FedENC: Federated Adaptive CLIP Model via Energy Alignment and Neural Collapse

Yihang Wu[1] , **Ahmad Chaddad**[1,2,*] and **Sarah A. Alkhodair**[3]
[1]School of Artificial Intelligence, Guilin University of Electronic Technology, Guilin, China
[2]LIVIA, École de Technologie Supérieure, Montreal, Canada
[3]College of Computer and Information Sciences, King Saud University, Riyadh, Saudi Arabia
ahmad8chaddad@gmail.com

Abstract

Federated Learning (FL) faces two major challenges when deploying large vision-language models such as contrastive language-image pre-training (CLIP) model, namely high computation/communication overhead and performance degradation due to data heterogeneity. To address these, we propose FedENC, a novel federated adaptive CLIP framework. FedENC keeps the CLIP encoder frozen and introduces a lightweight communicable global expert adaptation module for learning general knowledge across clients and two local experts for client-specific adaptation to reduce computation and communication overhead. To mitigate data heterogeneity, we introduce neural collapse (NC) alignment to encourage the model to learn unified and balanced image feature representations. To enrich the performance, we propose a cosine and Kullback–Leibler (KL) divergence based alignment function to improve the prediction consistency between the global and local experts. Finally, we ensemble their outputs for robust inference. Extensive experiments on six datasets (4 natural + 2 medical) in challenging settings (e.g., domain shift) show that FedENC improves the test accuracy (e.g., by +5.37% on DomainNet compared to LoRA) with reasonable communication load (e.g., 15.2x lower than LoRA, with 3.3GB GPU memory usage). The code is available at https://github.com/AIPMLab/FedENC.

1 Introduction

Without sufficient training data, the performance of the deep learning (DL) model tends to decrease and overfit [Qi *et al.*, 2026; Choi and Ko, 2025]. One naive way is to collect multi-source data to support the training phase. However, due to data regularization laws such as GDPR [Voigt and Von dem Bussche, 2017], accessing source data is not practical since it introduces privacy leakage concerns.

Federated learning (FL) enables collaborative model training with various decentralized institutions (a.k.a., FL clients), providing a promising solution to solve data privacy [Wu *et al.*, 2026; Wu and Chaddad, 2026]. FL is widely used in many applications, such as private face recognition, image classification and segmentation [Li *et al.*, 2025; Qi *et al.*, 2025; Meng *et al.*, 2025]. However, recent DL models such as CLIP tend to have billions of parameters to achieve remarkable performance [Radford *et al.*, 2021], leading to considerable *computation and communication overhead* (e.g., 10^8 parameters in ViT-B-32, require $\sim$ 342 MB to transmit). This hinders the use of these models in the FL context. Furthermore, in the real-world, data from multiple clients exhibit non-independent and identically distributed (non-IID) characteristics due to variations in texture discrepancies and data imbalance [Chen *et al.*, 2025]. This *data heterogeneity* poses a challenge for the global model to learn unified decision boundaries, resulting in poor performance.

Recent studies proposed several techniques to solve these challenges. Basically, these methods can be classified into three groups: adapter-, prompt learning-, and linear probe-based approaches. However, adapter based methods, such as FedCLIP [Lu *et al.*, 2023], can not fully capture domain-invariant features as suggested in [Wu *et al.*, 2025]. Prompt-learning methods, although efficient in communication, require large GPU memory to adapt learnable prompts [Guo *et al.*, 2024; Phung *et al.*, 2025]. Linear probe based methods such as LP++ [Huang *et al.*, 2024] are efficient in both computation and communication, but linear projections limit their ability to learn task-specific features.

In light of previous challenges, we propose a novel CLIP-based FL framework for non-IID image classification task (FedENC). To reduce computation and communication overhead, the pre-trained CLIP encoder is frozen, while we introduce three lightweight modules (one global expert and two local experts) to adapt the model. Specifically, the global expert is used as a communication module to learn shared knowledge across clients, while the two local experts focus on learning client-specific features to fit the local tasks. To address data heterogeneity, we optimize the parameters of the global expert using NC alignment. The primary goal of NC loss is to enforce the feature representations learned by each client to align with a pre-defined prototypes, mitigating the impact of domain shifts and ensuring global feature consistency. To improve the prediction stability, we minimize the differences between the global expert and local expert out-

* Corresponding author.

puts using KL and cosine similarity based energy alignment. During inference, the outputs of global and local experts are ensembled to obtain the final prediction. Finally, the global expert of each client is aggregated using average aggregation. We evaluate FedENC using six image classification benchmarks. The results show that FedENC outperforms other FL approaches with reasonable computation and communication load. The contributions of this paper are as follows.

- *Algorithm.* We propose a novel CLIP-based FL framework for image classification tasks. Specifically, we propose adapting a frozen CLIP backbone with global and local experts to adapt the CLIP with lower computation and communication load. Furthermore, we introduce NC adaptation to solve data heterogeneity, and we design an energy alignment technique to improve prediction consistency between the global and local experts. Finally, global experts are aggregated and broadcast to each client for local training.
- *Empirical analysis.* We perform extensive experiments on six datasets (4 natural + 2 medical) to show the usefulness of the proposed approach in: (i) robustness to heterogeneity FL, (ii) efficiency of computation and communication, (iii) adaptability with different network architectures, (iv) scalability with different number of clients and (v) calibration ability during inference.

2 Related Work

Federated learning. For example, based on FedCLIP, FAA-CLIP [Wu *et al.*, 2025] extends it with an adversarial domain adaptation technique to solve the feature shifts among the clients. However, adversarial adaptation introduces extra learnable parameters, increasing the computational overhead. Unlike these methods that adapting the image features, PromptFL [Guo *et al.*, 2024] proposes fine-tuning the learnable prompts and share these prompts for aggregation. Despite its lower computational load, these learnable parameters will increase the communication overhead when the number of classes is large. This limits its potential for real-world applications. Federated adaptive prompt learning (FedAPT) adds a domain-specific prompt related to clients to the general prompts compared to PromptFL [Su *et al.*, 2024]. In [Motamedi *et al.*, 2025], they update the Pre-norm layer parameters in the vision encoder and aggregate these parameters to efficiently adapt the CLIP model (NormFit). In addition, FedDIM [Liao *et al.*, 2025] introduces a decision insight matrix to reduce domain shifts and improve the classifier performance. Results show that it achieves a higher accuracy compared to FedAVG. However, it requires more computation time under large-scale dataset.

CLIP model. In [Zanella and Ben Ayed, 2024], they introduce low-rank adaptation (LoRA) to adapt the CLIP backbone into downstream tasks with minimal costs. Similar to LoRA, in [Zhou *et al.*, 2022], they propose a conditional prompt learning technique to improve the text features (CoCoOp). However, it consumes large GPU memory as the batch size increases. LP++ [Huang *et al.*, 2024] explores the usefulness of linear probing and design a novel prediction strategy that uses both image and text information. In [Hakim *et al.*, 2025], they design a test-time adaptation technique to effectively adapt the CLIP model for image classification tasks in a transductive manner.

