基于多模态风格迁移的提示调优：实现高效联邦域泛化

Yuliang Chen[1, 2*], **Xi Lin**[1, 2*], **Jun Wu**[1,2], **Xiangrui Cai**[3†], **Qiaolun Zhang**[4], **Xichun Fan**[5], **Jiapeng Xu**[1,2], **Xiu Su**[6†]

[1]上海交通大学计算机科学学院 [2]上海市信息安全综合管理技术重点实验室[3]南开大学计算机科学学院[4]米兰理工大学电子、信息与生物工程系[5]上海纽约大学 [6]中南大学大数据研究院{chenyuliang, linxi234, junwuhn, xjp20021018}@sjtu.edu.cn, caixr@nankai.edu.cn, qiaolun.zhang@mail.polimi.it, xf731@nyu.edu, xiusu1994@csu.edu.cn

arXiv:2601.05955v1 [cs.DC] 9 Jan 2026

摘要

联邦域泛化（FDG）旨在通过分布式客户端协同训练全局模型，使其能够在未见域上实现良好的泛化。然而，现有联邦域泛化方法通常难以应对客户端间的数据异质性，并带来显著的通信和计算开销。为应对这些挑战，本文提出一种新的联邦域泛化框架 FaST-PT，可在分布式环境下实现本地特征增强和高效的未见域适应。首先，我们提出一种轻量级多模态风格迁移（MST）方法，在文本监督下转换图像嵌入，从而扩展训练数据分布并缓解域偏移。随后，我们设计了一个双提示模块，将提示分解为全局提示和域提示。具体而言，全局提示从各客户端增强后的嵌入中捕获通用知识，而域提示则从本地数据中捕获域特定知识。此外，我们引入域感知提示生成（DPG），为每个样本自适应地生成合适的提示，通过知识融合促进未见域适应。在四个跨域基准数据集（如 PACS 和 DomainNet）上的大量实验表明，FaST-PT 的性能优于 FedDG-GA 和 DiPrompt 等最先进的联邦域泛化方法。消融研究进一步验证了 FaST-PT 的有效性和效率。

引言

联邦学习（FL）（McMahan 等人，2017）是一种广泛采用的分布式机器学习范式，可在保护隐私的同时促进协同模型训练。与依赖单个集中式服务器不同，联邦学习将训练过程分散到多个客户端（例如移动设备或机构），让它们在不共享私有数据的情况下协同训练全局模型。它尤其适用于敏感数据分布在多个来源的现实世界应用（Guan 等人，2024年）。

尽管联邦学习（FL）已在多种任务中展现出良好的性能，但现有研究主要关注提升参与客户端的模型性能。然而，这些研究往往忽视了模型对参与客户端之外未见域的泛化能力。由于数据源、采集方法或特定域特征存在差异，在这些未见环境中遇到的数据可能与训练客户端的数据存在显著的分布差异。如何使联邦学习有效适应此类域偏移（Muandet、Balduzzi 和 Sch¨olkopf，2013年；Li 等，2018年），仍然是一个技术上颇具挑战的问题。

联邦域泛化被定义为在多个分布式源域上训练全局模型，旨在实现对完全未见目标域的稳健泛化。已有一些早期研究尝试解决联邦域泛化问题。例如，ELCFS（Liu 等人，2021年）允许客户端共享频域中的振幅谱，以用于联邦医学图像分割。类似地，CCST（Chen 等，2023）提取并交换风格信息，以执行风格迁移。此外，FPL（Huang 等，2023）促进了客户端之间原型的交换。这些方法的一个共同局限是，它们依赖于客户端之间共享与域相关的信息，这会损害联邦学习所固有的数据隐私并加剧通信约束。此外，提取特定域的信息通常需要辅助模块，从而增加计算开销。

为应对上述挑战，我们提出了一种名为“带提示调优的联邦多模态风格迁移”（FaST-PT）的新框架。在 FaST-PT 中，我们将 CLIP 引入 FL，并利用其文本-图像对齐能力和泛化能力来解决联邦域泛化问题。一方面，我们利用 CLIP 强大的图像-文本对齐能力，提出了多模态风格迁移（MST），使客户端侧外部领域风格迁移能够扩展训练数据分布。具体而言，MST 引入了一个

*这些作者贡献相同。

†通讯作者 版权所有 © 2026年，美国人工智能促进协会（www.aaai.org）。保留所有权利。

变换网络，该网络利用外部域的文本表示变换图像嵌入，从而实现风格迁移，同时保留原始图像中的类别特定信息。另一方面，我们采用 CLIP 的提示调优范式，以促进对未见域的高效适应。我们首先设计了一个双提示模块，其中包含全局提示和域提示。具体而言，全局提示在增强嵌入上训练，并在客户端之间聚合，以捕获通用知识；而本地提示仅利用本地数据训练，以保留域特定知识。此外，我们提出了一种域感知提示生成机制，基于域相似度进一步融入域知识，从而增强提示对未见域的适应能力。我们的贡献总结如下：

- 我们提出了 FaST-PT，这是一种新的联邦学习框架，通过客户端侧的分布对齐以及高效的未见域适应来解决联邦域泛化问题，并利用预训练 CLIP 模型的文本-图像对齐与泛化能力。
- 我们提出了一种多模态风格迁移方法，使客户端能够将图像嵌入转换为外部域的风格，同时在文本描述的监督下保留类别信息，从而缓解客户端之间的分布差异。
- 我们设计了一种双提示模块，通过提示划分简化传统域泛化中的特征解耦，并采用域感知提示生成来整合域知识，从而实现高效的未见域适应。

