隐私与学习性能之间的权衡。

B. 循环张量并行用于全参数边缘 LAM 训练

主流的 LAM 基本上是基于变压器的深度神经网络 (DNN)，拥有大量的神经元，并通过大量数据进行训练以实现上下文建模。与传统的学习任务（如 ResNet18）不同，后者可以在资源有限的边缘设备上执行（例如配备 4GB RAM 的 NVIDIA Nano），而像 BERT 这样的轻量级 LAM 的全参数训练通常每批至少需要 10GB 的 RAM，远远超过了单个设备的计算和存储能力。这要求使用并行训练机制（如流水线和平行数据处理），然而这些机制通常是为集中式云训练设计的 [4]。为了利用分布式资源进行轻量级 LAM 的训练，我们提出了一种循环张量并行框架，该框架采用一组一致的网络参数来表示计算密集型变压器块，并将网络参数分解成在边缘设备上执行的分布式矩阵计算。由于基于张量的通用矩阵乘法在轻量级 LAMs 的全参数训练中占据了计算负担的主要部分，张量并行性可以将注意力和前馈网络中的大规模矩阵计算分解为多个小规模矩阵计算。这些较小的任务随后由不同的边缘设备执行，从而减轻了计算开销。

为了规避原始数据共享，可以在边缘设备上通过分割权重矩阵对原始数据进行编码，然后由边缘服务器聚

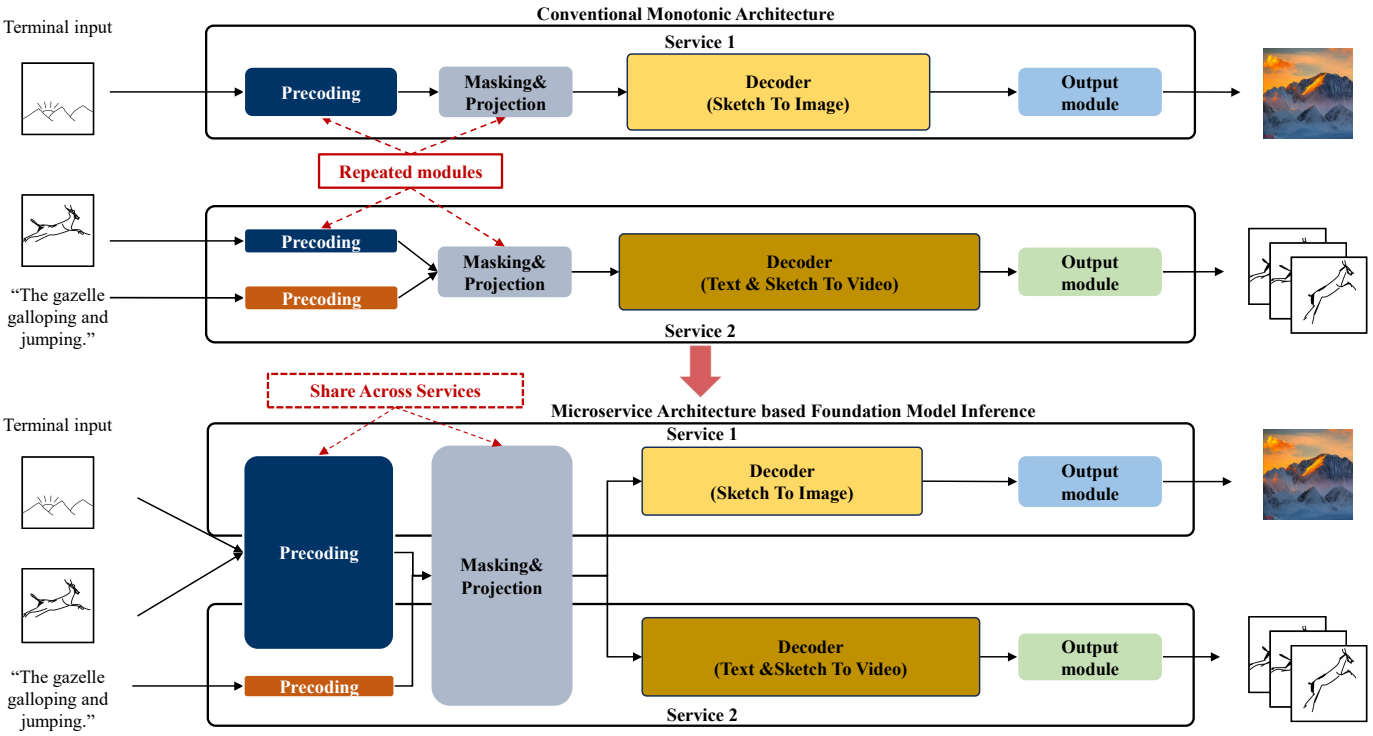

图 3. 所提议的边缘 LAM 微服务架构的示例。

合这些非线性激活。这是通过在计算中采用转置矩阵并在边缘服务器合并中间激活来实现的。为了缓解边缘设备的存储限制，可以应用跨层参数共享技术来分割注意力网络和前馈网络，其中不同深度级联的变压器块以循环方式重用相同的参数。这种创新的工作流程给匹配边缘资源和训练延迟带来了新的挑战。具体而言，训练延迟受到边缘设备异构计算能力、时变信道条件以及有限无线资源的影响。此外，并行张量合并的要求需要在设备之间保持矩阵维度的一致性，从而将任务大小限制为参与边缘设备的最小可用存储空间。因此，要最小化训练延迟，需要联合优化通信、计算和存储资源。

