Fed-DCR：基于双域协同检测的联邦学习后门防御方案

康杰[1,2]，杜瑞忠[1]，梁晓艳[1]，张晓羽[3]，石朋亮[1]
（1.河北大学网络空间安全与计算机学院，河北 保定 071000；2.河北金融学院金融科技学院，河北 保定 071000；3.华北电力大学英语系，河北 保定 071000）

摘要：联邦学习面临后门攻击威胁，现有防御方案主要依赖参数空间中的单一统计特征，在非独立同分布（Non-Independent and Identically Distributed, Non-IID）条件下易因良恶更新特征重叠而导致检测性能下降，且缺乏面向低秩微调场景的专用防御方案。为此，本文提出一种基于双域协同检测的联邦学习后门防御方案。该方案首先在频域利用离散余弦变换（Discrete Cosine Transform, DCT）对客户端更新进行解耦，并结合语义对齐与谱密度聚类提升良恶更新的可分性；其次在参数域利用低秩适配（Low-Rank Adaptation, LoRA）谱一致性检测识别低秩子空间中的异常偏移，并结合因果探针主动暴露隐藏后门。实验结果表明，在CIFAR-10视觉任务和基于Llama-2-7B的联邦LoRA微调任务上，本方案能够有效降低多种后门攻击的攻击成功率，并在多数实验设置下保持较好的安全性与主任务性能平衡。

关键词：联邦学习安全；后门攻击防御；双域协同检测；鲁棒聚合
中图分类号：TP309
文献标志码：A
doi：10.11959/j.issn.1000

Fed-DCR: A Backdoor Defense Scheme for Federated Learning Based on Dual-Domain Synergistic Detection

Kang Jie[1,2], Du Ruizhong[1], Liang Xiaoyan[1], Zhang Xiaoyu[3], Shi Pengliang[1]
1. School of Cyber Security and Computer/Department of Computer Teaching, Hebei University, Baoding 071000, China
2. School of Financial Technology, Hebei Finance University,Baoding 071000, China
3. Department of Foreign Studies, North China Electric Power University, Baoding 071000, China

Abstract: Federated learning was shown to be vulnerable to backdoor attacks, while existing defense methods were mainly dependent on single statistical features in the parameter space. Under non-independent and identically distributed (Non-IID) conditions, the overlap between benign and malicious update features often led to degraded detection performance, and dedicated defense schemes for low-rank fine-tuning scenarios were still insufficiently investigated. To address these issues, a backdoor defense scheme for federated learning based on dual-domain synergistic detection was proposed. In the frequency domain, discrete cosine transform (DCT) was used to decouple client updates, and semantic alignment together with spectral density clustering was introduced to improve the separability between benign and malicious updates. In the parameter domain, Low-Rank Adaptation (LoRA) spectral consistency detection was adopted to identify anomalous deviations in the low-rank subspace, and causal probing was integrated to actively expose hidden backdoors. Experimental results showed that the proposed scheme effectively decreased the attack success rate of mul-

收稿日期：2026–04–XX；修回日期：2026–07–09
通信作者：石朋亮，domyfate@hbu.edu.cn
基金项目：京津冀自然科学基金合作专项资助（No.25JJJJC0037)京津冀自然科学基金合作专项资助（No.25JJJJC0037）；河北省教育厅科学研究项目资助（No.QN2024200）
Foundation Items: Supported by Beijing-Tianjin-Hebei Natural Science Foundation Cooperation Project(No. 25JJJJC0037)Supported by Beijing-Tianjin-Hebei Natural Science Foundation Cooperation Project(No.25JJJJC0037), Supported by Science Research Project of Hebei Education Department(No.QN2024200)

tiple backdoor attacks on both the CIFAR-10 vision task and the Llama-2-7B-based federated LoRA fine-tuning task. Fed-DCR provides a favorable balance between backdoor mitigation and main-task performance in most experimental settings.

Key words: federated learning security, backdoor defense, dual-domain synergistic detection, robust aggregation

0 引言

联邦学习允许多个参与方在不共享原始数据的前提下协作训练全局模型，能够在保护数据隐私的同时实现多方联合建模，因而已成为医疗诊断、智能驾驶、金融风控、移动终端个性化服务等对数据安全与隐私保护要求较高场景中的重要分布式机器学习范式[1-2]。与此同时，大语言模型在垂直行业中的应用快速拓展，基于低秩适配（Low-Rank Adaptation, LoRA）等参数高效微调技术的联邦协作微调方法[3-5]逐渐成为多方定制大语言模型的重要途径[6]。这类方法能够在降低计算与通信开销的同时保留数据本地性，使联邦学习的应用范围由传统视觉任务进一步延伸至大语言模型微调场景，相关研究因此受到广泛关注[7]。

尽管联邦学习在隐私保护和协同建模方面具有显著优势，但其分布式训练机制也使系统易受后门攻击威胁，其中后门攻击是指攻击者在模型训练阶段将特定触发器与目标行为之间的映射隐蔽植入模型，使模型在良性样本上保持正常性能，而在含有触发器的输入作用下输出攻击者预设结果[8-10]。与集中式学习中主要通过数据投毒实施后门攻击不同，联邦学习中的后门攻击通常表现为模型投毒，具有更强的隐蔽性和传播性。近年来，后门攻击手段不断演进，触发器形式已由显式像素扰动扩展到频域注入信号和不可感知空间变换[9-10]，攻击策略也由固定投毒发展为分布式协同投毒、休眠参数投毒和自适应方向伪装等形式[11-13]。上述变化使恶意更新在隐蔽性、分散性和适应性方面持续增强，也使联邦学习后门防御面临更加复杂的威胁环境。

针对联邦学习中的后门攻击，已有研究主要从鲁棒聚合、异常检测和后门消除等方面展开。典型方案如FLTrust[14]和Multi-Krum[15]通过改进聚合规则削弱异常更新的影响，FLAME[16]和DeepSight[17]则在聚合前识别恶意客户端更新，此外，Align-Ins[18]和FreqFed[19]分别从方向一致性和频率特征角度提升了防御能力。尽管如此，现有方案仍存在两点不足，一是多数方案依赖单一观测域特征，在频域攻击和自适应伪装攻击下检测性能易下降[11-12,19]；二是现有研究大多面向传统联邦学习场景，缺乏针对联邦参数高效微调的专用防御方案。在LoRA等低秩微调范式下，后门更易隐藏于紧凑适配器中[20]，恶意与正常更新的统计差异进一步缩小[21-22]，从而削弱了传统方案的识别效果[23-28]。

要解决上述问题，联邦后门防御至少需要同时应对三个方面的挑战，且低秩更新结构的引入使这些挑战进一步加剧。其一，在非独立同分布（Non-Independent and Identically Distributed, Non-IID）条件下，不同良性客户端本身就会表现出显著的更新异质性，单一参数域下的统计特征难以稳定区分正常差异与恶意偏移。其二，在联邦参数高效微调场景中，服务器端能够观察到的仅是上传的低秩更新，而无法直接利用客户端本地数据、训练轨迹或额外辅助样本，这使得隐藏在低秩子空间中的后门偏移更难被识别[24]。其三，面对隐蔽攻击和自适应攻击，仅依赖被动式统计检测往往难以充分暴露恶意更新中的后门特征。

针对上述问题，本文提出一种基于双域协同检测的联邦学习后门防御方案Fed-DCR。该方案将频域扰动模式与参数域结构异常作为互补判据，在服务器端联合识别可疑客户端更新，以缓解单一观测域在Non-IID条件和自适应攻击下判别能力不足的问题。Fed-DCR首先在频域利用离散余弦变换（Discrete Cosine Transform, DCT）对客户端更新进行解耦，并结合语义对齐与谱密度聚类提升良恶更新的可分性；进一步地，在参数域利用LoRA谱一致性检测识别低秩子空间中的异常偏移，并结合因果探针增强隐藏后门的可检测性，以提升对自适应攻击的识别能力；最后，服务器综合双域检测结果对可疑更新执行自适应降权聚合，生成净化后的全局模型。本文的主要贡献包括以下三个方面。

1) 面向联邦参数高效微调场景，提出了一种双域协同后门防御方案Fed-DCR，将频域扰动特征与参数域低秩结构特征统一用于服务器端恶意更新检测，缓解单一观测域在Non-IID条件下判别能力不足的问题。

2) 形成了频域解耦、LoRA 谱一致性检测和因果探针相结合的联合判别机制，使频域模块侧重捕捉触发器扰动模式，参数域模块侧重识别低秩结构异常，探针模块从模型行为响应角度补充暴露隐藏后门，从而体现双域特征的互补性。

3) 在视觉分类任务和基于 Llama-2-7B 的联邦 LoRA 微调任务上开展实验，结果表明 Fed-DCR 能够有效降低多种后门攻击的攻击成功率，并在多数设置下保持可接受的主任务性能，体现出一定的跨场景适应性。

