Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Decoupled Text-Guided Distillation for Efficient Federated Learning on Edge Devices

Younghan Kim
Gachon University
Yongjae Park
Sejong University
Jae Won Cho
Konkuk University
Jungchan Cho
thinkai@gachon.ac.kr

Gachon University

Article

Keywords:

Posted Date: March 4th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8883453/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Decoupled Text-Guided Distillation for Efficient Federated Learning on Edge Devices

Younghan Kim[1], Yongjae Park[2], Jae Won Cho[3,*], and Jungchan Cho[1,*]

[1]College of IT Convergence, Gachon University, Seongnam, South Korea
[2]Department of Artificial Intelligence, Sejong University, Seoul, South Korea
[3]Department of Electrical and Electronics Engineering, Konkuk University, Seoul, South Korea
[*]Corresponding authors: chojw@konkuk.ac.kr, thinkai@gachon.ac.kr

ABSTRACT

Federated Learning (FL) enables the collaborative training of models across heterogeneous edge devices while preserving data privacy; however, its performance degrades significantly under domain shift. While integrating Vision-Language Models (VLMs) can mitigate this, existing prompt-tuning methods typically remain coupled to the massive VLM backbone during inference, rendering them impractical for resource-constrained edge devices. To address this challenge, we propose CLIP-assisted Domain-Invariant Federated Learning (CDIFed), which decouples the VLM from the deployment model to enhance robustness without incurring high inference latency. This framework integrates a Text-Guided Domain Adapter, implemented as a parameter-efficient bottleneck module, which aligns visual features with invariant text-based anchors to filter domain-specific noise while maintaining class-discriminative semantics. CDIFed operates through a communication-efficient two-phase framework: clients first adapt a frozen CLIP teacher, and then the adapted teacher supervises the training of a lightweight student network via feature knowledge distillation. Unlike previous approaches, the heavy VLM is discarded after adaptation, and only the student model parameters are transmitted to the server for aggregation. Experiments on the Digits and Office-Caltech benchmarks demonstrate that CDIFed significantly outperforms state-of-the-art methods in federated domain generalisation while maintaining the inference efficiency required for heterogeneous edge devices.

1 Introduction

The proliferation of data generated on resource-constrained edge devices has created immense opportunities for training powerful machine learning models. However, centralising this data can be unworkable because of strict privacy regulations, communication overheads, and the bandwidth limitations inherent to distributed edge networks. To mitigate these constraints, Federated Learning (FL) has been established as a pivotal framework. Within this paradigm, numerous local clients participate in the joint optimisation of a global model orchestrated by a central node, while retaining their raw datasets locally to ensure confidentiality. The standard process, demonstrated by the FedAvg[1] algorithm, involves clients training a model on their local data and transmitting only the updated model parameters to the server for aggregation.

Despite its efficacy, FL faces substantial hurdles due to the intrinsic non-IID nature of local data. As reviewed by Bhanbhro *et al.*[2], these hurdles are particularly acute in heterogeneous edge computing environments, where differences in device capabilities, data distribution, and model requirements significantly degrade performance. A predominant challenge lies in statistical heterogeneity, characterised by divergent data distributions among participants (e.g., variations in $P(y)$). To address this in resource-constrained edge systems, recent methods such as those by Wu *et al.*[3] have employed strategies like prototype-based knowledge distillation and two-stage feature extraction to manage non-IID data. In the context of image classification, this implies that a specific client might exclusively hold samples from a subset of categories, whereas others contain disjoint label sets. Methods such as FedProx[4] have tried to mitigate the "client drift" caused by this type of heterogeneity by regularising local training. Recent advancements also include joint exploration of drift and forgetting[5], privacy-preserving AI frameworks for smart cities[6], and adaptive hierarchical clustering for resource-constrained IoT devices[7].

Furthermore, a distinct yet critical obstacle in practical deployments is domain shift. This phenomenon manifests as covariate shift, wherein the marginal distribution of input features, denoted as $P(x)$, exhibits significant discrepancies across clients, notwithstanding the alignment of the label space.[9]. For example, in real-world IoT deployments, a smart home security system must recognise objects (e.g., 'person', 'package') across a fleet of heterogeneous cameras. One user may deploy a high-definition indoor camera with warm lighting, while another uses a doorbell camera with a fisheye lens and night-vision infrared capabilities. Similarly, in mobile computing, image signal processors (ISPs) from different smartphone manufacturers produce distinct colour profiles and noise patterns for the same scene. In these cases, a model trained on data dominated by one specific device type or environmental condition learns features specific to that domain's visual style (e.g., the green tint of night

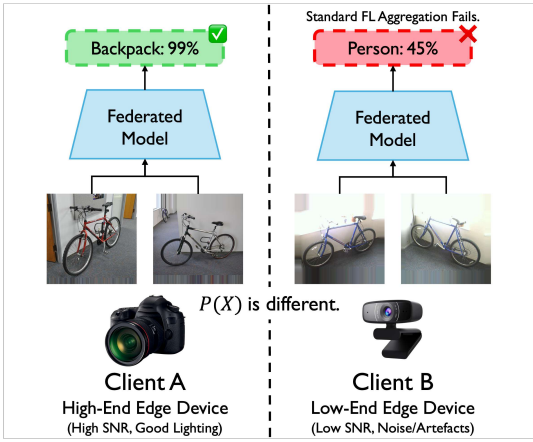


Figure 1. In a Federated Learning setup, a model often learns features specific to high-quality data (left). When deployed to heterogeneous edge devices (right), the standard global model fails to recognise the same object due to statistical divergence in visual styles (covariate shift), despite the semantic label remaining the same.

vision). Standard aggregation methods, such as FedAvg, fail in this setting because averaging the parameters of models trained on such visually distinct sensor domains results in a global model that generalises poorly across the ecosystem.

To directly address this issue, the field of Federated Domain Generalisation (FDG) has emerged. Previous work in FDG has largely focused on three categories of methods. Some methods employ regularisation techniques to encourage the model to learn domain-invariant representations by suppressing domain-specific correlations[10,11]. Others propose architectural modifications, such as disentangling the model into shared domain-invariant and private domain-specific components[12,13]. A third line of work attempts to explicitly align the feature distributions from different client domains[14]. While these approaches have advanced the field, they are often applied to networks trained from scratch. Their effectiveness can be limited when the domain discrepancy across clients is large, as learning robust, generalisable features from scratch under these conditions is fundamentally difficult[15].

This limitation has motivated a shift towards leveraging large-scale pre-trained models. In particular, Vision-Language Models (VLMs) such as Contrastive Language-Image Pre-training (CLIP)[16] offer a promising foundation. By utilising a massive dataset of hundreds of millions of image-text pairs, CLIP establishes a shared embedding space that directly connects visual imagery with human language. This process compels the model to learn abstract, semantic features that are inherently more robust to domain variations, as it learns the core concept of an object rather than its context-specific appearance[17].

In this study, we propose a novel framework, CLIP-assisted Domain Invariant Federated Learning (CDIFed), to explicitly leverage the domain generalisation capabilities of CLIP within a communication-efficient FL setting. An overview of the proposed framework is illustrated in Fig. 2. Directly federating a large model such as CLIP is infeasible due to excessive communication costs. Therefore, CDIFed employs a two-phase knowledge transfer process. First, on each client, a pre-trained CLIP model is adapted to the local data domain using a parameter-efficient adapter module. This creates a powerful, specialised "teacher" model without updating the entire network. The adapter's bottleneck structure acts as a regulariser, filtering out domain-specific noise while retaining semantic information. Second, the rich, domain-invariant knowledge from this teacher is distilled into a much smaller, lightweight "student" model (e.g., a ResNet[18]). Crucially, only the parameters of these compact student models are transmitted to the server for aggregation. This decoupled approach ensures training stability by providing a fixed teacher target and improves efficiency by allowing the computationally heavy text encoder to be offloaded after the initial adaptation. Crucially, this method preserves the low communication overhead necessary for practical FL, while still enriching the final global model with CLIP's robust features.

Our contributions are described as follows:

1. We propose CDIFed, a federated learning framework which suppresses domain-specific visual noise using a text-guided domain filter and through fine-tuning of a local CLIP-based teacher with knowledge distillation to a lightweight student model.