Neural collapse. Neural collapse (NC) proves that for a well-trained classifier on a balanced dataset, the features produced by the network backbone for all samples within each class converge to a highly symmetric structure, and within-class variability effectively vanishes [Papyan *et al.*, 2020]. For example, in [Li *et al.*, 2023], they design a synthetic and fixed classifier during local training in FL. The experimental results show that the optimal classifier structure enables all clients to learn unified and optimal feature representations under heterogeneous data. In [Du and Li, 2025], they introduce a simplex learnable embedding module guided by NC to learn local features, while use a knowledge decoupling module to extract general knowledge to align local features with the global embeddings.

Unlike these methods, FedENC addresses these limitations through a systematic framework that uses global and local experts for efficient adaptation of a pre-trained CLIP model, NC-based alignment to align image features of each client to a balanced feature space to solve data heterogeneity, and energy-based consistency regularization between experts.

3 Methods

This study considers a standard FL setting, consisting of $\mathcal{N}$ clients and a global server for aggregation. Each client has a private dataset $\mathcal{D}_i$. Each data set $\mathcal{D}_i$ consists of three non-overlapping parts, namely a training $\mathcal{D}_i^{train} = \{(\mathbf{x}_{i,j}^{train}, y_{i,j}^{train})\}_{j=1}^{n_i^{train}}$, a validation $\mathcal{D}_i^{val} = \{(\mathbf{x}_{i,j}^{val}, y_{i,j}^{val})\}_{j=1}^{n_i^{val}}$ and a test set $\mathcal{D}_i^{test} = \{(\mathbf{x}_{i,j}^{test}, y_{i,j}^{test})\}_{j=1}^{n_i^{test}}$. Figure 1 shows the general framework of the proposed method. It consists of two parts, namely 1) local training and 2) global aggregation.

3.1 Local Training and Inference

Feature extraction with CLIP. We use a pre-trained CLIP model to extract both image and text features. Let e_I and e_T be the image and text encoder, respectively. Given an image sample $\mathbf{x}$, we denote $\mathbf{I}_j = e_I(\mathbf{x}_j) \in \mathrm{R}^D$ as the D-dimensional vector of image features. For text encoder, we use a standard text prompt (A PICTURE OF A $\{$CLASS$\}$) to obtain the text features $\mathbf{T}_j = e_T(\mathbf{x}_j) \in \mathrm{R}^D$.

Feature adaptation via expert module. Despite CLIP encoders can extract a rich set of features, however, these features should be refined for a specific task (i.e., learn dominant feature representations). Since most heterogeneity exists in images, we only adapted image features in this study. To efficiently adapt the image features, the CLIP encoder is frozen and we design three lightweight modules for each client, consisting of a global expert (apt_g^i) and two local experts (apt_1^i, apt_2^i). We consider two local experts because it introduces less computation complexity. Specifically, the global expert focuses on learning cross-client knowledge by aggregation, while the local experts are used to fit client-specific characteristics. The goal of the global expert is to generate a mask $\mathbf{M} \in [0,1]^D$ based on $\mathbf{I}$ using a Softmax function with

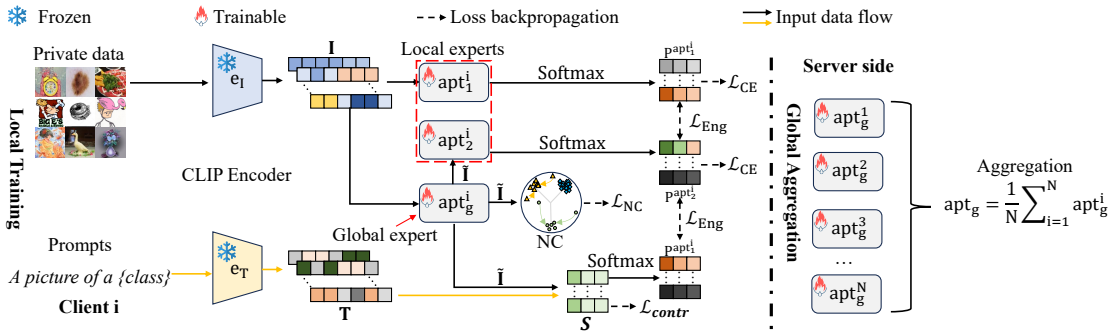

Figure 1: Pipeline of FedENC. It consists of two parts, namely local training and global aggregation. During local training, global expert apt_g^i and local experts apt_1^i, apt_2^i are optimized using contrastive loss $\mathcal{L}_{contr}$, classification loss $\mathcal{L}_{CE}$, feature alignment loss $\mathcal{L}_{NC}$ and energy alignment $\mathcal{L}_{Eng}$. For aggregation, the apt_g^i are aggregated into one shared expert apt_g.

linear layers. This mask is used to obtain an adapted image feature $\tilde{\mathbf{I}} = apt_g^i(\mathbf{I}) \otimes \mathbf{I}$, where $\otimes$ is the Hadamard product (element wise). For local experts, $\mathbf{I}$ and $\tilde{\mathbf{I}}$ are used as the input for apt_1^i and apt_2^i, respectively. This complementary feature input will encourage local experts to learn robust and diverse features. Finally, local experts will provide predicted probability based on these inputs. To refine the adapted image features $\tilde{\mathbf{I}}$, we use the standard contrastive loss in CLIP. Let $\mathbf{S}$ be the $B \times B$ matrix consisting of cosine similarities between the image features $\tilde{\mathbf{I}}_j$ and $\mathbf{T}_{j'}$. We compute an image probability matrix $\mathbf{P} = Softmax(\mathbf{S}/\tau) \in [0,1]^{B\times B}$ and a text probability matrix $\mathbf{Q} = Softmax(\mathbf{S}^\top/\tau) \in [0,1]^{B\times B}$. The contrastive loss is then formulated as follows:

$$\mathcal{L}_{contr} = -\frac{1}{B}\sum_{j=1}^{B}\frac{1}{2}\Big(\log p_{j,j} + \log q_{j,j}\Big), \tag{1}$$

where τ is a pre-defined scaling parameter in CLIP. To train the local experts, we use standard cross-entropy loss as follows.

$$\mathcal{L}_{CE} = -\frac{1}{B}\sum_{j=1}^{B}\mathcal{L}_{CrossEntropy}(o_j, y), \tag{2}$$

where o_j is the output from local expert.