III. 微服务启用的多模态边缘 LAM 推理

与传统的独立神经网络边缘 AI 推理不同，多模态边缘 LAM 推理作为一个涉及多个专业模块的顺序过程运行，每个模块执行一个不同的任务，例如预编码或去噪。然而，将它们部署的传统单体方法导致冗余部署和计算，降低了资源利用率 [10]。为了解决这些问题，我们提出了一种基于微服务的方法，该方法将 LAM 服务分解为功能组件，通过将大型模型拆分为轻量级的微服务并重用共享组件来支持多样化的多模态下游推理任务，如图 3 所示。鉴于资源可用性和用户请求均为随机的，我们进一步研究了通过实时状态学习和决策制定来部署策略、鲁棒编排和在线迁移。

A. 边缘 LAM 推理的微服务部署

主流 LAM 包含模态编码器、输入投影器、骨干计算器、输出投影器和模态解码器，通过这些模块对请求的数据进行顺序处理，从而实现各种推理任务。

然而，现有单个 LAM 支持的模态有限。

例如，DALL-E 仅支持文本和图像模态，而 VideoGPT 支持视频和文本模态。

为了满足复杂的模态需求，需要在边缘设备上部署不同的 LAM，导致投影器和模态编码器的重复部署。

例如，语音和视觉转换器使用与 VisualBERT 相同的视觉编码器，同时集成了独特的语音编码器。

这种部署导致视觉编码器在边缘设备上重复出现，引入冗余，导致资源利用率低下。为了提高资源利用率，我们提出了一种基于微服务的推理框架，将多模态 LAM 的功能模块（例如，模式编码器、输入投影器）虚拟化为一系列独立的微服务，每个微服务专门执行特定功能。这种架构将过程转换为由不同微服务组成的单向无环图，其中节点表示不同的微服务，边表示它们之间的数据依赖关系。该设计将基于不同多模态 LAM 的推理任务转化为在边缘设备之间共享的独立微服务流，可以进一步优化以提高整体效率、可扩展性和对用户请求动态变化的响应能力。