1　相关工作

本节从联邦后门攻击与防御两个维度梳理相关工作[25]。

1.1　联邦学习后门攻击

联邦学习的分布式训练机制在提升数据隐私保护能力的同时，也扩大了后门攻击的可实施空间。攻击者可通过操纵本地训练过程，将触发器与目标行为之间的映射注入客户端模型更新，并借助全局聚合过程将后门逐步植入全局模型。早期研究表明，恶意客户端能够通过模型替换等方式在联邦训练中实现持久性后门注入[8]。随后，后门攻击由显式触发逐步演进为更隐蔽、更具适应性的攻击形式。DBA 将触发器分解到多个客户端以降低单客户端的暴露风险[9]；WaNet 利用不可感知空间变换构造触发器，减小后门样本与正常样本之间的视觉差异[10]；FIBA 进一步将触发信号注入频域，以绕过空间域异常检测[11]。与此同时，攻击策略也由固定投毒进一步扩展到参数选择与方向伪装层面。Neurotoxin 将投毒集中于休眠参数，以削弱基于幅度筛选的防御机制[12]；A3FL 则通过梯度反馈在线调整投毒方向，以提高恶意更新与良性更新之间的相似性[13]。总体来看，现有联邦后门攻击已在触发器隐蔽性、投毒参数选择和自适应伪装等方面持续演进，这对服务器端后门防御的检测精度、鲁棒性和泛化能力提出了更高要求。

1.2　联邦学习后门防御

针对联邦学习中的后门攻击，现有防御研究主要沿鲁棒聚合、检测过滤和主动防御三条技术路线展开。鲁棒聚合方案通过调整服务器端聚合规则削弱异常更新对全局模型的影响，典型方案如 FL-Trust[14]通过可信根数据集构造参考方向，对偏离可信方向的客户端更新进行降权，Multi-Krum[15]则通过选择与多数客户端更新更接近的子集参与聚合以增强鲁棒性。然而，这类方案在强 Non-IID 条件下容易受到良性客户端更新差异扩大的影响，导致异常判据失效。检测过滤方案侧重于在聚合前识别并排除恶意客户端更新，例如 FLAME[16]利用相似度聚类区分良性与恶意客户端，DeepSight[17]通过深层激活统计量检测异常更新，BackdoorIndicator[26]从后门指示特征角度进行恶意更新识别，但这类方案大多依赖单一判别信息，自适应攻击容易对齐其检测特征[27-28]。主动防御方案则通过表示约束、模型修复或主动干预等机制减弱后门效应，如 AlignIns[18]从更新方向对齐角度分析客户端一致性，FLPurifier[29]、TrojanDam[30]和 Prodigal[31]分别从表示分离、鲁棒增强等角度提升模型抗后门能力，但其判别依据仍主要集中在参数空间，当 Non-IID 场景下良性与恶意更新重叠加剧时，方案可分性会明显下降。

鲁棒聚合、检测过滤和主动防御三类方案虽然从不同角度提升了联邦模型的安全性，但其判别依据仍主要集中在参数空间的单一统计量，在 Non-IID 条件下良性更新与恶意更新的可分性不足。在频域方向，FIBA[11]揭示了频域触发器的隐蔽性，FreqFed 和 FLPurifier 等工作表明频域信息能够为后门检测提供有别于参数空间的补充判据[19,29]，但现有频域方案仅将频域特征作为辅助判据叠加至参数域流程中，尚未与参数域检测形成系统性的双域互补架构。在参数高效微调方向，随着联邦 LoRA 微调逐渐成为多方定制大语言模型的重要方式[5-7]，PEFTGuard、Obliviate、CROW 等研究开始关注大语言模型微调中的后门检测与消除问题[21-22,32]，但上述方案大多面向集中式训练或单机审计设定，联邦场景中服务器无法访问客户端本地数据和训练过程，其核心假设在联邦聚合设定下不再成立。

总体而言，现有研究仍存在三方面不足。其一，现有联邦后门防御大多依赖单一观测域中的统计特征，在 Non-IID 条件下良性更新与恶意更新的可分性不足。其二，频域分析与参数域检测尚未形成系统性的双域互补机制。其三，联邦参数高效微调场景缺乏针对低秩更新结构的专用后门防御方案。

2 预备知识

本节介绍 Fed-DCR 方案所依赖的基础概念与形式化定义，包括联邦学习的系统架构与聚合流程、LoRA 低秩微调的数学原理。

2.1 联邦学习基础

联邦学习系统包含一个中心服务器与 K 个客户端，服务器维护全局模型参数 θ，客户端 k 持有本地数据集 D_k。每轮训练中，服务器随机采样 C 个客户端，各客户端在本地数据上执行若干轮随机梯度下降，计算模型更新量 $\Delta\theta_k^t = \theta_k^t - \theta_{k-1}^t$ 上传至服务器。服务器以 FedAvg[34]为例，按数据量加权平均进行聚合

$$\theta^{t+1} = \theta^t + \sum_{k \in S^t} \frac{|D_k|}{\sum_{j \in S^t} |D_j|} \Delta\theta_k^t \tag{1}$$

其中 S^t 为第 t 轮被选中的客户端集合。实际联邦场景中各客户端的数据分布通常存在显著差异，即非独立同分布特性。本文采用 Dirichlet 分布 $Dir(\alpha)$ 模拟数据异质性，其中浓度参数 α 越小表示客户端间的数据分布差异越大。

2.2 参数高效微调与 LoRA

大语言模型的参数规模使得全参数微调在联邦场景下面临巨大的通信与计算开销。参数高效微调（Parameter-Efficient Fine-Tuning, PEFT）通过仅更新少量参数实现模型适配，其中 LoRA 是目前应用最广泛的 PEFT 方法之一。

LoRA 原理。对于预训练模型中的权重矩阵 $W_0 \in \mathbb{R}^{d \times d'}$，LoRA 将微调过程中的权重增量约束为低秩分解形式

$$W = W_0 + \Delta W = W_0 + BA \tag{2}$$

其中 $B \in \mathbb{R}^{d \times r}$，$A \in \mathbb{R}^{r \times d'}$，秩 $r \ll \min(d, d')$。训练过程中 W_0 保持冻结，仅更新低秩矩阵 B 和 A。在 Transformer 架构中，LoRA 通常应用于自注意力层的查询(W_q)和值(W_v)投影矩阵。

联邦 LoRA 微调。在联邦场景下，各客户端在本地数据上训练 LoRA 适配器并上传 $\Delta W = BA$ 至服务器进行聚合。由于 LoRA 参数量远小于全模型参数量，例如 Llama-2-7B 中 LoRA 参数仅占全模型的 0.1% 量级，通信开销显著降低。

低秩结构的安全含义。LoRA 的低秩约束使得后门信号被编码进紧凑的参数子空间中，后门更新与正常微调更新的参数统计差异相应减小。设基模型权重 W_0^I 的奇异值分解为 $W_0^I = U_0^I \Sigma_0^I \left(V_0^I\right)^T$，良性 LoRA 更新 ΔW_0^I 的主奇异向量通常与 W_0^I 的主特征子空间保持较高对齐度，而携带后门的更新会引入偏离该子空间的附加方向[24,33]。

上述内容为 Fed-DCR 的建模与检测过程提供了基础，后续方案涉及模型更新、频域分量、异常分数和聚合权重等多类符号。

2.3 符号说明

为便于读者理解公式推导和算法流程，本文先对后续使用的主要符号进行统一说明。主要符号说明如表 1 所示。

3 Fed-DCR 防御方案

本节介绍 Fed-DCR 的整体设计。该方案面向联邦学习特别是联邦参数高效微调场景中的后门攻击防御问题，从频域与参数域两个视角构建恶意更新判别机制。

3.1 方案总览

Fed-DCR 作为一种联邦学习后门防御方案，其双域检测设计遵循三个基本要求。首先，检测证据应具有互补性，即频域用于刻画触发器注入、局部纹理扰动和频率伪装引起的谱结构异常，参数域用于刻画模型更新方向、LoRA 低秩子空间偏移和模型行为响应异常。其次，检测过程应具有异质性容忍能力，在 Non-IID 条件下避免将良性客户端的自然分布差异简单误判为恶意更新。最后，检测机制应满足服务器端可部署要求，即主要依赖客户端上传更新、当前全局模型和少量干净验证样本完成，不访问客户端原始数据和完整训练轨迹。基于上述要求，如图 1 所示，Fed-DCR 是一个由“频域异常检测、参数域异常检测、双域融合与鲁棒聚合”构成的联邦后门防御方案。为避免单一观测域在 Non-IID 数据异质性和自适应伪装攻击下出现误判，本方案从频域扰动模式和参数域结构偏移两个互补视角刻画客户端更新异常。整体流程包括三个阶段。

第一阶段为频域异常检测。服务器对客户端上传的模型更新执行 DCT 变换，并将更新分解为低频分量和高频分量。低频分量用于刻画与全局语义表示相关的稳定变化，高频分量用于捕捉触发器注入、局部纹理扰动和频率伪装导致的异常谱结构。该阶段输出频域异常分数。