2. We introduce a decoupled two-phase training strategy that utilises the domain-invariant properties of text supervision for adaptation, while ensuring communication efficiency and training stability during aggregation.

3. We conduct experiments on the Digits and Office-Caltech benchmarks, demonstrating that CDIFed significantly outperforms existing state-of-the-art FL methods.

2 Related Work

2.1 Federated Learning with Statistical Heterogeneity

Statistical heterogeneity in federated learning (FL), often termed Non-IID data, arises when client data distributions differ substantially, which can impede convergence of methods such as FedAvg[1]. This motivates two main research directions: mitigating heterogeneity to learn a single robust global model, or learning personalised models for each client.

The global model direction targets client drift, where local optimisation moves away from the global objective. A common remedy is to constrain local updates: FedProx[4] adds a proximal term that penalises large deviations from the global parameters, stabilising training without extra communication. Other methods explicitly correct drift: SCAFFOLD[19] uses control variates to adjust local gradients, while FedNova[20] normalises updates to account for heterogeneous computation and data sizes. FedDyn[21] further improves stability via dynamic regularisation. Complementary work improves aggregation; FedMA[22] constructs the global model by matching and averaging hidden units with similar feature signatures, avoiding naive averaging of permutable layers.

Personalised FL (pFL) instead argues that a single global model can be suboptimal under strong heterogeneity, and seeks client-specific models that still benefit from collaboration. In a multi-task framing, Ditto[23] learns a personalised model regularised towards a shared global model, enabling local specialisation. Model decoupling personalises only selected components; FedBN[24] keeps Batch Normalisation layers private to handle client-specific feature statistics while federating the remaining parameters.

However, these approaches primarily address label-space variation and are less suited to domain shift, where visual style differs markedly across clients. In contrast, our framework explicitly targets domain (covariate) shift by leveraging the semantic priors of Vision–Language Models to extract domain-invariant features.

2.2 Federated Domain Generalisation

The primary objective of Federated Domain Generalisation (FDG) is to train a robust global model using diverse, decentralised source domains, ensuring it maintains high performance on unseen target domains.

One major line of work focuses on applying regularisation techniques to encourage the model to learn domain-invariant representations. For instance, FedSR[10] employs a dual-regularisation strategy: it utilises L2-norm regularisation to synchronise marginal distributions among clients, alongside a conditional mutual information regulariser to filter out domain-specific correlations. Similarly, the Domain Adaptive Regularisation (DAR) module in DapperFL[11] applies a regularisation term to the model's encoder to learn robust features. Extending this idea to an unsupervised context, FedGaLA[25] identifies a link between domain shift and gradient misalignment, proposing a method that aligns gradients at both the local and global levels to promote generalisation from unlabelled data.

A second strategy involves modifying the model architecture or the aggregation process to separate domain-specific knowledge from domain-invariant knowledge explicitly. Methods like Disentangled Federated Learning (DFL)[12] and Federated Domain Shift Eraser (FDSE)[13] propose decomposing the model into two parts: a domain-invariant component that is shared and aggregated globally, and a domain-specific component that is kept private to each client. This disentanglement allows the model to learn generalisable features while still personalising for local data characteristics.

A third approach is to align the feature distributions across different domains directly. FedADG[14] accomplishes this through federated adversarial learning. Each client attempts to align its local data distribution with a common reference distribution. A key contribution in FedADG is that this reference distribution is not fixed but is adaptively generated to be the centroid of all source domains, which helps to minimise information loss during the alignment process.

2.3 Federated Learning with Pre-trained Models and CLIP

Incorporating large vision–language models (VLMs) such as CLIP into federated learning (FL) introduces practical concerns: transmitting the full VLM parameters is communication-inefficient. Recent work therefore emphasises parameter-efficient adaptation and knowledge transfer, aiming to preserve foundation-model generalisation whilst controlling compute, memory, and per-round communication.

A dominant approach is Parameter-Efficient Federated Tuning (PEFT), where clients freeze the VLM backbone and federate limited task-specific parameters. Adapter-based methods, such as FedCLIP, update only compact adapters. Prompt-based tuning further reduces trainable state; FedRSCLIP employs dual prompts for global and personal knowledge, ZeroDFL uses peer-to-peer prompt sharing, and DiPrompT learns disentangled prompts.

Despite their training efficiency, prompt and adapter methods generally remain coupled to the heavy VLM at inference: edge devices must still execute the CLIP backbone to apply learned prompts or adapters. Consequently, inference latency and memory footprint remain dictated by the underlying VLM architecture (e.g., ViT-B/ViT-L), which is prohibitive for many low-power edge devices and excludes weaker clients from participation.

To mitigate deployment cost, hybrid frameworks increasingly treat CLIP as a teacher and distil its knowledge into lightweight students. CLIP2FL adopts a client-side teacher–student design and, on the server, leverages the public CLIP text encoder to build semantic class prototypes for data-free re-balancing. TriplePlay proposes a related framework that integrates CLIP-derived supervision to improve robustness under non-IID distributions. However, these approaches typically distil from a static teacher or strictly bind student training to the server-side VLM, lacking a mechanism to specialise the teacher for local domains prior to distillation.

Our work follows this PEFT-and-distillation trajectory but introduces a strictly decoupled architecture. Each client first adapts a pre-trained CLIP model via a parameter-efficient adapter to obtain a domain-specialised teacher. Subsequently, the client trains a lightweight student through feature distillation. Only the student participates in the federation and is deployed. This decouples inference cost from VLM scale whilst retaining foundation-model supervision during training, enabling deployment on constrained devices without invoking the VLM.

3 Preliminaries

3.1 Contrastive Language-Image Pre-training (CLIP)

CLIP[16] functions as a multi-modal system aimed at aligning visual and linguistic information within a shared latent space. The architecture utilises a dual-encoder structure, consisting of an image encoder $E_I(\cdot)$ and a text encoder $E_T(\cdot)$. Given a batch of N image-text pairs $\{(x_i,t_i)\}_{i=1}^N$, where x_i is an image and t_i is the corresponding text description, the model generates normalised feature vectors $\mathbf{v}_i = E_I(x_i)/\|E_I(x_i)\|$ and $\mathbf{t}_i = E_T(t_i)/\|E_T(t_i)\|$.

The core learning goal is to enforce high cosine similarity between corresponding image-text pairs (x_i,t_i), while simultaneously distancing mismatched pairs within the same batch. This is achieved by optimising a symmetric cross-entropy loss over the similarity matrix, scaled by a learnable temperature parameter τ. This pre-training paradigm enables the model to learn robust, transferable visual features aligned with natural language concepts, facilitating effective zero-shot generalisation.

3.2 Parameter-Efficient Fine-Tuning (PEFT) with Adapters

Retraining all parameters of massive pre-trained foundations like CLIP is computationally prohibitive and memory-intensive for downstream applications. Also, fully fine-tuning can cause the catastrophic forgetting phenomenon, which causes the model to forget previously trained data. Parameter-Efficient Fine-Tuning (PEFT) strategies mitigate these issues by modifying only a fraction of the model's weights while keeping the bulk of the network static.

Adapters[33] represent a specific PEFT approach involving the injection of compact neural modules between the pre-existing layers of a frozen backbone. During the fine-tuning process, the original model weights remain frozen, and only the parameters of these newly added adapters are trained.

The adapter's signal is merged with the frozen layer's output via a skip connection. Typically, adapters utilise a bottleneck design that first projects features into a lower-dimensional space and then expands them back to the original dimension. This "down-up" projection mechanism significantly limits the trainable parameter count while serving as an information bottleneck to preserve general knowledge. By forcing the features through a compressed latent representation, the adapter is encouraged to retain essential semantic information while discarding high-frequency noise or domain-specific artefacts.

4 Methods

Our framework employs CLIP[16], as a source of domain-invariant knowledge. On each client, this knowledge is first fine-tuned to the local data distribution and then distilled into a lightweight student model. The student models, which are trained to mimic the feature representations of the fine-tuned CLIP, are the sole participants in the federated aggregation process. This approach enhances the generalisation capability of the final global model without incurring the high communication and computation costs associated with federating large models.

