
# 无线网络中的分布式边缘智能

## —从联邦学习到拆分联邦学习

常征 教授

电子科技大学 计算机（网安）学院

# 目录

CONTENTS

一、研究背景与需求

二、分布式边缘智能

三、优化方案与技术

四、总结和未来研究

# 目 录

CONTENTS

一、研究背景与需求

二、分布式边缘智能

三、优化方案与技术

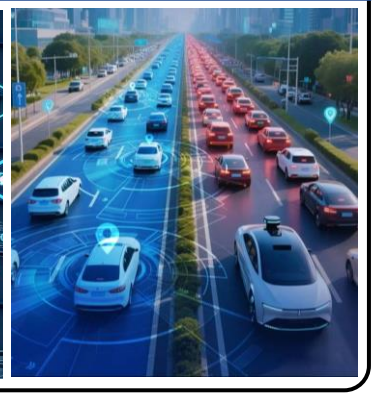
四、总结和未来研究

# 研究背景——分布式边缘智能

**重大需求** **分布式边缘智能是国家算力网络与数据安全战略的核心支撑**


在 “人工智能 +” 行动、“东数西算” 工程及《新一代人工智能发展规划》等国家战略牵引下，由相关部委协同推动的边缘智能发展政策，旨在通过构建安全、协同、开放的边缘智能基础设施与服务生态，强化自主创新、深化边云协同、完善标准体系，为制造业数字化转型与经济社会智能化发展提供关键支撑，进而保障全国一体化算力网络与数字中国建设的核心需求

**智慧交通** 车路云协同与自动驾驶



**工业智造** 大规模实时感知与决策

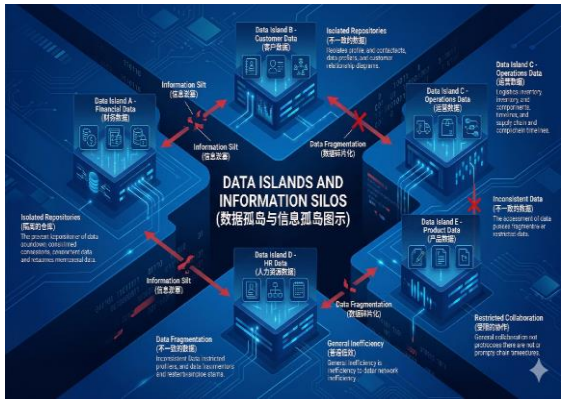
**具身智能** 多机交互跨域协同

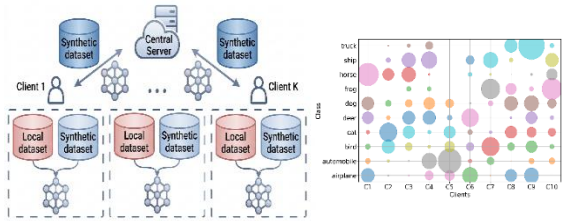
**分布式边缘智能深化赋能数智化转型突破，是实现数据安全流转、算力高效协同的核心与基础 。**

# 研究背景与挑战

**挑战** 分布式边缘智能面临“分不均”、“算不动”、“传不快”、“学不精”的挑战。

## 分不均



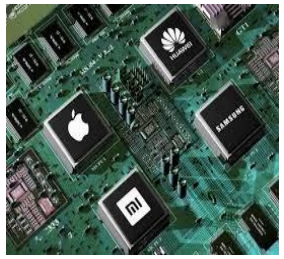






边缘设备之间**数据异构**分布与全局模型收敛的冲突

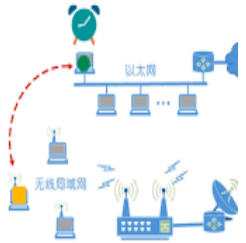
## 算不动





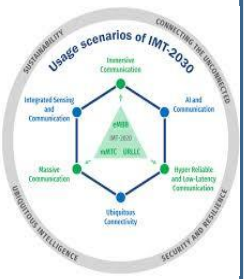
**资源受限**终端的硬件承载能力与**模型训练开销**的冲突

## 传不快




**有限带宽**环境下高维特征频繁**交互**与**时效性**的冲突

## 学不精


**有限带宽**环境下**Token选择传输**与**任务性能**的冲突

**以数据为基、计算为核、通信为桥，是实现分布式系统效能跃迁的关键范式。**

# 研究思路

| | | | | |
|---|---|---|---|---|
| 科学问题 | 边缘智能的通信与计算瓶颈 | | | |
| 研究挑战 | 分不均 | 算不动 | 传不快 | 学不精 |
| 关键难题 | 边缘端数据分布差异大 | 终端算力无法支撑模型 | 高维特征传输造成拥塞 | 任务语义保留与传输开销难均衡 |
| 技术创新 | AIGC辅助的数据增强 | 自适应模型拆分 | 高效传输激活量化 | 基于语义的Token选择传输 |
| 核心贡献 | 构建了一套数据-计算-通信-任务多维协同的分布式边缘智能基座 | | | |
| 核心技术 | 数据层 (Data)<br>• AIGC生成<br>• 异构度量<br>• 增强聚合 | 计算层 (Computation)<br>• 模型拆分<br>• 计算卸载<br>• 拆分选择 | 通信层 (Communication)<br>• 激活量化<br>• 特征压缩<br>• 带宽优化 | 任务层 (Task)<br>• 特征提取<br>• 语义度量<br>• Token选择 |

# 目录

CONTENTS

一、研究背景与需求

二、分布式边缘智能

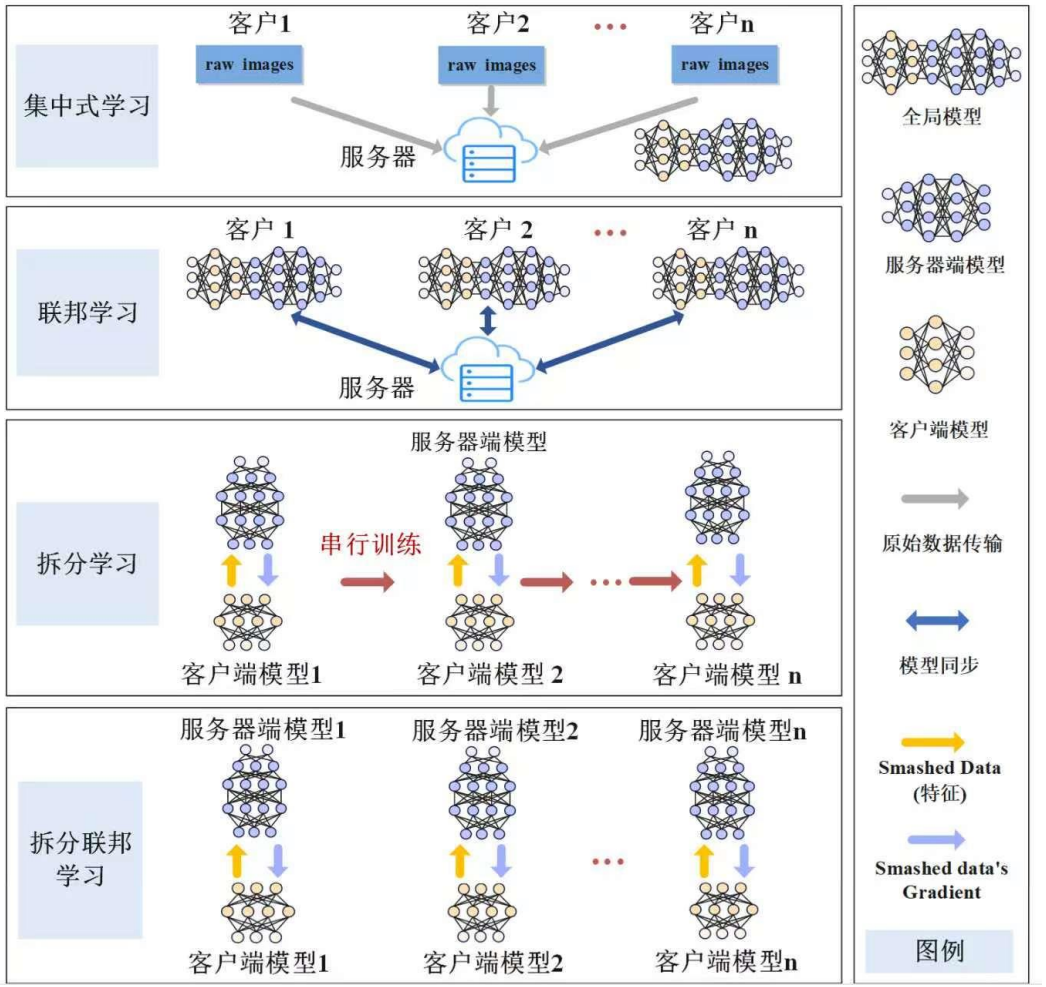
三、优化方案与技术

四、总结和未来研究

# 分布式架构——从联邦学习到拆分联邦学习

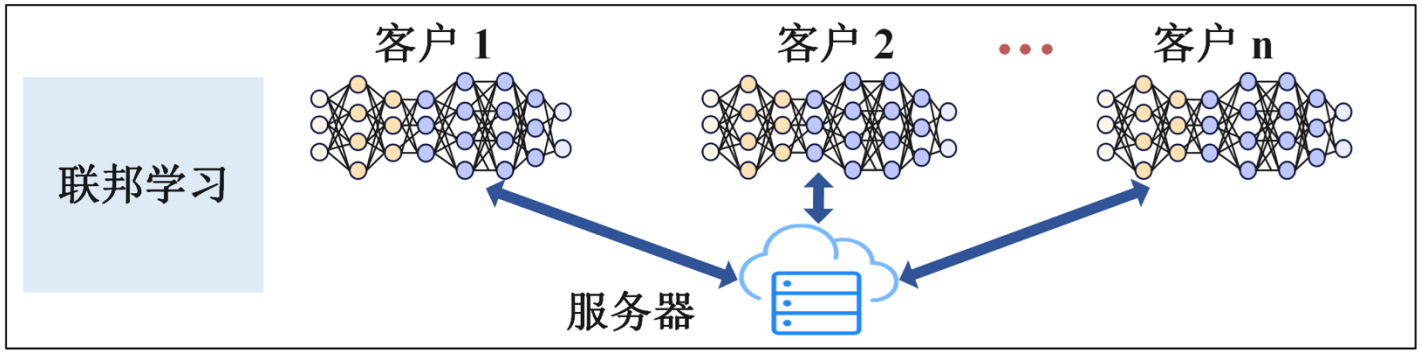
**分布式架构** 联邦学习（FL）、拆分学习（SL）、拆分联邦学习（SFL）

- 联邦学习(并行)
  - 本地全模型训练
  - 服务器端聚合
- 拆分学习(串行)
  - 设备端模型训练
  - 服务器端模型训练
- 拆分联邦学习(并行)
  - 设备端模型训练
  - 服务器端模型训练
  - 服务器端模型聚合




FL, SL, SFL分布式框架

# 分布式架构——联邦学习




| | |
|---|---|
| 服务器 | 聚合 |
| 无线网络 | 模型权重交互 |
| 客户端 | 模型计算 |

## 训练流程

1. 服务器初始化全局模型 $\omega_t$
2. 各客户端在本地数据集上训练本地模型
3. 客户端上传本地模型更新
4. 服务器聚合得到新一轮全局模型 $\omega_{t+1}$
5. 重复迭代直到收敛

## 本地损失函数

对第n个客户端，其本地数据集为 $D_n$ ,本地损失函数可写为

$$L_n(\omega)=\frac{1}{|D_n|}\sum_{i=1}^{|D_n|}\ell(\omega,x_n^i)$$

本地数据量　　损失函数　　输入

## 全局优化目标

$$\min_{\omega} \quad L(\omega)=\sum_{n=1}^{N}\rho_n L_n(\omega)$$

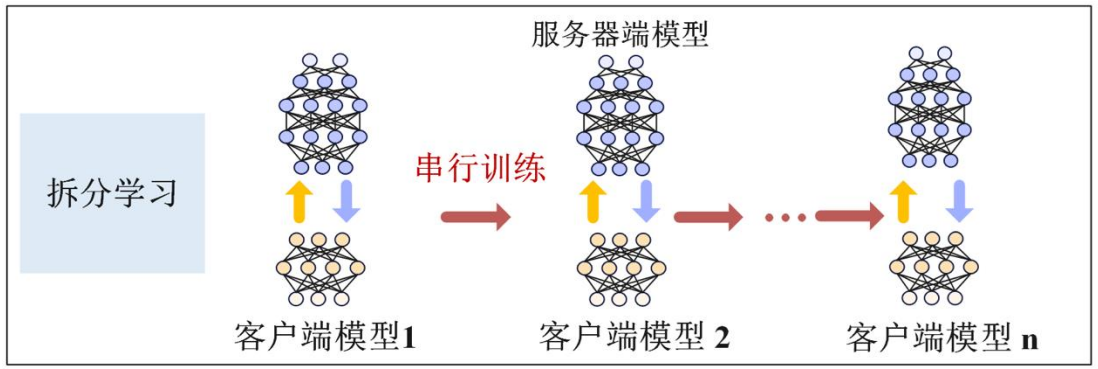
权重

其中 $\rho_n=|D_n|/\sum_{n=1}^{N}|D_n|$ ，表示客户端n的权重。聚合公式为：

$$\omega^{t+1}=\sum_{n=1}^{N}\rho_n\omega_n^{t+1}$$

**数据不出本地，训练并行进行，服务器只负责模型聚合。**

# 分布式架构——拆分学习




| | |
|---|---|
| 服务器 | 服务器端计算 |
| 无线网络 | 客户端模型权重交互 |
| | 激活及梯度交互 |
| 客户端 | 客户端模型计算 |

## 训练流程

1. 在拆分层处将模型切分
2. 客户端执行前向传播，得到激活
3. 激活上传到服务器
4. 服务器完成后续前向传播、计算损失，并进行反向传播
5. 服务器将拆分层对应梯度返回客户端
6. 客户端继续反向传播更新本地前部模型

## 模型拆分公式化

拆分后模型权重表示：客户端模型

$$\omega = \{\omega^C, \omega^S\}$$

服务器端模型

客户端输出的激活表示：

$$\mathbf{a}_n^i = f_C(\omega^C, x_n^i)$$

输入

服务器预测结果输出表示：

$$\hat{y}_n^i = f_S(\omega^S, \mathbf{a}_n^i)$$

激活

## 全局优化目标

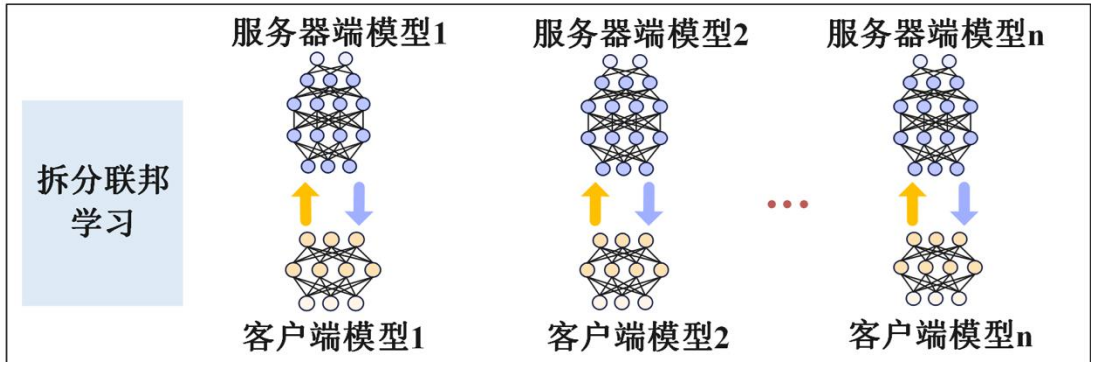
客户端本地损失为：服务器端输出

$$L_n(\omega^C, \omega^S) = \frac{1}{|D_n|}\sum_{i=1}^{|D_n|} \ell\left(f(\omega^S, \mathbf{a}_n^i), y_n^i\right)$$

在经典 SL 中，各客户端按顺序与服务器协同完成前向与反向传播，模型参数在串行过程中持续更新，而不是像 FL 那样在每轮结束后进行显式聚合。

**模型拆开训，计算卸载到服务器，但客户端与服务器之间需要串行交互。**

# 分布式架构——拆分联邦学习




| | |
|---|---|
| 服务器 | 服务器端计算 |
| | 客户端模型聚合 |
| 无线网络 | 客户端模型权重交互 |
| | 激活及梯度交互 |
| 客户端 | 客户端模型计算 |

## 训练流程

1. 在拆分层处将模型划分为客户端侧与服务器侧子模型
2. 服务器向参与客户端下发客户端侧子模型
3. 各客户端并行前向传播，得到中间激活
4. 客户端上传激活，服务器完成后续前向、损失计算与反向传播
5. 服务器将拆分层梯度返回客户端，客户端完成本地反向更新
6. 客户端上传更新后的客户端侧模型参数

