Neural Collapse-Inspired Multi-Label Federated Learning under Label-Distribution Skew

Can Peng[1], Yuyuan Liu[1], Yingyu Yang[1], Pramit Saha[1], Qianye Yang[1], and J. Alison Noble[1]

[1]University of Oxford, Oxford, United Kingdom

Abstract

Federated Learning (FL) enables collaborative model training across distributed clients while preserving data privacy, but remains challenging when client data are highly heterogeneous. These challenges are further amplified in multi-label scenarios, where inter-label dependencies and mismatches between local and global label relationships introduce additional optimization conflicts. While most FL studies focus on single-label classification, many real-world applications are inherently multi-label and often exhibit severe label skew across clients. To address this important yet underexplored problem, we propose FedNCA-ML, a novel FL framework that aligns client representations and learns discriminative, well-clustered features inspired by Neural Collapse (NC) theory. NC describes an ideal latent geometry where each class's features collapse to their mean, forming a maximally separated simplex. FedNCA-ML further introduces an attention-based module to extract class-specific representations, enabling more balanced learning under heavy label imbalance. These class-wise representations are then aligned via a shared NC-inspired structure, mitigating inter-client conflicts induced by heterogeneous local data and inconsistent label dependencies. In addition, we design regularisation losses to encourage compact and consistent feature clustering in the latent space. Experiments on five benchmark datasets under nine FL settings demonstrate the effectiveness of the proposed method, achieving improvements of up to 3.92% in class-wise AUC and 4.93% in class-wise F1 score.

1 Introduction

Federated Learning (FL) enables collaborative model training on sensitive data, particularly in medical imaging, where privacy regulations and legal constraints prohibit data sharing. However, most existing FL methods primarily target standard multi-class classification and overlook more realistic scenarios in which multiple objects or conditions co-occur within a single sample. For instance, many diseases are interrelated, and patients frequently present with multiple concurrent conditions. Meanwhile, each client (e.g., hospital) may have a uniquely skewed data distribution due to variations in geography, population demographics, medical facilities, and clinical expertise. In this paper, we study multi-label FL under label-skew settings, where each client's local data is highly imbalanced and can substantially deviate from the global distribution. Figure 1 illustrates this scenario, showing that clients possess heterogeneous data distributions and distinct label dependency patterns. Without sharing raw data, our goal is to collaboratively train a global model that performs well across all target classes.

The multi-label, label-skewed FL setting poses three key challenges. **(1) Imbalanced data distribution.** Local datasets often exhibit severe label imbalance, containing majority, minority, or even missing classes. Such skew drives each client to optimize toward its local distribution, typically overfitting dominant labels while under-training rare ones. Consequently, clients learn locally biased features that mainly benefit their own data, leading to a global model with impaired generalization. Moreover, imbalance is not limited to individual clients, since the overall label distribution is often imbalanced as well, further amplifying the difficulty of balanced learning. **(2) Multi-label co-occurrence bias.** Multi-label data further exacerbates the imbalance due to label co-occurrence. Frequent labels often appear alongside many others and dominate the training signal, suppressing the learning of discriminative features for minority labels and making rare conditions harder to recognize. **(3) Cross-client inconsistency in label distributions and relationships.** Under FL, clients differ in both label frequencies and label dependency structures. These mismatches lead to optimization conflicts across clients, complicating collaborative training and hindering global convergence and generalization.

To address these challenges, we propose Federated Neural Collapse Alignment for Multi-Label Learning (**FedNCA-ML**). FedNCA-ML is a unified representation learning framework for multi-label FL, inspired by Neural Collapse (NC) theory [1], that promotes a consistent and discriminative feature geometry across heterogeneous clients. It is designed with two goals in mind. First, it encourages the model to learn balanced, class-discriminative representations, so that tail labels receive comparable attention to head labels. Second, it mitigates client drift by

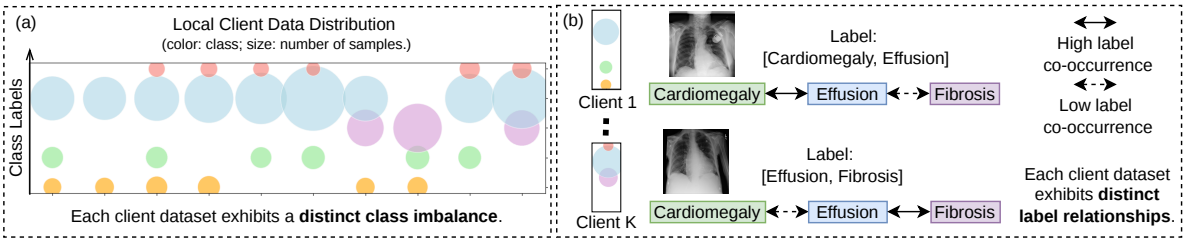


Figure 1: Multi-label, label-skewed federated learning poses three key challenges: (1) heterogeneous client data with distinct class-imbalance patterns, (2) amplified heterogeneity due to multi-label co-occurrence structure, and (3) cross-client inconsistency in both data distributions and label relationships.