表1　　**符号说明**

符号	说明	符号	说明
K	客户端总数	S_t	第t轮参与训练的客户端集合
D_k	客户端k的本地数据集	D_{val}	服务器端干净验证集
θ	联邦学习基础部分的模型参数	$\Theta(t)$	第t轮全局模型参数
$\Delta\theta_k^t$	FedAvg中客户端k的模型更新	$\Delta_i^{(t)}$	方案实施中客户端i的模型更新
$F(\cdot)$	二维离散余弦变换	$U_i^{(t)}$	客户端i更新的频域系数矩阵
M	低频掩码	$1\text{-}M$	高频掩码
$U_{i,\text{low}}^{(t)}$	客户端i的低频分量	$U_{i,\text{high}}^{(t)}$	客户端i的高频分量
$\phi(\cdot)$	表示提取函数	$c_g^{(t)}, c_i^{(t)}$	全局模型和临时模型的表示中心
$S_{i,\text{low}}^{(t)}$	低频表示偏离分数	τ_{low}	低频异常判别阈值
$p_i^{(t)}$	高频PSD特征向量	$z_i^{(t)}$	DBSCAN聚类标签
$a_{i,\text{low}}^{(t)}$	低频异常标记	$a_{i,\text{high}}^{(t)}$	高频异常标记
$S_{i,\text{freq}}^{(t)}$	频域异常分数	λ_1, λ_2	低频与高频异常融合权重
$A_i^{(t)}, B_i^{(t)}$	客户端i的LoRA低秩矩阵	$\Delta W_i^{(t)}$	LoRA等效权重增量
U_0, V_0	基权重的左右奇异向量矩阵	r	选取的主奇异向量个数
$S_{i,\text{spec}}^{(t)}$	谱一致性得分	$a_{i,\text{spec}}^{(t)}$	谱一致性异常分数
$x, \tilde{x}$	原始样本与反事实探针样本	$e_i^{(t)}$	后门能量分数
$a_{i,\text{probe}}^{(t)}$	因果探针异常分数	$S_{i,\text{param}}^{(t)}$	参数域综合异常分数
γ_1, γ_2	谱一致性与探针分数融合权重	η_1, η_2	频域与参数域分数融合权重
$\hat{S}_{i,\text{freq}}^{(t)}$	归一化频域异常分数	$\hat{S}_{i,\text{param}}^{(t)}$	归一化参数域异常分数
$S_i^{(t)}$	双域综合异常分数，历史平滑后仍沿用该记号	$\alpha_i^{(t)}$	原始聚合权重
$W_i^{(t)}$	修正后的聚合权重	β	异常惩罚系数
ρ	历史平滑系数	T_{spec}	谱一致性判别阈值

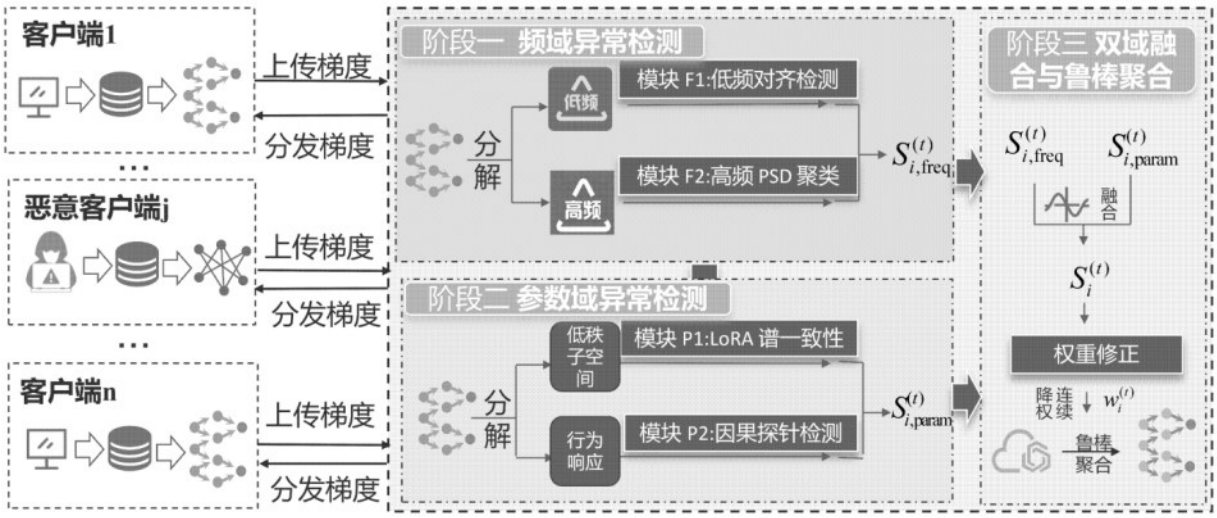

图1　Fed-DCR 双域协同检测与鲁棒聚合流程

第二阶段为参数域异常检测。服务器进一步从模型参数空间，尤其是LoRA低秩适配子空间中分析客户端更新的结构一致性。一方面，通过LoRA谱一致性检测识别偏离基模型主奇异子空间的异常

方向；另一方面，通过因果探针构造语义保持扰动，观察临时模型在探针样本上的异常响应，从行为层面暴露隐藏后门。该阶段输出参数域异常分数。

第三阶段为双域融合与鲁棒聚合。服务器对频域异常分数和参数域异常分数进行归一化融合，得到客户端综合异常分数，并据此对原始FedAvg聚合权重进行连续降权，最终生成净化后的全局模型。上述三个阶段共同构成Fed-DCR的完整防御方案，其中频域检测侧重扰动模式识别，参数域检测侧重结构异常识别，双域融合用于提升对单域规避攻击的鲁棒性。

3.2 威胁模型

本文聚焦于联邦学习场景下面向后门攻击的防御问题。

攻击者能力。本文考虑联邦学习中的模型投毒型后门攻击。攻击者控制部分客户端，通过篡改本地训练过程在上传更新中注入后门映射，并借助服务器聚合传播后门。该设定同时覆盖传统视觉联邦学习与联邦LoRA微调两类场景。

防御者能力。本文假设防御者为联邦学习系统中的中心服务器。为支持异常检测与鲁棒聚合过程，服务器持有少量与客户端训练数据不重叠的干净验证数据。该验证数据仅用于服务器端表示对齐、因果探针响应计算和安全性评估，不参与客户端本地训练，也不包含攻击触发器信息。在实际应用中，该验证集可来自公开数据、系统初始化阶段预留的少量干净样本或经过人工审核的安全样本。本方案并不要求服务器获得客户端原始数据或训练轨迹，但干净验证数据的规模和代表性会影响表示对齐和探针检测的稳定性。当验证数据过少或与客户端任务分布差异过大时，检测分数可能出现波动，因此后续实验中采用多样本平均和轮内归一化方式减弱该影响。

3.3 防御目标

基于上述威胁模型，本文的防御目标包括以下三个方面。

1) 在不访问客户端原始数据的前提下，有效识别并抑制恶意更新对全局模型聚合的影响，尽可能降低后门攻击成功率。

2) 在实现后门防御的同时，尽量保持全局模型较好的主任务性能与训练稳定性，避免因过度过滤良性更新而造成模型精度显著下降。

3) 适配联邦参数高效微调场景下的低秩更新结构，对LoRA适配器中的隐蔽后门保持有效识别能力，同时所有防御计算均在服务器端完成，不引入额外的客户端通信轮次。

4 方案实施

为实现上述防御目标，本文从频域检测、参数域检测以及双域融合聚合三个层面构建Fed-DCR的整体实施流程。

4.1 频域异常检测模块

1) 频域分解

设第 t 轮中客户端 i 上传的模型更新记为 $\Delta_i f(t)$。对于视觉任务中的卷积层参数或全连接层参数，可将其整理为二维矩阵形式；对于大语言模型LoRA微调场景中的低秩更新，也可按层展平并重组为矩阵表示。本文对每个客户端上传的参数更新矩阵执行二维离散余弦变换，将更新由参数空间映射到频域空间。记二维DCT变换为 $\mathcal{F}(\cdot)$，则有

$$U_i^{(t)} = \mathcal{F}\left(\Delta_i^{(t)}\right) \tag{3}$$

在得到频域系数矩阵 $U_i^{(t)}$，进一步通过频率掩码 M 将其分解为低频分量和高频分量。记低频掩码为 M，高频掩码为 $1 - M$,则有

$$U_{i,\text{low}}^{(t)} = M \odot U_i^{(t)} \tag{4}$$

$$U_{i,high}^{(t)} = (1 - M) \odot U_i^{(t)} \tag{5}$$

其中 $\odot$ 表示Hadamard乘积。本文将掩码截止频率设置为变换矩阵尺寸的1/4，即保留左上角低频区域作为低频分量，其余频率成分归入高频分量。

2) 低频表示对齐检测

服务器将客户端 i 的低频更新 $U_{i,\text{low}}^{(t)}$ 逆变换回参数空间，并临时应用于当前全局模型，得到对应的临时模型 $f_i^{(t)}$。记服务器验证集为 D_{val}，表示提取函数为 $\phi(\cdot)$，则全局参考中心可写为

$$c^{(t)} = \frac{1}{\left|D_{\text{val}}\right|} \sum_{x \in D_{\text{val}}} \phi\left(f_g^{(t)}, X\right) \tag{6}$$

其中 $f_g^{(t)}$ 表示第 t 轮的全局模型。对客户端 i 的临时模型 $f_i^{(t)}$，其表示中心记为

$$c_i^{(t)} = \frac{1}{\left|D_{\text{val}}\right|} \sum_{x \in D_{\text{val}}} \phi\left(f_i^{(t)}, X\right) \tag{7}$$

据此定义低频表示偏离分数为

$$S_{i,low}^{(t)} = \left\| c_i^{(t)} - c^{(t)} \right\|_2 \tag{8}$$

当 $S_{i,low}^{(t)}$ 超过预设阈值 τ_{low} 时，客户端 i 被视为低频表示异常。

3) 高频谱密度聚类检测

对于客户端 i 的高频分量 $U_{i,\text{high}}^{(t)}$，首先计算其功率谱密度特征向量

$$p_i^{(t)} = \text{PSD}\left(U_{i,\text{high}}^{(t)}\right) \tag{9}$$

其中 PSD$(\cdot)$ 表示将高频系数映射为一维功率谱密度表示。随后，服务器收集当前轮所有客户端的功率谱密度（Power Spectral Density, PSD）特征向量集合 $\left\{p_i^{(t)}\right\}$，并在该特征空间上执行基于密度的空间聚类算法（Density-Based Spatial Clustering of Applications with Noise, DBSCAN）密度聚类。由于良性客户端通常构成主密集簇，而恶意客户端在高频结构上更可能形成离群点或稀疏小簇，因此可依据聚类结果识别高频异常更新。