Our framework is built upon a standard federated learning architecture involving a central server and a collection of K distinct clients, denoted by $k \in \{1,\dots,K\}$. Each client k possesses a private local dataset $\mathcal{D}_k = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N_k}$, where $N_k = |\mathcal{D}_k|$ is the number of local samples. Each data point consists of an input $\mathbf{x}_i \in \mathcal{X}$ from an input space and a corresponding label $y_i \in \mathcal{Y}$ from a label space. The clients' data is not shared with the server or other clients.

The core challenge we address is federated domain generalisation. In this setting, the data distribution is heterogeneous across the clients. In this setup, the local dataset $\mathcal{D}_k$ possessed by each client is sampled from a unique probability distribution, represented as $P_k(\mathbf{x},y)$. We operate under the assumption of a covariate shift, which is characteristic of domain shift problems. This means that the marginal distribution of the inputs varies across clients, while the conditional distribution of labels given the inputs remains unchanged. Formally, for any two distinct clients k and j:

$$P_k(\mathbf{x}) \neq P_j(\mathbf{x}), \quad \text{but} \quad P_k(y|\mathbf{x}) = P_j(y|\mathbf{x}). \tag{1}$$

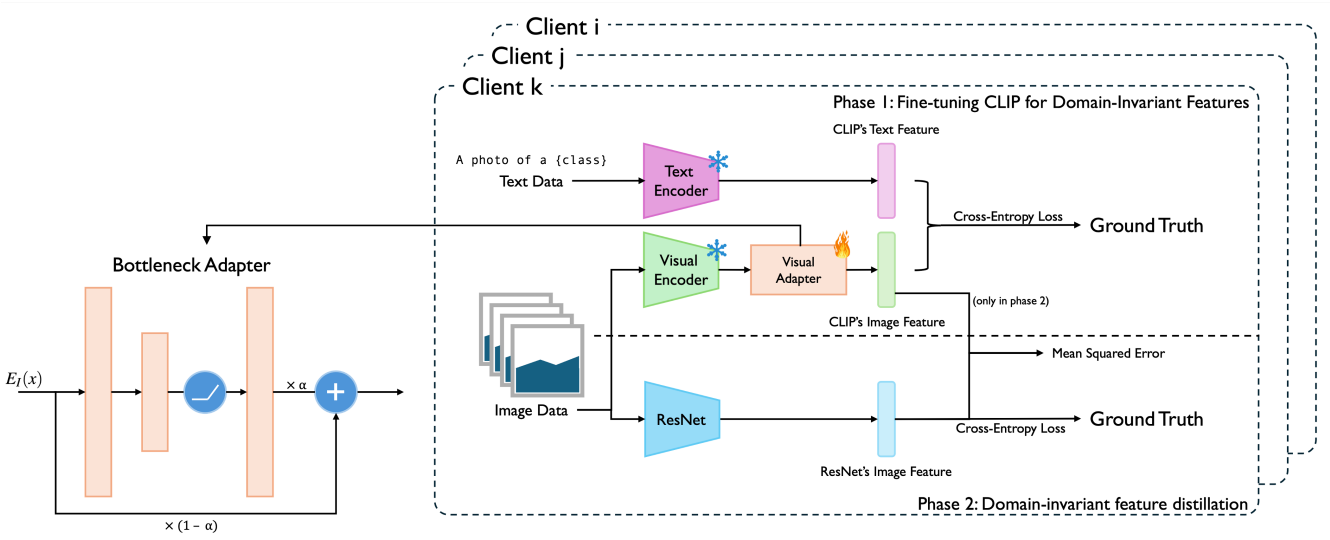


Figure 2. Architectural overview of the proposed CDIFed framework. On each client, a pre-trained CLIP model is adapted once (Phase 1). Subsequently, this adapted teacher model guides the training of a smaller student model via feature distillation (Phase 2). Only the parameters of the student model are shared with the central server for global aggregation.

Our objective is to learn a single global model $\mathbf{w}$ that generalises well across all K client domains. Let $f_{\mathbf{w}} : \mathcal{X} \to \mathcal{Y}$ denote the model and $\ell(\cdot,\cdot)$ be a task-specific loss function (e.g., cross-entropy). The local objective for each client k is to minimise its empirical risk on its own data:

$$L_k(\mathbf{w}) = \frac{1}{N_k} \sum_{(\mathbf{x}_i, y_i) \in \mathcal{D}_k} \ell(f_{\mathbf{w}}(\mathbf{x}_i), y_i). \tag{2}$$

The primary objective of this federated framework is to identify the optimal model parameters, $\mathbf{w}^*$, which effectively minimise the global loss function, $\mathcal{L}(\mathbf{w})$, which is the average of the local empirical risks:

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \mathcal{L}(\mathbf{w}) \quad \text{where} \quad \mathcal{L}(\mathbf{w}) = \sum_{k=1}^{K} \frac{N_k}{N} L_k(\mathbf{w}), \tag{3}$$

and $N = \sum_{k=1}^{K} N_k$ is the total number of samples across the network.

4.1 Framework Overview

We propose CLIP-assisted Domain Invariant Federated Learning (CDIFed), a framework designed to mitigate domain shifts by leveraging the robust representations of a pre-trained Vision-Language Model. The method decouples the extraction of domain-invariant features from the communication-constrained aggregation process. The local training procedure for each client k consists of two distinct phases (illustrated in Figure 2):

1. **Phase 1: Fine-tuning CLIP for Domain-Invariant Features.** Before joining the federated rounds, a local CLIP-based teacher model is fine-tuned on the private dataset $\mathcal{D}_k$.

2. **Phase 2: Domain-Invariant Feature Distillation.** In subsequent communication rounds, the adapted teacher guides the training of a lightweight student model via feature distillation.

This decoupled design is motivated by two key factors: stability and efficiency. First, end-to-end training (simultaneously optimising the adapter and the student) introduces instability, as the student model is forced to mimic a moving target whose feature space is constantly shifting. By fixing the teacher's parameters after Phase 1, we provide a stable objective for the distillation process. Second, the CLIP text encoder is only required to provide semantic supervision during the adaptation phase. By completing the adaptation in Phase 1, clients can offload the text encoder from memory during the recurring communication rounds of Phase 2, significantly reducing the computational burden on edge devices.

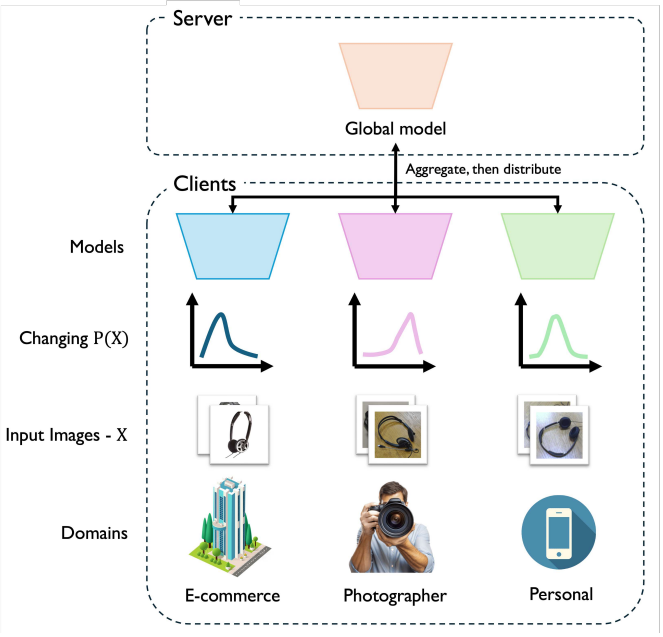


Figure 3. Each client sends their trained model to the server in every communication round. The server creates a global model by aggregating the transmitted models by federated averaging, then distributes the global model to all clients.

4.2 Phase 1: Fine-tuning CLIP for Domain-Invariant Features

To adapt the model to local domains without the computational cost of full fine-tuning, we employ a PEFT strategy using a residual adapter.

Adapter Architecture. Let $E_I(\cdot)$ and $E_T(\cdot)$ denote the frozen visual encoder and text encoder of the pre-trained CLIP model, respectively. We introduce a trainable adapter module $\mathcal{A}(\cdot)$, comprising a down-projection layer W_{down}, a non-linear activation σ (ReLU), and an up-projection layer W_{up}.