## 模型拆分公式化

拆分后模型权重表示：

$$\omega = \{\omega^C, \omega^S\}$$

客户端输出的激活表示：

$$\mathbf{a}_n^i = f_C(\omega^C, x_n^i)$$

服务器预测结果输出表示：

$$\hat{y}_n^i = f_S(\omega^S, \mathbf{a}_n^i)$$

## 全局优化目标

$$\min_{\omega} \quad L(\omega) = \sum_{n=1}^{N} \rho_n L_n(\omega^C, \omega^S)$$

$\rho_n$ 权重

其中 $\rho_n = |D_n| / \sum_{n=1}^{N} |D_n|$，表示客户端n的模型权重。聚合公式为：

$$\omega^{C,t+1} = \sum_{n=1}^{N} \rho_n \omega_n^{C,t+1}$$

$\rho_n$ 模型权重　　$\omega_n^{C,t+1}$ 客户端模型

**模型拆分训练与联邦聚合相结合，在降低客户端计算负担的同时支持多客户端并行训练。**

# 目 录

CONTENTS

一、研究背景与需求

二、分布式边缘智能

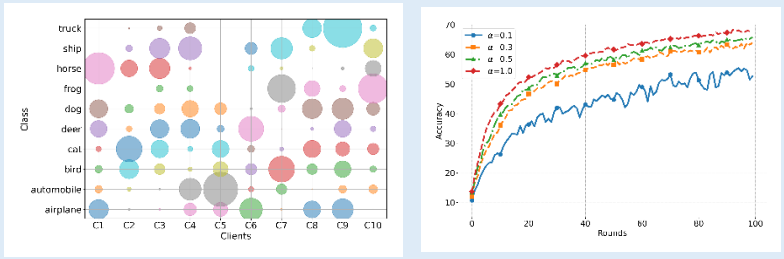
三、优化方案与技术

四、总结和未来研究

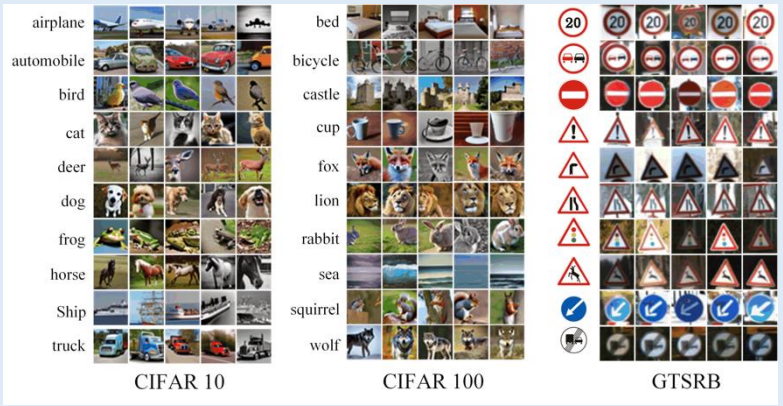
# 内容1： AIGC 辅助的联邦学习增强方案

**问题** 联邦学习中数据异构性极大制约了模型性能提升，“分不均”

## 挑战 数据越异构，性能越差



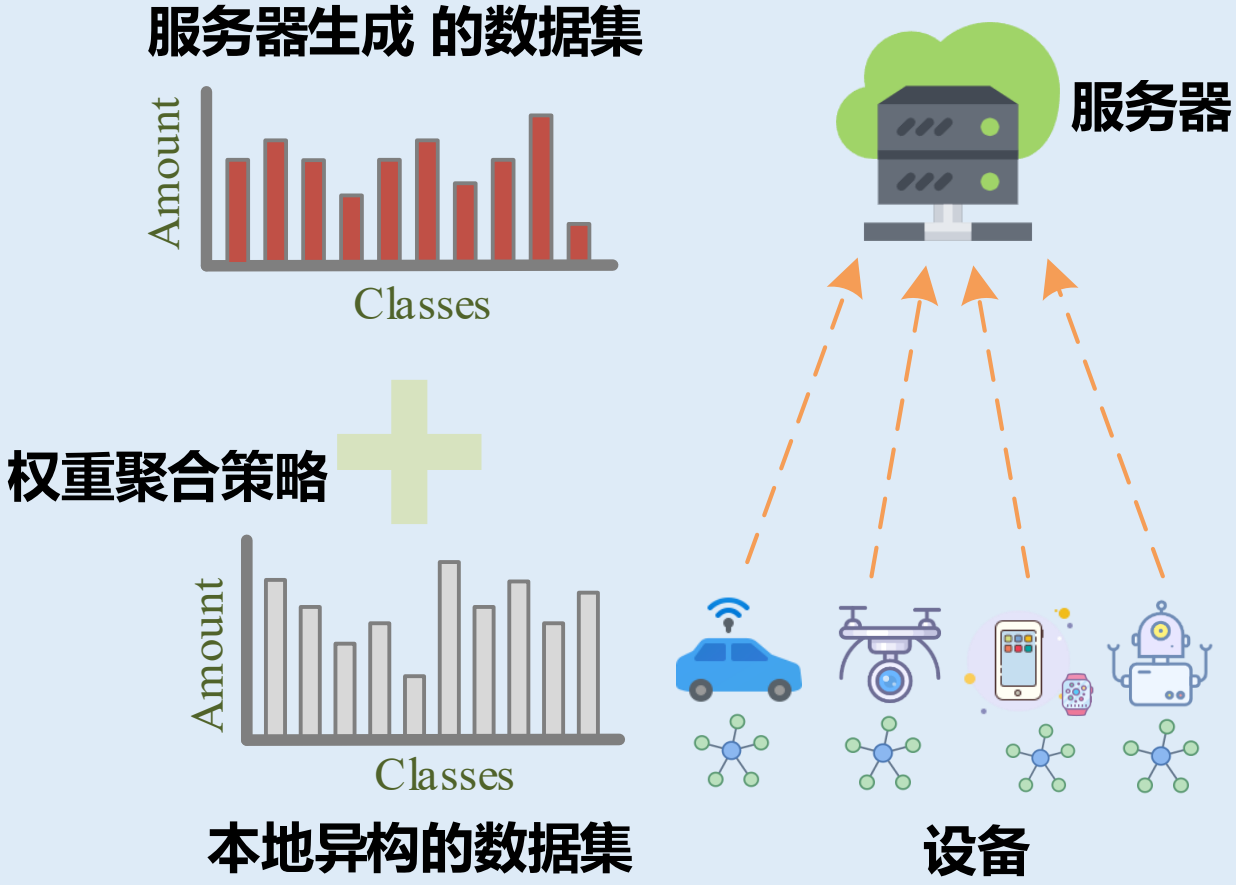

## 机遇 AIGC生成高质量图片




Diffusion model基于提示词生成图片

生成数据是否能弥补数据异构分布？

## 研究思路 AIGC辅助FL解决数据异构问题



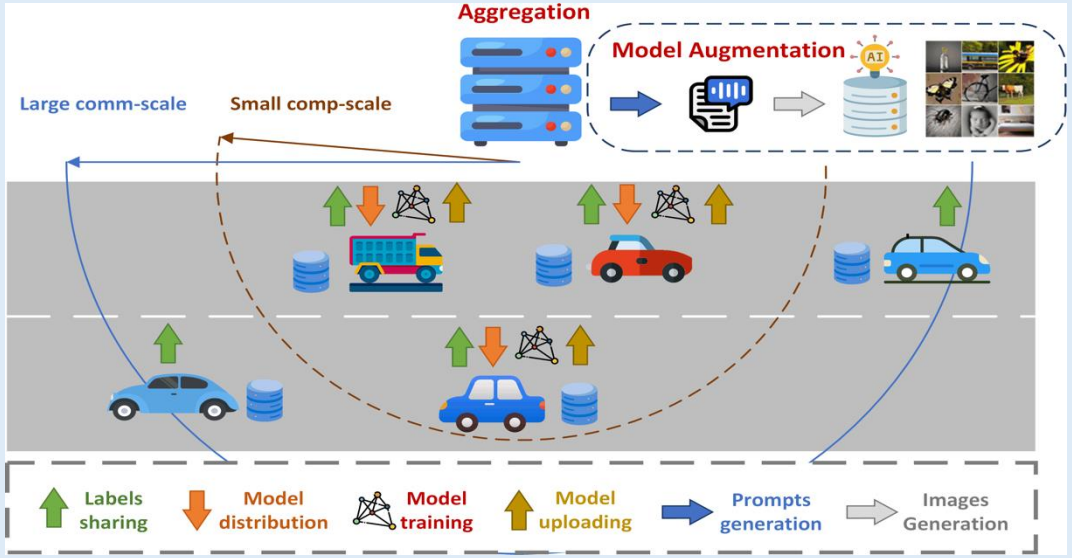


- X. Qiang, Z. Chang and G. Min, "AIGC-Assisted Federated Learning for Vehicular Edge Intelligence: Vehicle Selection, Resource Allocation and Model Augmentation," in *IEEE Transactions on Mobile Computing*, vol. 24, no. 11, pp. 11896-11909, Nov. 2025, doi: 10.1109/TMC.2025.3581983.
- X. Qiang, Z. Chang and Y.-C. Liang, "AIGC-assisted Federated Learning for Edge Intelligence: Architecture Design, Research Challenges and Future Directions," arXiv preprint arXiv:2503.20166, 2025.

# 内容1： AIGC 辅助的联邦学习增强方案

**核心思想** 服务器端使用AIGC生成数据集弥补数据异构带来的模型差异

## GenFV系统流程




1. 设备端向服务器发送低隐私标签元数据
2. 服务器选择稳定设备并分配资源、广播模型权重
3. 被选择的设备执行本地训练；服务器利用AIGC生成数据，并基于生成数据训练增强模型
4. 服务器将本地更新与其增强模型一并聚合

## 模型权重聚合策略

$$\boldsymbol{\omega}^t = \kappa_1 \sum_{\forall n \in \mathcal{N}^t} \rho_n \omega_n^t + \kappa_2 \omega_a^t$$

其中 $\kappa_1 = 1 - \left(\frac{\bar{EMD}}{2}\right)^2, \kappa_2 = \left(\frac{\bar{EMD}}{2}\right)^2$

- 数据异质性通过EMD度量：

$$EMD_n = \sum_{i=1}^{Y} |p_n(y=i) - p(y=i)|.$$

- EMD决定了聚合时联邦模型（$\kappa_1$）与AIGC增强模型（$\kappa_2$）的权重。
- EMD越大 $\Rightarrow$ 数据越异构 $\Rightarrow$ FL模型越差 $\Rightarrow$ 需更多AIGC增强模型支持。

# 内容1： AIGC 辅助的联邦学习增强方案

**问题建立** 构建关于用户选择，带宽分配，传输功率和数据生成量的联合优化问题

## 时延能耗分析

1. 本地模型计算，以GPU为例：
   - 时延：$$T_n^{cp} = t_n^0 + \frac{c_1 b_n \theta_n^{mem}}{f_n^{mem}} + \frac{c_2 b_n \theta_n^{core}}{f_n^{core}},$$
     开辟显存的时间　　计算任务的时间
   - 能耗：$$E_n^{cp} = p_n^{cp} \cdot T_n^{cp}$$
     计算功率　　计算功率
2. 模型上传：
   - 时延：$$T_n^{mu} = \frac{s(\omega)}{r_n^U},$$
   - 能耗：$$E_n^{mu} = \phi_n \frac{s(\omega)}{r_n^U},$$
3. 服务器端数据生成及增强模型训练时延：
   - 数据生成时延：$$T_s^{inf} = b \sum_{t=1}^{I} \frac{d_{m,t}}{f_{rsu}} = b^t t_0,$$
   - 累计生成数据训练时延：$$T_s^{cp} = t_s^0 + \frac{c_s^1 b_s^t \theta_s^{mem}}{f_s^{mem}} + \frac{c_s^2 b_s^t \theta_s^{core}}{f_s^{core}}$$

## 问题建立

$$\begin{aligned}
\mathcal{P}: \quad & \min_{\boldsymbol{\alpha},\mathbf{l},\boldsymbol{\phi},b} \max_{n \in \mathcal{N}^t} T_n \\
\text{s.t.} \quad & E_n^{cp} + E_n^{mu} \le \bar{E}, \quad \forall n \in \mathcal{N}^t, \\
& l_n^t \in \{0,1\}, \quad \forall n \in \mathcal{N}^t, \\
& \sum_{n=1}^{|\mathcal{N}^t|} l_n^t \le M, \\
& \phi_{\min} < \phi_n^t < \phi_{\max}, \quad \forall n \in \mathcal{N}^t, \\
& b^t \in \mathbb{N}, \\
& T_s^{inf} + T_s^{cp} \le \bar{T}.
\end{aligned}$$

**优化目标及约束**

最小化用户最大时延

- 能耗约束
- 用户选择约束1
- 用户选择约束2
- 发射功率约束
- 数据生成量约束
- **服务器生成及模型增强训练时间约束**

**优化变量**

- 用户选择指数 $\alpha$
- 带宽分配 $l$
- 发射功率 $\phi$
- 数据生成量 $b$

**问题分析**

混合整数非线性规划，且非凸，是一个NP-hard优化问题

# 内容1： AIGC 辅助的联邦学习增强方案

**问题求解** **两阶段联合优化求解策略**

## 大通信规模阶段

**大通信范围：** RSU覆盖的通信最大范围

**子问题1：移动用户选择**

$$\mathcal{SUBP}1: \min_{\boldsymbol{\alpha}} \max \quad T_n$$

约束条件1: 确保移动用户在通信范围驻留时间内完成训练任务

$$T_n^{cp} + T_n^{mu} \leq \bar{T}_n, \forall n \in \mathcal{N}.$$

本地模型计算时间　模型上传时间　移动用户最长驻留时间 $T_n = \min(t_n^{hold}, t^{max})$,

约束条件2: 选择本地数据不过于异构的用户

$$EMD_n^t \leq \hat{EMD}, \forall n \in \mathcal{N},$$

本地数据异构指标值　预设上限

## 小计算规模阶段

**小计算范围：** 符合要求的被选择用户范围

**解决带宽选择、发射功率、生成数量优化问题：**

$$\begin{aligned}
\mathcal{P}: \quad & \min_{\boldsymbol{\alpha},\mathbf{l},\boldsymbol{\phi},b} \bar{T} \\
\text{s.t.} \quad & T_n^{cp} + T_n^{mu} \leq \bar{T}, \quad \forall n \in \mathcal{N}^t, \\
& E_n^{cp} + E_n^{mu} \leq \bar{E}, \quad \forall n \in \mathcal{N}^t, \\
& l_n^t \in \{0,1\}, \quad \forall n \in \mathcal{N}^t, \\
& \sum_{n=1}^{|\mathcal{N}^t|} l_n^t \leq M, \\
& \phi_{\min} < \phi_n^t < \phi_{\max}, \quad \forall n \in \mathcal{N}^t, \\
& b^t \in \mathbb{N}, \\
& T_s^{inf} + T_s^{cp} \leq \bar{T}.
\end{aligned}$$

**子问题2：带宽分配策略**

$$\begin{aligned}
\mathcal{SUBP}2: & \min_{\{l_n\}_{n\in\mathcal{N}^t}} \bar{T} \\
\text{s.t. } & l_n^t \in \{0,1\}, \forall n \in \mathcal{N}^t, \\
& \sum_{n=1}^{\mathcal{N}^t} l_n^t \leq M, \\
& A + \frac{B}{l_n} \leq \bar{T}, \forall n \in \mathcal{N}^t, \\
& C + \frac{D}{l_n} \leq \bar{E}, \forall n \in \mathcal{N}^t.
\end{aligned}$$

通过KKT得到闭式解: $l_n^* = \sqrt{\frac{\lambda_{1,n} B_n + \lambda_2 D_n}{\lambda_3}}$.