Neural collapse alignment. As suggested in [Papyan *et al.*, 2020], they show that for deep networks trained on a balanced dataset, let $M = [M_1, M_2, ..., M_C]$ (C is the number of classes) be the weight matrix of the linear classifier, it exists a simplex Equiangular Tight Frame (ETF) M^* in $\mathbb{R}^D$ for C classes as follows:

$$M^* = \alpha R\sqrt{\frac{C}{C-1}}\left(I_C - \frac{1}{C}\mathbf{1}_C\mathbf{1}_C^\top\right), \tag{3}$$

where $I_C \in \mathbb{R}^{C\times C}$ is the identity matrix, $\mathbf{1}_C \in \mathbb{R}^C$ is the all-ones vector, $R \in \mathbb{R}^{d\times C}$ ($d \geq C$) is any orthogonal projection, and $\alpha > 0$ is a scaling constant.

NC suggests that well-trained networks exhibit class-wise feature means that converge to the vertices of a Simplex ETF. Suppose $F_{ETF} = [M_1, ..., M_C]$ is the reference geometry for all clients, each column M_i represents the ideal feature prototype for class C and satisfies the ETF properties:

$$\|M_i\|_2 = 1, \quad \langle M_i, M_j\rangle = -\frac{1}{C-1}, \quad \forall i \neq j. \tag{4}$$

This globally fixed frame is shared across all clients, providing a consistent and maximally symmetric target geometry. To align the clients features with this fixed geometry, we introduce a NC alignment loss that encourages the adapted image features $\tilde{\mathbf{I}}$ to move toward their correct prototype while repelling them from others:

$$\mathcal{L}_{\text{NC}} = w_k^c\left(-\frac{\tilde{\mathbf{I}}_k^c \cdot M_c}{\kappa} + \frac{1}{C-1}\sum_{j\neq c}\frac{\tilde{\mathbf{I}}_k^c \cdot M_j}{\kappa}\right), \tag{5}$$

where κ is a scaling constant, w_k^c is the weight for class c in the client k. The w_k^c can be formulated as follows.

$$w_k^c = \begin{cases} \left(\frac{n_k^c}{n_k}\right)^{-\gamma}, & n_k^c > 0, \\ 0, & n_k^c = 0, \end{cases} \tag{6}$$

where n_k^c is the number of samples for class c in client k, γ is a scaling parameter. The first term in Eq. 5 pulls the feature toward its correct prototype, while the second term pushes it away from other classes. A lower γ indicates more weights on rare class samples.

Energy alignment. Since the global expert focuses on general information across clients, while the local experts focus on local details, we propose an energy function-based technique to improve the prediction consistency of the global and local experts as follows.

$$\mathcal{L}_{Eng} = \text{Softplus}\left(\phi\left(\text{KL}(O_1 \| O_2),\ 1 - \cos(O_1, O_2)\right)\right), \tag{7}$$

where $O_1 = \text{Softmax}(o_1/\tau)$, $O_2 = \text{Softmax}(o_2/\tau)$, $\phi(\cdot)$ is a MLP. $o_{1|2}$ represents the outputs from $apt_{1|2}^i$ and apt_g^i. A smaller value of $\mathcal{L}_{Eng}$ indicates the outputs of the shared expert and local experts are similar, and vice versa.

Finally, the loss used for local training can be formulated as:

$$\mathcal{L} = \underbrace{\mathcal{L}_{contr} + \mathcal{L}_{CE}}_{\text{Classification}} + \underbrace{\lambda_1\mathcal{L}_{NC} + \lambda_2\mathcal{L}_{Eng}}_{\text{Regularization}}, \tag{8}$$

where λ_1 and λ_2 are two hyper-parameters.

Ensemble prediction. To enrich the prediction with both global and locall learned knowledge, we design an ensemble prediction strategy as follows:

$$p^{ens} = (1-\beta)\cdot p^{apt_g^i} + \beta\cdot(p^{apt_1^i} + p^{apt_2^i}), \tag{9}$$

where p^{ens}, $p^{apt_g^i}$ and $p^{apt_{1|2}^i}$ are the predicted probability of the ensemble prediction, apt_g^i and $apt_{1|2}^i$, respectively. β is defined using an entropy function $\mathcal{H}(\cdot)$ as follows.

$$\beta = \frac{\mathcal{H}(p^{apt_g^i})}{\mathcal{H}(p^{apt_1^i} + p^{apt_2^i}) + \mathcal{H}(p^{apt_g^i})}. \tag{10}$$

Specifically, β dynamically balances global and local predictions, assigning higher weight to the prediction branch with lower uncertainty.

3.2 Global Aggregation

To provide an unbiased global model, we use a simple average aggretaion technique as follows:

$$apt_g = \frac{1}{N}\sum_{i=1}^{N} apt_g^i. \tag{11}$$

Method	Client				AVG
OfficeHome	Ar	Cl	Pr	Re	
CLIP$_{zs}$	76.29	65.64	85.23	86.31	78.36
Individual	61.86	75.72	81.62	70.72	72.48
FedAVG	76.91	85.11	89.06	86.68	84.44
FedDIM	75.00	85.65	89.00	85.07	83.68
FAA-CLIP	78.76	75.26	90.30	87.83	83.04
LoRA	74.64	83.96	89.18	84.84	83.16
PromptFL	74.43	81.10	91.77	88.75	84.01
CocoOp	77.73	74.11	88.84	88.52	82.3
FedCLIP	78.54	74.77	89.55	88.31	82.79
LP++	75.88	73.42	87.71	87.14	81.04
FedAPT	77.73	75.83	91.66	88.63	83.46
NormFit	78.54	67.25	85.88	86.69	79.59
Ours	81.04	84.26	94.21	87.96	86.87
ModernOffice31	amazon	webcam	dslr	synthetic	
CLIP$_{zs}$	85.08	82.83	49.03	76.10	73.26
Individual	87.92	94.95	91.13	91.82	91.45
FedAVG	95.20	100	95.81	98.74	97.44
FedDIM	92.28	96.88	88.82	100	94.5
FAA-CLIP	94.32	94.95	74.52	85.53	87.33
LoRA	92.83	93.75	92.60	95.31	93.62
PromptFL	95.38	94.95	87.74	95.60	93.42
CocoOp	91.83	93.94	65.48	85.53	84.19
FedCLIP	93.61	92.93	70.48	83.65	85.17
LP++	93.43	95.96	67.74	88.05	86.3
FedAPT	96.09	96.97	91.45	97.48	95.49
NormFit	87.32	82.29	61.35	78.12	77.27
Ours	97.43	100	95.89	96.88	97.55
PACS	art	cartoon	photo	sketch	
CLIP$_{zs}$	96.55	97.92	99.60	83.80	94.47
Individual	83.22	86.78	83.57	49.79	75.84
FedAVG	96.46	95.76	98.49	94.09	96.2
FedDIM	96.41	96.13	97.82	94.20	96.14
FAA-CLIP	96.13	97.63	99.60	90.73	96.02
LoRA	97.37	98.35	99.70	93.62	97.26
PromptFL	97.45	98.20	99.70	91.62	96.74
CocoOp	96.63	97.99	99.60	89.67	95.97
FedCLIP	97.04	98.13	99.60	91.24	96.5
LP++	95.97	98.20	99.50	88.35	95.5
FedAPT	98.04	99.15	99.70	87.64	96.13
NormFit	96.38	97.84	99.60	84.44	94.56
Ours	97.20	98.06	99.70	91.96	96.73

Table 1: Test accuracy (%) in OfficeHome, ModernOffice31 and PACS. Red means the best, while blue indicates the second best. AVG represents the average value of all clients.