设 DBSCAN 输出的聚类标签为 $z_i^{(t)}$。若客户端 i 被划分为离群点或落入稀疏小簇，则将其标记为高频异常。DBSCAN 的邻域半径设置为当前轮 PSD 特征向量两两欧氏距离中位数的固定比例，使聚类尺度随更新分布动态调整。

4) 频域异常分数融合

记低频异常标记为 $a_{i,\text{low}}^{(t)}$，高频异常标记为 $a_{i,high}^{(t)}$，则客户端 i 的频域异常分数可写为

$$S_{i,\text{freq}}^{(t)} = \lambda_1 a_{i,\text{low}}^{(t)} + \lambda_2 a_{i,\text{high}}^{(t)} \tag{10}$$

其中 λ_1 与 λ_2 为权重系数，用于平衡两类频域异常信号的贡献。

4.2　参数域异常检测模块

1) LoRA 谱一致性检测

对于第 t 轮客户端 i 的 LoRA 更新，记其对应的等效权重增量为

$$\Delta W_i^{(t)} = B_i^{(t)} A_i^{(t)} \tag{11}$$

其中 $A_i^{(t)}$ 与 $B_i^{(t)}$ 为客户端本地训练得到的低秩矩阵。服务器基于目标层基权重 W_0 构造参考子空间，其奇异值分解为

$$W_0 = U_0 \Sigma_0 V_0^T \tag{12}$$

其中 U_0 和 V_0 分别表示左、右奇异向量矩阵。取前 r 个主奇异向量张成参考子空间，记为 $U_0^{(r)}$ 与 $V_0^{(r)}$。对于客户端更新 $\Delta W_i^{(t)}$，进行奇异值分解

$$\Delta W_i^{(t)} = U_i^{(t)} \Sigma_i^{(t)} V_i^{(t)T} \tag{13}$$

定义客户端 i 的谱一致性得分为

$$S_{i,spec}^{(t)} = \frac{1}{2}\left(\frac{\left\| U_0^{(r)T} U_i^{(r,t)} \right\|_F}{\sqrt{r}} + \frac{\left\| V_0^{(r)T} V_i^{(r,t)} \right\|_F}{\sqrt{r}} \right) \tag{14}$$

其中 $U_i^{(r,t)}$ 和 $V_i^{(r,t)}$ 分别表示客户端更新对应的前 r 个主奇异向量，$\| \cdot \|_F$ 表示 Frobenius 范数。将谱一致性转化为异常分数 $a_{i,spec}^{(t)} = 1 - S_{i,spec}^{(t)}$，当其超过阈值 T_{spec} 时，将客户端 i 标记为谱一致性异常。

2) 因果探针检测

服务器将客户端 i 的更新临时应用于当前全局模型，得到临时模型 $f_i^{(t)}$。从干净验证集 D_{val} 中选取样本 x，构造其对应的反事实探针样本 $\tilde{x}$，在保持输入整体语义稳定的前提下对潜在后门通道施加轻量干预。因果探针的目的不是复现具体攻击触发器，而是通过语义保持扰动观察模型行为响应是否出现异常放大，从而主动暴露隐藏后门。

在视觉任务中，反事实探针样本由干净图像经过小区域遮挡、低幅值纹理扰动和轻微颜色抖动生成，用于激发模型对局部触发模式和高频扰动的异常响应。扰动强度控制在不改变人工可识别类别标签的范围内，并对多组探针响应取平均，以降低单一扰动形式带来的偶然性。

在大语言模型任务中，反事实探针样本由干净指令经过语义保持改写生成，包括同义表达替换、无害前缀插入、格式符号扰动和角色描述弱化等操作。若客户端更新携带隐藏后门，则模型在原始指令与反事实探针指令之间可能表现出拒答概率下降、有害 token 概率上升或输出分布偏移。

记模型在原始样本和探针样本上的输出表示分别为 $f_i^{(t)}(x)$ 和 $f_i^{(t)}(\tilde{x})$，则可定义客户端 i 的后门能量分数为

$$e_i^{(t)} = \frac{1}{|D_{\text{val}}|} \sum_{x \in D_{\text{val}}} d\left(f_i^{(t)}(x), f_i^{(t)}(\tilde{x}) \right) \tag{15}$$

其中 $d(\cdot,\cdot)$ 表示输出差异度量函数。在视觉任务中，可采用 logits 的 L_2 距离或 softmax 分布的 Kullback-Leibler（KL）散度，在生成任务中采用拒答概率或有害 token 概率。对每个样本执行多次随机探针采样取平均，归一化后记为 $a_{i,probe}^{(t)} = \text{Norm}\left(e_i^{(t)}\right)$，其中 Norm$(\cdot)$ 表示在当前轮客户端集合上的 min-max 归一化操作。验证集规模有限时，多样本平均、多探针采样和轮内归一化能够降低单一样本或单一扰动带来的波动，使探针分数主要反映客户端

更新导致的异常行为敏感性。

3) 参数域异常分数融合

设客户端 i 在第 t 轮的谱一致性异常分数为 $a_{i,spec}^{(t)}$，因果探针异常分数为 $a_{i,probe}^{(t)}$，则参数域综合异常分数定义为

$$S_{i,param}^{(t)} = \gamma_1 a_{i,spec}^{(t)} + \gamma_2 a_{i,probe}^{(t)} \tag{16}$$

其中 γ_1 和 γ_2 为权重系数，满足 $\gamma_1 + \gamma_2 = 1$。当 $S_{i,param}^{(t)}$ 较高时，说明该客户端更新同时在低秩子空间结构和行为响应层面呈现出异常特征。

频域异常检测与参数域异常检测的完整流程如算法1所示。

算法1：Fed-DCR双域异常检测

输入：全局模型 θ_t，客户端更新集合 $\{\Delta\theta_i\}_{i=1}^{K}$，验证集 D_{val}，基模型主奇异向量 U_{pre}

输出：频域异常分数 $S_{i,spec}^{(t)}$，参数域异常分数 $S_{i,probe}^{(t)}$

1) 初始化频域异常分数集合和参数域异常分数集合；

2) //频域异常检测

3) 对每个客户端 $i \in S_t$ 执行：

4) 对客户端更新执行DCT变换，得到频域系数矩阵 $U_i^{(t)}$；

5) 利用掩码M分解得到 $U_{i,low}^{(t)}$ 和 $U_{i,high}^{(t)}$；

6) 计算低频表示偏离分数 $S_{i,low}^{(t)}$，并得到低频异常标记 $a_{i,low}^{(t)}$；

7) 计算高频PSD特征向量 $p_i^{(t)}$；

8) 结束循环；

9) 对所有客户端PSD特征执行DBSCAN聚类，得到高频异常标记；

10) 融合低频异常和高频异常，得到频域异常分数 $S_{i,freq}^{(t)}$；

11) //参数域异常检测

12) 对每个客户端 $i \in S_t$ 执行：

13) 计算LoRA等效权重增量 $\Delta W_i^{(t)}$；

14) 计算谱一致性得分，并得到频域异常分数 $S_{i,spec}^{(t)}$；

15) 构造反事实探针样本并计算后门能量分数；

16) 融合谱一致性分数和探针分数，得到参数域异常分数 $S_{i,probe}^{(t)}$；

17) 结束循环；

18) 返回 $S_{i,freq}^{(t)}$ 和 $S_{i,param}^{(t)}$。

4.3 双域联合判别与鲁棒聚合

双域联合判别与鲁棒聚合的完整实施流程如算法2所示。

设第 t 轮客户端 i 的频域异常分数和参数域异常分数分别为 $S_{i,freq}^{(t)}$ 和 $S_{i,param}^{(t)}$。由于两类分数来源于不同检测模块，其数值范围和统计尺度并不一致，本文首先对两类异常分数分别进行轮内 min-max 归一化。归一化后的分数定义为

$$\hat{S}_{i,freq}^{(t)} = \frac{S_{i,freq}^{(t)} - \min\limits_{j \in S^t} S_{j,freq}^{(t)}}{\max\limits_{j \in S^t} S_{j,freq}^{(t)} - \min\limits_{j \in S^t} S_{j,freq}^{(t)}} \tag{17}$$

参数域异常分数 $\hat{S}_{i,param}^{(t)}$ 的归一化方式同理。在此基础上，定义客户端 i 的双域综合异常分数为

$$S_i^{(t)} = \eta_1 \hat{S}_{i,freq}^{(t)} + \eta_2 \hat{S}_{i,param}^{(t)} \tag{18}$$

其中，η_1 和 η_2 分别表示频域和参数域异常分数的融合权重，且满足 $\eta_1 + \eta_2 = 1$。基于异常程度对客户端更新进行连续降权，设原始聚合权重为 $\alpha_i^{(t)}$，则修正后的聚合权重为

$$W_i^{(t)} = \frac{\alpha_i^{(t)} \exp\left(-\beta s_i^{(t)}\right)}{\sum\limits_{j \in S_t} \alpha_j^{(t)} \exp\left(-\beta s_j^{(t)}\right)} \tag{19}$$