- 我们在四个基准数据集上使用 FaST-PT 进行了大量实验，结果一致表明其性能优于基线方法。全面的消融研究进一步验证了 FaST-PT 的有效性和效率。

相关工作

联邦学习

联邦学习（FL）将训练过程分散到多个设备或计算节点上，而非集中于单一中央服务器，从而在保护数据隐私的同时实现模型的全局训练。最典型的联邦学习框架是 FedAvg（McMahan 等人，2017；Koneˇcnˇy 等人，2016），它从本地训练和全局平均两个方面对分布式学习过程进行分解。然而，传统联邦学习在域偏移情况下表现不佳。现有方法主要通过对客户端侧的本地训练施加正则化，以获得性能良好的全局模型。Fed-Prox（Li 等人，2020）在损失函数中添加了一个额外项，用于约束模型权重之间的 L_2 距离。FedBN（Li 等人，2021）将批量归一化层的参数保留在本地，并将其排除在聚合过程之外。

联邦域泛化

联邦域泛化是针对联邦学习场景设计的域泛化变体，旨在缓解客户端之间的领域偏移，并提升模型对未见域的泛化能力（Liu 等人，2021；Nguyen、Torr 和 Lim，2022；Wu 和 Gong，2021；Chen 等，2023；Zhang 等人，2021a；Huang 等，2023；Le 等，2024）。ELCFS（Liu 等人，2021）是首个解决联邦域泛化问题的方法，它在频域中交换振幅谱，并采用情景式学习进一步提升泛化性能。CCST（Chen 等，2023）允许客户端共享风格信息并进行本地风格迁移，从而使源客户端的分布更加一致。除共享域信息的方法外，其余方法侧重于提取域不变特征。例如，FedSR（Nguyen、Torr 和 Lim，2022）利用 L2 范数和条件互信息正则化，学习简单且具有良好泛化能力的数据表示。COPA（Wu 和 Gong，2021）则同时采用域不变表示提取器和由多个域特定分类器组成的集成模型。

提示调优

提示调优（Lester、Al-Rfou 和 Constant，2021）是一种参数高效微调技术，它将少量可学习的提示词元作为额外输入，同时冻结基础模型的参数。作为开创性工作，CoOp（Zhou 等人，2022b）通过将文本编码器中的输入上下文词元转换为可学习提示，将 CLIP（Radford 等人，2021）适配到下游图像识别任务。CoCoOp（Zhou 等人，2022a）通过轻量级网络生成输入依赖提示，从而扩展了 CoOp。MaPLe（Khattak 等人，2023）进一步通过引入视觉提示，与文本提示相结合，增强了 CLIP 中文本和视觉表征之间的对齐。近年来，鉴于提示调优天然适合高效通信，它已被引入联邦学习。例如，PromptFL（Guo 等人，2023）和 FedPrompt（Zhao 等人，2023）分别在每个本地客户端上学习文本提示，并在服务器上聚合优化后的提示。

方法

我们假设 K 个客户端，其中每个客户端 i 拥有来自域 $\mathcal{D}_i$的数据集。主要目标是联合训练一个全局提示模块，使其能够有效泛化到未见过的目标域 $\mathcal{D}_u$。我们框架中的一个关键假设是，每个客户端都可以访问外部训练领域 t_i 的简单文本描述（例如，“一张……的剪贴画”），这些描述被视为非敏感信息，不会损害客户端隐私。FaST-PT的框架如图1所示。

多模态风格迁移

现有的联邦域泛化方法通过融入来自多样领域的知识来提升本地模型的泛化能力，但这需要在客户端之间共享部分域知识。虽然在客户端之间共享此类敏感信息可能带来数据泄露风险，我们提出了一种轻量且安全的本地风格迁移方案 MST。该方案利用 CLIP 丰富的域知识和强大的文本-图像对齐能力，促进在嵌入层面进行潜在特征增强，

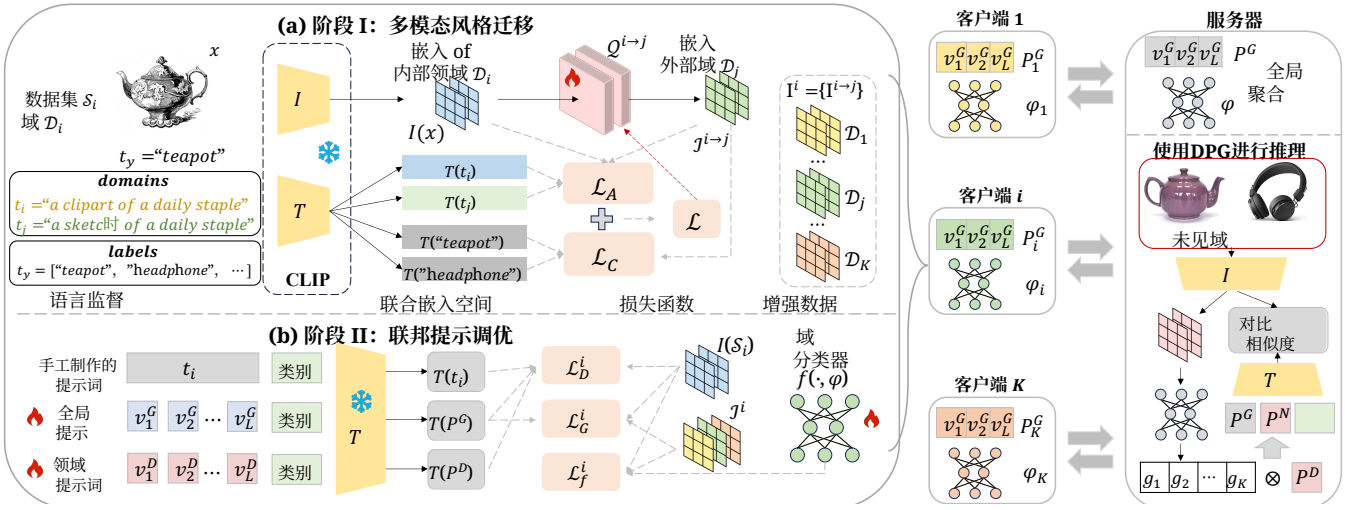


图1：我们提出的 FaST-PT 框架。左侧展示本地训练过程，右侧展示客户端与服务器之间的通信。客户端训练分为两个阶段：(a) MST 和 (b) FPT。(a) 在 MST 阶段，客户端使用冻结的 CLIP 模型以及来自不同领域的文本描述来训练变换网络，从而在嵌入层面实现风格迁移。(b) 在 FPT 阶段，客户端使用增强数据训练全局提示、域提示和域分类器。服务器端也展示了推理过程。