**子问题3：功率控制策略**

$$\begin{aligned}
\mathcal{SUBP}3: & \min_{\boldsymbol{\phi}} \bar{T} \\
\text{s.t. } & \phi_{min} < \phi_n^t < \phi_{max}, \forall n \in \mathcal{N}^t, \\
& A + \frac{s(\omega)}{W log_2(1 + \frac{\phi_n h_0 d_n^{-\gamma}}{N_0})} <= \bar{T}, \forall n \in \mathcal{N}^t, \\
& G + \phi_n \frac{s(\omega)}{W log_2(1 + \frac{\phi_n h_0 d_n^{-\gamma}}{N_0})} \leq \bar{E}, \forall n \in \mathcal{N}^t,
\end{aligned}$$

通过SCA迭代得到近似解

**子问题4：数据生成策略**

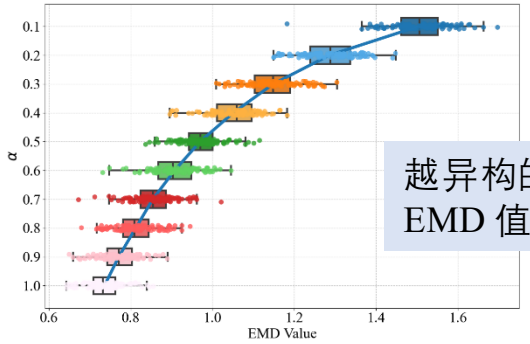
$$\begin{aligned}
\mathcal{SUBP}4: & \min_{\boldsymbol{b}} \bar{T} \\
& \text{s.t. } b^t \in \mathbb{N}, \\
& T_s^{inf} + T_s^{cp} \leq \bar{T}, \forall n \in \mathcal{N}^t.
\end{aligned}$$

得到闭式解: $\left\lfloor \frac{\max_{n\in\mathcal{N}^t}\{T_n^{cp} + T_n^{mu}\} - T_s^{cp}(\boldsymbol{b}^{t-1})}{\sum_{i=1}^{I} \frac{d_{m,i}}{f_{rm}}} \right\rfloor$,

# 内容1： AIGC 辅助的联邦学习增强方案

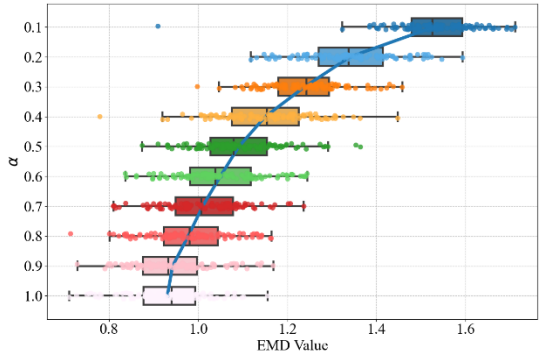
## 性能评估 两阶段优化算法性能对比

### EMD值与异构分布关系



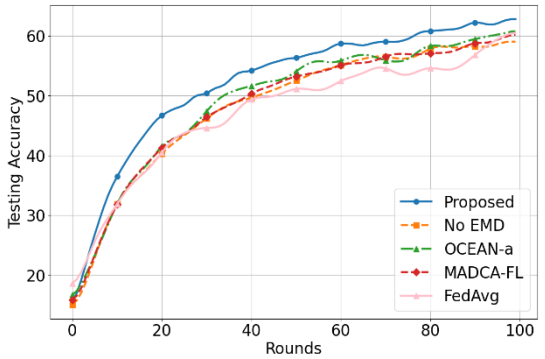

越异构的数据，EMD 值 越 大 。

(a) Cifar100数据集

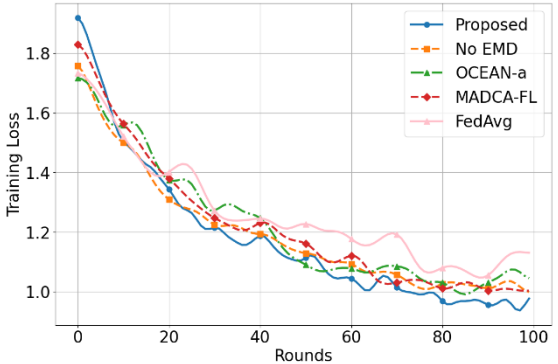



(b) GTSRB数据集

### 用户选择算法对比

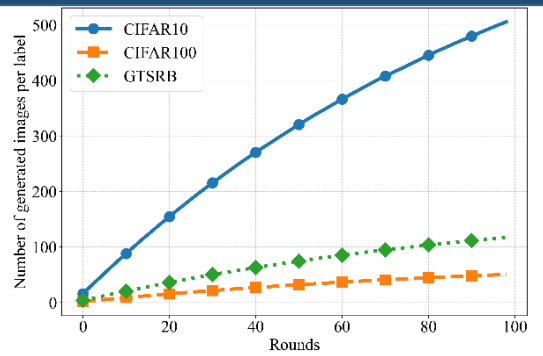



(a) 训练准确度

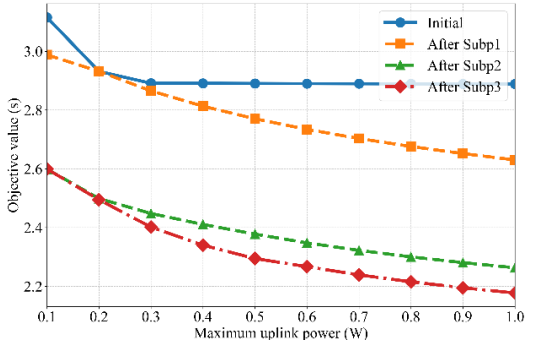



(b) 训练损失

### 两阶段优化算法对比




(a) 累计生成图片数量随着轮数的趋势

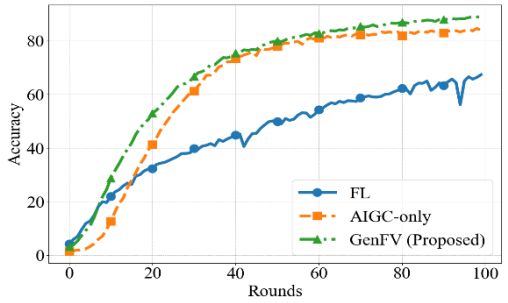



(b) 优化目标在各阶段随最大功率变化趋势

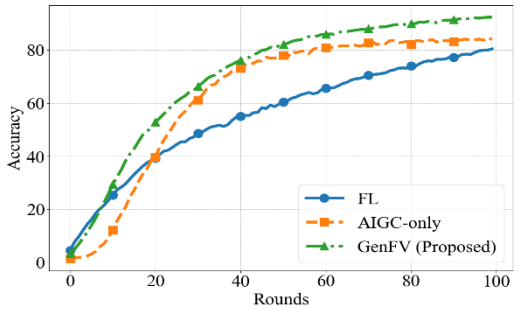

- X. Qiang, Z. Chang and G. Min, "AIGC-Assisted Federated Learning for Vehicular Edge Intelligence: Vehicle Selection, Resource Allocation and Model Augmentation," in IEEE Transactions on Mobile Computing, vol. 24, no. 11, pp. 11896-11909, Nov. 2025, doi: 10.1109/TMC.2025.3581983.
- X. Qiang, Z. Chang and Y.-C. Liang, "AIGC-assisted Federated Learning for Edge Intelligence: Architecture Design, Research Challenges and Future Directions," major revision, IEEE Communication Magazine, arXiv preprint arXiv:2503.20166, 2025.

# 内容1： AIGC 辅助的联邦学习增强方案

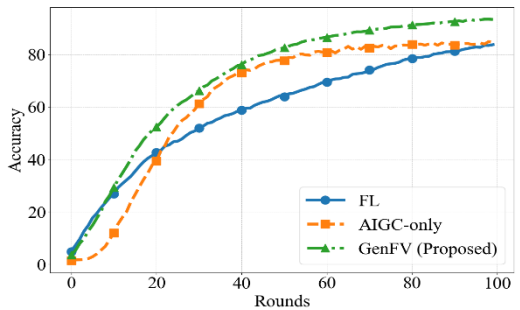
**性能评估** 不同数据集和异构分布下收敛速度与模型性能对比



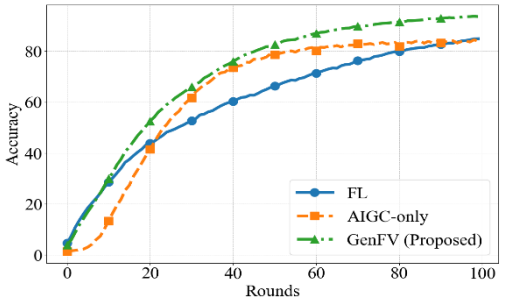

(a) Dir(0.1)




(b) Dir(0.3)

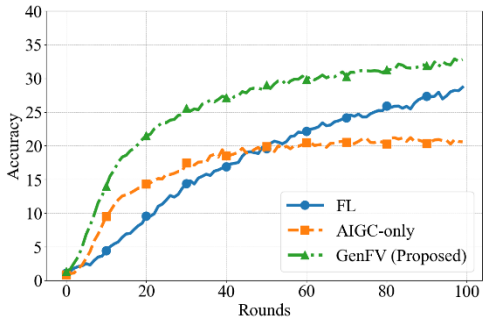



(c) Dir(0.5)

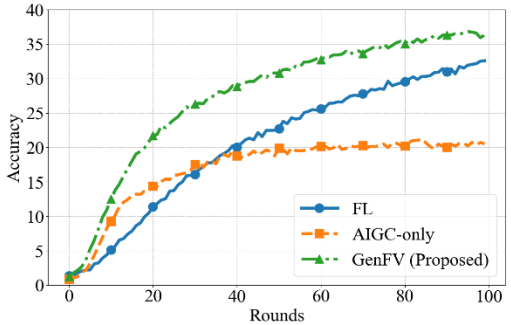



(d) Dir(1.0)

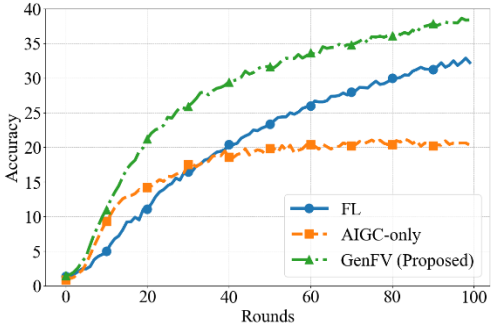
GTSRB数据集



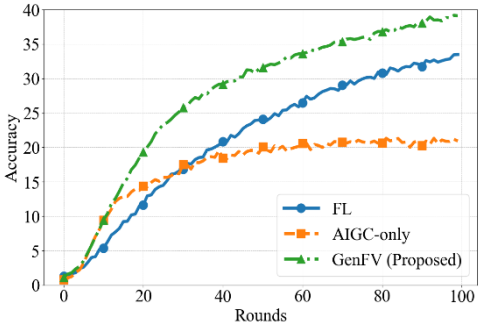

(a) Dir(0.1)




(b) Dir(0.3)




(c) Dir(0.5)




(d) Dir(1.0)

CIFAR100数据集

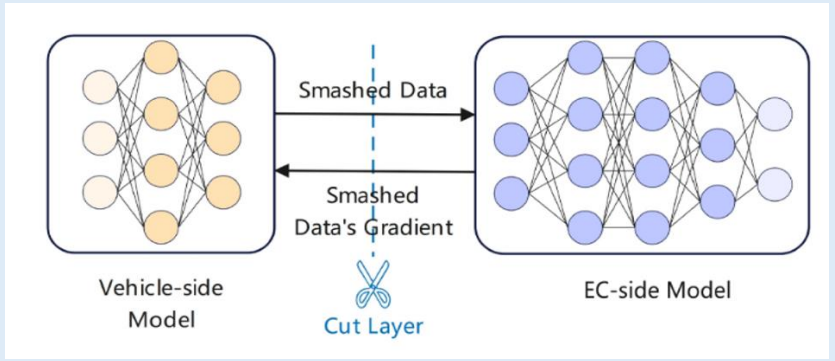
**GenFV利用AIGC增强数据，显著提升联邦学习精度与效率**

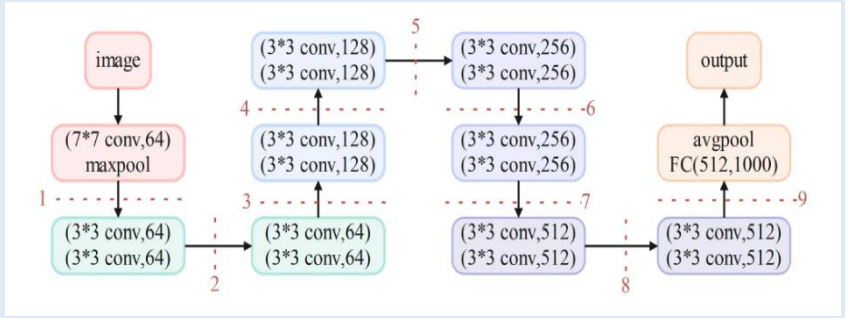
# 内容2：异构资源约束下的自适应模型拆分

**关键挑战** 边缘资源高度异构受限，导致本地模型训练载荷失衡。

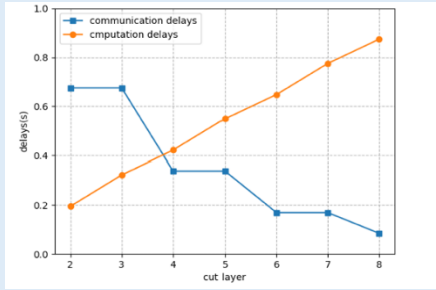
## 研究思路


**显存瓶颈：**
**终端难以支撑模型训练**




**模型拆分：**
**卸载计算压力到服务器端**




| 拆分层 | 通信时延 | 计算时延 | 总时延 |
|---|---|---|---|
| + | - | + | ? |

拆分选择　信道状态　计算频率

**tradeoff**

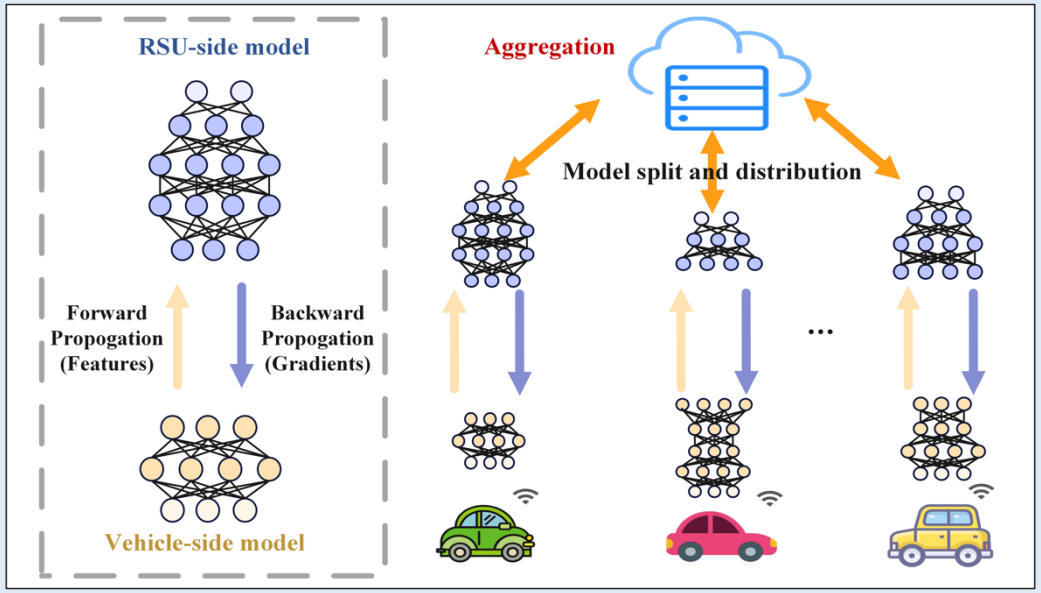
**自适应拆分：**
**均衡计算和通信开销**


- X. Qiang, Z. Chang, Y. Hu, L. Liu and T. Hämäläinen, "Adaptive and Parallel Split Federated Learning in Vehicular Edge Computing," *IEEE Internet of Things Journal*, vol. 12, no. 5, pp. 4591-4604, 1 Mar. 1, 2025.
- X. Qiang, Z. Chang, C. Ye, T. Hämäläinen and G. Min, "Split Federated Learning Empowered Vehicular Edge Intelligence: Concept, Adaptive Design, and Future Directions," *IEEE Wireless Communications*, vol. 32, no. 4, pp. 90-97, Aug.2025.