4 Experiments

4.1 Datasets

OfficeHome. OfficeHome is a large image classification benchmark that contains 65 classes [Venkateswara *et al.*, 2017]. This dataset has four sub-domains, namely Art (Ar), Clipart (Cl), Product (Pr), and Real World (Re) with about 15500 images.

Method	Client						AVG
	C	I	P	Q	R	S	
CLIP$_{zs}$	66.38	37.55	60.79	12.20	79.50	56.48	52.15
Individual	65.67	26.01	52.85	64.38	68.02	53.73	55.11
FedAVG	74.17	30.51	61.73	61.93	74.83	60.42	60.6
FedDIM	74.59	30.10	62.46	60.12	75.12	61.55	60.66
FAA-CLIP	70.26	37.94	65.20	22.01	80.82	61.19	56.24
LoRA	75.06	33.64	64.12	63.40	77.17	61.64	62.5
PromptFL	73.94	42.06	66.63	30.95	81.82	63.30	59.78
FedCLIP	69.99	39.26	63.64	31.82	81.03	59.20	57.49
LP++	40.13	29.32	41.21	8.53	45.30	38.40	33.82
FedAPT	70.15	43.69	63.69	15.21	79.19	57.15	54.85
NormFit	66.36	36.90	61.00	12.09	79.68	56.41	52.07
Ours	77.54	47.13	72.64	56.19	85.87	67.84	67.87

Table 2: Test accuracy (%) in DomainNet. Red means the best, while blue indicates the second best. AVG represents the average value of all clients.

ModernOffice31. It is a modern benchmark with 31 classes covering four domains: amazon, webcam, dslr, and synthetic [Ringwald and Stiefelhagen, 2021].

PACS. PACS has four domains, namely art, cartoon, photo and sketch [Li *et al.*, 2017]. It has seven classes, with about 9991 images.

DomainNet. DomainNet contains 6 domains (i.e., clipart, C; infograph, I; painting, P; quickdraw, Q; real world, R; sketch, S), with more than 0.6 million images, covering 345 classes [Peng *et al.*, 2019].

ISIC2019. ISIC2019 is a skin cancer classification dataset with various characteristics (e.g., age) [Tschandl *et al.*, 2018; Codella *et al.*, 2018; Combalia *et al.*, 2019]. To simulate data heterogeneous, it was divided based on the "anatom site (AS)" metadata, resulting in seven clients: C_1 holds the data whose AS is "anterior torso", while C_2, C_3, C_4, C_5, C_6 and C_{glo} hold the data whose AS is "head or neck", "lower extremity", "palms or soles", "posterior torso", "upper extremity" and "Nan", respectively. It has six classes.

Brain_ICH. The Brain_ICH dataset [Flanders *et al.*, 2020], which contains five ICH subtypes, is used for experiments. The same pre-processing strategies as in [Wu *et al.*, 2023] are applied. To simulate heterogeneous data, following [Wu *et al.*, 2023], Dirichlet distribution is used to divide the dataset to {5,10,15} clients.

For OfficeHome, ModernOffice31, PACS, DomainNet and ISIC2019, each domain is considered as a client. For each client in these datasets, the data is further divided into three groups, namely a training set (60%), a validation set (20%) and a test set (20%).

4.2 Implementation Details

Pre-trained CLIP (ViT-B-32 as the vision encoder) provided by OpenAI is used as the backbone [Radford *et al.*, 2021]. The AdamW optimizer is used [Loshchilov and Hutter, 2017]. The learning rate is set to 5×10^{-5} for OfficeHome, ModernOffice31 and PACS datasets, and to 5×10^{-4} for DomainNet, ISIC2019 and Brain_ICH datasets. The λ_1 and λ_2 are set to the optimal values based on sensitivity analysis, respectively. The local training epoch is set to one,

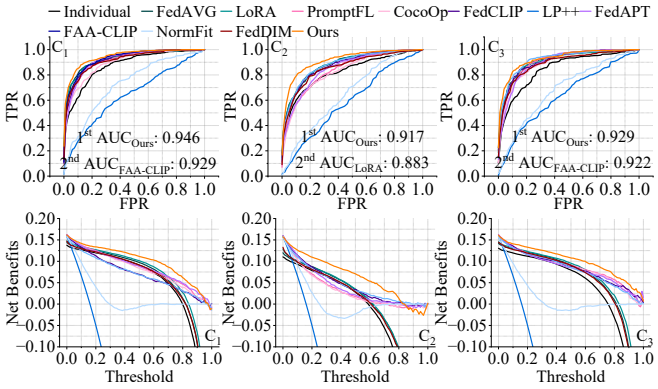

Figure 2: ROC and DCA curves for FedENC and baselines on ISIC2019 dataset.

while the global communication epoch is set to 50. The batch size is 32, and the random seed is initialized to 0 for all methods. For comparison, we use both non-federated and federated techniques. Non-federated includes CLIP$_{zs}$ (zero-shot) and Individual (no aggregation). Federated approaches include FedAVG (full-parameter tuning), FAA-CLIP, LoRA with FedAVG (rank is set to three), PromptFL, CocoOp with FedAVG, LP++ with FedAVG, FedAPT, NormFit and FedDIM. The experiments are based on an RTX 4090 GPU, featuring an Intel 13900KF CPU with 128 GB memory. *More details can be found in codes.*

4.3 Results

OfficeHome. Table 1 shows the test accuracy (ACC) for the baselines and FedENC. FedENC provides an average (AVG) ACC of 86.87%, outperforming traditional FL such as FedAVG by 2.43% and PEFT techniques such as PromptFL by 2.86%. The improvement is notable on client C (84.26%), where pre-trained CLIP has limited domain knowledge (65.64% ACC with CLIP$_{zs}$).

ModernOffice31. Similar to OfficeHome, FedENC yields the highest AVG ACC of 97.55% among all methods, as summarized in Table 1. We note that FedENC shows a lower test ACC on client synthetic compared to FedAVG. This suggests that pre-trained CLIP fails to extract more representative features (i.e., 76.10% with CLIP$_{zs}$) than full-parameter tuning (FedAVG).