使客户端 k 能够根据其域 $\mathcal{D}_k$ 生成外部域 $\mathcal{D}_{\{1,2\ldots K\}\backslash k}$ 的图像嵌入。

MST 主要学习一个变换网络 $\mathcal{Q}^{i\to j}$，将输入图像的图像嵌入从 $\mathcal{D}_i$ 变换为$\mathcal{D}_j$，其目标由以下两个方面构成：

- $\mathcal{D}_i$ 与 $\mathcal{D}_j$之间的对齐
- 类别一致性特征的保留

令 x 表示输入图像，$I(x)$ 表示 x的图像嵌入。此外，令 t_k 表示域 $\mathcal{D}_k$ 的领域相关描述，t_y 表示类别标签。我们将这两部分组合，得到文本输入$t=[t_k,t_y]$，再对其进行编码以获得文本嵌入 $T(t)$。例如，若 t_i = 为"a clipart of a {}"，t_j = 为"a sketch of a {}"，且 t_y 可以是"老虎"，则组合后的文本 t =为"a clipart of a tiger"。

为实现从 $\mathcal{D}_i$ 到 $\mathcal{D}_j$ 的嵌入变换，同时保留类别一致性特征，我们设计了两个损失函数，用于训练每个不同于本地域 $\mathcal{D}_i$的域 $\mathcal{D}_j$ 对应的 $\mathcal{Q}^{i\to j}$：对齐损失和一致性损失。

对齐损失 对齐损失 $\mathcal{L}_{\mathrm{A}}(\mathcal{Q}^{i\to j})$旨在确保增强图像的嵌入能够更忠实地反映各自领域的显著属性。$\mathcal{L}_{\mathrm{A}}(\mathcal{Q}^{i\to j})$由与不同源域相关联的文本描述提供监督。尽管 CLIP 能有效捕捉文本嵌入与图像嵌入之间的关系，但要直接获得二者的精确对应关系仍然颇具挑战。为此，我们将图像和文本嵌入空间中的变化方向作为优化目标进行对齐。具体而言，该变化方向定义为 $\mathcal{D}_i$与 $\mathcal{D}_j$中的嵌入之间的归一化差值。因此，对齐损失可表示为：

$$\mathcal{L}_{\mathrm{A}}(\mathcal{Q}^{i\to j})=\sum_{(x,y)\in\mathcal{S}_i}\left(1-\frac{\mathcal{Q}^{i\to j}(I(x))-I(x)}{\|\mathcal{Q}^{i\to j}(I(x))-I(x)\|}\cdot\frac{T(t_j,t_y)-T(t_i,t_y)}{\|T(t_j,t_y)-T(t_i,t_y)\|}\right). \tag{1}$$

一致性损失 一致性损失 $\mathcal{L}_{\mathrm{C}}$ 旨在保留变换后嵌入中的类别知识。设计 $\mathcal{L}_{\mathrm{C}}$ 的直觉在于，仅使用对齐损失或许能够确保嵌入的风格迁移，但也可能抹去部分类别信息。我们利用 CLIP 的零样本推理能力来评估增强后的图像嵌入所包含的类别信息。具体而言，我们通过将这些增强后的嵌入与类别名称的文本嵌入进行比较来执行分类。为此，我们采用 CLIP 的标准对比损失，其形式如下：

$$\mathcal{L}_{\mathrm{C}}(\mathcal{Q}^{i\to j})=\mathbb{E}_{(x,y)\in\mathcal{S}_i}\ell_{ce}([\mathcal{Q}^{i\to j}(I(x))\cdot T(t_y)],y), \tag{2}$$

其中，$[I\cdot T]$ 表示基于 Softmax 的对比相似度计算和 Top-1预测过程。

训练变换网络 $\mathcal{Q}^{i\to j}$ 的一般目标函数 $\mathcal{L}_{\mathrm{MST}}(\mathcal{Q}^{i\to j})$由上述两个损失函数组合而成。为平衡对齐与一致性的效果，我们还引入了一个超参数 λ，以实现线性组合：

$$\mathcal{L}_{\mathrm{MST}}(\mathcal{Q}^{i\to j})=\lambda\mathcal{L}_{\mathrm{A}}(\mathcal{Q}^{i\to j})+(1-\lambda)\mathcal{L}_{\mathrm{C}}(\mathcal{Q}^{i\to j}). \tag{3}$$

具有域 $\mathcal{D}_i$ 的客户端 i 在本地为其他 $K-1$ 个客户端学习 $(K-1)$ 变换网络，并生成相应的增强嵌入 $\mathcal{I}^{i\to j}=\mathcal{Q}^{i\to j}(x)\quad x\in\mathcal{S}_i$。因此，每个客户端维护一组增强嵌入 $\mathcal{I}_i=\{\mathcal{I}^{i\to j}\},\ j\in\{1,2\ldots K\}\backslash i$。

联邦提示调优

然而，传统的领域泛化方法通常涉及复杂的特征提取技术。如何利用增强嵌入，在尽可能少消耗资源的情况下实现域泛化，仍然是一个悬而未决的挑战。基于 CLIP 强大的特征提取能力，我们引入了 CLIP 的提示调优，以实现高效的域泛化。受传统领域泛化中关于特征解耦的思想（Bui 等，2021年；Zhang 等，2022）的启发，我们提出了一个双提示模块，将提示划分为全局提示和领域提示。