# 内容2：异构资源约束下的自适应模型拆分

**核心思想** 通过自适应拆分选择平衡通信和计算来最小化训练延迟

## ASFL系统流程



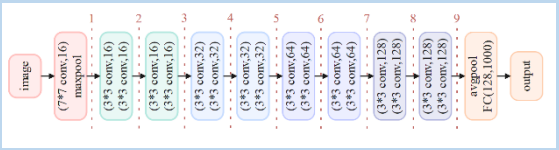

1. 服务器决策拆分层
2. 设备端模型计算并上传激活
3. 服务器端继续训练，反馈激活梯度
4. 设备端基于返回的激活梯度更新模型
5. 服务器收集客户端和服务器端模型进行聚合更新

## 自适应拆分选择

### 基于移动性的用户选择

$$\alpha_n^t = \begin{cases} 1, & t_n \le \bar{t}, \\ 0, & otherwise, \end{cases}$$

### 基于ResNet18的模型拆分




### 收敛性分析

**Assumption 1 ($\ell$-smooth):** 限制梯度变化率，确保函数平滑。
**Assumption 2 ($\mu$-strongly convex):** 保证存在唯一的全局最优解。
**Assumption 3 (Bounded Variance):** 限制局部随机梯度的噪声。
**Assumption 4 (Bounded Gradients):** 保证梯度不会无限增长。
**Assumption 5 (Uniform Sampling):** 描述分布数据non-IID。

收敛性分析结果：

$$\mathrm{E}\left[L(\boldsymbol{\omega}_T)\right] - L^* \le \frac{\nu}{\iota + T - 1}\left(\frac{2\Gamma}{\mu} + \frac{\mu\iota}{2}\mathrm{E}\left\|\boldsymbol{\omega}_1 - \boldsymbol{\omega}^*\right\|^2\right),$$

其中：$\Gamma = \sum_{n=1}^{N} p_n^2 \delta_n^2 + 6\ell\gamma_v + 8G^2 + \left(\frac{N}{K} - 1\right)\frac{N}{N-1}G^2$，$\gamma_v = L^* - \sum_{n=1}^{K} p_n L_n^*$。

# 内容2：异构资源约束下的自适应模型拆分

**问题建立** 构建关于带宽分配率、计算频率、发射功率、拆分选择的联合优化问题

## 时延能耗分析

1. 设备端模型广播：
   - 时延：$t_{d,n}=\frac{s(\omega_{n,\epsilon_n^t})}{R^{DL}}$，客户端模型大小
2. 设备端前向计算：
   - 时延：$t_{e,n}=\frac{|\mathcal{D}_n|\gamma_v^F(\epsilon_n^t)}{f_n\kappa}=\frac{|\mathcal{D}_n|c_v^F}{f_n}$，前向计算算力大小（Flops）
   - 能耗：$e_{e,n}=\frac{\zeta}{2}|\mathcal{D}_n|c_v^F(\epsilon_n^t)f_n^2$，
3. 激活上行传输：
   - 时延：$t_{s,n}=\frac{s(A_{n,\epsilon_n^t})}{R_n^{UL}}$，激活大小
   - 能耗：$e_{s,n}=\phi_n t_{s,n}$
4. 服务器端模型计算：
   - 时延：$t_R=\frac{|\mathcal{D}_n|(\gamma_r^F(\epsilon_n^t)+\gamma_r^B(\epsilon_n^t))}{f_s\kappa}$，前后向计算算力大小(Flops)
5. 激活梯度下行传输：
   - 时延：$t_{g,n}=\frac{s(g(A_{n,\epsilon_n^t}))}{R^{DL}}$，激活梯度大小
6. 设备端后向计算
   - 时延：$t_{u,n}=\frac{|\mathcal{D}_n|\gamma_v^B(\epsilon_n^t)}{f_n\kappa}=\frac{|\mathcal{D}_n|c_v^B}{f_n}$，后向计算算力大小（Flops）
   - 能耗：$e_{u,n}=\frac{\zeta}{2}|\mathcal{D}_n|c_v^B(\epsilon_n^t)f_n^2$，
7. 设备端模型上传
   - 时延：$t_{w,n}=\frac{s(\omega_{n,\epsilon_n^t})}{R_n^{UL}}$，设备端模型大小
   - 能耗：$e_{w,n}=\phi_n t_{w,n}$

## 问题建立

$$\mathcal{P}: \min_{\beta_n^t,f_n^t,\phi_n^t,\epsilon_n^t} T(\{\beta_n^l,f_n^l,\phi_n^l,c_n^l\}_{n\subset\mathcal{N}_t})$$

$$\text{s.t. C1: } \sum_{n=1}^{\mathcal{N}_t}\beta_n^l\le 1$$

$$\text{C2: } 0\le\beta_n^l\le 1,$$

$$\text{C3: } e_n^l\le\hat{E},$$

$$\text{C4: } c_n^l\in\mathcal{E},$$

$$\text{C5: } f_{min}\le f_n^l\le f_{max},$$

$$\text{C6: } \phi_{min}\le\phi_n^l\le\phi_{max}.$$

**优化目标及约束**

最小化系统时延

- 带宽约束1
- 带宽约束2
- 能耗约束
- 拆分层约束
- 计算功率约束
- 发射功率约束

**优化变量**

- 带宽分配率 $\beta$
- 计算频率 $f$
- 发射功率 $\phi$
- 拆分层 $\epsilon$

**问题分析**

混合整数非线性规划，且非凸，是一个NP-hard优化问题

# 内容2：异构资源约束下的自适应模型拆分

## 问题求解　联合迭代优化求解问题

$$\mathcal{P}: \min_{\beta_n^t, f_n^t, \phi_n^t, \epsilon_n^t} T(\{\beta_n^t, f_n^t, \phi_n^t, c_n^t\}_{n\subset\mathcal{N}_t})$$

$$\text{s.t. C1: } \sum_{n=1}^{\mathcal{N}_t} \beta_n^t \le 1$$

$$\text{C2: } 0 \le \beta_n^t \le 1,$$

$$\text{C3: } e_n^t \le \hat{E},$$

$$\text{C4: } c_n^t \in \mathcal{E},$$

$$\text{C5: } f_{min} \le f_n^t \le f_{max},$$

$$\text{C6: } \phi_{min} \le \phi_n^t \le \phi_{max}.$$

### 子问题1：自适应拆分选择

$$\mathcal{SUBP}1: \min_{\epsilon_n^t} \{\frac{|\mathcal{D}_n| c_v(\epsilon_n^t)}{f_n} + \frac{|\mathcal{D}_n| c_r(\epsilon_n^t)}{f_r} + \frac{\overline{s_g(\epsilon_n^t)}}{R^{DL}} + \frac{\overline{s_a(\epsilon_n^t)}}{R_n^{UL}}\},$$

$$\text{s.t.} \quad e_n^t \le \hat{E},$$

$$\epsilon_n^t \in \mathcal{E},$$

拆分层变量范围小

由于变量范围小且为整数，采取逐个遍历也能快速得到结果

### 子问题2：优化传输功率

$$\mathcal{SUBP}2: \min_{\phi_n^t} \quad T,$$

$$\text{s.t.} \quad e_n^t \le \hat{E},$$

$$\phi_{min} \le \phi_n^t \le \phi_{max},$$

$$\frac{A}{f_n} + \frac{\overline{s_a(\epsilon_n)}}{\beta_n W ln(1 + \frac{h_n \phi_n d_n^{-\gamma}}{\sigma_0^2})} + C \le \bar{T}.$$

$$\hat{e}(\phi_n^i, \phi_n) = e(\phi_n^i) + e'(\phi_n^i)(\phi_n - \phi_n^i),$$

关于发射功率是非凸的，使用泰勒展开进行替代，替换后，使用SCA迭代优化得到近优解

### 子问题3：联合优化计算频率和无线资源分配

$$\mathcal{SUBP}3: \min_{f_n, \beta_n} \quad \bar{T}$$

$$\text{s.t.} \sum_{n\in\mathcal{N}_t} \beta_n \le 1,$$

$$0 < \beta_n \le 1,$$

$$Df_n^2 + \frac{F}{\beta_n} \le \hat{E}, \forall n \in \mathcal{N}_t,$$

$$f_{min} \le f_n \le f_{max},$$

$$t_n = \frac{A}{f_n} + \frac{B}{\beta_n} + C \le \bar{T}, \forall n \in \mathcal{N}_t.$$

**定理1**：SUBP 3 是凸的。

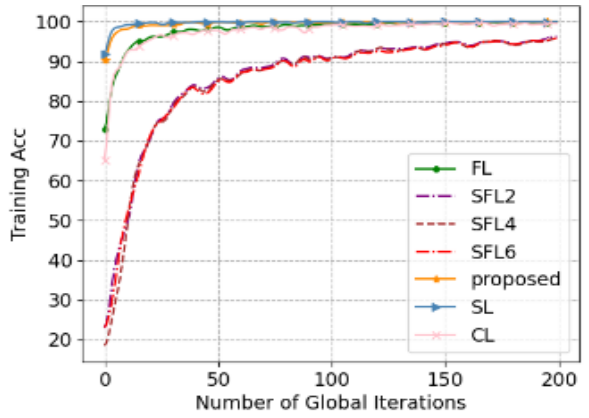
**证明**：子问题SUBP 3的目标函数可分解为三部分：$\frac{A}{f_n}$、$\frac{B}{\beta_n}$ 和 $Df_n^2 + \frac{F}{\beta_n}$。由于上述各项在其定义域内均为凸函数，且该子问题的约束均为仿射形式，因此可以判定SUBP 3是一个凸优化问题。

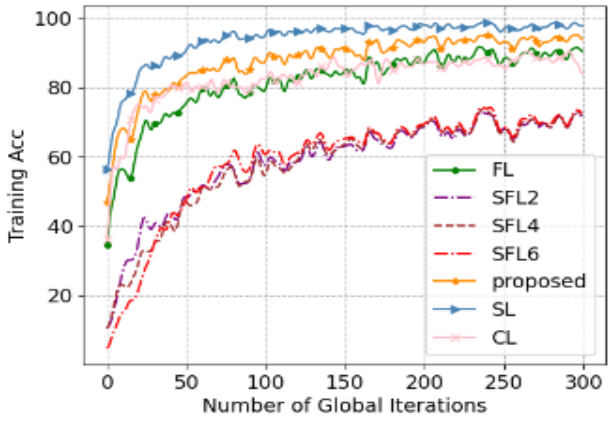
利用KKT得到闭式解

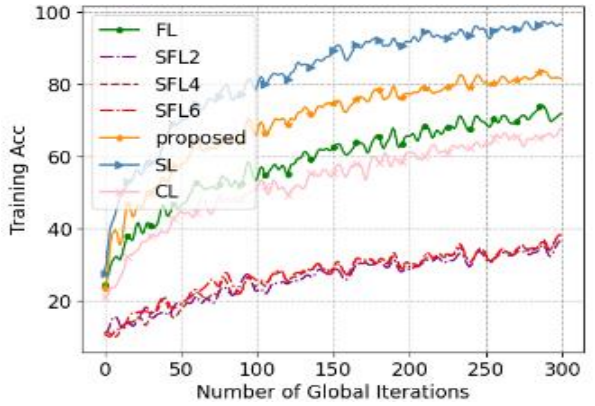
$$\beta_n^* = \frac{\sigma_n^{\frac{1}{2}}(B + \frac{A}{2Df_n^3}F)^{\frac{1}{2}}}{\sum_{n\in\mathrm{N}_t} \sigma_n^{\frac{1}{2}}(B + \frac{A}{2Df_n^3}F)^{\frac{1}{2}}}$$

# 内容2：异构资源约束下的自适应模型拆分

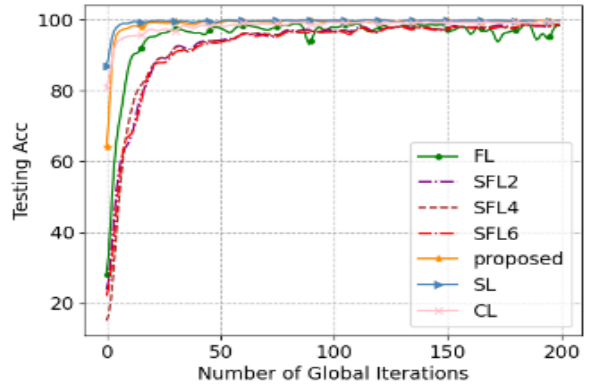
**性能评估** **模型性能显著增强、系统开销显著降低**

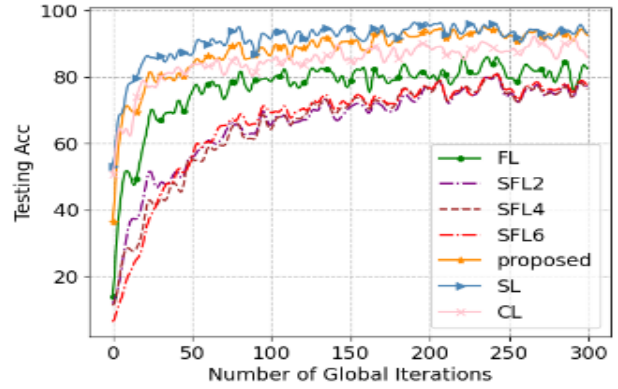


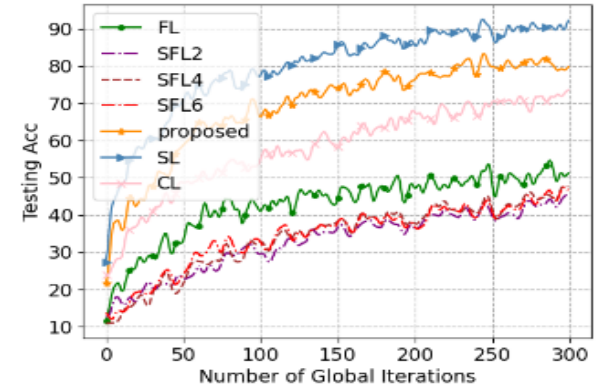




训练准确度




测试准确度

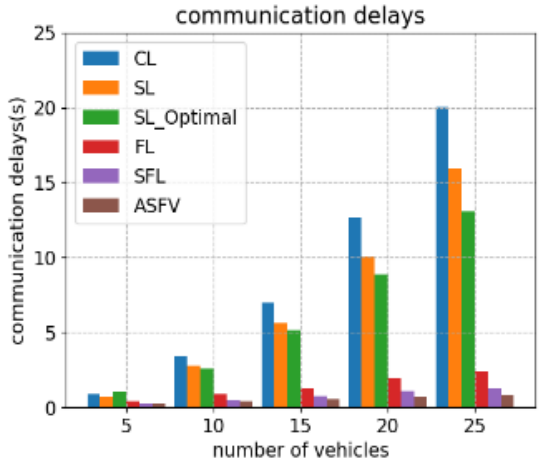
(a) Mnist (b) F-Mnist (c) Cifar10


- X. Qiang, Z. Chang, Y. Hu, L. Liu and T. Hämäläinen, "Adaptive and Parallel Split Federated Learning in Vehicular Edge Computing," IEEE Internet of Things Journal, vol. 12, no. 5, pp. 4591-4604, 1 Mar. 1, 2025.
- X. Qiang, Z. Chang, C. Ye, T. Hämäläinen and G. Min, "Split Federated Learning Empowered Vehicular Edge Intelligence: Concept, Adaptive Design, and Future Directions," IEEE Wireless Communications, vol. 32, no. 4, pp. 90-97, Aug.2025.