The adapter encoder for each client is formally defined as $\phi(\cdot)$, which represents a weighted aggregation of the frozen encoder's output and the adapter's transformation. For an input image x, the adapted feature representation $\phi(x)$ is formulated as:

$$\phi(x) = (1-\alpha)E_I(x) + \alpha \cdot \mathcal{A}(E_I(x)), \tag{4}$$

where $\mathcal{A}(\mathbf{z}) = W_{\text{up}}(\sigma(W_{\text{down}}\mathbf{z}))$. Here, $\mathbf{z} = E_I(x)$ represents the intermediate embedding from the frozen CLIP encoder, and $\alpha \in [0,1]$ is a hyperparameter controlling the residual blending ratio.

Beyond parameter efficiency, this bottleneck architecture serves as an information filter. By compressing the high-dimensional CLIP features into a low-rank latent space, the adapter acts as a structural regulariser[34,35]. It encourages the model to retain dominant semantic features while suppressing high-frequency domain artefacts or noise that do not contribute to the core concept.

Text-Guided Filtering Objective. During Phase 1, the parameters of the image encoder E_I and the text encoder E_T remain frozen; only the adapter parameters $\theta_{\mathcal{A}}$ are optimised using the local dataset $\mathcal{D}_k$. To construct the classification targets, we generate a text prompt T_c for each class $c \in \mathcal{C}$ (e.g., "a photo of a {class}"). These prompts are encoded and normalised to form static semantic anchors $\{\mathbf{t}_c\}_{c\in\mathcal{C}}$:

$$\mathbf{t}_c = \frac{E_T(T_c)}{\|E_T(T_c)\|}. \tag{5}$$

Similarly, we define the normalised adapted visual feature as $\mathbf{v} = \phi(x)/\|\phi(x)\|$. The loss function aims to maximise the cosine similarity score computed between the adapted image features and their corresponding ground-truth text anchors. The probability of the ground-truth label y is given by:

$$p(y|x) = \frac{\exp(\langle \mathbf{v}, \mathbf{t}_y \rangle / \tau)}{\sum_{c\in\mathcal{C}} \exp(\langle \mathbf{v}, \mathbf{t}_c \rangle / \tau)}, \tag{6}$$

where τ is the temperature parameter. Finally, we minimise the negative log-likelihood:

$$\mathcal{L}_{\text{adapter}}^{(k)} = \mathbb{E}_{(x,y)\sim\mathcal{D}_k}[-\log p(y|x)]. \tag{7}$$

This objective serves as the Text-Guided Domain Filter. The text embeddings $\mathbf{t}_c$ (e.g., representing "a photo of a backpack") act as static, domain-agnostic anchors. These embeddings encapsulate the semantic concept of the object but contain no information regarding lighting, camera noise, or background style.

To minimise $\mathcal{L}_{\text{adapter}}$, the adapter $\mathcal{A}$ must modify the visual features $\phi(x)$ to maximise their correlation with $\mathbf{t}_y$. Since $\mathbf{t}_y$ is invariant to the visual domain, the adapter is implicitly forced to suppress any information in $E_I(x)$ that is uncorrelated with $\mathbf{t}_y$—specifically, the domain-specific pixel statistics. Consequently, the bottleneck structure of the adapter combined with the text-alignment objective filters out high-frequency domain artefacts, retaining only the semantic structure required for classification.

4.3 Phase 2: Domain-Invariant Feature Distillation

In the federated learning rounds, we transfer knowledge from the frozen, adapted teacher to a trainable student network $f_{\text{student}}^{(k)}$. The student comprises a feature extractor h_k and a classifier head C_k,

$$f_{\text{student}}^{(k)}(x) = C_k(h_k(x)). \tag{8}$$

Feature Space Alignment. The feature dimension of the teacher, d_{teacher}, typically differs from that of the lightweight student, d_{student}. To facilitate feature-based distillation, we introduce a learnable linear projector $\mathcal{P}_k : \mathbb{R}^{d_{\text{teacher}}} \to \mathbb{R}^{d_{\text{student}}}$ to map the teacher's representations into the student's feature space. The projector is not aggregated and trained throughout the federation process.

Joint Objective. The student model and projector are optimised jointly. The total local objective $\mathcal{L}_{\text{total}}^{(k)}$ is a weighted sum of the task classification loss and the distillation loss:

$$\mathcal{L}_{\text{total}}^{(k)} = \mathcal{L}_{\text{cls}}^{(k)} + \lambda \mathcal{L}_{\text{distil}}^{(k)}, \tag{9}$$

where λ regulates the distillation strength.

The classification loss $\mathcal{L}_{\text{cls}}^{(k)}$ is the standard cross-entropy between the student's predictions and ground-truth labels:

$$\mathcal{L}_{\text{cls}}^{(k)} = \mathbb{E}_{(x,y)\sim\mathcal{D}_k}\left[\ell_{CE}(C_k(h_k(x)), y)\right]. \tag{10}$$

The distillation loss $\mathcal{L}_{\text{distil}}^{(k)}$ minimises the Mean Squared Error (MSE) between the student's features and the projected teacher's features, encouraging the student to mimic the domain-invariant structure of the CLIP embeddings:

$$\mathcal{L}_{\text{distil}}^{(k)} = \mathbb{E}_{x\sim\mathcal{D}_k}\left[\|h_k(x) - \mathcal{P}_k(\phi_k(x))\|_2^2\right]. \tag{11}$$

We explicitly select feature-based distillation over the more conventional logit-based distillation (e.g., minimising Kullback-Leibler divergence). We posit that logits primarily convey class probability distributions, which may not fully encapsulate the domain-invariant semantics learned by the teacher in Phase 1. In contrast, the intermediate feature embeddings $\phi_k(x)$ contain the rich, structural information aligned with the text anchors, effectively filtering domain-specific noise. By enforcing alignment in the feature space, the student is compelled to replicate the teacher's internal representation mechanism rather than merely mimicking its final decision confidence. This hypothesis is empirically validated in our ablation studies, where feature distillation demonstrates superior robustness compared to logit distillation.

After local updates, only the parameters of the student model $\{h_k, C_k\}$ are transmitted to the server for aggregation via Federated Averaging (Figure 3). Crucially, this ensures that the heavy CLIP backbone and the adapter are never required during deployment, as the student model becomes a standalone entity capable of domain-invariant inference. The full algorithm is described in Algorithm 1.

5 Experiments

To validate the proposed CDIFed framework, we performed extensive empirical tests focusing on its ability to handle domain shifts. This study addresses two primary research questions: (1) Can CDIFed achieve superior performance compared to existing state-of-the-art FL methods tailored for domain generalisation? (2) What is the specific impact of each individual module within our method on the final accuracy?

Algorithm 1: CLIP-assisted Domain Invariant Federated Learning (CDIFed)

Input : Pre-trained CLIP $\{E_I, E_T\}$, K Clients with data $\mathcal{D}_k$, Global Rounds T
Output : Global Student Model parameters $\mathbf{w}^T$

Server: Initialise global student model $\mathbf{w}^0$;
`/* Phase 1: One-time Teacher Adaptation */`
for *client* $k = 1, \ldots, K$ **do**
  Construct class text embeddings $\{\mathbf{t}_c\}$ using E_T;
  Initialise local Adapter $\mathcal{A}_k$;
  for *epoch* $m = 1, \ldots, M_{local1}$ **do**
    Sample batch $(x, y) \sim \mathcal{D}_k$;
    Update $\mathcal{A}_k$ by minimising $\mathcal{L}^{(k)}_{\text{adapter}}$ (Eq. 7);
  Fix Teacher $\phi_k(\cdot)$ with trained $\mathcal{A}_k$;
  Discard/Offload Text Encoder E_T;
`/* Phase 2: Federated Distillation Loop */`
for *round* $t = 0, \ldots, T-1$ **do**
  Server: Distribute $\mathbf{w}^t$ to all clients;
  for *client* $k = 1, \ldots, K$ **do**
    $\mathbf{w}^t_k \leftarrow \mathbf{w}^t$;
    Initialise/Load Projector $\mathcal{P}_k$;
    for *epoch* $m = 1, \ldots, M_{local2}$ **do**
      Sample batch $(x, y) \sim \mathcal{D}_k$;
      Compute teacher features $\mathbf{z} = \phi_k(x)$;
      Compute student features and logits using $\mathbf{w}^t_k$;
      Update $\mathbf{w}^t_k$ and $\mathcal{P}_k$ by minimising $\mathcal{L}^{(k)}_{\text{total}}$ (Eq. 9);
    Send updated $\mathbf{w}^{t+1}_k$ to Server;
  Server: Aggregation $\mathbf{w}^{t+1} \leftarrow \sum_{k=1}^{K} \frac{1}{K} \mathbf{w}^{t+1}_k$;
return $\mathbf{w}^T$

5.1 Experimental Setup

To ensure a fair comparison, we adopt an evaluation protocol from recent work in federated domain generalisation[11]. The details of using datasets and the federation protocol are described below.