全局提示P^G 旨在捕获下游任务的通用知识，并应在各客户端之间共享。 P^G 是一组 L 可学习向量 $P^G = \{v_1^G, v_2^G, \ldots, v_L^G\}$，其维度与文本嵌入 d相同。我们将类别标签记为 t_y。对于第 j类，我们将带有全局提示的文本编码器输入定义为$t = [P^G, t_y^j]$。我们将图像 x 的特征图记为 $I(x)$，文本特征图记为 $T([P^G, t_y^j])$。因此，利用对比相似度计算的第 j类预测概率可表示为：

$$p_G(y=j|x) = \frac{exp(\mathcal{H}(T([P^G, t_y^j]), I(x))/\tau)}{\sum_{i=1}^{C} exp(\mathcal{H}(T([P^G, t_y^i]), I(x))/\tau)}, \quad (4)$$

其中， $\mathcal{H}$ 表示余弦相似度， C 表示类别数量， τ 表示Softmax函数的温度。我们通过最小化增强嵌入和原始数据集上的交叉熵损失来优化全局提示。具体而言，客户端 i 上的训练损失定义如下：

$$\mathcal{L}_{\mathrm{G}}^i = \mathbb{E}_{(x,y)\in\mathcal{S}_i\cup\mathcal{I}_i}[\ell_{ce}(y, p_G(y|x))]. \quad (5)$$

领域提示词P_i^D旨在捕获客户端 i的领域特定知识。同样， P_i^D是一组向量 $P_i^D = \{v_1^D, v_2^D, \ldots, v_L^D\}$，其长度为 L、维度为 d。我们将 P^G 与领域提示词拼接，形成新的复合提示 $P^C = [P^G, P_i^D]$。具体而言，引入领域提示词后，通过用 P^C替换 P^G，相应地更新预测概率。领域特定提示词的核心思想是提升 P_i^D在本地数据 $\mathcal{S}_i$ 上的性能。我们同样使用交叉熵损失 $\mathcal{L}_{\mathrm{Cla}}^i = \mathbb{E}_{(x,y)\in\mathcal{S}_i}[\ell_{ce}(y, p_C(y|x))]$对其进行优化。此外，$P_i^D$的训练还应与 P^G 保持正交，以确保提取出的知识具有唯一性。为此，我们还引入如下对比损失项：

$$\mathcal{L}_{\mathrm{D}}^i = \mathcal{L}_{\mathrm{Cla}}^i + \mathcal{L}_{\mathrm{Con}}^i = \mathbb{E}_{(x,y)\in\mathcal{S}_i}[\ell_{ce}(y, p_C(y|x))] - \log\frac{exp(sim(P_i^D, t_i))}{exp(sim(P_i^D, t_i)) + exp(sim(P_i^D, P^G))}, \quad (6)$$

其中， t_i 是域 $\mathcal{D}_i$ 的手工制作的提示词，$p_C(y|x)$ 是使用提示词 P^C预测的概率。

值得注意的是，训练期间客户端共享全局提示以进行聚合，而领域特定提示则保留在本地，仅在训练结束后共享。最终的领域提示列表构造为 $P^D = [P_1^D, P_2^D, \ldots, P_K^D]$。

算法 1：FaST-PT

输入： 数据集 $\mathcal{S}_i$、客户端数量 K、训练轮次 R，训练周期数 E_G 和 E_L
输出： 全局提示 P^G、域提示列表 P^D，域分类器 $f(\cdot, \varphi)$
阶段 1：多模态风格迁移
对于 客户端 $i = 1, 2, \ldots, K$ **执行**
  对于 域 $\mathcal{D}_j \in \{\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_K\} \backslash \mathcal{D}_i$ **执行**
    使用 $\mathcal{L}_{\mathrm{MST}}(\mathcal{Q}^{i\to j})$训练 $\mathcal{Q}^{i\to j}$ （公式 3）
  结束循环
结束循环
客户端在本地增强图像嵌入 $\mathcal{I}^{i\to j}$
阶段 2：联邦提示调优
对于轮次 $r = 1, 2, \ldots, R$ **do**
  对于客户端 $i = 1, 2, \ldots, K$ **执行**
    对于训练周期 $e = 1, 2, \ldots, E_G$ **执行**
      使用 $\mathcal{L}_{\mathrm{G}}^i$ 和 $\mathcal{L}_{\mathrm{F}}^i$训练 P_i^G和 φ_i
    结束循环
  结束循环
  上传 P_i^G和 φ_i
  服务器聚合 P^G 和 φ 并广播
  针对客户端 $i = 1, 2, \ldots, K$ **do**
    对于训练周期 $e = 1, 2, \ldots, E_D$ **执行**
      使用 $\mathcal{L}_{\mathrm{D}}^i$训练 P_i^D
    结束循环
  结束循环
结束循环
服务器收集 P^D 并进行广播

域感知提示生成 考虑到训练域与未见域之间的潜在相似性，我们进一步设计了一种域感知提示生成方案，整合来自不同领域的知识，以提升泛化性能，从而更好地适应目标域 $\mathcal{D}_u$。我们使用域分类器 $f(\cdot, \varphi)$ 来促进域相似度分析。 $f(\cdot, \varphi)$ 将图像特征作为输入，并输出具有 K 维的独热格式结果：$\mathbf{G} = [g_1, g_2 \ldots g_K]$。

我们将领域提示列表 P^D 与伪领域标签 $\mathbf{G}$ 逐元素加权，以生成适用于任意图像 x 的动态提示 P^N，如下所示：

$$P^N = f(I(x), \varphi) \otimes P^D. \quad (7)$$

P^N 整合多个源域的知识，生成近似目标域的知识。在推理阶段，将 P^N与 P^G拼接后输入文本编码器。对于 $f(\cdot, \varphi)$ 的优化，我们利用增强数据集 $\mathcal{S}_i \cup \mathcal{I}_i$ 进行训练，并将域索引记作标签。我们采用交叉熵损失，目标函数如下：

$$\mathcal{L}_{\mathrm{F}}^i = \mathbb{E}_{(x,y,j)\in\mathcal{S}_i\cup\mathcal{I}_i}[\ell_{ce}(j, f(x, \varphi_i))], \quad (8)$$