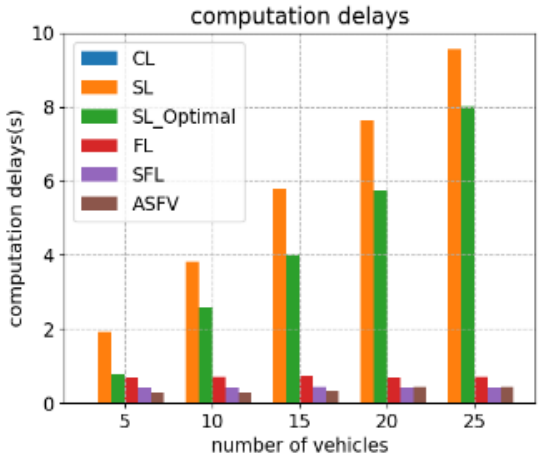
# 内容2：异构资源约束下的自适应模型拆分

**性能评估** **模型性能显著增强、系统开销显著降低**

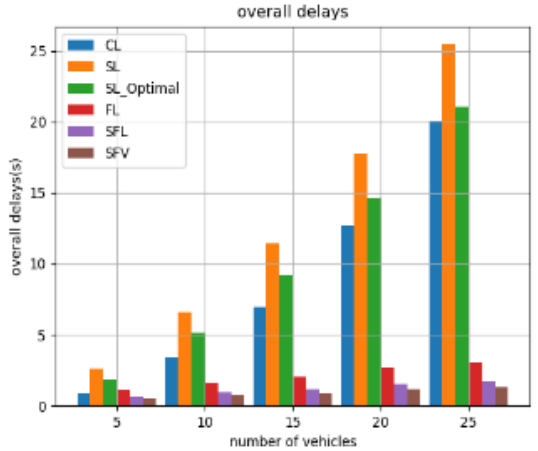
通信时延




计算时延

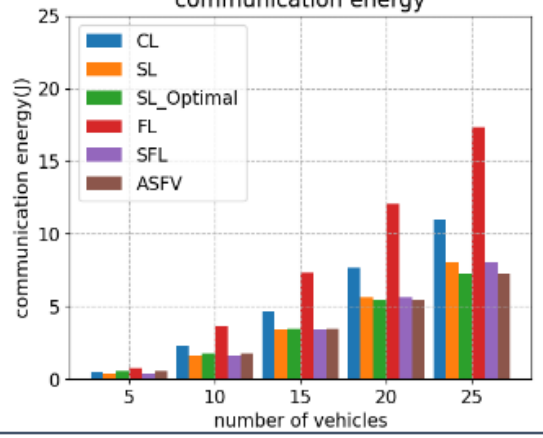



总时延

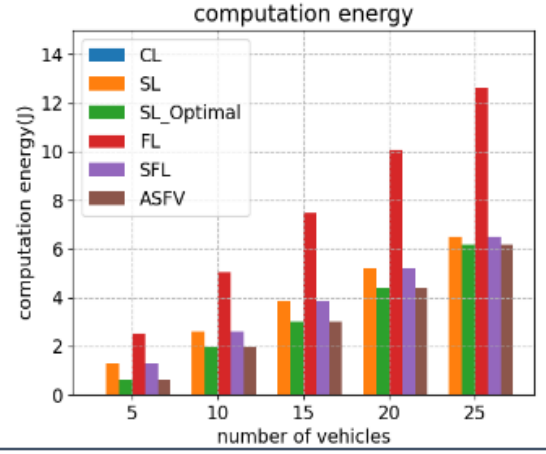



时延开销

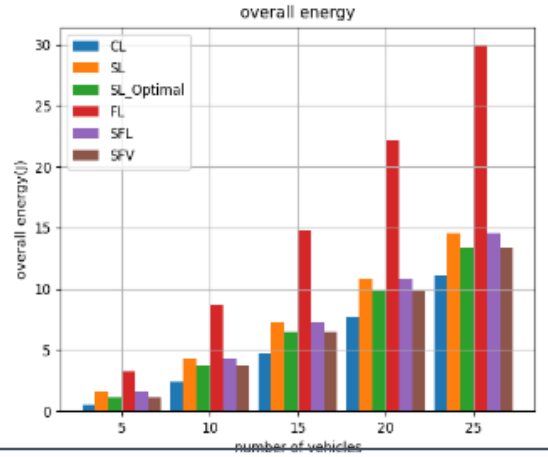
通信能耗




计算能耗




总能耗




能耗开销

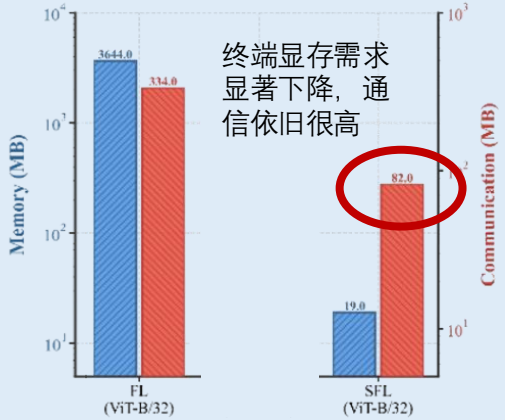
**GenFV利用AIGC增强数据，显著提升联邦学习精度与效率**

# 内容3：带宽受限下的自适应激活量化传输

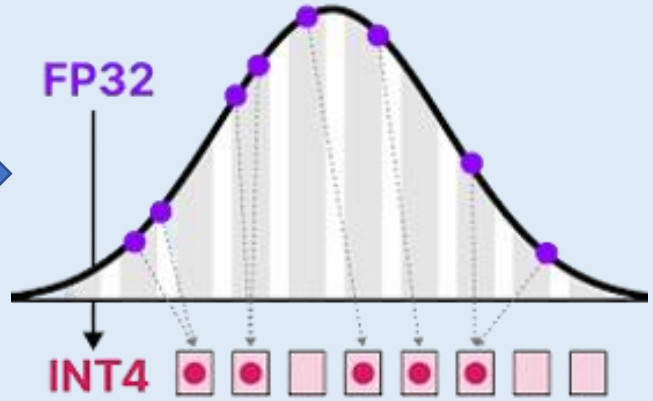
**关键挑战** 无线带宽资源受限，难以承载高维激活特征的频繁交互。

## 研究思路

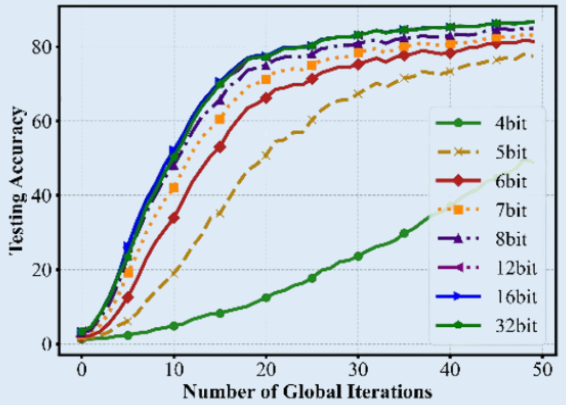
**挑战1：高维激活上传带来通信瓶颈**

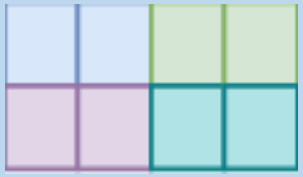



**解决方案：激活量化**




**挑战2：激活量化带来的精度损失**



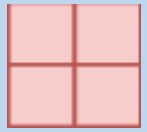
量化前

准确度

功率分配

带宽分配

**自适应选择激活量化位数**



量化后

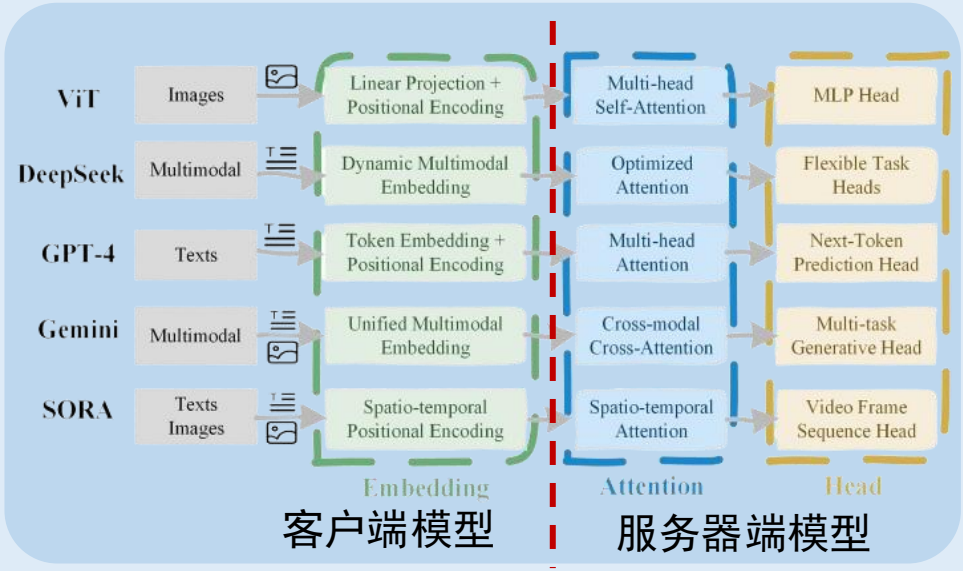
X. Qiang, H. Liu, X. Zhang, Z. Chang, Y. C. Liang, "Deploying Large AI Models on Resource-Limited Devices with Split Federated Learning." IEEE Transaction on Mobile Computing, 2026, arXiv preprint arXiv:2504.09114

# 内容3：带宽受限下的自适应激活量化传输

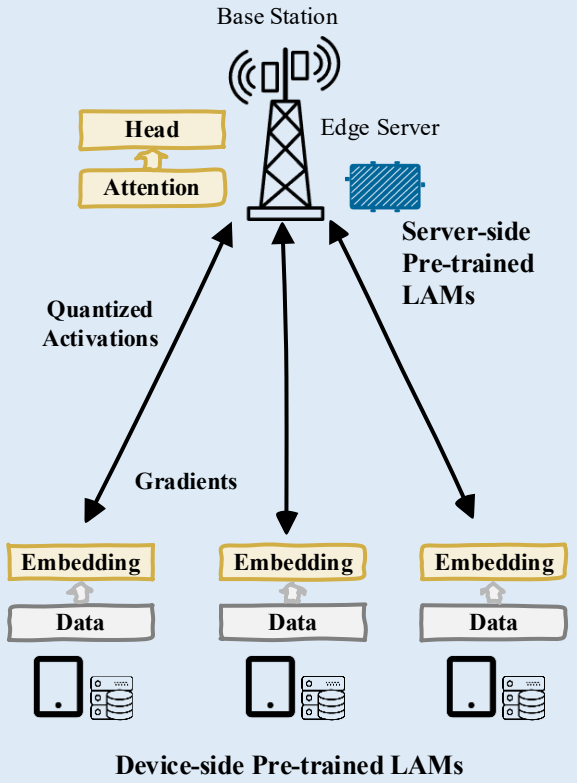
**核心思想** 通过自适应量化均衡模型准确度与通信开销

## SFLAM流程

从传统CNN小模型到基于Transformer的大模型




基于Transformer的模型拆分



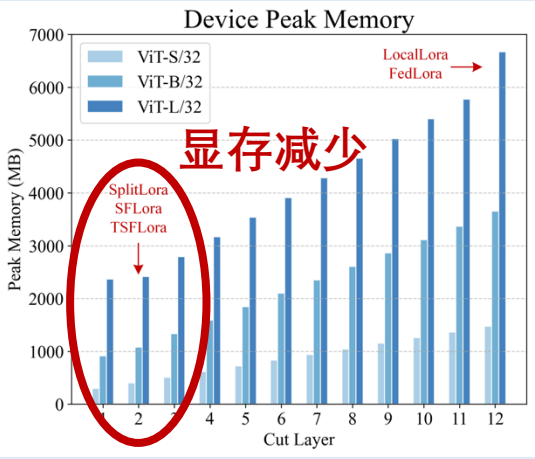

1. 服务器决策量化位数
2. 设备端模型计算并上传量化后的激活
3. 服务器端继续训练，反馈激活梯度
4. 设备端基于返回的激活梯度更新模型
5. 服务器收集客户端和服务器端模型进行聚合更新

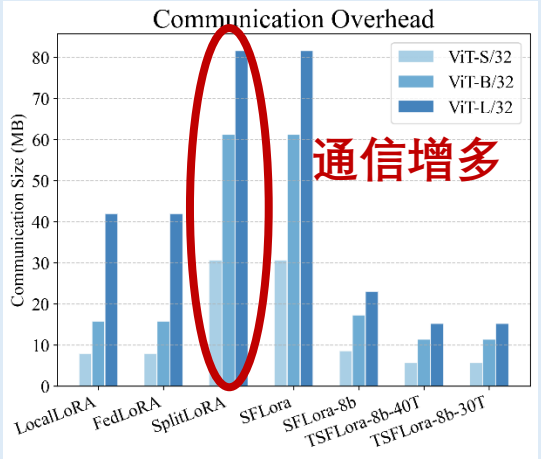
# 内容3：带宽受限下的自适应激活量化传输

**核心思想** 通过自适应量化均衡模型准确度与通信开销

## 模型拆分

从传统CNN小模型到基于Transformer的大模型



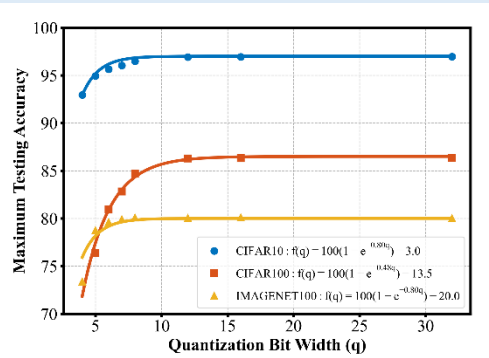

**观察1**：以ViT模型为例，客户端仅需部署 Embedding，其显存占用相比于加载完整模型显著降低。

**观察2**：频繁的激活传输带来了很大的通信开销。

**通过拆分学习突破存储瓶颈**
**利用激活量化加速训练进程。**

## 自适应量化

**观察3**：随着激活量化位数的减少，测试准确度也逐渐降低




**定义**准确度与量化位数的关系：

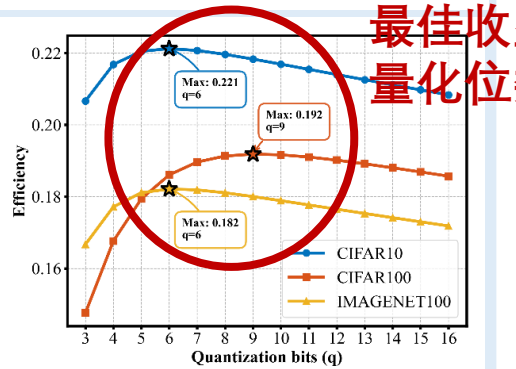
$$O_{n,k} = a\left(1 - e^{-bq_{n,k}}\right) + c,$$

$O_{n,k}$：准确度　$q_{n,k}$：量化位数

**定义**通信能耗与量化位数的关系：

$$E_n^{mu} = p_{n,k} \frac{s_0 \cdot q_{n,k} / q^{\max}}{R_{n,k}^{\mathrm{UL}}}$$

$E_n^{mu}$：激活上传能耗




**效率指标**
单位能量消耗所带来的学习效益

$$\mathcal{E}_{n,k} = \frac{O_{n,k}(q)}{E_{n,k}(q)},$$

$q\uparrow$ → $Accuracy\ O\uparrow$
$q\uparrow$ → $Comm.\ Energy\ E\uparrow$

# 内容3：带宽受限下的自适应激活量化传输

**问题建立** 构建关于用户选择，带宽分配，传输功率和数据生成量的联合优化问题

## 时延能耗分析

1. 本地模型计算：

- 时延：$T_n^{\text{cmp}} = \dfrac{b\gamma_d}{f_n C_n D_n}$，
- 能耗：$E_n^{\text{cmp}} = \dfrac{\kappa_1 f_n^2 b\gamma_d}{C_n D_n}$.

2. 设备服务器之间的传输：

$$s_{n,k} = s_0 \frac{q_{n,k}}{q^{\max}}$$

传输开销与量化位数成正比

- 时延：$T_{n,k}^{\text{com}} = \dfrac{s_{n,k}}{R_{n,k}^{\text{UL}}}$，
- 能耗：$E_n^{mu} = p_{n,k} \dfrac{s_{n,k}}{R_{n,k}^{\text{UL}}}$

3. 服务器端计算：

- 时延：$T_s^{\text{cmp}} = \dfrac{b\gamma_s}{f_s C_s D_s}$，

设备n的总时延：

$$T_{n,k} = T_n^{\text{cmp}} + T_s^{\text{cmp}} + T_{n,k}^{\text{com}}$$

设备n的总能耗：

$$E_{n,k} = E_n^{\text{cmp}} + E_{n,k}^{\text{com}}$$

## 问题建立

$$\begin{aligned}
\mathcal{P}: \quad & \max_{\mathbf{z},\mathbf{p},\mathbf{q}} \sum_{n\in\mathcal{V}} a_{n,k}\mathcal{E}_{n,k} \\
\text{s.t.} \quad & T_{n,k} \le T_{\max}, \quad \forall n \in \mathcal{V}, \\
& \sum_{n\in\mathcal{V}} z_{n,k}^{(m)} \le 1, \quad \forall m = 1, \ldots, M, \\
& \sum_{m=1}^{M} z_{n,k}^{(m)} \le 1, \quad \forall n \in \mathcal{V}, \\
& z_{n,k}^{(m)} \in \{0,1\}, \quad \forall n \in \mathcal{V}, \forall m, \\
& p^{\min} \le p_{n,k} \le p^{\max}, \quad \forall n \in \mathcal{V}, \\
& q^{\min} \le q_{n,k} \le q^{\max}, \quad q_{n,k} \in \mathbb{Z}, \quad \forall n \in \mathcal{V}, \\
& a_{n,k} = \sum_{m=1}^{M} z_{n,k}^{(m)}, \quad \forall n \in \mathcal{V}.
\end{aligned}$$

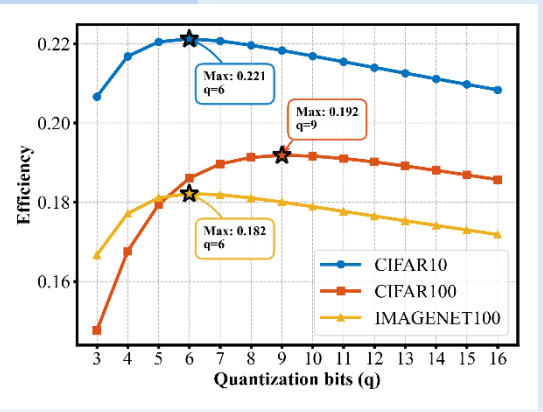
**优化目标及约束**

最大化系统效率

- 时间约束
- 带宽分配约束1
- 带宽分配约束2
- 带宽分配约束3
- 发射功率约束
- 量化约束1

**优化变量**

- 带宽分配 $z$
- 发射功率 $p$
- 量化位数 $q$



# 内容3：带宽受限下的自适应激活量化传输

## 问题求解　联合迭代优化求解问题

$$
\begin{aligned}
\mathcal{P}: \quad & \max_{\mathbf{z},\mathbf{p},\mathbf{q}} \quad \sum_{n\in\mathcal{V}} a_{n,k}\mathcal{E}_{n,k} \\
\text{s.t.} \quad & T_{n,k} \le T_{\max}, \quad \forall n \in \mathcal{V}, \\
& \sum_{n\in\mathcal{V}} z_{n,k}^{(m)} \le 1, \quad \forall m = 1,\ldots,M, \\
& \sum_{m=1}^{M} z_{n,k}^{(m)} \le 1, \quad \forall n \in \mathcal{V}, \\
& z_{n,k}^{(m)} \in \{0,1\}, \quad \forall n \in \mathcal{V}, \forall m, \\
& p^{\min} \le p_{n,k} \le p^{\max}, \quad \forall n \in \mathcal{V}, \\
& q^{\min} \le q_{n,k} \le q^{\max}, \quad q_{n,k} \in \mathbb{Z}, \quad \forall n \in \mathcal{V}, \\
& a_{n,k} = \sum_{m=1}^{M} z_{n,k}^{(m)}, \quad \forall n \in \mathcal{V}.
\end{aligned}
$$