PACS. As shown in Table 1, unlike OfficeHome, FedENC provides a comparable AVG ACC of 96.73% compared to LoRA (97.26%) and PromptFL (96.74%). Furthermore, since pre-trained CLIP lacks knowledge for sketch images, a similar conclusion can be drawn compared to ModernOffice31.

DomainNet. Table 2 reports the test ACC on DomainNet. Traditional FL such as FedAVG provides a higher AVG ACC of 60.6% compared to PEFT-based FL techniques such as FedCLIP (57.49%), FAA-CLIP (56.24%) and PromptFL (59.78%). Unlike these methods, FedENC yields an AVG ACC of 67.87%, outperforming the second-best (LoRA) by 5.37%. These results suggest that in a large-domain shift setting, the proposed method can provide feasible performance.

ISIC2019. Table 3 presents the test ACC and F1 score for FedENC and the baselines. Fine-tuning the entire CLIP encoder leads to a higher test metric (e.g., 71.72% AVG ACC) compared to CLIP$_{zs}$, while using FedAVG further improves the AVG ACC to 73.41%. However, most of the PEFT techniques, such as PromptFL and FedAPT yield lower test performance compared to FedAVG. Unlike these methods, FedENC provides an AVG ACC of 81.53%, with an AVG F1 score of 69.78%, outperforming the second-best approach by 8% and 13.29%, respectively. In addition, Figure 2 illustrates the Receiver Operating Characteristic (ROC) curves and Decision Curve Analysis (DCA) [Vickers and Elkin, 2006] on client C_1 to C_3. For example, FedENC yields an AUC value of 0.946 on C_1, outperforming the second-best approach (FAA-CLIP) by 0.017. The DCA curves show that FedENC provides a higher net benefit compared to other methods across a wide range of thresholds (0, 0.8).

Method	Client							AVG
Accuracy	C_1	C_2	C_3	C_4	C_5	C_6	C_7	
CLIP$_{zs}$	32.05	23.96	24.31	17.71	20.12	33.54	15.83	23.93
Individual	77.23	58.18	73.61	65.62	73.44	66.25	87.71	71.72
FedAVG	79.11	60.27	77.20	58.33	77.93	71.67	89.38	73.41
FedDIM	77.05	57.74	76.74	55.21	75.78	71.25	88.96	71.82
LoRA	80.45	59.97	78.47	54.17	78.71	73.96	88.96	73.53
PromptFL	74.91	50.30	75.58	65.62	77.73	71.25	85.83	71.6
CocoOp	74.46	49.26	74.88	62.50	78.12	69.79	84.17	70.45
FedCLIP	68.84	56.10	74.88	57.29	67.77	71.46	81.88	68.32
LP++	67.50	29.32	67.48	51.04	74.22	58.12	83.75	61.63
FedAPT	76.11	51.37	73.23	60.78	80.35	70.2	88.31	71.47
FAA-CLIP	68.57	56.10	75.58	60.42	68.55	67.08	80.21	68.07
NormFit	28.43	21.08	31.25	16.67	21.29	28.63	33.67	25.86
Ours	83.48	71.43	83.45	83.33	80.27	77.92	90.83	81.53
F1 score	C_1	C_2	C_3	C_4	C_5	C_6	C_7	
CLIP$_{zs}$	21.32	17.57	14.15	7.81	8.65	19.81	8.19	13.93
Individual	52.00	55.41	49.49	57.96	36.97	41.33	63.22	50.91
FedAVG	64.30	57.22	57.55	24.51	51.23	57.50	75.98	55.47
FedDIM	60.30	56.06	55.47	23.98	50.15	51.11	72.28	52.76
LoRA	65.49	58.06	59.27	22.47	51.34	61.50	77.33	56.49
PromptFL	53.59	47.53	47.86	38.60	50.47	54.47	68.25	51.54
CocoOp	48.04	44.34	43.71	26.19	42.41	46.08	37.73	41.21
FedCLIP	52.12	52.36	52.65	40.03	40.85	58.74	53.53	50.04
LP++	20.64	13.30	20.00	21.42	15.92	18.48	17.44	18.17
FedAPT	58.2	50.59	42.84	43.73	48.02	46.98	69.39	51.39
FAA-CLIP	53.19	52.36	55.36	45.77	44.24	49.52	49.42	49.98
NormFit	17.31	11.98	15.95	9.00	10.19	15.75	12.43	13.23
Ours	71.14	70.96	71.80	79.34	50.62	68.74	75.84	69.78

Table 3: Test metrics (%) in ISIC2019.

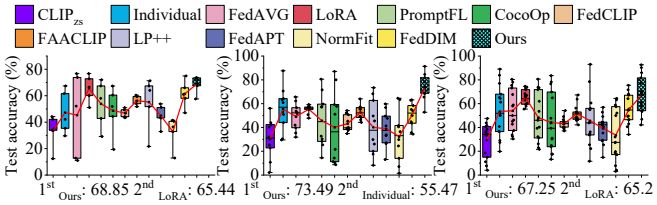

Figure 3: Test accuracy on Brain_ICH dataset under different client settings. Red line connects the average accuracy of each method.

Brain_ICH. Figure 3 shows the test ACC for FedENC and baselines using box chart. For example, FedENC provides the highest AVG ACC of 68.84% under 5 clients, outperforming the second-best (LoRA) by 3.4%. Furthermore, the box plot for FedENC is narrower compared to other PEFT approaches such as PromptFL, indicating low variance across clients. Similar conclusions can be drawn from other client settings.

Computation and communication overhead. Table 4 re-

Method	Computation						Communication						GPU memory					
	OfficeHome / ModernOffice31 / PACS / DomainNet / ISIC2019 / Brain_ICH (5 clients)																	
Individual	76.13	65.64	43.61	402.67	128.07	83.45	0	0	0	0	0	0	9.5					
FedAVG	36.52	57.91	27.75	770.68	105.63	5.28	7.52	14.36	7.52	16.75	13.33	1.02	9.5					
FedDIM	47.31	13.22	10.67	755.01	101.42	55.80	2.46	1.197	1.197	5.50	4.45	4.61	6.1					
FAA-CLIP	87.16	62.97	35.91	674.63	128.31	53.91	0.088	0.098	0.064	0.094	0.098	0.072	3.8					
LoRA	35.99	64.94	28.13	962.38	96.83	43.28	0.131	0.285	0.273	0.285	0.215	0.139	5.4					
PromptFL	66.73	72.49	45.91	1050.07	132.73	78.61	0.093	0.107	0.016	1.269	0.020	0.017	4.5	4.3	4.2	10.5	5.5	4.2
CocoOp	153.14	89.32	108.06	✘	194.43	204.61	0.15	0.084	0.022	✘	0.017	0.022	19.4	18.0	6.6	✘	4.5	4.1
FedCLIP	70.28	60.51	45.75	554.29	98.58	56.59	0.096	0.096	0.086	0.094	0.076	0.082	3.0					
LP++	25.64	32.77	13.36	183.12	60.64	35.51	0.0002	0.0002	0.0001	0.169	0.0001	0.0001	2.9	2.8	2.7	4.1	2.9	2.7
FedAPT	89.06	29.34	13.91	1155.60	114.25	107.60	0.295	0.145	0.039	0.451	0.142	0.038	3.3	2.6	2.0	8.1	2.2	2.0
NormFit	73.04	53.84	49.40	178.14	129.75	13.66	<0.0001						2.8					
Ours	63.11	25.05	9.64	581.17	118.10	16.49	0.088	0.040	0.018	0.096	0.092	0.086	3.1	3.1	3.1	3.3	3.1	3.1

Table 4: Total computation time to reach its best validation metric, communication overhead (GB) and GPU memory usage (GB) for baselines and FedENC. Note that the time is measured on a single RTX4090.