鉴于推理任务的统计需求，微服务部署成为一个关

键问题 [11]，这指的是将计算和通信任务战略性地分配以解决严格的延迟要求与推理任务之间的不匹配。其目标是在异构边缘设备之间有效利用有限资源的同时最小化总延迟。因此，微服务部署问题可以被表述为一个涉及离散值部署指示符的 NP 难组合优化问题。为了高效解决这个组合优化问题，训练一个近似最优部署策略所需分布的似然模型是一种有前景的方法。该模型允许根据边缘设备的状态以端到端的方式实时生成部署指标。然而，使用潜在变量模型的边际样本概率构建似然模型会由于潜在变量引入不可计算的上界。为了解决这个问题，可以将关于潜在变量的后验概率映射到一个预定义的平稳分布上，该分布可以通过数据驱动的方式利用扩散模型进行学习。

B. 微服务编排与迁移

用户移动性决定了已部署微服务系统的效用。用户请求的不确定性也会影响已部署微服务的系统延迟和运营成本，从而降低边缘 LAM 推理的可靠性，并强调了可靠微服务编排的迫切需求。在边缘 LAM 推理中的可靠微服务编排被定义为与设计稳健的服务路由方案以响应用户请求相关的计算和通信任务协调。不同于微服务部署，微服务编排假设相同的微服务至少部署在两个或更多不同的边缘设备上以响应用户请求，在这里的目标是在动态用户请求下找到最大化长期系统效用的鲁棒路由策略。这可以被视为一个马尔可夫决策过程，其中由于用户移动性导致奖励函数不确定，并由之前的决策轮次中的网络状态和路由设计确定。因此，可以采用鲁棒对抗强化学习来建立所需的编排政策。这是通过利用主代理在动态请求和移动性下学习最大化长期系统效用的编排策略，而对抗代理生成旨在挑战主代理以最小化效用请求策略来实现的。可以通过参数化的 DNN 以数据驱动的方式学习所需的编排政策。

微服务迁移是解决用户移动性问题的另一种方法。具体来说，边缘服务器覆盖范围有限，加上高移动性的场景（例如，在高速公路上或高速列车上使用边缘 LAM 推理服务），可能导致显著较高的推理延迟。资源可用性和服务请求的这种动态波动要求将微服务和用户数据从之前的服务边缘服务器迁移到更合适的服务器上。由于用户的地理位置存在空间和时间上的相关性，因此推理延迟直接由微服务迁移决策决定。为了最大化系统效用的时间平均期望值，微服务迁移问题是一个长期优化问题，需要考虑迁移成本（例如带宽、存储和功率），同时考虑到延迟、抖动和数据包丢失的要求。通过采用 Lyapunov 优化方法，可以将满足长期约束的目标转换为一个虚拟队列，以表示资源效用与预算之间的差距。受 Lyapunov 优化的漂移加惩罚算法启发，这个长期优化问题可以转化为多个一次性迁移决策问题。可以通过开发交替优化算法或在线学习算法来动态分配计算和通信资源，从而解决微服务迁移成本最小化的问题。

IV. 空口设计的边缘 LAMs

无线通信系统的基础长期以来依赖于空中接口技术。虽然传统的基于优化的方法利用领域专家知识来实现理想的网络性能，但在大规模网络中它们难以扩展。通过利用 DNN 直接建立系统状态和优化参数之间的非线性映射，基于 AI 的方法可以解决各种网络优化和管理问题。然而，它们通常是为特定场景和特定问题设计的，因此泛化能力较差。特别是，频繁的重新训练是为了适应快速变化的网络条件，这对效率和可扩展性提出了重大挑战。相比之下，边缘 LAM 在提取训练数据的复杂时空特征以及学习上下文嵌入方面表现出色，展示了增强的泛化能力。使用大规模无线数据训练边缘 LAM 使其能够作为通用特征提取器，有效捕捉可用于信道预测和波束成形 [12], [13] 的复杂模式和表示。要充分发挥边缘 LAM 在空中接口的全部潜力，必须研究预编码器设计以有效处理无线数据，这是一种与文本和图像不同的独特模式，专门针对特定任务的设计生成器以解决领域特定的约束条件和不断变化的网络维度，并利用来自无线环境的专业知识进行微调。