其中， j 为域索引。

在提示模块的训练过程中，我们采用两阶段交替更新策略。首先，客户端更新全局参数 P^G 和 φ，并上传梯度

方法	PACS（%）					VLCS（%）				
	A	C	P	S	Avg	C	L	S	V	Avg
FedAvg	81.27	76.78	92.33	63.54	78.48	92.78	62.39	73.02	74.98	75.79
FedProx	80.22	77.17	96.08	64.28	79.44	93.22	62.80	74.64	74.96	76.41
ELCFS	85.89	74.02	92.14	69.16	80.30	93.28	61.64	73.42	76.06	76.10
FedSR	84.97	85.13	91.37	67.13	82.15	95.55	63.29	71.93	75.43	76.55
FedADG	80.16	81.92	91.28	68.43	80.45	95.96	65.63	70.29	76.70	77.15
FedDG-GA	94.03	83.19	76.85	82.93	84.25	95.14	62.82	74.41	77.49	77.47
PromptFL	93.19	94.72	99.10	83.76	92.70	99.26	66.08	79.38	78.56	80.82
FedAPT	93.78	95.52	99.63	85.31	93.56	99.01	67.91	78.66	81.53	81.78
DiPrompt	95.72	96.39	99.24	88.07	94.86	**99.70**	68.97	84.72	82.93	84.08
FaST-PT	**97.32**	**99.31**	**99.93**	**96.09**	**98.16**	99.17	**70.59**	**86.05**	**86.26**	**85.52**

表1：不同方法在 PACS 数据集和 VLCS 数据集上的性能比较。

更新结果，随后由服务器聚合并广播回所有客户端。其次，客户端更新本地参数 P_i^D，并在最后一轮后上传这些参数。由于 P^G 和 $f(\cdot,\ \varphi)$ 的训练不会相互干扰，因此同时训练它们是合理的。完整的训练过程如算法 1 所示。

讨论

FaST-PT 能够实现高效的局部特征增强和稳健的域泛化。具体而言，尽管 FaST-PT 利用 MST 缓解域偏移，并采用提示调优实现域泛化，但这两个组件均基于 CLIP 模型构建。这种统一的骨干网络设计无需在本地客户端部署额外模型，从而降低了内存消耗和计算开销。此外，不同于需要共享完整模型参数的传统联邦学习方法，FaST-PT 仅在客户端之间交换提示及其关联的轻量级生成模块，大幅降低了通信成本（我们的实验结果也进一步证实了这一点）。最后，与传输域特定特征的方法不同，MST 仅要求客户端共享各自域的文本描述——这类信息本身敏感性较低，因此能够提供更强的隐私保护。

实验

实验设置

数据集 我们在四个跨域数据集上进行了大量实验：**PACS**（Li 等人，2017）包含四个领域：照片（P）、艺术绘画 （A）、卡通 （C）和素描 （S）。**VLCS** （Fang、Xu 和 Rockmore，2013）是域泛化中广泛使用的基准，包含四个领域：*VOC2007* （V）、*LabelMe* （L）、*Caltech101*（C）和*SUN09* （S）。**DomainNet** （Peng 等人，2019）是一个涵盖六个领域的大规模数据集：剪贴画 （C）、信息图（I）、绘画 （P）、*QuickDraw* （Q）、真实图像 （R）和素描 （S）。**Office-Home** （Venkateswara 等人，2017）是一个中等规模的基准数据集，包含四个领域：艺术 （A）、剪贴画 （C）、产品 （P）和真实世界 （R）。遵循（Li 等人，2021；Yang、Wang 和 Wang，2023），我们选取 DomainNet 和 Office Home 中排名前 10 的类别进行实验。我们采用留一域协议，将一个领域作为目标域，其余领域作为源域。

基线方法 我们将所提方法与三类最先进的方法进行比较：1）传统的联邦学习框架，包括**FedAvg**（McMahan 等人，2017）和**FedProx**（Li 等人，2020），使客户端能够从头开始协同训练全局模型。2）将域泛化与联邦学习相结合，以协同训练能在未见领域上表现良好的泛化模型的传统 FedDG 方法，包括**ELCFS** （Liu 等人，2021）、**FedSR**（Nguyen、Torr 和 Lim，2022）、**FedADG** （Zhang 等人，2021b）和**FedDG-GA** （Zhang 等人，2021a）。3）将提示调优引入联邦学习，以增强 CLIP 泛化能力的基于提示调优的方法，包括**PromptFL** （Guo 等人，2023）、**FedAPT**（Su 等人，2024）和**DiPrompt**（Bai 等人，2024）。

实现细节 我们使用 Python 3.11 在 NVIDIA RTX A40 GPU 上进行实验。我们使用 OpenAI CLIP 模型，其骨干网络为 ViT-L/14，可将图像和文本编码为 768 维特征向量。在 MST 阶段，我们使用一个两层多层感知机作为变换网络 $\mathcal{Q}^{i\to j}$，隐藏维度为 384。我们使用 Adam优化器，学习率为 $1e^{-3}$，权重衰减为 0.05。损失权重 λ 设置为 0。5。在提示训练阶段，我们使用一个全连接层作为域分类器 $f(\cdot,\ \varphi)$。我们使用 SGD 优化器，提示和域分类器训练的学习率分别为 0。005 和 0。01。通信轮次 R 设为 20，本地训练轮次$E_G = E_L = 5$。我们将批量大小设为 16，提示长度设为 4。为确保公平比较，所有实验均使用相同的超参数和骨干网络。

评估

表1和表2展示了FaST-PT及其他方法的实验结果，涵盖PACS、VLCS、OfficeHome和DomainNet上的18项域泛化任务。所有实验结果均为使用不同随机种子运行三次后的平均值。总体而言，我们观察到，尽管传统联邦域泛化方法的泛化性能明显优于传统联邦学习方法，