Components					OfficeHome					ISIC2019							
$\mathcal{L}_{contr}$	$\mathcal{L}_{CE}$	$\mathcal{L}_{NC}$	$\mathcal{L}_{Eng}$	Agg.	A	C	P	R	AVG	C_1	C_2	C_3	C_4	C_5	C_6	C_7	AVG
✓	✘	✘	✘	✓	78.12	75.69	90.62	87.15	82.9	59.00	56.73	57.27	45.21	46.49	58.77	56.83	54.33
✓	✓	✘	✘	✓	77.5	78.04	90.0	87.33	83.22	70.91	66.50	67.02	66.75	48.11	63.68	78.01	65.85
✓	✓	✘	✓	✓	79.38	83.68	94.21	88.31	86.39	71.22	69.52	65.66	78.30	50.24	66.77	75.03	68.11
✓	✓	✓	✘	✓	80.42	83.56	93.98	88.77	86.68	68.27	68.06	68.82	76.43	45.95	58.15	78.84	66.36
✓	✓	✓	✓	✘	72.07	82.90	93.72	85.40	83.52	67.71	64.89	72.43	70.84	43.99	63.31	73.83	65.29
✓	✓	✓	✓	✓	81.04	84.26	94.21	87.96	86.87	71.14	70.96	71.80	79.34	50.62	68.74	75.84	69.78

Table 5: Test accuracy on OfficeHome and F1 score on ISIC2019 with different components. Agg. indicates global aggregation.

ports the total computation time, communication load and GPU memory usage for baselines and FedENC. Compared to strong baselines like LoRA, it achieves higher accuracy (e.g., +5.37% on AVG) with lower computation and communication costs (e.g., 39.6% less time and 2.96× lower data transmission than LoRA). While FedENC requires more computation time than some PEFT methods (e.g., FedCLIP, FedAPT) on the ISIC2019 dataset, it outperforms these methods by up to 13.21% on AVG. Notably, FedENC maintains a consistently lower GPU memory footprint (e.g., 3.1GB).

4.4 Model Analysis

Ablation study. We validate the contribution of each component using the OfficeHome and ISIC2019 datasets. As summarized in Table 5, introducing local experts improves the test ACC compared to using global expert alone (e.g., AVG increased from 82.9% to 83.22% on OfficeHome). Furthermore, using $\mathcal{L}_{NC}$ further reduces the impact of data heterogeneity, increasing the ACC and F1 score to 86.68% and 66.36% on OfficeHome and ISIC2019, respectively. Finally, combining all techniques leads to the best AVG of 86.76% and 69.78%, respectively.

Calibration. We evaluate calibration performance using Expected Calibration Error (ECE) and reliability diagrams [Vaicenavicius *et al.*, 2019] on OfficeHome and ISIC2019. The reliability plots are shown in Figure 4 and Figure 5. We observe that: 1) On OfficeHome, FedENC achieves a calibration error comparable to that of FedCLIP on clients A and C. 2) On ISIC2019, FedENC yields lower ECE than both PromptFL and FedCLIP on C_3, C_4 and C_6, while achieving similar error to them on clients C_1 and C_5.

Sensitivity. We explore the sensitivity of the hyperparameter λ_1 and λ_2 on OfficeHome and ISIC2019 datasets. As shown in Figure 6, a higher λ_1 is beneficial for client A, while a lower λ_2 is useful for client A. The optimal result can be

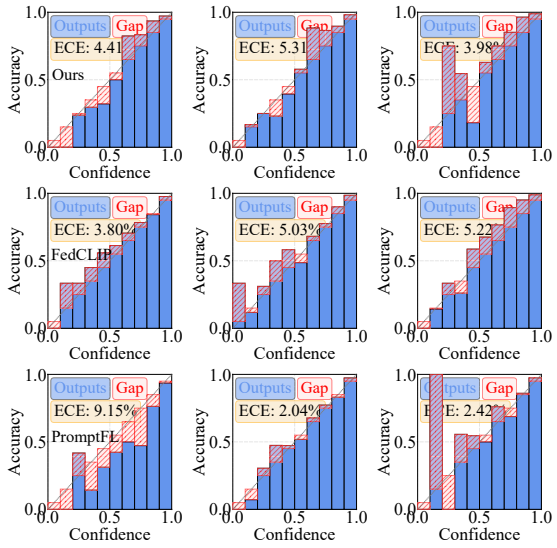

Figure 4: Reliability plots on OfficeHome dataset. Each column and row represent one client (A, C, P) and one method, respectively.

achieved by setting λ_1 to 0.5 and λ_2 to 0.1, respectively. For ISIC2019, a higher λ_1 leads to a better AVG F1 score, while an increase of λ_2 can reduce prediction diversity, leading to a lower test F1 score. Finally, the optimal result can be obtained by setting λ_1 to 1.0 and λ_2 to 0.1, respectively.

Adaptability. We replace the CLIP encoder with other backbones to verify its adaptability. Specifically, SigLIP [Zhai *et al.*, 2023] and EVA02-CLIP [Fang *et al.*, 2024] are used for experiments in the OfficeHome and ISIC2019 datasets. As illustrated in Figure 7, FedENC improves the performance of SigLIP and EVA02-CLIP on OfficeHome dataset by 26.53% and 2.9% in AVG, respectively. Furthermore, in the ISIC2019 dataset, FedENC yields an AVG ACC of 76.82% and 83.17%