Datasets and Data Partition We evaluate our method on two standard domain generalisation benchmarks: **Digits** and **Office-Caltech** as depicted in Figure 4.

- **Digits**: We utilised this standard evaluation suite for domain generalisation tasks, which encompasses four unique digit recognition subsets: MNIST[41], USPS[42], SVHN[43], and Synthetic Digits (SYN)[44]. While MNIST and USPS are comprised of simple grayscale handwritten samples, SVHN introduces greater complexity through coloured, real-world imagery of house numbers set against cluttered backgrounds. Finally, SYN offers synthesised character data, introducing a significant stylistic shift from the other real-world datasets, making it challenging for generalisation.

- **Office-Caltech**[45]: This dataset serves as a benchmark for real-world object recognition across four distinct domains: Caltech (C), which contains standard object photography; Amazon (A), characterised by products on clean white backgrounds commonly found in online stores; Webcam (W), which consists of low-fidelity images captured via typical web cameras; and DSLR (D), offering high-resolution photography with distinct lighting and focus effects. The domain shifts in this benchmark are representative of real-world variations caused by different camera sensors, lighting conditions, and image quality, providing a practical test of our method's capabilities.

The simulation environment involves a central server coordinating with 10 distinct clients. Adopting the protocol described by Jia et al.[11], we randomly distribute the 10 clients among the domains of a given benchmark, ensuring that each domain is assigned at least one client and each client is assigned to exactly one domain. To simulate a challenging non-IID and domain-shifted scenario, we generate local client datasets by randomly sampling a small proportion of the data from their

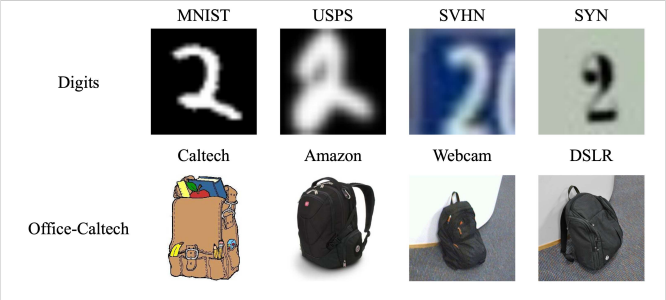


Figure 4. Example samples of each domain.

assigned domain. This proportion is set to 1% for the Digits benchmark and 20% for Office-Caltech, reflecting the varying complexity and scale of the datasets.

Implementation Details All experiments are implemented in PyTorch[46]. For the student models that participate in federated aggregation, we use a ResNet-10 model for the Digits benchmark and a ResNet-18 model[18] for the Office-Caltech benchmark. For our proposed CDIFed framework, the teacher model is a pre-trained CLIP model with a ResNet-50 visual backbone. For the text-guided adaptation in Phase 1, we utilise the standard prompt template "a photo of a {class}" to generate class embeddings.

The federated learning process is simulated for 100 communication rounds, with all 10 clients participating in each round. During local training for all frameworks, we use the Adam optimiser with a batch size of 64, a learning rate of 1e-3, and a weight decay of 1e-5. For CDIFed, the parameter-efficient adapter tuning (Phase 1) uses a learning rate of 1e-3. The feature distillation process (Phase 2) employs a student learning rate of 1e-4, and the distillation loss weight, λ, is set to 1.0. The adapter follows a bottleneck design with a reduction ratio of 4 (i.e., channel dimension $c_{\text{in}} \to c_{\text{in}}/4 \to c_{\text{in}}$). The adapter output is combined with the original post-encoder feature via residual blending with coefficient $\alpha = 0.2$. The values of M_{local1} and M_{local2} are 10 and 20, respectively.

5.2 Comparison with State-of-the-Art Methods

We evaluate all frameworks on the Digits and Office-Caltech benchmarks to assess their domain generalisation capabilities. The results are detailed in Table 1 and Table 2.

Performance on Digits Benchmark As shown in Table 1, our method, CDIFed, achieves a state-of-the-art global accuracy of 80.27%. This represents a significant improvement over all baseline methods, surpassing the next best-performing framework, DapperFL, by a margin of 5.97 percentage points. A key observation is the dramatic performance gain on the most challenging synthetic (SYN) domain, where CDIFed achieves 61.26% accuracy, nearly 24 points higher than the runner-up. This suggests that the domain-invariant features learned via CLIP distillation are particularly effective at generalising to artificial, out-of-distribution data. While maintaining highly competitive performance on traditional domains like MNIST and USPS, the exceptional result on SYN underscores our method's robustness against severe domain shifts.

Table 1. Comparison of model accuracy (%) on the Digits benchmark. Our method, CDIFed, demonstrates superior global accuracy, with a particularly strong performance on the challenging SYN domain.

Framework	MNIST	USPS	SVHN	SYN	Global
FedAvg[1]	95.89	86.84	78.39	33.63	71.81
MOON[36]	93.03	78.38	**84.45**	25.97	69.44
FedSR[10]	**96.77**	86.15	81.48	31.64	73.89
FPL[37]	95.54	87.69	83.74	34.73	74.17
FedDrop[38]	89.48	82.51	72.98	29.35	66.85
FedProx[4]	96.68	83.96	76.69	30.95	70.74
FedMP[39]	94.16	85.30	81.37	35.12	72.29
NeFL[40]	84.98	**88.49**	78.41	36.02	67.64
DapperFL[11]	96.25	86.30	82.45	37.26	74.30
CDIFed (Ours)	96.12	87.05	76.67	**61.26**	**80.27**

Table 2. Comparison of model accuracy (%) on the Office-Caltech benchmark. CDIFed achieves the highest accuracy on all individual domains and outperforms the runner-up by 7.92% globally.

Framework	Caltech	Amazon	Webcam	DSLR	Global
FedAvg[1]	66.07	76.84	65.52	56.67	64.54
MOON[36]	65.62	75.79	72.41	53.33	61.86
FedSR[10]	62.95	78.95	75.86	50.00	65.47
FPL[37]	63.84	**82.63**	65.52	60.00	65.45
FedDrop[38]	66.07	79.47	56.90	53.33	60.58
FedProx[4]	61.61	71.05	68.97	46.67	62.08
FedMP[39]	65.62	75.79	56.90	66.67	62.34
NeFL[40]	54.91	71.05	**77.59**	66.67	62.26
DapperFL[11]	64.73	81.58	74.14	66.67	67.75
CDIFed (Ours)	**71.88**	80.00	74.14	**76.67**	**75.67**

Performance on Office-Caltech Benchmark The superiority of CDIFed is further solidified on the more complex real-world images of the Office-Caltech benchmark, with results presented in Table 2. Our method achieves a global accuracy of 75.67%, outperforming all baselines. Notably, CDIFed not only achieves the best overall performance but also secures the highest accuracy on each of the four individual domains.

5.3 Ablation Studies

To validate the contributions of the core components within CDIFed, we conduct a series of ablation studies on the Office-Caltech benchmark. We analyse the performance impact of removing the adapter fine-tuning, replacing the CLIP teacher, and removing the student feature extractor. The results are summarised in Table 3.

Table 3. Ablation study of CDIFed's components on the Office-Caltech benchmark. Results show that each component is critical to achieving the final performance.