方法	OfficeHome (%)					DomainNet (%)						
	A	C	P	R	Avg	C	I	P	Q	R	S	Avg
FedAvg	82.56	79.20	88.28	91.15	85.30	78.04	62.69	80.13	45.78	82.49	84.17	72.22
FedProx	83.19	77.91	90.28	91.60	85.74	79.11	64.77	82.82	51.65	84.65	83.94	74.49
ELCFS	88.72	82.64	89.37	91.75	88.12	83.28	67.18	85.78	58.65	88.31	86.37	78.26
FedSR	88.17	84.47	88.89	90.65	88.04	85.17	66.57	86.01	60.95	87.26	90.10	79.34
FedADG	86.51	87.25	91.52	92.30	89.40	83.95	69.68	84.05	61.51	90.78	86.79	79.46
FedDG-GA	87.86	86.90	92.42	93.25	90.11	86.69	70.93	85.69	60.72	89.82	87.18	80.17
PromptFL	92.35	88.73	95.20	98.80	93.77	89.92	72.75	87.46	60.85	93.54	90.47	82.50
FedAPT	94.28	90.76	97.08	98.95	95.27	91.44	72.04	90.62	63.88	95.27	94.23	84.58
DiPrompt	95.48	91.26	97.90	99.10	95.94	93.89	75.34	93.72	64.16	96.31	97.62	86.84
FaST-PT	**98.57**	**92.72**	**99.24**	**99.40**	**97.48**	**97.48**	**78.14**	**97.17**	**66.88**	**97.90**	**98.46**	**89.17**

表2：不同方法在 OfficeHome 数据集和 DomainNet 数据集上的性能比较。

但与近期基于提示的联邦学习方法相比，性能仍存在显著差距。泛化能力的大幅提升主要归功于CLIP强大的跨模态能力和泛化能力。即使所有方法都基于相同的预训练视觉骨干网络（ViT-L/14），提示调优也能更高效地适应各个域的分布特征。具体而言，FaST-PT取得了最先进的性能，在PACS、VLCS、OfficeHome和DomainNet上的平均准确率分别达到98.16%、85.52%、97.48%和89.17%；在PACS上比最先进方法DiPrompt高出 +3.30%，在DomainNet上高出 +2.33%。在所有基准数据集的18项域泛化任务中，FaST-PT在17项任务上取得最佳性能，表现出卓越的一致性。与现有基于提示的联邦学习方法相比，FaST-PT仍展现出显著的性能优势，凸显了所提出的MST和DAP策略的有效性；这两者对于增强跨域适应性、减轻联邦学习中分布偏移带来的不利影响都起着关键作用。

消融实验

组件的有效性为探究 FaST-PT 中不同组件的有效性，我们在 PACS 数据集上评估了 FaST-PT 的各种变体，如图所示

见表3。在变体 1 和2 中，我们仅采用全局提示或与 DPG 机制结合的域提示。与变体 4 相比，性能明显下降，这证明了融合全局知识与域知识对泛化任务的有效性。此外，通过两组对比分析（1 与3，以及4 与 FaST-PT），我们发现，引入 MST 策略能够缓解联邦域泛化中的域偏移问题，并有效提升模型性能，从而增强全局模型的泛化能力。与 FaST-PT 相比，变体 5 的性能有所下降，这也反映出我们采用的对比损失能够强化全局知识与域知识之间的区分，进而间接增强模型对目标域的适应能力。在变体 6 中，我们通过补充目标域的文本描述开展实验，使客户端能够生成目标域的图像特征，以辅助全局模型的训练。

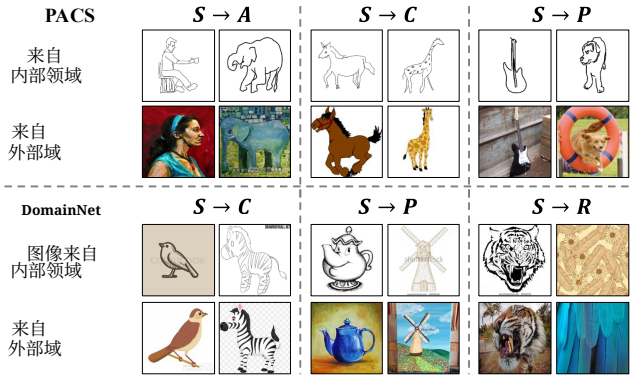


图2：PACS 和 DomainNet 上的最近邻。

其性能仅比 FaST-PT 高出 0.24%。

MST 的可视化 我们评估了 MST 策略的风格迁移效果，结果如图2所示。由于现有方法不支持将图像特征转换回图像，我们采用最近邻（NN）方法评估生成的增强特征。具体而言，我们使用扩展域中的所有数据作为参考集。对于每个增强嵌入，我们计算其与参考集中各嵌入的余弦相似度，并选择相似度最高的嵌入作为其最近邻。在图2中，顶行展示了用于风格迁移的原始图像，底行展示了与增强特征对应的最近邻。我们观察到，尽管顶行与底行图像的风格存在明显差异，但它们始终属于同一语义类别。这一观察结果与训练期间采用的双部分损失函数的目标一致，并证明了 MST 策略的有效性。

提示长度的影响 我们分析了提示长度的影响，并在图3中展示了结果。这里，我们将全局提示和领域提示的长度设为相同。在 PACS 和 DomainNet 数据集上，我们观察到，在大多数任务中，随着提示长度增加，模型性能呈现先上升后下降的趋势。提示过短会限制性能，因为提示所传达的信息有限。相反，较长的