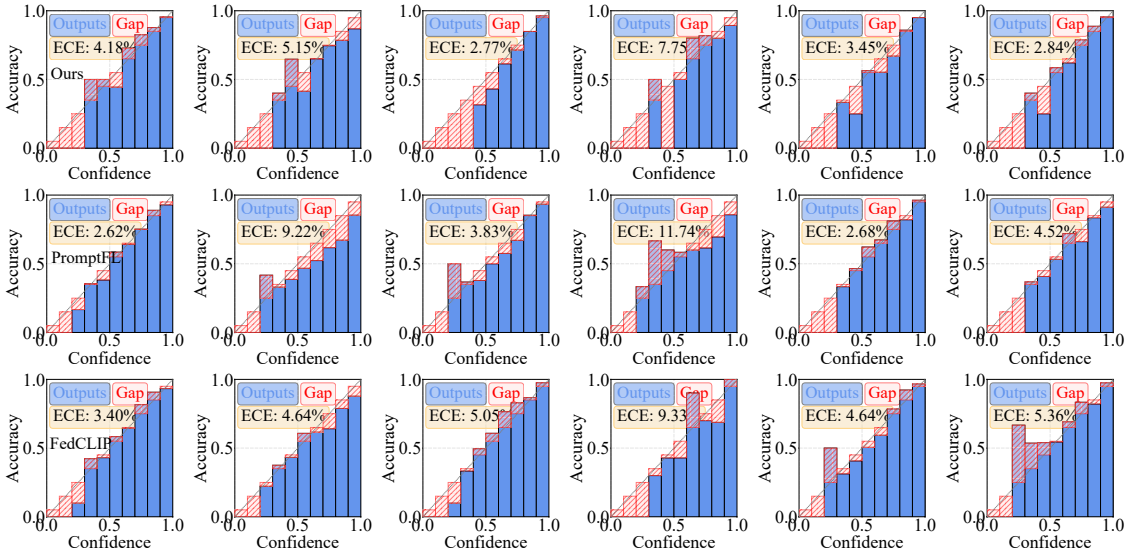

Figure 5: Reliability plots on ISIC2019 dataset. Each column and row represent one client (C_1 to C_6) and one method, respectively.

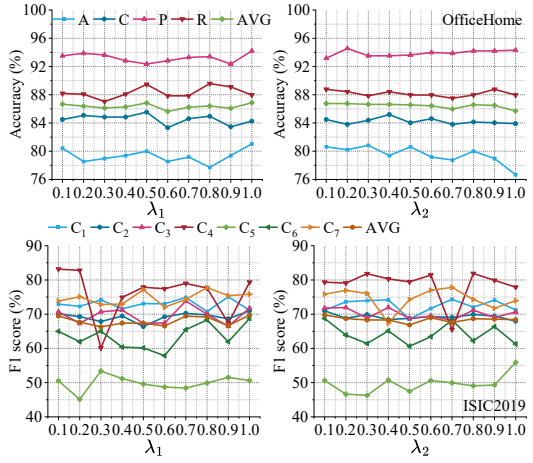

Figure 6: Test accuracy and F1 score on OfficeHome and ISIC2019 datasets with different λ_1 and λ_2 values.

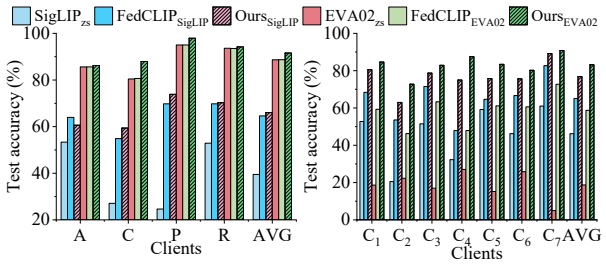

Figure 7: Test accuracy on OfficeHome and ISIC2019 with different backbones.

with SigLIP and EVA02-CLIP, outperforming the zero-shot baseline by 30.6% and 64.44%, respectively.

Scalability. To evaluate the scalability of FedENC on DomainNet, each domain is divided into five subdomains using a Dirichlet distribution following [Su *et al.*, 2024], resulting in a total of 30 clients. Figure 8 (left) illustrates the test ACC for FedENC and the baselines. FedENC achieves an AVG ACC of 66.91% on DomainNet, outperforming LoRA and PromptFL by 3.8% and 7.11%, respectively.

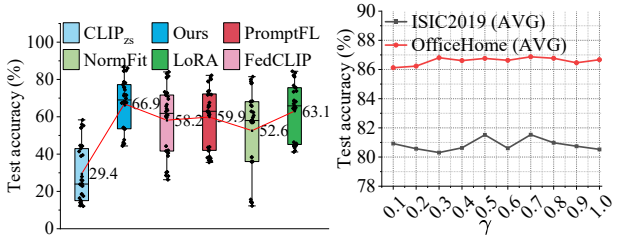

Figure 8: **Left:** Test accuracy under different client scales on DomainNet. **Right:** Test accuracy with different γ.

Impact of γ. We validate the impact of γ used in Eq. 6 on OfficeHome and ISIC2019 datasets. As illustrated in Figure 8 (right), using a lower γ value shows a performance drop (i.e., putting more weights on common classes), while setting γ to 0.7 leads to the highest AVG for both datasets. We note that a γ value higher than 0.7 leads to a lower performance, this finding suggests that a moderate weight on rare classes is important for feasible performance and vice versa.

5 Conclusion

This study presented FedENC, a novel CLIP-based FL framework to address data heterogeneity. FedENC learns shared knowledge across clients using a global adaptation module, while preserves local knowledge through two local expert modules. Furthermore, NC is introduced to mitigate heterogeneity, with a KL and cosine similarity based alignment technique to improve prediction consistency. Experimental results show that FedENC achieves remarkable performance with reasonable computation and communication overhead.

Acknowledgments

This research was funded by the Guangxi Science and Technology Base and Talent Project (#2022AC18004 and #2022AC21040), and the National Natural Science Foundation of China (Grant #82260360).

References

[Chen *et al.*, 2025] Weihang Chen, Cheng Yang, Jie Ren, Zhiqiang Li, and Zheng Wang. Optimizing personalized federated learning through adaptive layer-wise learning. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pages 1–9, 2025.

[Choi and Ko, 2025] Kyuri Choi and Youngjoong Ko. Robust fine-tuning for low-resource nlp: Combining adversarial and metric-based learning to mitigate overfitting. *Expert Systems with Applications*, page 127737, 2025.

[Codella *et al.*, 2018] Noel CF Codella, David Gutman, M Emre Celebi, Brian Helba, Michael A Marchetti, Stephen W Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, et al. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*, pages 168–172. IEEE, 2018.

[Combalia *et al.*, 2019] Marc Combalia, Noel CF Codella, Veronica Rotemberg, Brian Helba, Veronica Vilaplana, Ofer Reiter, Cristina Carrera, Alicia Barreiro, Allan C Halpern, Susana Puig, et al. Bcn20000: Dermoscopic lesions in the wild. *arXiv preprint arXiv:1908.02288*, 2019.

[Du and Li, 2025] Haizhou Du and Pengfei Li. Decoupling feature entanglement for personalized federated learning via neural collapse. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, pages 615–624, 2025.

[Fang *et al.*, 2024] Yuxin Fang, Quan Sun, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva-02: A visual representation for neon genesis. *Image and Vision Computing*, 149:105171, 2024.