Variant	Global Accuracy (%)
CDIFed (Full Model)	**75.67**
Without Adapter Fine-tuning	73.14
ImageNet pre-trained Teacher	72.33
Federated Linear Classifier	12.32

Effect of Adapter Fine-tuning In this variant, we remove the adapter module and distil knowledge directly from a frozen, off-the-shelf CLIP model. The global accuracy drops from 75.67% to 73.14%. This degradation demonstrates the importance of Phase 1 of our method. By performing parameter-efficient fine-tuning, we create a fine-tuned teacher for each client, which provides more relevant and effective guidance to the student model than a generic, non-adapted CLIP model.

Effect of CLIP as Teacher To quantify the benefit of using a Vision-Language Model, we replace the CLIP teacher with an ImageNet-pretrained ResNet-50 model, as it is standard practice in knowledge distillation. This change results in a more significant performance degradation, with accuracy falling to 72.33%. This highlights that general-purpose representations learned by CLIP provide a substantially better foundation for domain-invariant feature distillation compared to models trained solely on image classification tasks like ImageNet.

Effect of Student Feature Extractor We analyse a variant where the ResNet student architecture is removed, and the system aggregates only a simple linear classifier. In this case, the student model becomes $f^{(k)}_{\text{student}}(x) = C_k(x)$, without the feature extractor $h(x)$ in Equation (8). This experiment isolates the contribution of the student's feature extraction capabilities. Performance degrades to 12.32%, a result comparable to random guessing. This confirms that aggregating a lightweight feature extractor is essential for retaining the domain-invariant knowledge distilled from the CLIP teacher, whereas a simple linear probe fails to capture the necessary semantic structure in a federated setting.

5.4 Scalability and Interpretability Analyses

Scalability with Number of Clients To evaluate the scalability of our approach, we conducted an experiment on the Digits benchmark, varying the number of participating clients from 10 to 50. As the number of clients increases, the global model

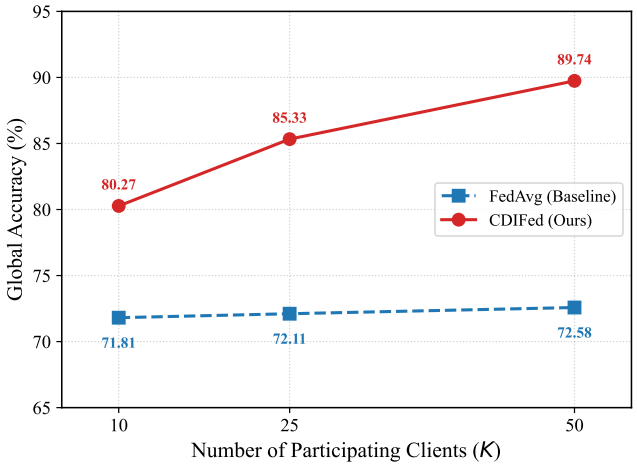


Figure 5. Scalability analysis on the Digits benchmark with varying numbers of participating clients ($K \in \{10, 25, 50\}$). While the performance of FedAvg stagnates due to its inability to resolve increased domain heterogeneity, CDIFed exhibits a significant performance gain (+17.16%) over FedAvg at $K = 50$, demonstrating its ability to effectively leverage larger, more diverse federated datasets.

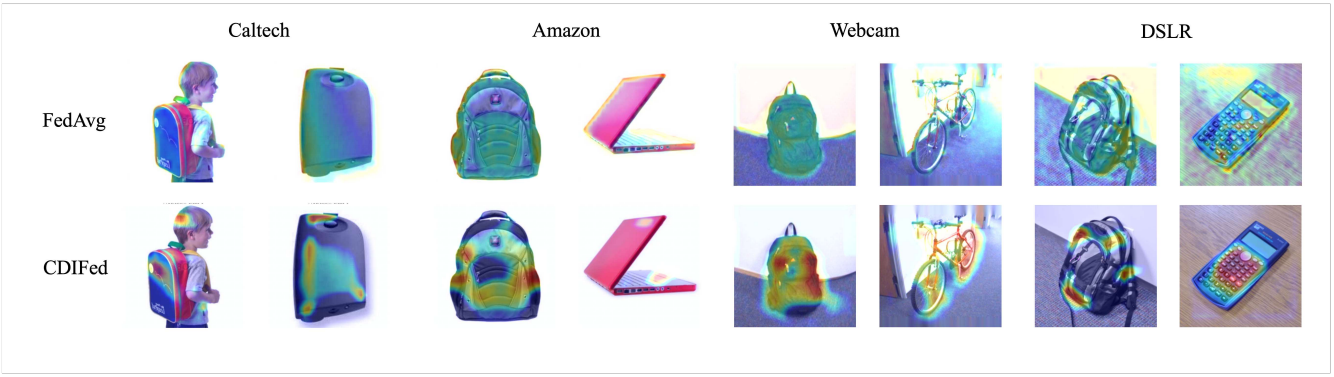


Figure 6. GradCAM visualisations for the backpack class across four domains from the Office-Caltech dataset. The top row (FedAvg) shows inconsistent and diffuse attention, often focusing on background elements. The bottom row (CDIFed) demonstrates consistent localisation of the object, suggesting its ability to learn domain-invariant features.

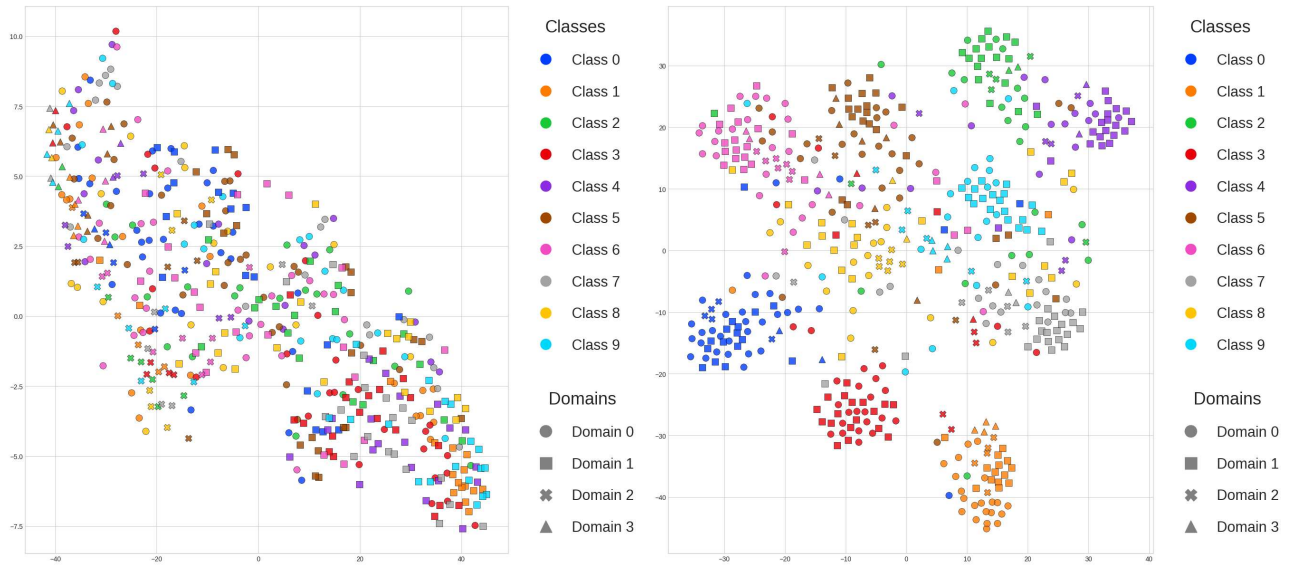
gains access to a broader distribution of data across different domains. Figure 5 presents the results of this experiment.

FedAvg shows only marginal improvement as more clients are added. This is because the standard aggregation scheme cannot effectively reconcile the conflicting updates from clients with disparate data distributions. Without a mechanism to learn shared representations, the global model fails to benefit from increased data diversity and cannot extract domain-invariant features.