变体	P^G	P^D	$\mathcal{L}_{\text{Con}}$	DPG	MST	t_{target}	A	C	P	S	平均值
①	✓						91.21	90.94	94.54	93.61	92.58
②		✓		✓			85.48	83.16	90.79	91.25	87.67
③	✓				✓		92.43	94.14	95.81	95.13	94.38
④	✓	✓	✓	✓			96.03	97.44	98.98	95.77	97.06
⑤	✓	✓		✓	✓		97.74	99.09	99.48	95.48	97.95
⑥	✓	✓	✓	✓	✓	✓	97.93	99.68	99.85	96.27	98.42
FaST-PT	✓	✓	✓	✓	✓		97.32	99.31	99.93	96.09	98.16

表3：Fast-PT 不同组件在 PACS 上的效果。

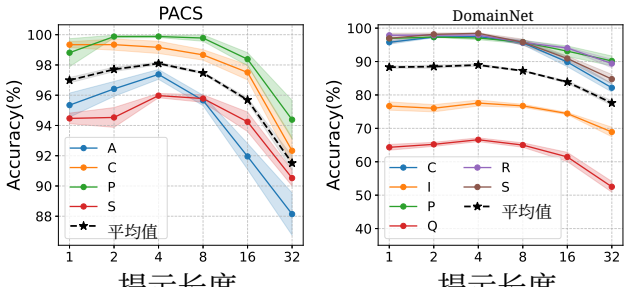


图3：在不同随机种子下运行三次，提示长度对 PACS 和 DomainNet 数据集的影响。

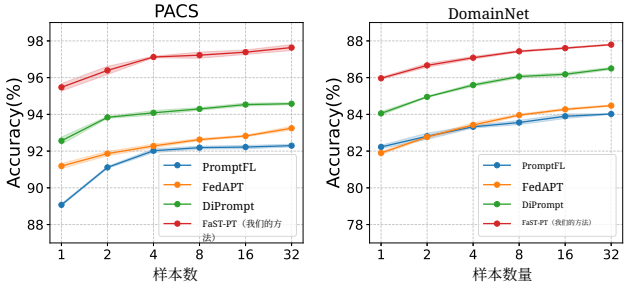


图 4：在 PACS 和 DomainNet 数据集上，使用不同随机种子运行三次后的小样本性能。

提示可能会引入噪声或与任务无关的信息，从而对图像和文本之间的对齐产生负面影响。此外，使用较长提示时，FaST-PT采用的提示加权策略可能会导致语义稀释、可解释性降低等问题。我们观察到，当提示长度设为 4 时，大多数任务以及平均准确率均达到最佳表现。因此，我们在实验中采用 4 作为默认提示长度。

少样本场景下的结果 为考察我们的框架在少样本设置下的有效性，我们在数据极其有限的不同条件下（shots = 1、2、4、8、16、32），将 FaST-PT 的泛化能力与现有的基于提示的方法进行比较。如图4所示，随着训练样本数量增加，所有方法的性能均有所提升。在 PACS 和 DomainNet 数据集的所有设置中，FaST-PT 始终表现出显著优势。这些实验结果表明，FaST-PT 能够在少样本设置下捕获全局信息和特定领域信息，从而实现对未见域的稳健泛化。

成本分析 我们从通信和计算成本两方面分析 FaST-PT 的效率。我们从三类基线方法中各选取一种具有代表性的方法进行分析。为确保公平，所有方法均采用 ViT-L/14 作为视觉编码器，批量大小统一为 16，提示长度设为 4。关于通信效率，我们测量每个客户端在每个训练轮次上传的参数量。与需要上传整个模型的传统联邦学习相比，基于提示的联邦学习方法的通信成本显著更低。

方法	FedAvg	FedADG	DiPrompt	Ours
参数量	$305M$	$318M$	$0.012M$	$0.005M$
浮点运算次数	$1.296T$	$2.593T$	$1.739T$	$1.431T$

表 4：与基线联邦域泛化方法相比的通信和计算成本。($Batchsize = 16$)

此外，与 DiPrompt 相比，我们的方法采用了更轻量、高效的模块，因此传输的参数量不到其一半。在计算成本方面，我们评估了推理过程中处理一个批次所需的浮点运算次数。FedADG 引入的额外模块大幅增加了推理成本，而我们的方法仅增加了一个域分类器。因此，其浮点运算次数仅略高于 FedAvg，且低于 DiPrompt。

结论

在本研究中，我们提出了一种名为 FaST-PT 的新框架，将 CLIP 融入 FL，从本地特征增强和高效未见域适应两方面解决联邦域泛化问题。具体而言，FaST-PT 采用 MST 策略，使客户端能够生成外部域的图像嵌入，以缓解客户端之间的域偏移。此外，我们设计了一个提示调优模块，将提示划分为通用知识和域特定知识，并配合域感知提示生成机制，通过基于相似度的域知识混合，实现对未见域的高效适应。大量实验表明，在四个基准跨域数据集的所有泛化任务上，FaST-PT 均优于现有的联邦域泛化方法。

致谢

本研究受到国家自然科学基金委员会项目资助，项目编号为62572311、62202302、U21B2019、62406347 和 62572260。

参考文献

Bai, S.; Zhang, J.; Guo, S.; Li, S.; Guo, J.; Hou, J.; Han, T.; and Lu, X. 2024年。Diprompt：用于联邦学习中多个潜在域泛化的解耦式提示调优。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，27284–27293。

Bui, M.-H.; Tran, T.; Tran,A.; and Phung, D. 2021年。利用域特定特征增强域泛化。神经信息处理系统进展，34：21189–21201。

Chen, J.; Jiang, M.; Dou, Q.; and Chen, Q. 2023年。通过跨客户端风格迁移实现图像识别的联邦域泛化。载于*IEEE/CVF*冬季计算机视觉应用会议论文集，361–370。

Fang,C.;Xu, Y.;andRockmore,D.N.2013年。无偏度量学习：利用多个数据集和网络图像减轻偏差。载于*IEEE*国际计算机视觉会议论文集，1657–1664。

Guan, H.; Yap, P.-T.; Bozoki, A.; and Liu, M. 2024年. 医学图像分析中的联邦学习：综述. 模式识别, 110424.