[Flanders *et al.*, 2020] Adam E Flanders, Luciano M Prevedello, George Shih, Safwan S Halabi, Jayashree Kalpathy-Cramer, Robyn Ball, John T Mongan, Anouk Stein, Felipe C Kitamura, Matthew P Lungren, et al. Construction of a machine learning dataset through collaboration: the rsna 2019 brain ct hemorrhage challenge. *Radiology: Artificial Intelligence*, 2(3):e190211, 2020.

[Guo *et al.*, 2024] Tao Guo, Song Guo, Junxiao Wang, Xueyang Tang, and Wenchao Xu. Promptfl: Let federated participants cooperatively learn prompts instead of models – federated learning in age of foundation model. *IEEE Transactions on Mobile Computing*, 23(5):5179–5194, 2024.

[Hakim *et al.*, 2025] Gustavo A Vargas Hakim, David Osowiechi, Mehrdad Noori, Milad Cheraghalikhani, Ali Bahri, Moslem Yazdanpanah, Ismail Ben Ayed, and Christian Desrosiers. Clipartt: Adaptation of clip to new domains at test time. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 7092–7101. IEEE, 2025.

[Huang *et al.*, 2024] Yunshi Huang, Fereshteh Shakeri, Jose Dolz, Malik Boudiaf, Houda Bahig, and Ismail Ben Ayed. Lp++: A surprisingly strong linear probe for few-shot clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23773–23782, 2024.

[Li *et al.*, 2017] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE international conference on computer vision*, pages 5542–5550, 2017.

[Li *et al.*, 2023] Zexi Li, Xinyi Shang, Rui He, Tao Lin, and Chao Wu. No fear of classifier biases: Neural collapse inspired federated learning with synthetic and fixed classifier. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5319–5329, 2023.

[Li *et al.*, 2025] Ming Li, Pengcheng Xu, Junjie Hu, Zeyu Tang, and Guang Yang. From challenges and pitfalls to recommendations and opportunities: Implementing federated learning in healthcare. *Medical Image Analysis*, page 103497, 2025.

[Liao *et al.*, 2025] Tianchi Liao, Binghui Xie, Sheng Huang Lele Fu, Bowen Deng, Chuan Chen, and Zibin Zheng. Federated domain generalization with decision insight matrix. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pages 5689–5697, 2025.

[Loshchilov and Hutter, 2017] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[Lu *et al.*, 2023] Wang Lu, Xixu Hu, Jindong Wang, and Xing Xie. Fedclip: Fast generalization and personalization for clip in federated learning. *arXiv preprint arXiv:2302.13485*, 2023.

[Meng *et al.*, 2025] Lei Meng, Zhuang Qi, Lei Wu, Xiaoyu Du, Zhaochuan Li, Lizhen Cui, and Xiangxu Meng. Improving global generalization and local personalization for federated learning. *IEEE Transactions on Neural Networks and Learning Systems*, 36(1):76–87, 2025.

[Motamedi *et al.*, 2025] Azadeh Motamedi, Jae-Mo Kang, and Il-Min Kim. Normfit: A lightweight solution for few-shot federated learning with non-iid data. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, pages 1–24, 2025.

[Papyan *et al.*, 2020] Vardan Papyan, XY Han, and David L Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40):24652–24663, 2020.

[Peng *et al.*, 2019] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of*

the IEEE/CVF international conference on computer vision, pages 1406–1415, 2019.

[Phung *et al.*, 2025] Thu Hang Phung, Duong M Nguyen, Thanh Trung Huynh, Quoc Viet Hung Nguyen, Trong Nghia Hoang, and Phi Le Nguyen. Federated prompt-tuning with heterogeneous and incomplete multimodal client data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3936–3946, 2025.

[Qi *et al.*, 2025] Zhuang Qi, Sijin Zhou, Lei Meng, Han Hu, Han Yu, and Xiangxu Meng. Federated deconfounding and debiasing learning for out-of-distribution generalization. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pages 6084–6092, 2025.

[Qi *et al.*, 2026] Xin Qi, Tao Xu, Chengrun Dang, Zhuang Qi, Lei Meng, and Han Yu. Federated learning in oncology: bridging artificial intelligence innovation and privacy protection. *Information Fusion*, page 104154, 2026.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[Ringwald and Stiefelhagen, 2021] Tobias Ringwald and Rainer Stiefelhagen. Adaptiope: A modern benchmark for unsupervised domain adaptation. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 101–110, 2021.

[Su *et al.*, 2024] Shangchao Su, Mingzhao Yang, Bin Li, and Xiangyang Xue. Federated adaptive prompt tuning for multi-domain collaborative learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 15117–15125, 2024.

[Tschandl *et al.*, 2018] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*, 5(1):1–9, 2018.

[Vaicenavicius *et al.*, 2019] Juozas Vaicenavicius, David Widmann, Carl Andersson, Fredrik Lindsten, Jacob Roll, and Thomas Schön. Evaluating model calibration in classification. In *The 22nd international conference on artificial intelligence and statistics*, pages 3459–3467. PMLR, 2019.

[Venkateswara *et al.*, 2017] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5018–5027, 2017.

[Vickers and Elkin, 2006] Andrew J Vickers and Elena B Elkin. Decision curve analysis: a novel method for evaluating prediction models. *Medical Decision Making*, 26(6):565–574, 2006.

[Voigt and Von dem Bussche, 2017] Paul Voigt and Axel Von dem Bussche. The eu general data protection regulation (gdpr). *A practical guide, 1st ed., Cham: Springer International Publishing*, 10(3152676):10–5555, 2017.

[Wu and Chaddad, 2026] Yihang Wu and Ahmad Chaddad. Federated clip for resource-efficient heterogeneous medical image classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 26992–27000, 2026.

[Wu *et al.*, 2023] Nannan Wu, Li Yu, Xin Yang, Kwang-Ting Cheng, and Zengqiang Yan. Fediic: Towards robust federated learning for class-imbalanced medical image classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 692–702. Springer, 2023.

[Wu *et al.*, 2025] Yihang Wu, Ahmad Chaddad, Christian Desrosiers, Tareef Daqqaq, and Reem Kateb. Faa-clip: Federated adversarial adaptation of clip. *IEEE Internet of Things Journal*, 12(12):21091–21102, 2025.

[Wu *et al.*, 2026] Yihang Wu, Ahmad Chaddad, Sarah A Alkhodair, Tareef Daqqaq, and Reem Kateb. Fedclipot: Federated clip model via parameter reusing and optimal transport. *Information Fusion*, page 104324, 2026.

[Zanella and Ben Ayed, 2024] Maxime Zanella and Ismail Ben Ayed. Low-rank few-shot adaptation of vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1593–1603, 2024.

[Zhai *et al.*, 2023] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986, 2023.

[Zhou *et al.*, 2022] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16816–16825, 2022.