In contrast, CDIFed's accuracy improves substantially with more clients, increasing by nearly 10 percentage points from 10 to 50 clients. This demonstrates that CDIFed successfully extracts domain-invariant knowledge and fully benefits from increased data exposure, highlighting its scalability and robustness in heterogeneous federated environments.

GradCAM Visualisation To qualitatively assess the learned representations, we employ Gradient-weighted Class Activation Mapping (GradCAM)[47] to visualise the model's attention. Figure 6 illustrates the class activation maps for the "backpack" class from the Office-Caltech dataset across four domains.

The top row shows that the baseline FedAvg's global model often fails to localise the object of interest. Its attention is diffused and frequently focuses on the background, indicating that the FedAvg captures domain-specific correlations. In contrast, the bottom row shows that the global model trained with CDIFed consistently focuses its attention on the backpack, regardless of the domain. The activation maps are concentrated on the object, which confirms that CDIFed learns domain-invariant and more generalisable representations.



(a) FedAvg Global Model **(b)** CDIFed Global Model

Figure 7. t-SNE visualisation of feature embeddings from the global models trained on the Office-Caltech benchmark. Colour denotes the class label, and marker shape denotes the source domain. (a) The feature space of the FedAvg model shows significant entanglement of class and domain features, indicating poor generalisation. (b) The CDIFed model learns features that are well-clustered by class, irrespective of domain.

Feature Space Visualisation To qualitatively assess the domain-invariance of the learned representations, we visualise the feature space of the final global models trained by FedAvg and our proposed CDIFed. We extract the feature embeddings from the penultimate layer of each global model for all samples in the Office-Caltech test sets and project them into a 2D space using t-SNE[48].

The results are presented in Figure 7. In these visualisations, each point represents an image sample, where the colour indicates its true class and the marker shape (e.g., circle, square, triangle) indicates its source domain. Figure 7a shows the feature space of the model trained with FedAvg. The features are highly scattered and intermingled, with only weak class-wise separation. Points of the same colour fail to cluster, and samples from different domains overlap significantly. This pattern reflects a form of representation collapse and indicates that the FedAvg global model is unable to learn stable, domain-invariant representations; its features remain entangled with domain-specific information and do not form meaningful clusters.

In contrast, Figure 7b illustrates the feature space of the CDIFed global model. We observe distinct, well-separated clusters, where each cluster predominantly consists of a single colour. Critically, these class-based clusters contain a mix of different marker shapes, signifying that samples from different domains are mapped to the same region of the feature space if they belong to the same class.

5.5 Analysis of Inference Efficiency and VLM Dependency

A critical distinction between CDIFed and prompt-learning approaches is the computational requirement during deployment. Methods such as DiPrompT[26] typically retain the heavy VLM backbone (e.g., ResNet-50 or ViT) at inference time to process learned prompts. In contrast, CDIFed is designed to decouple the inference process from the VLM, deploying only the lightweight student model. To quantify this trade-off, we conducted a comparative analysis on the Office-Caltech benchmark,

Table 4. Comparison of Accuracy and Inference Cost on Office-Caltech.

Method	Inference Model	GFLOPs	# of param.	Accuracy
DiPrompT[26]	CLIP (ResNet-50)	4.7	38.0 M	89.93
CDIFed (Ours)	**ResNet-18**	**1.8**	11.2 M	**75.67**
DiPrompT (Decoupled)	*ResNet-18*	*1.8*	11.2 M	*63.16*
CDIFed (Logit Distill)	*ResNet-18*	*1.8*	11.2 M	*70.90*

focusing on inference cost (GFLOPs and number of parameters) and the impact of removing the VLM.

Table 4 presents the results. The standard DiPrompT method achieves a high global accuracy of 89.93%. However, this comes at the cost of running the full CLIP image encoder during inference, requiring 4.7 GFLOPs per image. When we enforce a realistic edge-device constraint by removing the VLM and distilling the DiPrompT model into a ResNet-18 (via logit distillation), the performance collapses to 63.16%. This drastic 26.8% degradation indicates that prompt-tuning methods rely heavily on the inherent capacity of the frozen VLM rather than explicitly transferring domain-invariant knowledge to a portable model.

Conversely, CDIFed maintains a robust accuracy of 75.67% using only a standalone ResNet-18, which requires just 1.8 GFLOPs—a 2.6× reduction in computational cost compared to the coupled baseline. Crucially, our method outperforms the decoupled DiPrompT variant by over 12 percentage points (75.67% vs. 63.16%). This demonstrates that CDIFed's Phase 2 effectively transfers semantic robustness to the student, whereas prompt tuning fails to retain performance when the VLM scaffolding is removed.

Feature vs. Logit Distillation. We further validated our choice of feature-based distillation. As shown in Table IV, replacing our feature alignment objective with standard logit distillation (minimising KL-divergence) reduces CDIFed's accuracy from 75.67% to 70.90%. This confirms that aligning intermediate representations captures richer, domain-invariant structural information than merely matching the final class probabilities.

Student Capacity Sensitivity. Finally, we analysed the impact of the student's capacity. Reducing the student to a ResNet-10 causes a significant drop to 60.18%, indicating a minimum capacity is required to inherit CLIP's knowledge. However, increasing the student to a ResNet-50 yields only marginal gains (76.12%), suggesting that performance is primarily driven by the quality of the teacher's domain-invariant supervision rather than the student's raw parameter count.

6 Conclusion

Most existing Federated Learning (FL) research focuses on statistical heterogeneity but does not address domain shift, which commonly arises across diverse edge devices deployed in real-world settings. To resolve this, the global model must learn feature representations that remain consistent across different client environments. This paper proposed CDIFed, a framework that leverages the robust semantic priors of the CLIP vision-language model to extract domain-invariant features from heterogeneous edge clients. By employing a parameter-efficient, text-guided domain filter and subsequent knowledge distillation, CDIFed effectively trains a lightweight student model. Crucially, this decoupled approach produces domain-invariant representations that enable efficient and robust global aggregation across resource-constrained edge devices, bypassing the prohibitive inference costs typically associated with massive foundation models. Experiments on the Digits and Office-Caltech benchmarks showed that CDIFed achieves higher global accuracy than current state-of-the-art FL methods, highlighting its practical applicability within IoT and edge computing ecosystems.

Data Availability

The datasets analysed during the current study are available in the MNIST repository, https://www.kaggle.com/datasets/hojjatk/mnist-dataset; the USPS repository, http://www.cad.zju.edu.cn/home/dengcai/Data/MLData.html; the SVHN repository, https://www.kaggle.com/datasets/stanfordu/street-view-house-numbers; the Synthetic Digits (SYN) repository, https://www.kaggle.com/datasets/prasunroy/synthetic-digits/data; and the Office-Caltech repository, https://github.com/ZhangJUJU/OfficeCaltechDomainAdaptation.

References

1. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proc. Int. Conf. Artif. Intell. Statist. (AISTATS)*, 2017, pp. 1273–1282.
2. J. Bhanbhro, S. Nisticò, and L. Palopoli, "Issues in federated learning: some experiments and preliminary results," *Sci. Rep.*, vol. 14, Art. no. 29 881, 2024.
3. S. Wu, J. Chen, X. Nie, Z. Wang, H. Wang, and Z. Cai, "Global prototype distillation for heterogeneous federated learning," *Sci. Rep.*, vol. 14, Art. no. 12 057, 2024.
4. T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," in *Proc. Mach. Learn. Syst. (MLSys)*, vol. 2, 2020, pp. 429–450.
5. N. Babendererde, M. H. Jeske, L. M. Reich, and R. Hirtz, "Jointly exploring client drift and catastrophic forgetting in dynamic learning," *Sci. Rep.*, vol. 15, Art. no. 89 873, 2025.