A. 联邦 LAM 用于信道预测

基于历史信道状态信息（CSI）数据的精确信道预测对于实现高度动态的 6G 网络中的高性能通信至关重要，由于短的信道相干时间，信道估计开销可能非常大。面对准确描述实际无线信道的困难，传统的稀疏性、Prony 和外推方法受到有限预测性能的影响。通过以数据驱动的方式学习时域和频域信道变化，基于 AI 的方法可以支持单天线和多天线系统的信道预测。然而，现有的基于 AI 的方法由于小尺寸神经网络导致的较差信道特征提取能力和特定场景设计导致的弱泛化能力而受到限制。借助庞大的参数数量和变压器架构，边缘 LAM 能够通过上下文泛化来建模复杂的时变无线网络，而不是构建特定场景的特征预测映射。因此，使用大量的信道状态信息数据集进行预训练的 LAM 可以

支持多种信道预测任务，而无需部署多个特定配置的模型 [14]。然而，应收集包含大量训练样本的各种三维 CSI 数据集并将其传输到中央服务器进行预训练，这会导致过高的通信开销和严重的隐私问题。

联合 LAM 有望在仅聚合可训练参数的梯度时实现有效的信道预测，同时促进通信高效且保护隐私的模型训练。具体而言，位于不同位置的边缘设备会感知到不同信道分布和网络配置下的不同信道样本，丰富数据多样性并提升联合 LAM 模型的一般化能力和适应性。为了捕捉连续信道状态之间的空间、时间和频率相关性，可以采用基于自回归神经网络的编码器来建模信道依赖性，提取嵌入特征并将它们映射为联邦 LAM 的令牌表示。此外，在联合 LAM 适配中可以采用低秩矩阵分解方法（如 LoRA），因为它们能够仅用 5%的参数实现与全参数调整相当的学习性能，显著降低计算成本和通信开销。通过进一步专门设计 CSI 数据预处理模块，经过联邦 LAM 良好训练的模型可以有效地支持跨多种场景和配置下的信道预测。

B. 用于波束成形的图 LAMs

波束成形设计对于提高多天线系统的频谱和能量效率至关重要，有助于实现定向通信、感知和定位。为了满足日益增长的高数据速率和低延迟通信需求，许多先进的多天线技术（例如，超大规模 MIMO、可重构智能表面和全息 MIMO）应运而生，提供了前所未有的容量，但同时也给波束成形设计带来了重大挑战。由于其高维特性和优化变量经常耦合的特点，波束成形设计问题通常是非凸的，无法通过基于优化的方法高效解决。将 AI 集成到波束成形设计中显示出巨大的潜力，但也面临几个关键挑战。首先，受限于模型大小，现有的基于 AI 的方法不能充分描述信道模型，限制了它们有效映射无线参数至最优波束成形设计的能力 [15]。其次，由于问题特定的模型训练，现有基于 AI 的方法表现出弱泛化能力，无法同时处理多个任务。第三，在没有充分利用底层网络拓扑和波束成形问题的置换性质的情况下，现有的基于 AI 的方法在可扩展性方面表现不佳，并且当网络规模发生变化时性能会显著下降。

因此，我们提出了一种基于图 LAM 的波束成形框架，该框架可以利用大量的神经元来实现丰富且语境化的信道特征提取，使用跨越不同信道环境的高质量数据集以确保强大的泛化能力，并利用网络拓扑和置换属性实现卓越的可扩展性。具体而言，针对各种场景（例如