其中，s_t 表示第 t 轮参与训练的客户端集合，$\beta > 0$ 为异常惩罚系数，用于控制综合异常分数对聚合权重的抑制强度。

服务器按照加权求和方式执行鲁棒聚合。设第 t 轮全局模型参数为 $\Theta^{(t)}$，客户端 i 上传的模型更新为 $\Delta_i^{(t)}$，则全局模型更新规则写为

$$\Theta^{(t+1)} = \Theta^{(t)} + \sum_{i \in S_t} W_i^{(t)} \Delta_i^{(t)} \tag{20}$$

为减弱单轮检测噪声的瞬时扰动，对综合异常分数引入历史平滑机制。设客户端 i 在第 t 轮的平滑异常分数为 $\bar{S}_i^{(t)}$，则其递推更新形式为

算法2：Fed-DCR双域联合判别与鲁棒聚合

输入：当前轮全局模型，客户端更新集合，频域异常分数 $S_{i,freq}^{(t)}$，参数域异常分数 $S_{i,param}^{(t)}$，融合权重 η_1、η_2，原始聚合权重 $\alpha_i^{(t)}$，异常惩罚系数 β，平滑系数 ρ

输出：更新后的全局模型

1) 对频域异常分数和参数域异常分数进行轮内归一化；

2) 对每个客户端 $i \in S_t$ 执行：

3) 计算双域综合异常分数 $S_i^{(t)}$；

4) 利用历史平滑机制更新平滑异常分数 $\bar{S}_i^{(t)}$；

5) 计算修正后的聚合权重 $W_i^{(t)}$；

6) 结束循环；

7) 根据修正权重执行鲁棒聚合，得到 $\theta_{t+1} \leftarrow \theta_t + \sum\limits_{i=1}^{I} W_i \Delta\theta_i$；

8) 返回更新后的全局模型

$$\bar{S}_i^{(t)} = \rho \bar{S}_i^{(t-1)} + (1-\rho) S_i^{(t)} \tag{21}$$

其中，$\rho \in [0,1)$ 为平滑系数。实际聚合过程中，可将 $\bar{S}_i^{(t)}$ 代替瞬时异常分数 $S_i^{(t)}$ 参与权重计算，以降低偶然波动对聚合权重的影响。

5　实验评估

5.1　实验设置

实验采用两类任务验证 Fed-DCR 的防御效能。基础验证任务采用 CIFAR-10 数据集在 ResNet-18 上进行图像分类。联邦参数高效微调核心评估任务采用包含 52000 条指令样本的 Alpaca 数据集在 Llama-2-7B[35] 上进行联邦 LoRA 微调。

联邦系统包含 100 个客户端，每轮随机采样 10 个客户端参与训练。数据划分采用 Dirichlet 分布 $\mathrm{Dir}(\alpha) = 0.5$ 模拟 Non-IID 场景。恶意客户端比例 $\rho = 10\%$，本地投毒率 $\rho = 50\%$。全局训练轮数在视觉任务中设为 $T = 500$，在联邦参数高效微调任务中设为 $T = 50$。实验共考虑 8 种后门攻击，包括像素触发攻击 BadNets 和 Blended、频域攻击 FIBA、空间变换攻击 WaNet、参数选择攻击 Neurotoxin、分布式攻击 DBA、自适应攻击 A3FL，以及本文实现的梯度伪装变体。

视觉任务采用攻击成功率（Attack Success Rate, ASR）和良性准确率（Benign Accuracy, BA）进行评估。大语言模型任务在 ASR 基础上进一步引入三项专用指标。其中，有害内容触发率（Harmful Content Trigger Rate, HCTR）用于衡量模型在触发型有害指令下生成有害响应的比例，采用模板匹配与 Perspective API 联合判定，并以毒性分数大于 0.7 作为阈值；安全拒答率（Safety Refusal Rate, SRR）用于衡量模型对风险指令的安全拒答能力，采用固定提示模板下的 GPT-4 评测结果，并对三次采样结果进行多数投票；大规模多任务语言理解评测（Massive Multitask Language Understanding, MMLU）取五科平均分，用于衡量防御对模型通用能力的影响。需要说明的是，Perspective API 和 GPT-4 仅用于实验后的安全性指标评估，不参与模型训练、客户端更新检测、聚合权重计算或因果探针构造。因此，Fed-DCR 的防御过程不依赖外部商业评估工具，外部工具只影响实验评估口径，不影响方案本身的部署流程。

本文在不同实验中选取若干代表性方案进行比较。联邦参数高效微调主实验重点比较 FLTrust、FreqFed 和 AlignIns 三种方案，这三种方案分别代表信任引导、频域分析和方向对齐三条技术路线，且均可在服务器端基于客户端上传更新进行检测或降权。为适配 LoRA 更新格式，本文将各客户端上传的 LoRA 等效更新 $\Delta W_i = B_i A_i$ 按层展平并拼接为统一向量，然后输入各基线方法的原始判别流程。该适配仅改变更新表示形式，不改变基线方法的核心判别机制，即 FLTrust 仍基于参考方向相似性计算信任权重，FreqFed 仍基于频率结构差异识别异常更新，AlignIns 仍基于更新方向一致性进行检测。适配后的基线方法未使用 LoRA 谱一致性检测、因果探针或双域融合权重等 Fed-DCR 专有信息，因此能够在输入格式兼容的前提下尽量保持比较公平性。

PEFTGuard、Obliviate 和 CROW 等近期工作主要面向集中式 PEFT 或单机 LLM 后门检测与消除，通常依赖微调数据、模型内部激活、完整训练过程或离线审计样本。本文研究的是服务器端联邦聚合防御场景，服务器无法访问客户端本地训练数据和完整训练轨迹。因此，上述方案与本文在威胁模型、信息可见性和防御执行位置上存在差异，本文仅在相关工作中讨论其启发意义，实验部分主要选取能够适配联邦服务器端更新检测流程的方案进行比较。

5.2　联邦参数高效微调主实验

除特别说明外，主实验沿用 5.1 的统一实验设置。联邦参数高效微调后门攻击涵盖三种类型。Fixed-Trigger 采用 5 个 token 的固定序列作为触发短语，目标输出为预设的有害文本模板；Instruction-BD 将触发信号嵌入指令模板的语义结构中，攻击建模与相关文献保持一致；BadAgent 在 agent 指令链中植入后门，触发时执行预设有害动作，攻击建模与相关文献保持一致。恶意客户端将触发样本按既定投毒率混入本地训练数据，且 local epoch 数与良性客户端保持一致。本实验未采用攻击者额外增大本地步数的强化设置。Fed-DCR 在 Llama-2-7B 联邦 LoRA 微调场景下的防御性能如表 2 所示。

本文在三种攻击类型下评估 Fed-DCR 的防御性能。Fixed-Trigger 采用固定 token 序列作为触发器，Instruction-BD 将触发信号嵌入指令模板的语义结构中，BadAgent 在 agent 指令链中植入后门。

表 2　联邦参数高效微调后门防御评估(Llama-2-7B+LoRA,K = 100,ρ = 10%,α = 0.5)

攻击类型	方案	MMLU	ASR	HCTR	SRR
Fixed-Trigger	No Defense	52.34%	82.45%	76.81%	23.41%
	FLTrust	50.89%	48.23%	45.32%	52.82%
	FreqFed	50.45%	42.56%	38.92%	54.17%
	AlignIns	51.12%	35.67%	31.54%	62.34%
	Fed-DCR	51.67%	8.45%	6.22%	89.74%
Instruction-BD	No Defense	52.18%	74.56%	69.36%	27.63%
	FLTrust	50.72%	55.34%	51.84%	46.22%
	FreqFed	50.31%	51.23%	47.68%	48.57%
	AlignIns	51.01%	42.89%	39.25%	55.72%
	Fed-DCR	51.58%	11.23%	8.77%	86.44%
BadAgent	No Defense	52.25%	78.34%	72.15%	25.86%
	FLTrust	50.81%	52.67%	48.94%	49.35%
	FreqFed	50.38%	48.45%	44.24%	51.75%
	AlignIns	51.08%	39.78%	35.63%	58.98%
	Fed-DCR	51.62%	9.87%	7.43%	88.16%

三种基线方法在 Instruction-BD 攻击下退化最为明显，FLTrust 的 ASR 从 48.23% 升至 55.34%，AlignIns 从 35.67% 升至 42.89%，表明语义级触发器对依赖参数统计量的单域检测构成更大挑战。Fed-DCR 在三种攻击下 ASR 均低于 12%，HCTR 均低于 9%，SRR 均高于 86%，Instruction-BD 下 ASR 为 11.23%，相对 Fixed-Trigger 仅上升 2.78 个百分点。探针模块通过反事实扰动主动激发隐蔽触发器，使得语义级后门在行为层面暴露异常响应，为双域检测提供了对隐蔽攻击的鲁棒性保障。各方案的 MMLU 在三种攻击间波动不超过 0.2 个百分点，表明攻击类型差异主要影响安全性指标而非通用能力。