6. M. Ragab, "Advanced artificial intelligence with federated learning framework for privacy-preserving cyberthreat detection in IoT-assisted sustainable smart cities," *Sci. Rep.*, vol. 15, Art. no. 88 843, 2025.
7. F. R. Mughal, J. He, B. Das, and M. S. Hadi, "Adaptive federated learning for resource-constrained IoT devices through edge intelligence and multi-edge clustering," *Sci. Rep.*, vol. 14, Art. no. 28 746, 2024.
8. C. Elvitigala, I. Khalil, and M. Atiquzzaman, "Hfedlmr: Personalized and hierarchical federated learning for consumer healthcare iot with non-iid data," *IEEE Trans. Consum. Electron.*, vol. 71, no. 3, pp. 8441–8452, 2025.
9. Y. Li, X. Wang, R. Zeng, P. K. Donta, I. Murturi, M. Huang, and S. Dustdar, "Federated domain generalization: A survey," *Proc. IEEE*, 2025.
10. A. T. Nguyen, P. Torr, and S. N. Lim, "FedSR: A simple and effective domain generalization method for federated learning," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 35, 2022, pp. 38 831–38 843.
11. Y. Jia, X. Zhang, H. Hu, K.-K. R. Choo, L. Qi, X. Xu, A. Beheshti, and W. Dou, "DapperFL: Domain adaptive federated learning with model fusion pruning for edge devices," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 37, 2024, pp. 13 099–13 123.
12. Z. Luo, Y. Wang, Z. Wang, Z. Sun, and T. Tan, "Disentangled federated learning for tackling attributes skew via invariant aggregation and diversity transferring," *arXiv preprint arXiv:2206.06818*, 2022.
13. Z. Wang, Z. Wang, X. Fan, and C. Wang, "Federated learning with domain shift eraser," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2025, pp. 4978–4987.
14. L. Zhang, X. Lei, Y. Shi, H. Huang, and C. Chen, "Federated learning with domain generalization," *arXiv preprint arXiv:2111.10487*, 2021.
15. J. Cha, K. Lee, S. Park, and S. Chun, "Domain generalization by mutual-information regularization with pre-trained models," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2022, pp. 440–457.
16. A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2021, pp. 8748–8763.
17. S. Addepalli, A. R. Asokan, L. Sharma, and R. V. Babu, "Leveraging vision-language models for improving domain generalization in image classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2024, pp. 23 922–23 932.
18. K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 770–778.
19. S. P. Karimireddy, S. Kale, M. Mohri, S. Reddi, S. Stich, and A. T. Suresh, "SCAFFOLD: Stochastic controlled averaging for federated learning," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2020, pp. 5132–5143.
20. J. Wang, Q. Liu, H. Liang, G. Joshi, and H. V. Poor, "Tackling the objective inconsistency problem in heterogeneous federated optimization," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 33, 2020, pp. 7611–7623.
21. D. A. E. Acar, Y. Zhao, R. M. Navarro, M. Mattina, P. N. Whatmough, and V. Saligrama, "Federated learning based on dynamic regularization," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2021.
22. H. Wang, M. Yurochkin, Y. Sun, D. Papailiopoulos, and Y. Khazaeni, "Federated learning with matched averaging," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2020.
23. T. Li, S. Hu, A. Beirami, and V. Smith, "Ditto: Fair and robust federated learning through personalization," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2021, pp. 6357–6368.
24. X. Li, M. Jiang, X. Zhang, M. Kamp, and Q. Dou, "FedBN: Federated learning on non-iid features via local batch normalization," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2021.
25. F. Pourpanah, M. Molahasani, M. Soltany, M. Greenspan, and A. Etemad, "Federated unsupervised domain generalization using global and local alignment of gradients," in *Proc. AAAI Conf. Artif. Intell.*, vol. 39, no. 19, 2025, pp. 19 948–19 958.
26. S. Bai, J. Zhang, S. Guo, S. Li, J. Guo, J. Hou, T. Han, and X. Lu, "DiPrompt: Disentangled prompt tuning for multiple latent domain generalization in federated learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2024, pp. 27 284–27 293.
27. W. Lu, X. Hu, J. Wang, and X. Xie, "FedCLIP: Fast generalization and personalization for CLIP in federated learning," *arXiv preprint arXiv:2302.13485*, 2023.

28. H. Lin, C. Zhang, D. Hong, K. Dong, and C. Wen, "FedRSCLIP: Federated learning for remote sensing scene classification using vision-language models," *IEEE Geosci. Remote Sens. Mag.*, vol. 13, pp. 2–18, 2025.

29. A. Masano, M. Pennisi, F. P. Salanitri, C. Spampinato, and G. Bellitto, "Zero-shot decentralized federated learning," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, 2025, pp. 1–8.

30. J. Shi, S. Zheng, X. Yin, Y. Lu, Y. Xie, and Y. Qu, "CLIP-guided federated learning on heterogeneity and long-tailed data," in *Proc. AAAI Conf. Artif. Intell.*, vol. 38, no. 13, 2024, pp. 14 955–14 963.

31. A. Imteaj, M. Z. Hossain, S. Zaman, and A. R. Shahid, "TriplePlay: Enhancing federated learning with CLIP for non-iid data and resource efficiency," in *Proc. Int. Conf. Mach. Learn. Appl. (ICMLA)*, 2024, pp. 1474–1480.

32. H. Zeng, Z. Yue, Y. Zhang, L. Shang, and D. Wang, "Fair federated learning with biased vision-language models," in *Findings Assoc. Comput. Linguist. (ACL)*, 2024, pp. 10 002–10 017.

33. P. Gao, S. Geng, R. Zhang, T. Ma, R. Fang, Y. Zhang, H. Li, and Y. Qiao, "CLIP-adapter: Better vision-language models with feature adapters," *Int. J. Comput. Vis. (IJCV)*, vol. 132, no. 2, pp. 581–595, 2024.

34. B. Li, Y. Shen, Y. Wang, W. Zhu, D. Li, K. Keutzer, and H. Zhao, "Invariant information bottleneck for domain generalization," in *Proc. AAAI Conf. Artif. Intell.*, vol. 36, no. 7, 2022, pp. 7399–7407.

35. S. Jung, K. Lee, Y. Jeong, H. Noh, J. Lee, and J. Choi, "Group-wise scaling and orthogonal decomposition for domain-invariant feature extraction in face anti-spoofing," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2025, pp. 13 372–13 381.

36. Q. Li, B. He, and D. Song, "Model-contrastive federated learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021, pp. 10 713–10 722.

37. W. Huang, M. Ye, Z. Shi, H. Li, and B. Du, "Rethinking federated learning with domain shift: A prototype view," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023, pp. 16 312–16 322.

38. S. Caldas, J. Konečny, H. B. McMahan, and A. Talwalkar, "Expanding the reach of federated learning by reducing client resource requirements," *arXiv preprint arXiv:1812.07210*, 2018.

39. Z. Jiang, Y. Xu, H. Xu, Z. Wang, C. Qiao, and Y. Zhao, "FedMP: Federated learning through adaptive model pruning in heterogeneous edge computing," in *Proc. IEEE 38th Int. Conf. Data Eng. (ICDE)*, 2022, pp. 767–779.

40. H. Kang, S. Cha, J. Shin, J. Lee, and J. Kang, "NeFL: Nested model scaling for federated learning with system heterogeneous clients," *arXiv preprint arXiv:2308.07761*, 2023.

41. Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

42. J. J. Hull, "A database for handwritten text recognition research," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 16, no. 5, pp. 550–554, 1994.

43. Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, A. Y. Ng *et al.*, "Reading digits in natural images with unsupervised feature learning," in *NIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2011.

44. P. Roy, S. Ghosh, S. Bhattacharya, and U. Pal, "Effects of degradations on deep neural network architectures," *arXiv preprint arXiv:1807.10108*, 2018.

45. B. Gong, Y. Shi, F. Sha, and K. Grauman, "Geodesic flow kernel for unsupervised domain adaptation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2012, pp. 2066–2073.

46. A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein *et al.*, "PyTorch: An imperative style, high-performance deep learning library," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 32, 2019.

47. R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2017, pp. 618–626.

48. L. v. d. Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 11, pp. 2579–2605, 2008.

Funding Declaration

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (RS-2025-22442984).

Author Contributions

Y.Kim conceived the study, designed the framework, performed the experiments, and wrote the original draft of the manuscript. Y.Park contributed to the formal analysis and validation of the experimental results. J.W.Cho and J.Cho supervised the project, provided computational resources, and reviewed the manuscript. All authors discussed the results and approved the final version of the manuscript.

Additional Information

Competing Interests

The author(s) declare no competing interests.