功率控制 → 量化管理 → 带宽分配

### 子问题1：功率控制

$$
\begin{aligned}
\mathcal{SUBP}1: \quad & \min_{p_{n,k}} \quad E_n^{\text{cmp}} + \frac{p_{n,k}s_{n,k}}{R_{n,m,k}^{\text{UL}}(p_{n,k})} \\
\text{s.t.} \quad & T_n^{\text{cmp}} + T_{n,k}^{\text{com}} \le T_{\max}, \\
& p^{\min} \le p_{n,k} \le p^{\max},
\end{aligned}
$$

SUBP1是非凸的,考虑使用一阶泰勒展开替代非凸优项，再通过SCA迭代求解近似解

$$
\hat{E}_{n,k}^{\text{com}}(p_{n,k}^{(i)}, p_{n,k}) = E_{n,k}^{\text{com}}(p_{n,k}^{(i)}) + \frac{dE_{n,k}^{\text{com}}}{dp_{n,k}^{(i)}} \cdot (p_{n,k} - p_{n,k}^{(i)}),
$$

$$
\hat{T}_{n,k}^{\text{com}}(p_{n,k}^{(i)}, p_{n,k}) = T_{n,k}^{\text{com}}(p_{n,k}^{(i)}) + \frac{dT_{n,k}^{\text{com}}}{dp_{n,k}^{(i)}} \cdot (p_{n,k} - p_{n,k}^{(i)}),
$$

### 子问题2：量化管理

$$
\begin{aligned}
\mathcal{SUBP}2': \quad & \max_{\mathbf{q}} \quad \frac{O_{n,k}}{E_{n,k}} \\
\text{s.t.} \quad & T_n^{\text{cmp}} + \frac{s_0}{R_{n,m,k}^{UL} q^{\max}} q_{n,k} \le T_{\max}, \\
& q^{\min} \le q_{n,k} \le q^{\max}.
\end{aligned}
$$

变量范围有限

系统最优等价于每个设备最优。由于变量范围有限，逐个遍历得到每个设备在当前的通信环境下的最优解

**Algorithm 2** Optimal Quantization Management for Device $n$

**Require:** Set the maximum time constraint $T_{\max}$, iteration round $i = 0$.
**for** each $q_{n,k} \in [q^{\min}, q^{\max}]$ **do**
Compute $\mathcal{E}_{n,k}$ and update optimal $q_{n,k}^*$.
**end for**
**Ensure:** Optimal quantization level $q_{n,k}^*$.

### 子问题3：带宽分配

$$
\begin{aligned}
\mathcal{SUBP}3: \quad & \max_{\mathbf{z}_k} \quad \sum_{n=1}^{N}\sum_{m=1}^{M} z_{n,k}^{(m)} \frac{O_{n,k}}{E_{n,k}} \\
\text{s.t.} \quad & \sum_{n\in\mathcal{V}} z_{n,k}^{(m)} \le 1, \quad \forall m = 1,\ldots,M, \\
& \sum_{m=1}^{M} z_{n,k}^{(m)} \le 1, \quad \forall n \in \mathcal{V}, \\
& z_{n,k}^{(m)} \in \{0,1\}, \quad \forall n \in \mathcal{V}, \forall m, \\
& p^{\min} \le p_{n,k} \le p^{\max}, \quad \forall n \in \mathcal{V},
\end{aligned}
$$

转化

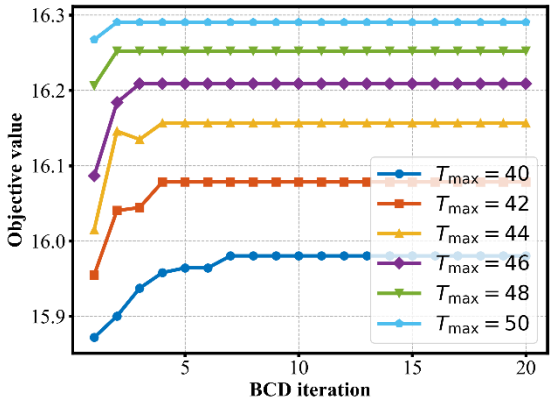
$$
\begin{aligned}
\overline{\mathcal{SUBP}}3: \quad & \max_{\theta} \quad \sum_{n=1}^{N}\sum_{m=1}^{M} \theta_{n,m}\Delta_{n,m} \\
\text{s.t.} \quad & \sum_{n=1}^{N} \theta_{n,m} = 1, \forall m \in \overline{\mathcal{M}}, \\
& \sum_{m=1}^{\overline{M}} \theta_{n,m} \le 1, \forall n \in \mathcal{V}, \\
& 0 \le \theta_{n,m} \le 1, \forall n \in \mathcal{V}, \forall m \in \overline{\mathcal{M}}.
\end{aligned}
$$

$\overline{SUBP}3$可视为的线性规划松弛问题。考虑到该问题共有NM个变量，故可利用内点法高效求解。

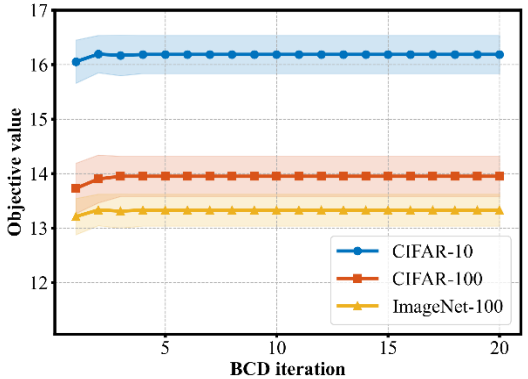
# 内容3：带宽受限下的自适应激活量化传输

**性能评估** 优化算法收敛分析、敏感性分析、不同优化算法对比

## 优化算法收敛情况

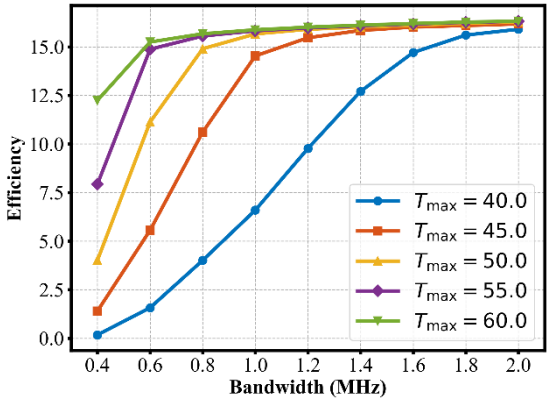



在不同约束下的目标收敛情况

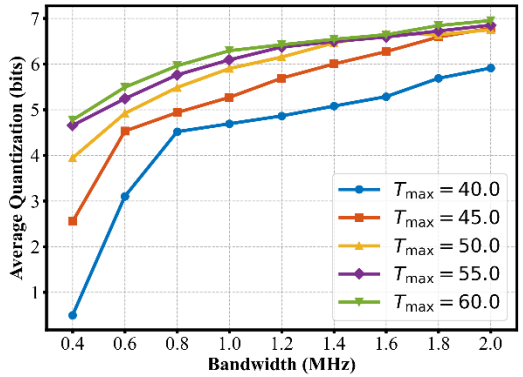



不同数据集下目标收敛情况

## 关于时间约束的敏感性分析

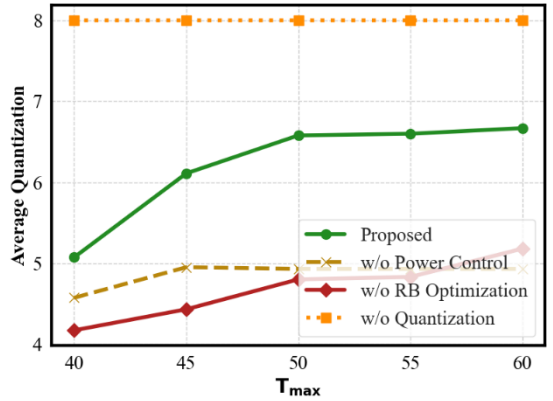



系统效率与带宽/时延约束分析

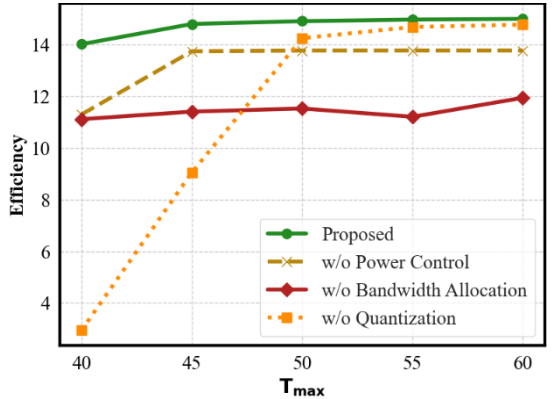



量化位数与带宽/时间约束分析

## 不同优化算法对比




不同优化算法下量化位数分析



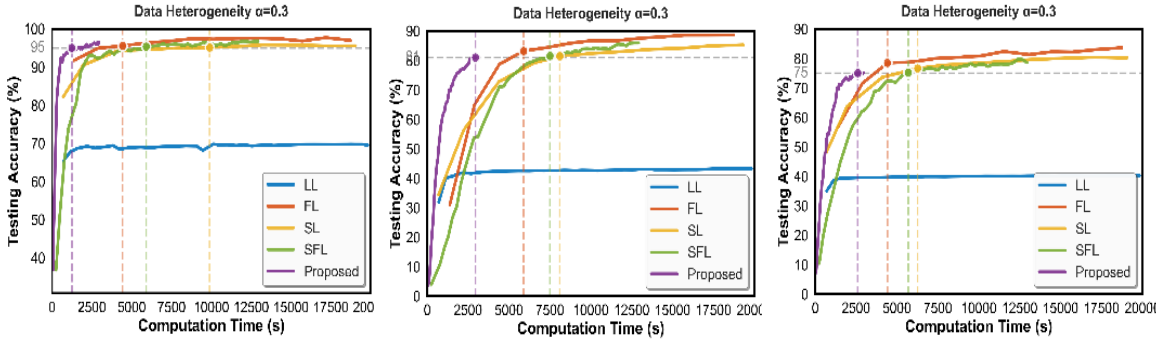

不同优化算法下系统效率分析

**SFLAM自适应均衡模型性能与系统开销，有效提高模型训练效率**

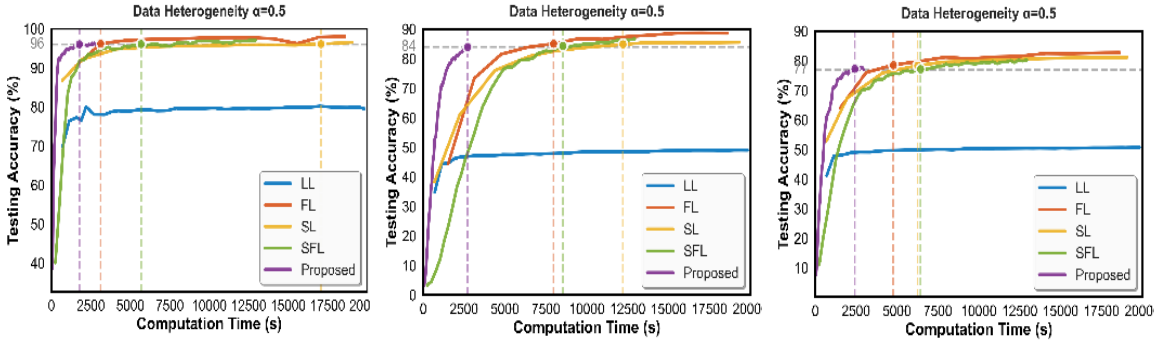
# 内容3：带宽受限下的自适应激活量化传输

**性能评估** **模型训练效率显著增强、系统开销显著降低**

测试准确度 Dir(0.3)



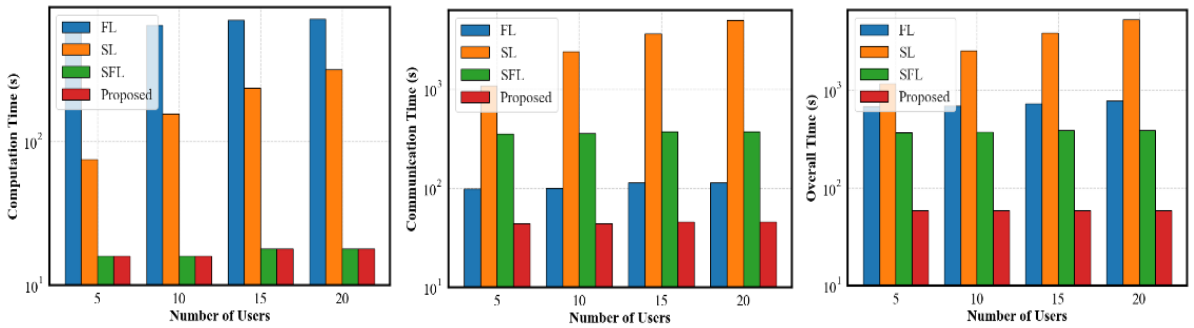

测试准确度 Dir(0.5)



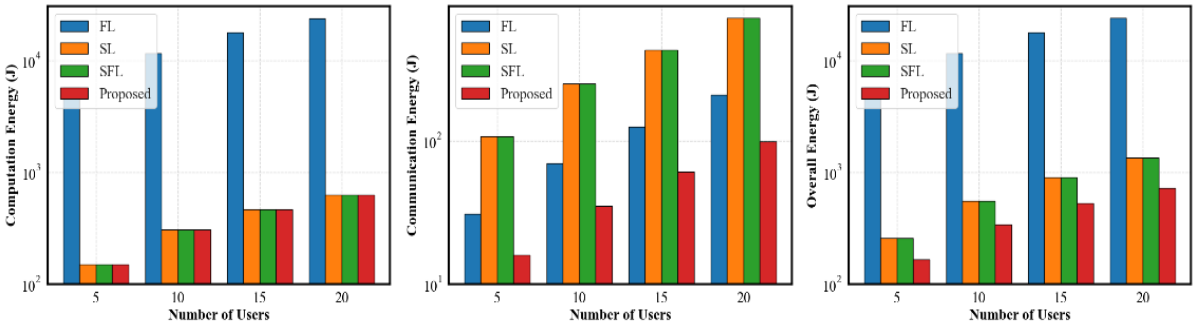

(a) CIFAR-10 (b) CIFAR-100 (c) ImageNet-100

**模型效果**

时延开销




能耗开销




通信 计算 总和

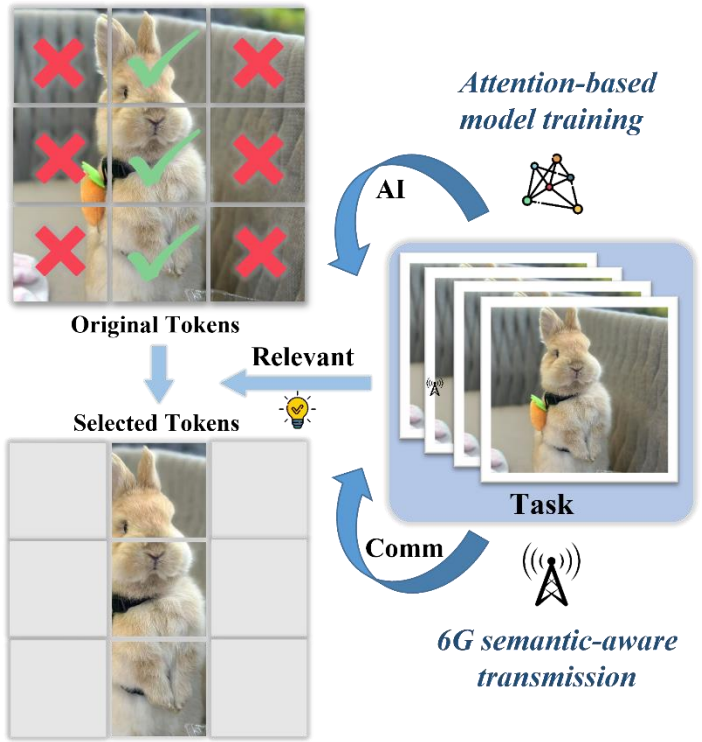
**系统开销**

- X. Qiang, H. Liu, X. Zhang, Z. Chang, Y. C. Liang, "Deploying Large AI Models on Resource-Limited Devices with Split Federated Learning." IEEE Transaction on Mobile Computing, 2026, arXiv preprint arXiv:2504.09114 .