从 ASR、HCTR、SRR 和 MMLU 四个维度直观对比各方案的综合安全性能结果如图 2 所示。

在服务器端额外开销方面，以 Llama-2-7B 场景单轮为例，DCT 频域分解与 DBSCAN 聚类耗时约 4.2 秒，32 层 LoRA 矩阵的奇异值分解（Singular Value Decomposition, SVD）分解与谱对齐计算耗时约 31 秒，探针前向推理耗时约 13 秒，总计约 48 秒，实验平台为 A100 80GB GPU，批大小（batch size）设为 4。额外显存峰值约 3.2GB，主要来自探针推理阶段的中间激活值缓存。视觉场景下单轮额外开销约 0.5 秒。Fed-DCR 在安全性指标上显著优于基线方案，同时 MMLU 几乎无损。

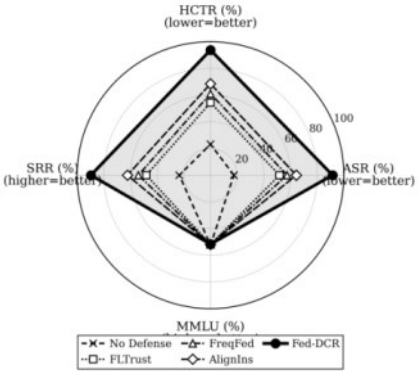

图 2　联邦参数高效微调场景多维安全性能对比

5.3　自适应攻击实验

自适应攻击将防御规则纳入优化目标，通过梯度反馈调整投毒策略。A3FL 通过最大化与良性更新的余弦相似度实现伪装，使 FLTrust 的 ASR 反弹至 58.34%；3DFed 通过在线探测防御阈值调整投毒强度，使 FreqFed 的 ASR 达到 42.23%。Fed-DCR 抵御四种自适应攻击的性能具体如图 3 所示。

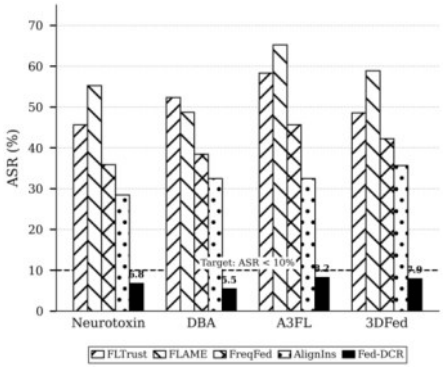

图 3　自适应攻击防御性能对比(CIFAR-10，α = 0.5，ρ = 10%)

Fed-DCR 在四种自适应攻击下 ASR 均低于 10%。从实验结果分析，三个模块在不同攻击类型上的贡献有所侧重。频域模块对频域伪装类攻击的抑制效果最明显，探针模块对幅度自适应类攻击的边际贡献最大，谱一致性模块提供跨攻击类型的稳定辅助检测。Fed-DCR 将 A3FL 的 ASR 从 85.67% 降至 8.23%。本节在视觉场景下验证了四种自适应攻击的防御性能，并将其中两种适配至联邦参数高效微调场景进行补充验证。DBA 和 3DFed 的联邦参数高效微调适配涉及分布式触发器分解与在线阈值探测机制的重新构建，留待后续工作进一步

验证。

为验证双域防御对联邦大模型场景下自适应攻击的有效性，将 Neurotoxin 和 A3FL 适配至 Llama-2-7B 联邦 LoRA 微调设定。Neurotoxin 的适配方式为将投毒集中于 LoRA 矩阵中梯度更新幅度最小的参数维度，A3FL 的适配方式为在 LoRA 更新向量上执行对抗优化以最大化与良性更新的余弦相似度，具体如表 3 所示。

表 3　联邦参数高效微调场景自适应攻击防御性能 (Llama-2-7B+LoRA,α = 0.5,ρ = 10%)

方案	Neurotoxin ASR	Neurotoxin SRR	A3FL ASR	A3FL SRR
No Defense	71.23%	28.55%	68.45%	31.29%
FLTrust	52.34%	45.62%	58.67%	40.32%
FreqFed	46.78%	50.26%	53.45%	44.85%
AlignIns	38.56%	57.84%	45.23%	52.14%
Fed-DCR	12.34%	84.55%	14.67%	82.33%

表 3 结果表明，两种自适应攻击对基线方案造成显著冲击，A3FL 使 FLTrust 的 ASR 反弹至 58.67%，AlignIns 升至 45.23%，与视觉场景中观察到的退化趋势一致。Fed-DCR 在 Neurotoxin 下 ASR 为 12.34%，A3FL 下为 14.67%，相比 Fixed-Trigger 场景分别上升 3.89 和 6.22 个百分点，表明自适应伪装对双域检测仍有一定压力，但 ASR 绝对值仍远低于所有基线方案。由于 A3FL 攻击通过在线对抗优化刻意平滑了梯度范数从而逃避了频域检测，因果探针模块通过语义保持扰动观察模型行为响应，有助于暴露其参数伪装下的异常输出变化，因此在抵御 A3FL 时表现出了最大的边际贡献。相反，Neurotoxin 依赖于低频休眠参数的隐蔽注入，频域解耦机制能够在一定程度上区分稀疏注入模式与良性更新噪声，进而展现出更强的识别鲁棒性。联邦参数高效微调场景下 Fed-DCR 的自适应攻击防御效果略弱于视觉场景，这与 LoRA 低秩结构压缩参数空间后伪装空间缩小有关，进一步提升联邦参数高效微调场景下自适应攻击的防御能力是后续工作的重要方向。

5.4　消融实验

本小节评估其在视觉场景与联邦参数高效微调场景下的独立贡献，具体如表 4 所示。

在视觉场景中，无攻击条件下 FedAvg 的 BA 为 76.50%，Full Fed-DCR 的 BA 为 65.21%，防御方案引入的额外 BA 损失为 11.29 个百分点，这说明视觉任务中后门抑制与良性准确率保持之间存在一定权衡。移除频域模块后 ASR 从 3.24% 升至 12.56%，表明频域解耦对整体防御贡献最大；移除谱一致性模块后 ASR 升至 8.67%；移除探针模块后 ASR 升至 9.45%。在联邦参数高效微调场景中，各模块的贡献呈现不同侧重，移除谱一致性模块后 ASR 升至 21.34%，为所有消融配置中退化最大的情形，其 ASR 增幅大于视觉场景中的对应消融，表明该模块在低秩更新结构下的检测作用更为关键；移除频域模块后 ASR 升至 15.23%，验证了频域解耦在联邦参数高效微调场景同样具有实质贡献；移除探针模块后 ASR 升至 18.67%。与视觉任务相比，联邦 LoRA 微调场景中的 MMLU 变化相对较小，Full Fed-DCR 的 MMLU 为 51.67%，接近无攻击 FedAvg 的 52.50%，说明低秩适配场景下防御对通用能力的影响相对有限。总体来看，Fed-DCR 在显著降低

表 4　模块消融实验(CIFAR-10,α = 0.5,BadNets 攻击;联邦参数高效微调:Llama-2-7B+LoRA,α = 0.5)

配置	CIFAR-10 BA	CIFAR-10 ASR	联邦参数高效微调 MMLU	联邦参数高效微调 ASR
FedAvg(No Attack)	76.50%	—	52.50%	—
FedAvg(No Defense)	65.89%	90.69%	52.34%	82.45%
Full Fed-DCR	65.21%	3.24%	51.67%	8.45%
w/o 频域模块	64.89%	12.56%	51.38%	15.23%
w/o 谱一致性模块	65.34%	8.67%	51.45%	21.34%
w/o 探针模块	64.78%	9.45%	51.52%	18.67%
w/o 奇异值裁剪	65.12%	5.89%	51.61%	11.56%
w/o 参数平滑	65.45%	4.82%	51.63%	9.78%

ASR 的同时，不同任务中的主任务性能代价并不完全一致，实际部署时需要根据安全需求和性能容忍度调节异常惩罚系数与融合权重。

5.5　双域检测有效性分析

为进一步验证频域与参数域两个起始检测域在恶意客户端识别中的实际贡献，本文从检测层面对单域检测与双域融合检测进行比较分析。在实验评估过程中，执行投毒训练的客户端身份可由攻击设置和训练记录确定，因此本文以每轮参与训练客户端是否执行投毒训练作为真实标签，以对应异常分数作为预测分数，统计 TPR、FPR、F1 和 AUC 等检测指标。其中，TPR 用于衡量投毒客户端被正确识别或显著抑制的比例，FPR 用于衡量良性客户端被误判或过度抑制的比例，F1 综合反映检测精度与召回能力，AUC 则用于评价异常分数在不同阈值下区分良性更新与恶意更新的整体能力。

Fed-DCR 采用连续降权机制而非简单剔除策略，因此本文在计算二值化检测指标时不额外调整检测阈值，而是采用与聚合权重修正一致的异常判定规则。当客户端修正后的聚合权重低于其原始聚合权重的 50% 时，认为该客户端在当前轮被显著抑制，并将其判定为异常客户端；AUC 指标则直接基于连续异常分数计算，不依赖固定阈值。本节选取 FIBA 和 Instruction-BD 分别作为频域扰动主导型攻击和语义/低秩结构偏移主导型攻击，用于分析单域检测与双域融合检测的差异。其中，FIBA 可检验频域分解、低频表示对齐和高频 PSD 聚类对谱结构异常的识别能力；Instruction-BD 可检验 LoRA 谱一致性检测和因果探针对低秩结构偏移及行为响应异常的识别能力。实验结果如表 5 所示。