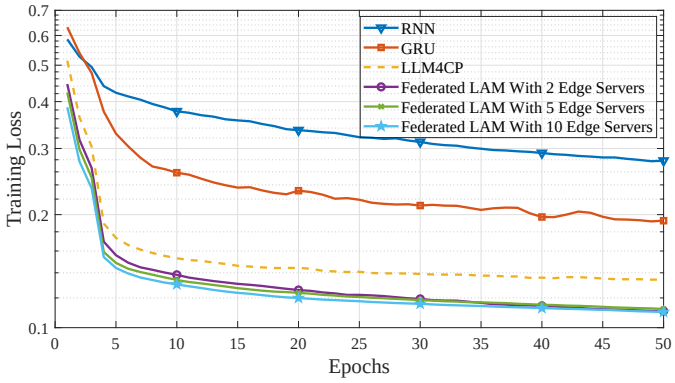

图 4. 联邦 LAM 用于信道预测。

混合波束成形、联合主动和被动波束成形）的波束设计可以被表述为图优化问题，在这种情况下，分布式通信设备（例如边缘设备、基站和可重构智能表面）表示为节点，并且相互之间的通信和/或感知依赖关系表示为边。为了弥合非欧几里得图形数据与具有拓扑信息以及适合预训练 LAM 的欧几里得令牌之间的差距，可以采用图神经网络预编码器来提取隐藏特征，随后使用轻量级神经网络将这些特征投影到令牌空间中。除了采用低秩矩阵分解进行特定于无线通信的适应外，投影和归一化模块还可以确保生成的波束成形设计满足领域特定的约束条件，例如传输速率、感知分辨率和学习性能。此外，绕过显式信道估计并直接利用噪声导频信号以实现端到端的设计可以进一步集成到基于图 LAM 的波束成形设计框架中，从而减轻由于波束形成设计与信道估计算法之间的目标不匹配而导致的性能下降。这可以通过开发复数值神经网络来编码导频序列成为令牌来实现。为了进一步减轻收集本地 CSI 和传播波束成形向量以进行集中学习所造成的通信开销，可以使用 FedFT 来实现不同网络架构和配置下的分布式和协作学习。

V. 案例研究

本小节评估所提出的联邦 LAM 的信道预测学习性能。

重要设置：我们采用 LLM 赋能的信道预测方法（LLM4CP）作为基础模型 [14]，并使用 QuaDRiGa 生成符合 3GPP 标准的信道实现作为训练数据集。我们考虑最多 10 个边缘服务器，每个服务器配备 5%独立同分布的信道样本。

评估结果：所提出方法的性能如图 4 所示。为了进行比较，我们考虑了三个基线：对 LLM4CP 进行微调、门控循环单元（GRU）以及仅使用局部数据并采用随机梯度下降（SGD）优化器的循环神经网络（RNN）。联

邦 LAM 的训练损失低于这些基准。其收敛性能最初随着设备数量的增加而提升，但随着时间推移趋于稳定。通过超越本地训练，联邦 LAM 实现了高效、保护隐私的信道预测，从而支持空中接口设计。

VI. 结论

在这篇文章中，我们探讨了边缘 LAM 作为实现原生 AI 6G 网络变革性技术的潜力。我们的研究集中在两个关键视角上："模型分解" 和 "资源管理"，这两个视角对于解决在网络边缘部署和利用 LAM 面临的挑战至关重要。我们提出了一种协作训练框架，该框架利用拆分 FL（联邦学习）和循环张量并行计算来支持通信高效且值得信赖的微调以及完整的参数训练。为了提供低延迟推理服务，我们开发了一个由微服务支持的架构，能够动态匹配微服务对计算需求与边缘设备能力之间的关系。此外，我们还研究了边缘 LAM 在信道预测和波束成形方面的应用。尽管有了这些进展，但边缘 LAM 仍面临重大挑战，包括由于边缘设备存储和计算资源有限而导致的模型大小限制、高能耗可能迅速耗尽电池寿命以及训练和推理过程中潜在的数据泄漏带来的隐私问题。解决这些问题将需要在模型压缩、节能计算和强大的隐私保护技术方面进行创新方法。

参考文献

[1] M. Xu *et al.*, "Unleashing the power of edge-cloud generative AI in mobile networks: A survey of AIGC services," *IEEE Commun. Surv. Tut.*, vol. 26, no. 2, pp. 1127–1170, 2024.