# 内容4： 任务驱动下的语义感知Token选择机制

**机遇** 6G与大模型微调/推理如何融合?

**AI训练** 基于attention机制的训练

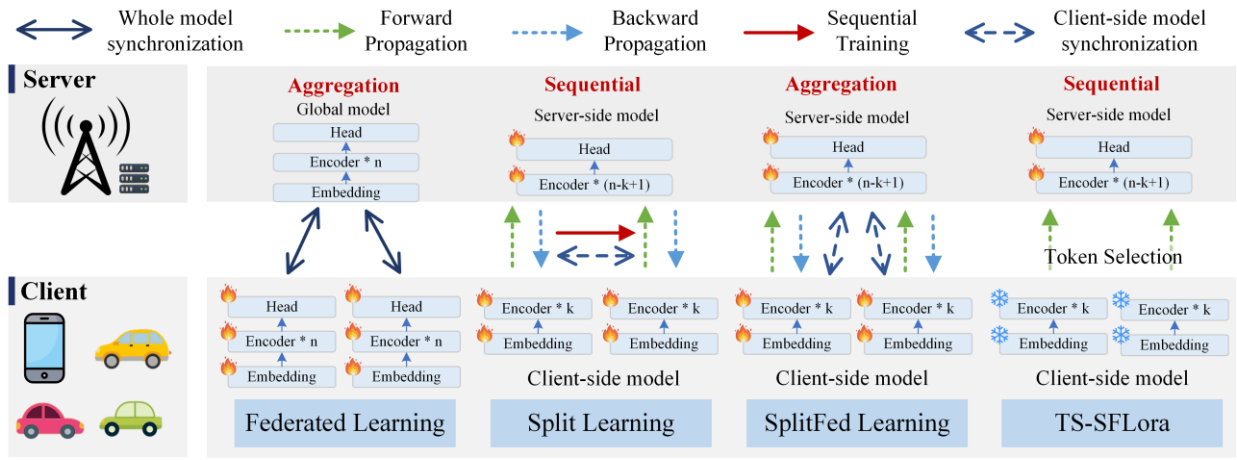



**6G视角** bit到语义的传输范式转变

## 基于transformer结构大模型分布式微调分析

FL微调： 权重 + 梯度
SFL微调： 权重 + 梯度 + 激活

- 含义：输入的特征
- 形式：**类Token序列**




## 6G从以bit传输到语义传输范式转变


**语义编码器**：通常为Encoder结构，用于将输入编码为Token级语义表示。

**语义解码器**：通常为decoder结构，将接收到的Token级语义转化为下游输出。


- X. Qiang, Z. Chang and G. Min, "Semantic-aware Token Selection and Resource Optimization for Communication-efficient Split Federated Fine-tuning in Edge Intelligence," submitted.
- X. Qiang, Z. Chang and Y. C. Liang, "TSFLora: Token-Compressed Split Fine-Tuning for Wireless Edge Networks", submitted.

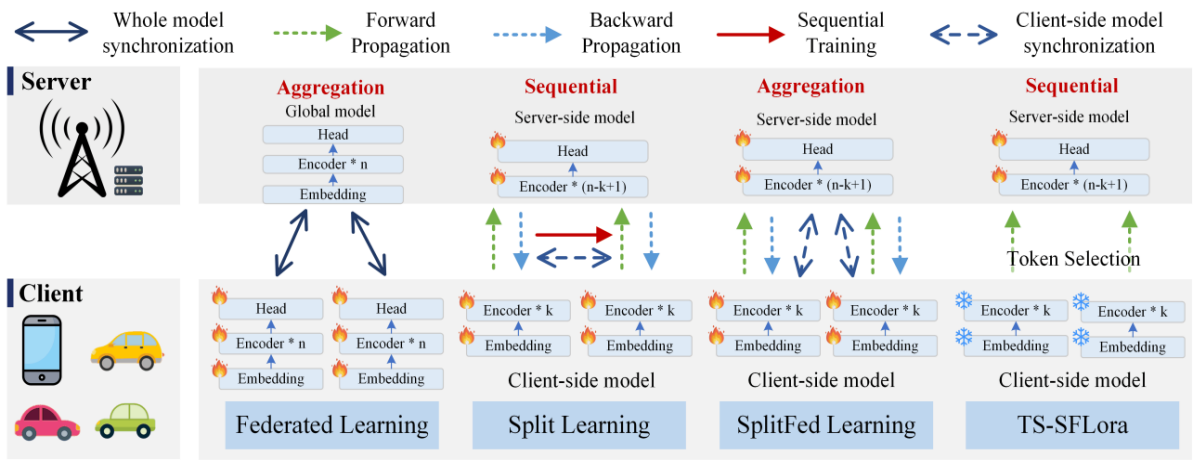
# 内容4： 任务驱动下的语义感知Token选择机制

**问题** 如何将语义思想与SFL微调通信压缩融合？

## 基于transformer结构大模型分布式微调分析

FL微调： 权重 + 梯度

SFL微调： 权重 + 梯度 + 激活 { 含义：输入的特征；形式：**Token序列**




## 6G从以bit传输到语义传输范式转变


**语义编码器**：通常为Encoder结构，用于将输入编码为Token级语义表示。

**语义解码器**：通常为decoder结构，将接收到的Token级语义转化为下游输出。

基于语义提取思想进行自适应token选择传输

**从微调角度**：SFL fine-tuning传输负担太重，尤其是频繁的激活传输

## 挑战

1. 如何定量描述token数量对激活传输开销？
2. 如何定量不同token对任务的重要程度/语义？
3. 如何设计基于token数量的联合优化问题？

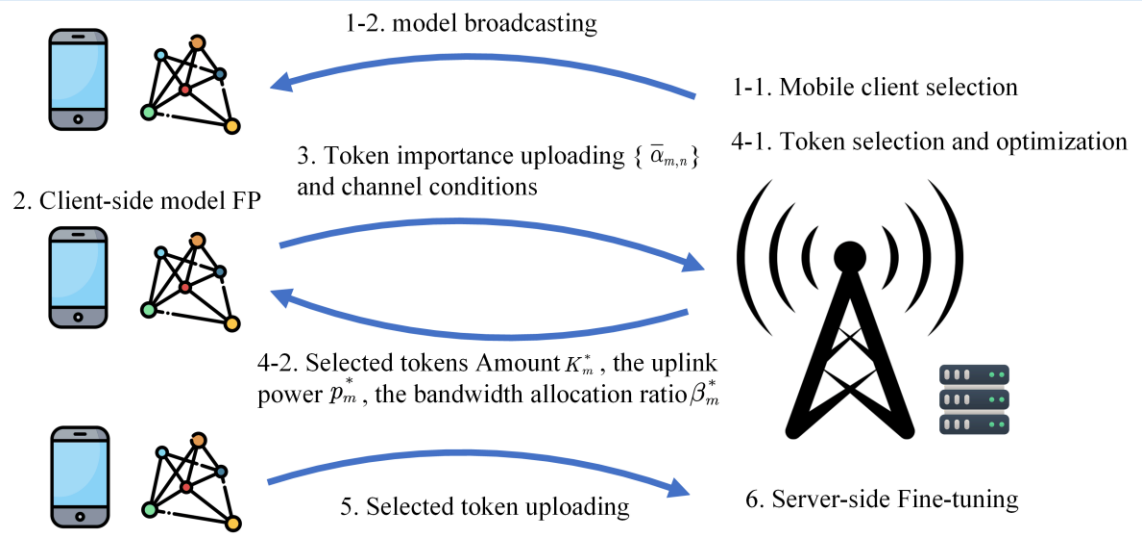
**从语义通信角度**：语义编码器与SFL中的客户端模型都是将输入token化


[1] X. Qiang, Z. Chang and G. Min, "Semantic-aware Token Selection and Resource Optimization for Communication-efficient Split Federated Fine-tuning in Edge Intelligence," submit.
[2] X. Qiang, Z. Chang, L.Wang and Y. C. Liang, "TSFLora: Token-Compressed Split Fine-Tuning forWireless Edge Networks", submit.

# 内容4： 任务驱动下的语义感知Token选择机制

**核心思想** **使用attention机制衡量token重要性进行拆分联邦微调通信压缩**

## ST-SFLora系统流程




1. 基于移动性的客户选择与模型广播
2. 客户端前向计算得到Token序列
3. 客户端上传Token重要性给服务器
4. 服务器进行token选择、资源分配优化和参数下发
5. 每个客户端上传选择后的Token序列
6. 服务器端进行模型微调

## 基于attention机制的token选择机制

基于attention机制的token评分标准

$$\mathbf{Att}_m^{(b)} = \mathrm{softmax}\left(\frac{\mathbf{Q}_m^{(b)}\mathbf{K}_m^{(b)\cdot}}{\sqrt{D}}\right), \quad \alpha_{m,n}^{(b)} = \frac{\exp\left(\mathbf{q}_{m,0}^{(b)}\cdot\mathbf{k}_{m,n}^{(b)}\right)}{\sum_{j=1}^{N}\exp\left(\mathbf{q}_{m,0}^{(b)}\cdot\mathbf{k}_{m,j}^{(b)}\right)},$$

- 选择Top-K Token

$$\mathbf{A}_m^{\mathrm{sel}} = \left\{\mathbf{a}_{m,n}^{(b)} \middle| n \in \text{Top-}K_m\left(\alpha_{m,1}^{(b)},\ldots,\alpha_{m,N}^{(b)}\right)\right\}_{b=1}^{B}.$$

- 剩下的token合并为1个token

$$\mathbf{a}_{m,\mathrm{merge}}^{(b)} = \frac{\sum_{n\in \mathrm{I}_m^{(b)}} \alpha_{m,n}^{(b)}\mathbf{a}_{m,n}^{(b)}}{\sum_{n\in \mathrm{I}_m^{(b)}} \alpha_{m,n}^{(b)}}.$$

N+1个token

↓

K+2个token

- 最终选择后传输的token序列

$$\mathbf{A}_m^{\mathrm{ref}} = \left[\mathbf{a}_{m,0}^{(b)}, \{\mathbf{a}_{m,n}^{(b)}\}_{n\in \mathrm{S}_m^{(b)}}, \mathbf{a}_{m,\mathrm{merge}}^{(b)}\right]_{b=1}^{B}$$

# 内容4： 任务驱动下的语义感知Token选择机制

**问题建立** 构建关于用户选择，带宽分配，传输功率和数据生成量的联合优化问题

## 指标：语义传输效率

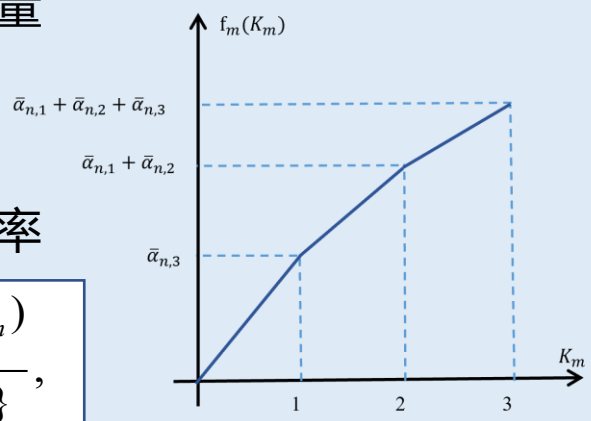
1. Token序列上传时间能耗分析

- 激活传输开销 $S_m = B\times(K_m^*+2)\times D\times q_0 \quad (\text{bits}),$
- 语义传输时间 $T_m^{\mathrm{U}} = \dfrac{S_m}{R_m^{UL}}$ $E_m^{\mathrm{U}} = p_m^* T_m^{\mathrm{U}}$

2. 语义传输效率

- 设备m的语义信息总量

$$f_m(K_m) = \sum_{n=1}^{K_m} \bar{\alpha}_{m,n}$$




- 一轮训练语义传输效率

$$\mathrm{E} = \frac{\Delta}{T} = \frac{\sum_{m\in\mathrm{U}} f_m(K_m)}{\max_{m\in\mathrm{U}}\{T_m^{\mathrm{U}}\}},$$

## 问题建立

$$\mathcal{P}0: \max_{\boldsymbol{K},\boldsymbol{W},\boldsymbol{p}} \quad \mathcal{E} = \frac{\sum_{m\in\mathcal{U}} f_m(K_m)}{\max_{m\in\mathcal{U}} T_m^{\mathrm{U}}}$$

$$\text{s.t.} \quad \mathrm{C1}: \ 0 \le p_m \le p^{\max}, \quad \forall m \subset \mathcal{U},$$

$$\mathrm{C2}: \ \sum_{m\in\mathcal{U}} W_m \le W_{\mathrm{tot}},$$

$$\mathrm{C3}: \ W_m \ge 0, \quad \forall m \subset \mathcal{U},$$

$$\mathrm{C4}: \ K^{\min} \le K_m \le N, \quad K_m \subset \mathbb{Z}, \quad \forall m \subset \mathcal{U},$$

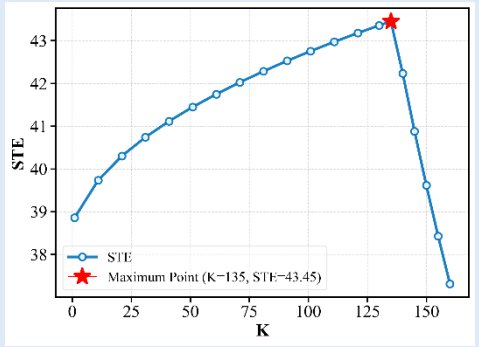
$$\mathrm{C5}: \ E_m^{\mathrm{U}}(K_m, W_m, p_m) \le E^{\max}, \quad \forall m \subset \mathcal{U},$$

$$\mathrm{C6}: \ T_m^0 + T_m^{\mathrm{U}}(K_m, W_m, p_m) \le T_m^{\text{standing}},$$

**优化目标及约束**

最大化系统语义效率

- 发射功率约束
- 带宽分配约束1
- 带宽分配约束2
- Token数量约束
- 能耗约束
- 时延约束




系统语义传输效率与token数量之间存在tradeoff

**优化变量**

· token数量 $k$
· 发射功率 $p$
· 量化位数 $q$

**问题分析**

混合整数非线性规划，且非凸，是一个NP-hard优化问题

# 内容4： 任务驱动下的语义感知Token选择机制

## 问题求解 联合迭代优化求解问题

$$\mathcal{P}0: \max_{\boldsymbol{K},\boldsymbol{W},\boldsymbol{p}} \quad \mathcal{E} = \frac{\sum_{m\in\mathcal{U}} f_m(K_m)}{\max_{m\in\mathcal{U}} T_m^{\mathrm{U}}}$$

$$\text{s.t.} \quad \text{C1}: \ 0 \le p_m \le p^{\max}, \quad \forall m \in \mathcal{U},$$

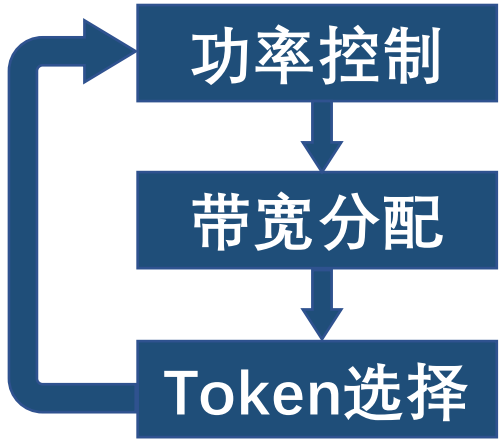
$$\text{C2}: \ \sum_{m\in\mathcal{U}} W_m \le W_{\mathrm{tot}},$$

$$\text{C3}: \ W_m \ge 0, \quad \forall m \in \mathcal{U},$$

$$\text{C4}: \ K^{\min} \le K_m \le N, \quad K_m \in \mathbb{Z}, \quad \forall m \in \mathcal{U},$$

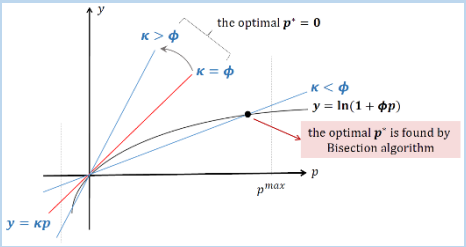
$$\text{C5}: \ E_m^{\mathrm{U}}(K_m, W_m, p_m) \le E^{\max}, \quad \forall m \in \mathcal{U},$$

$$\text{C6}: \ T_m^0 + T_m^{\mathrm{U}}(K_m, W_m, p_m) \le T_m^{\mathrm{standing}},$$