表 5　单域检测与双域融合检测性能比较

攻击类型	检测配置	TPR ↑	FPR ↓	F1 ↑	AUC ↑	ASR ↓
FIBA	仅频域	0.83	0.12	0.67	0.89	21%
FIBA	仅参数域	0.55	0.16	0.48	0.71	40%
FIBA	双域融合	0.91	0.09	0.75	0.94	9%
Instruction-BD	仅频域	0.48	0.18	0.43	0.69	45%
Instruction-BD	仅参数域	0.80	0.11	0.68	0.87	26%
Instruction-BD	双域融合	0.88	0.07	0.77	0.93	12%

由表 5 可见，不同攻击类型对两个检测域的敏感性存在差异。在 FIBA 攻击下，频域检测的 TPR 和 AUC 高于参数域检测，说明频域特征能够更直接地捕捉频率触发和局部扰动造成的谱结构异常；而在 Instruction-BD 攻击下，参数域检测优于频域检测，说明语义触发和低秩微调后门更容易表现为低秩结构偏移或模型行为响应异常。双域融合检测在两类攻击下均取得更稳定的检测性能，并进一步降低 ASR，表明频域和参数域提供的是互补证据，而非简单重复信息。

5.6　Non-IID 鲁棒性分析

联邦大模型场景下各参与方的领域数据差异可能远超传统视觉任务，频域解耦机制在极端 Non-IID 条件下的检测稳定性需要验证。通过将 Dirichlet 参数 α 从 1.0 推至极端的 0.1，在 CIFAR-10 数据集 BadNets 攻击下评估各方案的防御性能结果具体如图 4 所示。结果揭示了两个关键发现。第一，异质性会对现有方案产生显著的非线性恶化效应。FLTrust 的 ASR 由 $\alpha = 1.0$ 时的 12.34% 升至 $\alpha = 0.1$ 时的 42.56%，退化约 30 个百分点；AlignIns 的 ASR 由 6.23% 升至 15.67%，在极端 Non-IID 条件下明显超出理想阈值。第二，Fed-DCR 在 $\alpha = 1.0$ 到 $\alpha = 0.1$ 的变化过程中，ASR 仅由 2.12% 升至 6.78%，退化幅度约 5 个百分点，不足 AlignIns 的一半；BA 由 71.45% 降至 54.89% 的变化趋势与无防御基线基本保持一致，各 α 设置下的额外损失均不超过 2.5 个百分点。

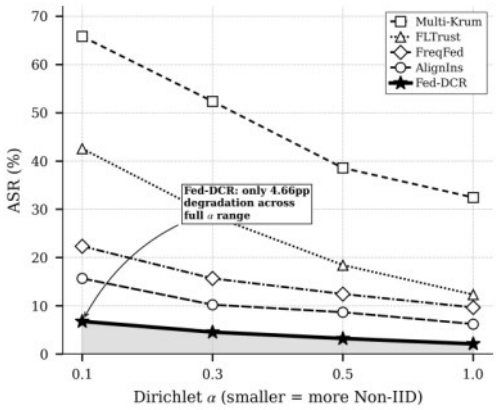

图 4　不同 Non-IID 程度下各方案 ASR 对比

该稳定性与频域解耦机制提供的补充判别信息有关。Non-IID 引起的良性异质性通常更多体现在语义相关的低频成分，而部分后门触发器特征更容易表现为局部高频扰动。频域分解能够在一定程度上增强两类信号的可分性，从而降低良性异质性变

化对异常判别的影响。

5.7　泛化能力分析

为验证 Fed-DCR 向更复杂任务迁移的能力，在 CIFAR-100 和 Tiny-ImageNet 上进行跨数据集泛化实验，实验条件为 $\alpha = 0.5$、BadNets 攻击。在 CIFAR-100 的 100 类细粒度分类任务上，Fed-DCR 的 ASR 为 4.89%，AlignIns 为 12.34%，FreqFed 为 18.34%；在 Tiny-ImageNet 的 200 类高分辨率任务上，Fed-DCR 的 ASR 为 5.67%，AlignIns 为 13.79%，FreqFed 为 21.23%。随任务复杂度增加基线方案退化加速，而 Fed-DCR 的退化幅度显著更小，真阳性率（True Positive Rate, TPR）保持在 92% 以上，假阳性率（False Positive Rate, FPR）低于 5%，表明频域解耦机制对任务复杂度具有良好的缩放特性。

以上实验中恶意客户端比例固定为 10%，而实际部署中攻击者可能控制更多参与方。进一步评估恶意比例从 10% 增至 30% 时各方案的防御性能变化具体如表 6 所示。

表 6　不同恶意客户端比例下的防御性能(CIFAR-10, BadNets,$\alpha = 0.5$)

ρ	Fed-DCR ASR	Fed-DCR BA	AlignIns ASR	FLTrust ASR
10%	3.24%	65.21%	8.67%	18.45%
20%	6.34%	63.46%	18.87%	35.37%
30%	11.15%	62.21%	34.59%	49.98%

随 ρ 从 10% 增至 30%，FLTrust 和 AlignIns 的 ASR 分别升至 49.98% 和 34.59%，而 Fed-DCR 仅升至 11.15%，退化幅度约为 AlignIns 的 1/3，且 BA 下降与无防御基线同步，表明防御模块未引入额外良性性能损失。在系统扩展性方面，客户端规模从 K =100 扩展至 K=1000 且每轮参与率保持 10%，ASR 从 3.24% 缓慢升至 4.87%，服务器单轮防御计算耗时增幅仅 5.4%；客户端掉线率达 30% 时 ASR 仍低于 6%、BA 相对于零掉线基准的损失不超过 1.6 个百分点，表明方案具备良好的实际部署鲁棒性。

综合以上实验，Fed-DCR 在联邦参数高效微调与视觉任务、多种攻击类型、自适应攻击、强 Non-IID 条件及中高恶意比例等设定下均保持有效防御，各模块贡献通过消融实验得到验证。

5.8　关键参数设定与影响分析

Fed-DCR 涉及的关键参数主要包括频域划分比例、DBSCAN 聚类尺度、谱一致性判别阈值以及双域融合权重。上述参数会影响异常分数的敏感性和聚合降权强度，因此本文对其设定依据及影响进行说明。

1) 频域划分比例。本文将 DCT 系数矩阵左上角 1/4 区域作为低频分量，其余部分作为高频分量。该设置主要考虑低频分量通常包含全局语义和主要结构信息，而高频分量更容易反映局部触发器、纹理扰动和异常注入模式。若低频区域过小，部分语义相关更新会被误划入高频区域，可能增加良性客户端误判；若低频区域过大，高频触发扰动会被稀释，可能降低对频域攻击和局部触发器的敏感性。因此，1/4 划分在语义保留和扰动分离之间取得了相对平衡。

2) 聚类尺度。高频谱密度聚类采用基于当前轮 PSD 特征两两距离中位数的自适应尺度，而非固定距离阈值。该设计使聚类尺度能够随不同通信轮次的更新分布自动调整。若聚类尺度过小，良性客户端可能被划分为多个稀疏簇，导致误报增加；若聚类尺度过大，恶意更新可能被并入主簇，导致漏报增加。因此，自适应聚类尺度有助于提升 Non-IID 条件下的稳定性。

3) 谱一致性判别阈值。LoRA 谱一致性检测用于衡量客户端低秩更新与参考子空间之间的偏离程度。阈值过低会使部分正常领域适配更新被误判为异常，影响主任务性能；阈值过高则可能放过隐藏在低秩子空间中的后门更新。为降低固定阈值对不同任务的敏感性，本文在每轮客户端集合内对谱一致性分数进行归一化，并结合参数域综合分数参与后续降权，而不是仅依赖单一硬阈值完成过滤。

4) 双域融合权重。频域分数和参数域分数分别反映扰动模式异常和低秩结构异常。视觉任务中，触发器往往体现为较明显的局部纹理或频率扰动，频域分数贡献相对更直接；联邦 LoRA 微调任务中，后门更可能隐藏于低秩适配器的结构偏移中，参数域分数的作用更为重要。本文采用双域分数融合而非单域硬判别，使不同攻击类型下的异常证据能够互相补充。若某一域权重过高，方案会退化为近似单域检测，降低对自适应攻击的鲁棒性；适当平衡两个域的权重有助于在攻击抑制和主任务

性能之间取得折中。

总体来看，Fed-DCR 的关键参数主要影响安全性与主任务性能之间的权衡。较严格的异常判别和较强的降权会进一步降低攻击成功率，但可能增加良性更新被抑制的风险；较宽松的判别策略有利于保持主任务性能，但可能降低对隐蔽后门的抑制能力。因此，在实际部署中可根据应用场景的安全需求调整频域划分、聚类尺度、谱一致性阈值和双域融合权重。

6 结束语

本文提出 Fed-DCR 双域后门防御方案，在频域与参数域两个观测域分别构建判别依据，使频域解耦、LoRA 谱一致性检测和因果探针三个模块分别锚定扰动模式、低秩结构异常和模型行为响应特征，并形成协同防御。实验从视觉基础任务到 Llama-2-7B 联邦 LoRA 微调场景进行递进验证，结果表明，Fed-DCR 能够有效降低多种后门攻击的攻击成功率，并在多数设置下保持可接受的主任务性能和系统开销。

同时，Fed-DCR 仍存在一定局限性。在极端 Non-IID 条件下，良性更新异质性可能干扰异常分数计算；当恶意客户端比例超过 30% 时，聚类、归一化和降权机制可能受到污染；在极低秩 LoRA 设置下，低秩子空间的可分性可能下降；面对未知新型后门或更强自适应攻击时，现有探针扰动空间可能难以覆盖全部潜在触发通道。后续工作将从弱验证集鲁棒检测、自适应融合权重学习、极低秩子空间建模和未知触发器探针自动生成等方面进一步改进。

参考文献：

[1] Obeidat H, Shuaieb W, Obeidat O, et al. A review of indoor localization techniques and wireless technologies[J]. Wireless Personal Communications, 2021, 119(1): 289-327.
Kairouz P, McMahan H B, Avent B, et al. Advances and open problems in federated learning[J]. Foundations and Trends in Machine Learning, 2021, 14(1-2): 1-210.