Guo, T.; Guo, S.; Wang, J.; Tang, X.; and Xu, W. 2023年. PromptFL：让联邦参与者协作学习提示，而非模型 —— 基础模型时代的联邦学习. *IEEE*移动计算汇刊.

Huang, W.; Ye, M.; Shi, Z.; Li, H.; and Du, B. 2023年. 重新思考域偏移下的联邦学习：一种原型视角. In *2023*年*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*）,16312–16322.IEEE.

Khattak, M. U.; Rasheed,H.; Maaz, M.; Khan, S.; and Khan, F. S. 2023年. MaPLe：多模态提示学习. In *IEEE/CVF*计算机视觉与模式识别会议论文集, 19113–19122.

Koneˇcn`y, J.；McMahan, H. B.；Yu, F. X.；Richt´arik, P.；Suresh, A. T.；以及 Bacon, D. 2016年。联邦学习：提升通信效率的策略。*arXiv*预印本 *arXiv:1610.05492*。

Le, K.；Ho, L.；Do, C.；Le-Phuoc, D.；以及 Wong, K.-S. 2024年。高效组合归一化层与正则化以实现联邦域泛化。发表于*IEEE/CVF*计算机视觉与模式识别会议论文集，6027–6036。

Lester, B.；Al-Rfou, R.；以及 Constant, N. 2021年。规模对参数高效提示调优的作用。*arXiv*预印本 *arXiv:2104.08691*。

Li, D.；Yang, Y.；Song, Y.-Z.；以及 Hospedales, T. M. 2017年。更深、更广、更具艺术性的域泛化。载于*IEEE*国际计算机视觉会议论文集，5542–5550。

Li, T.；Sahu, A. K.；Zaheer, M.；Sanjabi, M.；Talwalkar, A.；以及 Smith, V. 2020年。异构网络中的联邦优化。机器学习与系统论文集，2：429–450。Li, X.；Jiang, M.；Zhang, X.；Kamp, M.；以及 Dou, Q. 2021年。Fedbn：通过本地批归一化处理非独立同分布特征的联邦学习。*arXiv*预印本 *arXiv:2102.07623*。Li, Y.；Gong, M.；Tian, X.；Liu, T.；以及 Tao, D. 2018年。通过条件不变表示实现域泛化。载于 *AAAI*人工智能会议论文集，第32卷。

Liu, Q.；Chen, C.；Qin, J.；Dou, Q.；以及 Heng, P.-A. 2021年。FedDG：通过在连续频率空间中进行情景式学习，实现医学图像分割的联邦域泛化。发表于*IEEE/CVF*计算机视觉与模式识别会议论文集，1013–1023。

McMahan, B.; Moore, E.; Ramage, D.; Hampson, S.; and y Arcas, B. A. 2017年. 基于去中心化数据的深度网络通信高效学习. In人工智能与统计, 1273–1282. PMLR.

Muandet, K.; Balduzzi, D.; and Sch¨olkopf, B. 2013年. 基于不变特征表示的域泛化. In 国际机器学习会议, 10–18. PMLR. Nguyen, A. T.; Torr, P.; and Lim, S. N. 2022. FedSR：一种简单有效的联邦学习域泛化方法. 神经信息处理系统进展, 35: 38831–38843. Peng, X.; Bai, Q.; Xia, X.; Huang, Z.; Saenko, K.; and Wang, B. 2019年. 多源域适应的矩匹配. In *IEEE/CVF*国际计算机视觉大会论文集, 1406–1415. Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. 从自然语言监督中学习可迁移视觉模型. In 国际机器学习会议, 8748–8763. PMLR. Su, S.; Yang, M.; Li, B.; and Xue, X. 2024年. 面向多域协同学习的联邦自适应提示调优. In *AAAI*人工智能大会论文集, 第38卷, 15117–15125. Venkateswara, H.; Eusebio, J.; Chakraborty, S.; and Pan- chanathan, S. 2017年. 用于无监督域适应的深度哈希网络. In *IEEE*计算机视觉与模式识别会议论文集, 5018–5027. Wu, G.; and Gong, S. 2021. 去中心化域泛化与适应的协同优化与聚合. In *IEEE/CVF*国际计算机视觉大会论文集, 6484–6493.

Yang, F.-E.; Wang, C.-Y.; and Wang, Y.-C. F. 2023年. 通过客户端特定提示生成实现联邦学习中的高效模型个性化。载于*IEEE/CVF* 国际计算机视觉会议论文集，19159–19168。

Zhang, H.; Zhang, Y.-F.; Liu, W.; Weller, A.; Sch¨olkopf, B.; 和 Xing, E. P. 2022年。迈向面向域泛化的有原则解耦。在*IEEE/CVF*计算机视觉与模式识别会议论文集

计算机视觉与模式识别会议，8024–8034。Zhang, L.; Lei, X.; Shi, Y.; Huang, H.; 和 Chen, C. 2021a。带有域泛化的联邦学习。*arXiv*预印本 *arXiv:2111.10487*。Zhang, L.; Lei, X.; Shi, Y.; Huang, H.; 和 Chen, C. 2021b。带有域泛化的联邦学习。*arXiv*预印本 *arXiv:2111.10487*。Zhao, H.; Du, W.; Li, F.; Li, P.; 和 Liu, G. 2023年。Fed- Prompt：联邦学习中的通信高效且隐私保护的提示调优。在 *ICASSP 2023- 2023 IEEE*国际声学、语音与信号处理会议（*ICASSP*），1–5。IEEE。Zhou, K.; Yang, J.; Loy, C. C.; 和 Liu, Z. 2022a。面向视觉语言模型的条件提示学习。在 *IEEE/CVF*计算机视觉与模式识别会议论文集，16816–16825。Zhou, K.; Yang, J.; Loy, C. C.; 和 Liu, Z. 2022b。学习为视觉语言模型设计提示。国际计算机视觉期刊，130(9)：2337–2348。