### 子问题1：功率控制

$$\mathcal{SUBP}1: \min_{p_m} \ T_m^{\mathrm{U}}(p_m) \quad \text{s.t.} \ \text{C1, C5-C6}$$

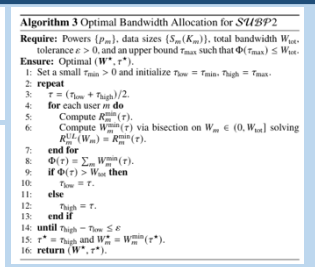



目标：在满足所有约束的前提下，把功率尽可能开到最大。SUBP1是非凸的，使用二分逼近最优解

### 子问题2：带宽分配

$$\tau \triangleq \max_{m\in\mathcal{U}} T_m^{\mathrm{U}}(K_m, W_m, p_m),$$

$$\min_{\boldsymbol{W},\tau} \ \tau \quad \text{s.t.} \ \text{C2, C3, C5, C6, C7.}$$




带宽分配子问题的本质，是把“最小化最大时延”转化成“搜索满足总带宽预算的最小时延门限”，然后用嵌套二分求解。最终得到唯一根。

### 子问题3：Token选择

$$\mathcal{SUBP}3: \max_{\boldsymbol{K}} \sum_{m\in\mathcal{U}} f_m(K_m) \quad \text{s.t.} \ \text{C4} - \text{C7}.$$

简化

$$K_m \le \frac{E^{\max} R_m^{\mathrm{UL}}}{p_m \beta_m} - 2,$$

$$K_m \le \frac{(T_m^{\mathrm{standing}} - T_0) R_m^{\mathrm{UL}}}{\beta_m} - 2,$$

$$K_m \le \frac{\tau R_m^{\mathrm{UL}}}{\beta_m} - 2.$$

闭式解

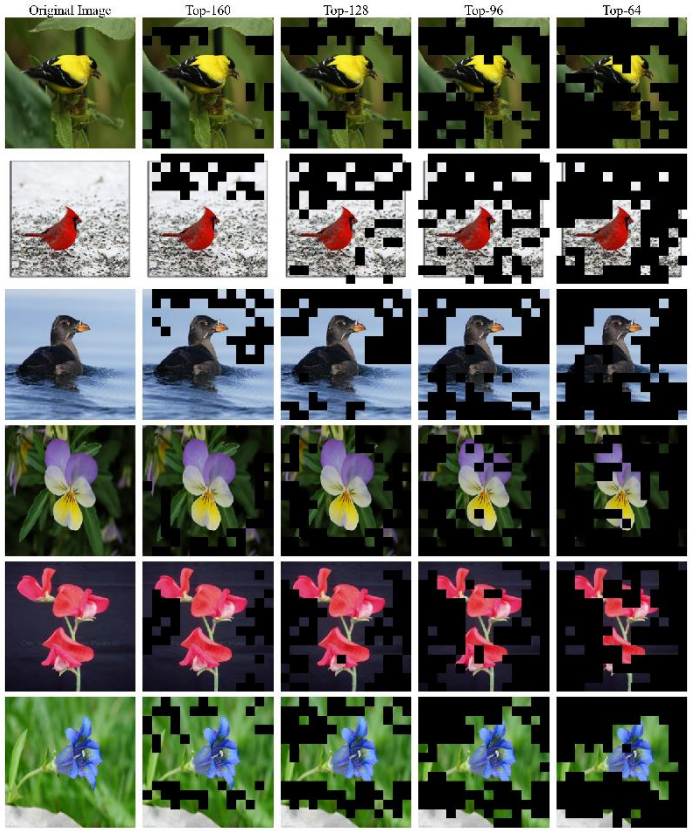
$$K_m^{\max} = \left\lfloor \min\left\{ N, \frac{E^{\max} R_m^{\mathrm{UL}}}{p_m \beta_m} - 2, \frac{(T_m^{\mathrm{standing}} - T_m^0) R_m^{\mathrm{UL}}}{\beta_m} - 2, \frac{\tau R_m^{\mathrm{UL}}}{\beta_m} - 2 \right\} \right\rfloor.$$

$$K_m^{\star} = \arg \max_{K^{\min} \le K_m \le K_m^{\max}} f_m(K_m).$$

目标：在所有约束允许的范围内，取最大的可行 token 数。

# 内容4： 任务驱动下的语义感知Token选择机制

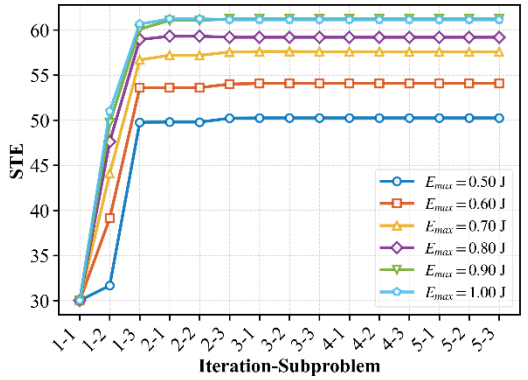
**性能评估** 优化算法收敛分析、敏感性分析、不同优化算法对比



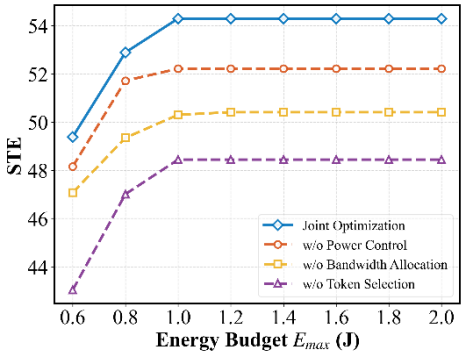

Token选择策略可视化

| Method | Computation (GB) | Communication (MB) | | |
|---|---|---|---|---|
| | GPU Mem. | Model | LoRA | Token |
| LocalLoRA | 9.0 | 335.3 | 7.9 | 0 |
| FedLoRA | 9.0 | 335.3 | 7.9 | 0 |
| SplitLoRA | 2.3 | 58.2 | 1.3 | $\frac{3}{16}N$ |
| SFLora | 2.3 | 58.2 | 1.3 | $\frac{3}{16}N$ |
| ST-SFLora-Full | 1.4 | 0 | 1.3 | $\frac{3}{16}N$ |
| ST-SFLora (top-$K$) | 1.4 | 0 | 1.3 | $\frac{3}{16}(K+2)$ |

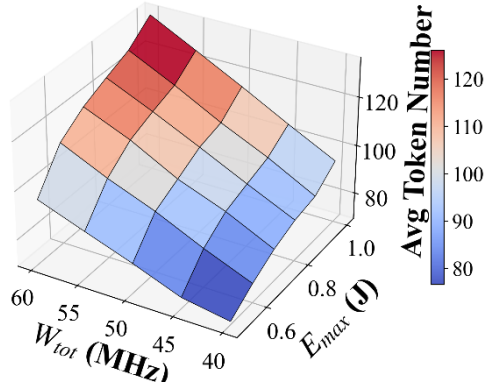
不同分布式微调策略下计算和通信开销




不同能耗约束下优化算法收敛情况




不同优化算法性能对比




不同约束下Token数变化趋势

**TS-SFLora自适应均衡模型性能与传输开销，有效提高模型训练效率**

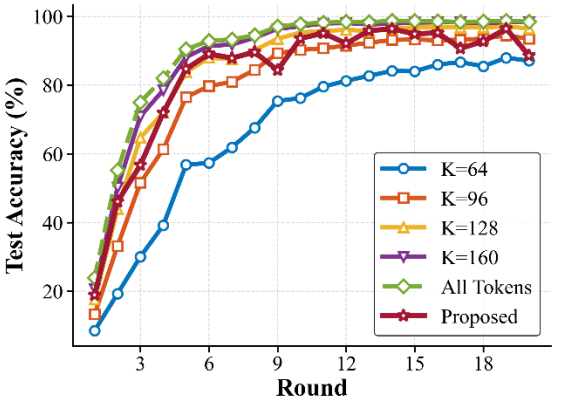
# 内容4： 任务驱动下的语义感知Token选择机制

## 性能评估 模型性能对比

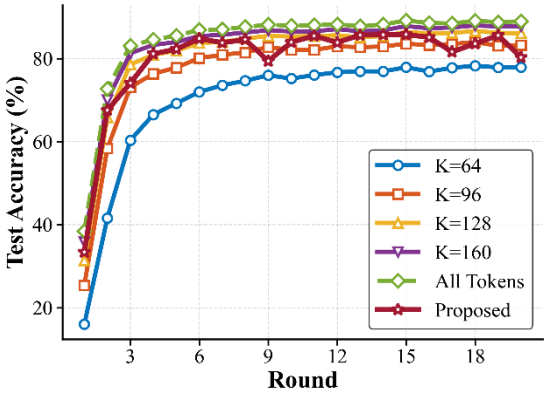
TABLE I: Top-1 Accuracy (%) on ImageNet100, Oxford Flowers-102, and CUB-200-2011.

| Backbone | Method | Datasets | | | | | |
|---|---|---|---|---|---|---|---|
| | | ImageNet100 | | Oxford Flowers-102 | | CUB-200-2011 | |
| | | *IID* | *Non-IID* | *IID* | *Non-IID* | *IID* | *Non-IID* |
| ViT-S/16 | LocalLoRA | 68.42 | 37.04 | 84.11 | 85.09 | 42.59 | 44.21 |
| | FedLoRA | 64.23 | 39.69 | 74.21 | 55.75 | 41.33 | 17.33 |
| | SplitLoRA | 85.16 | 84.34 | 98.00 | 99.61 | 82.21 | 75.91 |
| | SFLora | 84.20 | 83.24 | 98.85 | 99.19 | 81.62 | 74.49 |
| | ST-SFLora-Full | 84.36 | 83.13 | 98.83 | 99.10 | 81.39 | 74.53 |
| | ST-SFLora (Ours) | 80.77 | 79.47 | 97.26 | 97.67 | 78.21 | 68.86 |
| ViT-B/16 | LocalLoRA | 80.78 | 45.64 | 80.44 | 75.43 | 37.09 | 38.97 |
| | FedLoRA | 81.56 | 52.49 | 74.11 | 57.95 | 59.83 | 21.53 |
| | SplitLoRA | 90.08 | 89.47 | 99.25 | 99.29 | 84.79 | 80.65 |
| | SFLora | 89.55 | 89.09 | 98.97 | 98.69 | 84.55 | 79.84 |
| | ST-SFLora-Full | 89.48 | 89.14 | 99.00 | 98.74 | 84.57 | 79.91 |
| | ST-SFLora (Ours) | 87.10 | 85.81 | 96.72 | 96.43 | 80.02 | 73.69 |
| ViT-L/16 | LocalLoRA | 82.52 | 44.86 | 93.89 | 92.54 | 58.06 | 57.76 |
| | FedLoRA | 83.02 | 51.32 | 93.86 | 73.72 | 67.15 | 32.56 |
| | SplitLoRA | 90.47 | 89.79 | 99.61 | 99.73 | 87.96 | 84.92 |
| | SFLora | 90.39 | 89.50 | 99.61 | 99.73 | 87.70 | 84.39 |
| | ST-SFLora-Full | 90.49 | 89.51 | 99.61 | 99.73 | 87.75 | 84.11 |
| | ST-SFLora (Ours) | 87.20 | 85.45 | 99.00 | 99.14 | 81.36 | 75.83 |

不同数据集不同分布式微调方法下的准确度对比




Imagenet-100测试准确度




Oxford flower测试准确度

**TS-SFLora自适应均衡模型性能与传输开销，有效提高模型训练效率**

# 目 录

CONTENTS

一、研究背景与需求

二、分布式边缘智能

三、优化方案与技术

四、总结和未来研究

# 研究总结

**挑战** 分布式边缘智能面临“分不均”、“算不动”、“传不快”、“学不精”的挑战。

## 分不均

矛盾：边缘设备数据异构分布与全局模型聚合性能的冲突

↓

挑战：如何减轻联邦学习中数据异构带来的模型性能下降?

↓

解决方案：利用AIGC生成高质量图片训练增强模型性能

## 算不动

矛盾：资源受限终端的硬件能力与深度模型训练开销的冲突

↓

挑战：如何均衡拆分联邦学习中计算和通信最小化系统开销?

↓

解决方案：自适应拆分层选择均衡计算和通信最小化系统开销

## 传不快

矛盾：有限带宽环境下高维特征频繁交互与系统时效性的冲突

↓

挑战：如何均衡拆分联邦学习中bit级激活通信开销与系统性能?

↓

解决方案：自适应激活量化均衡通信开销与模型性能

## 学不精

矛盾：有限带宽环境下token选择传输与任务性能的冲突

↓

挑战：如何均衡拆分联邦学习中Token级激活压缩与任务性能?

↓

解决方案：自适应Token选择均衡通信开销与任务性能

**以数据为基、计算为核、通信为桥，是实现分布式边端协同效能跃迁的关键范式。**

# 未来方向

## LLM for Network

| | |
|---|---|
| **LLM辅助网络管理** | 1. 基于意图的配置<br>2. 自动编排 |
| **LLM辅助网络安全** | 1. 入侵检测<br>2. 威胁分析 |
| **LLM辅助网络优化** | 1. 信道估计<br>2. 资源优化 |
| **LLM辅助网络智能体** | 1. 单智能体<br>2. 多智能体 |

## Network for LLM

| | |
|---|---|
| **感知能力** | 更多模态感知辅助LLM训练/推理 |
| **传输能力** | 更加实时地分布式LLM训练/推理 |
| **智能网络** | 基于语义的通信辅助LLM训练/推理 |
| **绿色通信** | 更加绿色与可持续地边端LLM部署 |

X. Qiang, Z. Chang, J. Tang, W. Feng, C. Yang, Y. Zhang. Bridging Large Language Models and 6G Networks: Overview and Open Issues. China Communication，2026.

# 论文列表

| 主题 | 论文 | 链接 |
| --- | --- | --- |
| 数据异构 | [1] X. Qiang, Z. Chang and G. Min, "AIGC-Assisted Federated Learning for Vehicular Edge Intelligence: Vehicle Selection, Resource Allocation and Model Augmentation," IEEE Transactions on Mobile Computing, vol. 24, no. 11, pp. 11896-11909, Nov. 2025. | https://ieeexplore.ieee.org/document/11045879 |
| | [2] X. Qiang, Z. Chang and Y.-C. Liang, "AIGC-assisted Federated Learning for Edge Intelligence: Architecture Design, Research Challenges and Future Directions," submitted, arXiv preprint arXiv:2503.20166, 2025. | https://arxiv.org/abs/2503.20166 |
| | [3] X. Qiang, Z. Chang and A. Kliks, "GenFV: AIGC-Assisted Federated Learning for Vehicular Edge Intelligence," 2025 Joint European Conference on Networks and Communications & 6G Summit (EuCNC/6G Summit), Poznan, Poland, 2025. | https://ieeexplore.ieee.org/document/11037154 |
| 模型拆分 | [4] X. Qiang, Z. Chang, Y. Hu, L. Liu and T. Hämäläinen, "Adaptive and Parallel Split Federated Learning in Vehicular Edge Computing," IEEE Internet of Things Journal, vol. 12, no. 5, pp. 4591-4604, 2025. | https://ieeexplore.ieee.org/abstract/document/10714368 |
| | [5]X. Qiang, Z. Chang, C. Ye, T. Hämäläinen and G. Min, "Split Federated Learning Empowered Vehicular Edge Intelligence: Concept, Adaptive Design, and Future Directions," IEEE Wireless Communications, vol. 32, no. 4, pp. 90-97, August 2025. | https://ieeexplore.ieee.org/document/10839234 |
| 通信压缩 | [6] X. Qiang, H. Liu, X. Zhang, Z. Chang, Y. C. Liang, "Deploying Large AI Models on Resource-Limited Devices with Split Federated Learning.“ TMC, arXiv preprint arXiv:2504.09114. | https://ieeexplore.ieee.org/abstract/document/11475546 |
| 任务驱动 | [7] X. Qiang, Z. Chang and G. Min, "Semantic-aware Token Selection and Resource Optimization for Communication-efficient Split Federated Fine-tuning in Edge Intelligence." submitted. | |
| | [8] X. Qiang, Z. Chang, L. Wang and Y. C. Liang, “TSFLora: Token-Compressed Split Fine-Tuning for Wireless Edge Networks”, submitted to GC’26. | |
| 综述展望 | [9]X. Qiang, Z. Chang, J. Tang, W. Feng, Yang. C, Zhang. Y. Bridging Large Language Models and 6G Networks: Overview and Open Issues. China Communication, 2026. | https://www.preprints.org/manuscript/202603.0564 |


# 谢谢大家

常征 教授
电子科技大学 计算机（网安）学院
zheng.chang@uestc.edu.cn