[2] Li Q, Diao Y, Chen Q, et al. Federated learning on non-IID data silos: an experimental study[C]//Proceedings of the 38th IEEE International Conference on Data Engineering. Piscataway: IEEE Press, 2022: 965-978.

[3] Hu E J, Shen Y, Wallis P, et al. LoRA: low-rank adaptation of large language models[C]//Proceedings of the 10th International Conference on Learning Representations. Vienna: ICLR, 2022.

[4] Bai J, Chen D, Qian B, et al. Federated fine-tuning of large language models under heterogeneous tasks and client resources[C]//Proceedings of the 38th Annual Conference on Neural Information Processing Systems. Massachusetts: MIT Press, 2024.

[5] Yan Y, Feng C M, Zuo W, et al. Federated residual low-rank adaptation of large language models[C]//Proceedings of the 13th International Conference on Learning Representations. Vienna: ICLR, 2025.

[6] Ning W, Wang J, Qian Y, et al. Federated fine-tuning on heterogeneous LoRAs with error-compensated aggregation[J]. IEEE Transactions on Neural Networks and Learning Systems, 2025, 36(10): 17826-17840.

[7] 姜毅, 杨勇, 印佳丽, 等. 大语言模型安全与隐私风险综述[J]. 计算机研究与发展, 2025, 62(8): 1979-2018.
Jiang Y, Yang Y, Yin J L, et al. A survey on security and privacy risks of large language models[J]. Journal of Computer Research and Development, 2025, 62(8): 1979-2018.

[8] Bagdasaryan E, Veit A, Hua Y, et al. How to backdoor federated learning [C]//Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics. New York: PMLR, 2020: 2938-2948.

[9] Xie C, Huang K, Chen P Y, et al. DBA: distributed backdoor attacks against federated learning[C]//Proceedings of the 8th International Conference on Learning Representations. Vancouver: ICLR, 2020.

[10] Nguyen A, Tran A. WaNet: imperceptible warping-based backdoor attack[C]//Proceedings of the 9th International Conference on Learning Representations. Vienna: ICLR, 2021.

[11] Feng Y, Ma H, Liu X, et al. FIBA: frequency-injection based backdoor attack in medical image analysis[C]//Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2022: 20876-20885.

[12] Zhang Z, Panda A, Song L, et al. Neurotoxin: durable backdoors in federated learning[C]//Proceedings of the 39th International Conference on Machine Learning. New York: PMLR, 2022: 26429-26446.

[13] Zhang H, Chen E, Lin B, et al. A3FL: adversarially adaptive backdoor attacks to federated learning[C]//Proceedings of the 37th Annual Conference on Neural Information Processing Systems. Massachusetts: MIT Press, 2023.

[14] Cao X, Fang M, Liu J, et al. FLTrust: Byzantine-robust federated learning via trust bootstrapping[C]//Proceedings of the 2021 Network and Distributed System Security Symposium. Reston: Internet Society, 2021.

[15] Blanchard P, El Mhamdi E M, Guerraoui R, et al. Machine learning with adversaries: Byzantine tolerant gradient descent[C]//Proceedings of the 31st Annual Conference on Neural Information Processing Systems. Massachusetts: MIT Press, 2017: 119-129.

[16] Nguyen T D, Rieger P, Chen H, et al. FLAME: taming backdoors in federated learning[C]//Proceedings of the 31st USENIX Conference on Security Symposium. New York: ACM Press, 2022: 1415-1432.

[17] Rieger P, Nguyen T D, Miettinen M, et al. DeepSight: mitigating backdoor attacks in federated learning through deep model inspection[C]//Proceedings of the 2022 Network and Distributed System Security Symposium. Reston: Internet Society, 2022.

[18] Xu J, Zhang Z, Hu R. Detecting backdoor attacks in federated learning via direction alignment inspection[C]//Proceedings of the 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2025: 20654-20664.

[19] Fereidooni H, Pegoraro A, Rieger P, et al. FreqFed: a frequency analysis-based approach for mitigating poisoning attacks in federated

learning[C]//Proceedings of the 2024 Network and Distributed System Security Symposium. Reston: Internet Society, 2024.

[20] Wang Y, Xue D, Zhang S, et al. BadAgent: inserting and activating backdoor attacks in LLM agents[C]//Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2024: 9817-9839.

[21] Sun Z, Cong T, Liu Y, et al. PEFTGuard: detecting backdoor attacks against parameter-efficient fine-tuning[C]//Proceedings of the 2025 IEEE Symposium on Security and Privacy. Piscataway: IEEE Press, 2025: 1713-1731.

[22] Min N M, Pham L H, Li Y, et al. CROW: eliminating backdoors from large language models via internal consistency regularization[C]//Proceedings of the 42nd International Conference on Machine Learning. New York: PMLR, 2025.

[23] Liu H, Zhong S, Sun X, et al. LoRATK: LoRA once, backdoor everywhere in the share-and-play ecosystem[C]//Findings of the Association for Computational Linguistics: EMNLP 2025. Stroudsburg: ACL, 2025: 23009-23047.

[24] Liu S, Pan Y, Hong K, et al. Backdoor threats in large language models: a survey[J]. Science China Information Sciences, 2025, 68(9): 190101.

[25] 刘嘉浪, 郭延明, 老明瑞, 等. 基于联邦学习的后门攻击与防御算法综述[J]. 计算机研究与发展, 2024, 61(10): 2607-2626.
Liu J L, Guo Y M, Lao M R, et al. A survey on backdoor attack and defense algorithms based on federated learning[J]. Journal of Computer Research and Development, 2024, 61(10): 2607-2626.

[26] Li S, Dai Y. BackdoorIndicator: leveraging OOD data for proactive backdoor detection in federated learning[C]//Proceedings of the 33rd USENIX Conference on Security Symposium. New York: ACM Press, 2024.

[27] Huang W, Ye M, Shi Z, et al. Parameter disparities dissection for backdoor defense in heterogeneous federated learning[C]//Proceedings of the 38th Annual Conference on Neural Information Processing Systems. Massachusetts: MIT Press, 2024.

[28] 张佳乐, 朱诚诚, 成翔, 等. 基于对比训练的联邦学习后门防御方法[J]. 通信学报, 2024, 45(3): 182-196.
Zhang J L, Zhu C C, Cheng X, et al. Backdoor defense method in federated learning based on contrastive training[J]. Journal on Communications, 2024, 45(3): 182-196.

[29] Zhang J, Li S, Chen H, et al. FLPurifier: backdoor defense in federated learning via decoupled contrastive training[J]. IEEE Transactions on Information Forensics and Security, 2024, 19: 9303-9319.

[30] Dai Y, Li S, Gan Z, et al. TrojanDam: detection-free backdoor defense in federated learning through proactive model robustification utilizing OOD data[J]. arXiv Preprint, arXiv:2504.15674, 2025.

[31] Liu G, Lin W, Huang T, et al. Prodigal: backdoor defense for federated learning beyond robust aggregation[J]. Knowledge-Based Systems, 2026, 333: 114903.

[32] Li Y, Huang H, Zhao Y, et al. BackdoorLLM: a comprehensive benchmark for backdoor attacks and defenses on large language models[C]//Proceedings of the 39th Annual Conference on Neural Information Processing Systems: Datasets and Benchmarks Track. Massachusetts: MIT Press, 2025.

[33] 王柳, 王申奥, 侯心怡, 等. 大语言模型存储机制安全风险综述[J]. 计算机研究与发展, 2026, 63(3): 640-667.
Wang L, Wang S A, Hou X Y, et al. A survey on security risks of storage mechanisms in large language models[J]. Journal of Computer Research and Development, 2026, 63(3): 640-667.

[34] McMahan H B, Moore E, Ramage D, et al. Communication-efficient learning of deep networks from decentralized data[C]//Proceedings of the 20th International Conference on Artificial Intelligence and Statistics. New York: PMLR, 2017: 1273-1282.

[35] Touvron H, Martin L, Stone K, et al. Llama 2: open foundation and fine-tuned chat models[J]. arXiv Preprint, arXiv:2307.09288, 2023.

康杰（1989– ），女，河北保定人，河北大学在读博士，河北金融学院讲师，主要研究方向为后门防御、网络空间安全、信息安全治理。

杜瑞忠（1975– ），女，河北保定人，武汉大学信息安全专业博士，博士生导师，河北大学网络空间安全与计算机学院院长，研究方向为数据安全与隐私保护。

梁晓艳（1981– ），女，河北衡水人，北京航空航天大学计算机应用专业博士，硕士生导师，中文信息学会大数据安全与隐私计算专业委员会委员，河北大学副教授，主要研究方向为网络安全、隐私保护。

张晓羽（1986– ），男，河北保定人，北京外国语大学博士，华北电力大学讲师，主要研究方向为数据安全与隐私保护。

石朋亮（1992– ），男，河北保定人，河北大学实验师，主要研究方向为数据安全。