[2] L. Bariah *et al.*, "Large generative AI models for telecom: The next big thing?" *IEEE Commun. Mag.*, vol. 62, no. 11, pp. 84–90, Nov. 2024.

[3] M. Tao *et al.*, "Federated edge learning for 6G: Foundations, methodologies, and applications," *Proc. IEEE*, pp. 1–39, 2024.

[4] S. Deng *et al.*, "Cloud-Native computing: A survey from the perspective of services," *Proc. IEEE*, vol. 112, no. 1, pp. 12–46, 2024.

[5] K. B. Letaief *et al.*, "The roadmap to 6G: AI empowered wireless networks," *IEEE Commun. Mag.*, vol. 57, no. 8, pp. 84–90, 2019.

[6] G. Sun *et al.*, "Large language model (LLM)-enabled graphs in dynamic networking," *IEEE Netw.*, pp. 1–1, 2024.

[7] D. Wu *et al.*, "NetLLM: Adapting large language models for networking," in *ACM SIGCOMM*, Washington DC, USA, Sept. 2024.

[8] Z. Wang *et al.*, "Federated fine-tuning for pre-trained foundation models over wireless networks," *IEEE Trans. Wireless Commun.*, Jan. 2025.

[9] K. B. Letaief *et al.*, "Edge artificial intelligence for 6G: Vision, enabling technologies and applications," *IEEE J. Sel. Area. Commun.*, vol. 40, no. 1, pp. 5–36, Nov. 2022.

[10] Y. He *et al.*, "Large language models (LLMs) inference offloading and resource allocation in cloud-edge computing: An active inference approach," *IEEE Trans. Mobile Comput.*, vol. 23, no. 12, pp. 11 253–11 264, 2024.

[11] Z. Yang *et al.*, "Latency-Aware microservice deployment for edge AI enabled video analytics," in *IEEE WCNC*, Apri. 2024.

[12] Z. Chen *et al.*, "Big AI models for 6G wireless networks: Opportunities, challenges, and research directions," *IEEE Wireless Communications*, vol. 31, no. 5, pp. 164–172, 2024.

[13] J. Guo and C. Yang, "Recursive GNNs for learning precoding policies with size-generalizability," *IEEE Trans. Machine Learn. Commun. Netw.*, vol. 2, 2024.

[14] B. Liu *et al.*, "WiFo: Wireless foundation model for channel prediction," *Sci. China Inf. Sci.*, 2024.

[15] Y. Sheng *et al.*, "Beam prediction based on large language models," 2024.

Zixin Wang (eewangzx@ust.hk) received his Ph.D. degree from the University of Chinese Academy of Sciences. He is currently a Postdoctoral Fellow with The Hong Kong University of Science and Technology.

Yuanming Shi (shiym@shanghaitech.edu.cn) received his Ph.D. degree from The Hong Kong University of Science and Technology. He is currently a Full Professor with ShanghaiTech University and an IET fellow.

Yong Zhou (zhouyong@shanghaitech.edu.cn) received his Ph.D. degree from University of Waterloo. He is currently an Associate Professor with ShanghaiTech University.

Jingyang Zhu (zhujy2@shanghaitech.edu.cn) is pursuing his Ph.D. degree with ShanghaiTech University.

Khaled B. Letaief (eekhaled@ust.hk) received his Ph.D. from Purdue University. He has been with HKUST since 1993, where he was an acting provost and dean of engineering. He is now a Senior Advisor to the President and the New Bright Professor of Engineering. From 2015 to 2018, he joined HBKU in Qatar as Provost. He is an ISI Highly Cited Researcher and a recipient of many distinguished awards. He has served in many IEEE leadership positions, including ComSoc president, vice-president for technical activities, and vice-president for conferences. He is a member of the US National Academy of Engineering.