guiding local training toward a shared global geometry, reducing the tendency of each client to over-specialize to its own data distribution. NC theory provides a geometric lens for this purpose. It shows that, when a classifier is trained to saturation on a balanced multi-class dataset, class-mean features converge to a simplex configuration and align with an Equiangular Tight Frame (ETF). This provides a principled geometric prior and has motivated a growing line of NC-inspired methods that encourage such structured representations under non-ideal training conditions and downstream tasks [2–4]. FedNCA-ML introduces a shared global ETF prior to promote client-agnostic clustering and align local models within a common feature space that supports all classes, thereby reducing overfitting to client-specific biases.

Furthermore, the NC theory was originally developed for single-label classification, where each image representation corresponds to exactly one class. In multi-label classification, a single shared representation is often insufficient to capture class-specific evidence and complex label relationships. FedNCA-ML introduces an attention-based module that extracts class-wise representations from the shared image features, effectively reformulating multi-label learning into a set of per-class subproblems compatible with NC-style alignment. Consequently, ETF anchoring is applied only to these class-wise representations, rather than enforcing mutual exclusivity on the shared backbone space, allowing semantic proximity to remain naturally preserved in the backbone features. Finally, two complementary regularizers further improve compactness and robustness. A rejection loss suppresses noisy negative features, while a contrastive loss promotes tight intra-class clustering and clear inter-class separation. The key contributions of this paper are as follows:

- We formalize the problem of multi-label FL under label skew, where clients differ in both label frequencies and label co-occurrence patterns.
- We propose FedNCA-ML, an NC–inspired representation alignment framework that enforces a shared ETF geometry across clients to mitigate representation drift and improve balanced learning.
- We introduce a class-wise attention mechanism that enables NC alignment in multi-label settings while preserving semantic relationships in the shared backbone feature space.
- We introduce complementary rejection and contrastive regularizers that enhance intra-class compactness and inter-class separation under heterogeneous label distributions.

2 Related Work

Heterogeneous Federated Learning. FedAvg [5], which iteratively aggregates client models via weighted parameter averaging, remains the foundation of most FL methods. However, its performance often degrades under non-IID data. Such heterogeneity typically arises from quantity skew, label distribution skew, and feature distribution skew. To mitigate these issues in multi-class classification, many FL methods have been proposed. Existing methods generally address heterogeneity from three perspectives. **(1) Local learning phase.** These methods steer client-side optimization by incorporating regularization terms [6–8] and/or auxiliary objectives [9–12] to reduce the drift between local updates and the global model. **(2) Post-learning phase.** These methods correct the divergence among local models at the server during aggregation, either by exchanging additional information [13] or by using more robust aggregation strategies [14]. **(3) Pre-learning phase.** This direction predefines a shared, fixed classifier or decision boundary to align local training, encouraging more consistent representation learning without modifying the core optimization procedure [2, 15]. In this work, we study multi-label FL under quantity-skewed and label-skewed distributions. Directly extending existing strategies to multi-label FL is often less effective, particularly for pre-learning methods, because multiple co-occurring labels per sample make it difficult to enforce clear semantic clustering and class separation in the latent space. To cope with severe class imbalance, missing labels on some clients, and heterogeneous label co-occurrence patterns, we propose a pre-learning approach that

explicitly organizes the latent feature space, improving robustness and generalization under label-skewed multi-label FL.

Multi-Label Federated Learning. Multi-label FL poses unique challenges beyond single-label FL due to heterogeneous label co-occurrence across clients. FedMLP [16] targets partial class annotation under task heterogeneity, where each client observes only a subset of labels with incomplete annotations. To alleviate missing annotations, it exchanges local model parameters and class-wise prototypes with the server to enable pseudo-labeling. FedLGT [17] studies a closely related setting and leverages label text to model shared label structure across clients. Building on C-Tran [18], it adopts frozen CLIP [19] text embeddings to enforce a consistent label-embedding space, thereby reducing cross-client divergence in label dependencies.

Neural Collapse-Inspired Methods. Recent studies show that overparameterized networks trained to saturation on balanced multi-class datasets exhibit Neural Collapse (NC) in the terminal phase of training (TPT) [1]. In this regime, last-layer features concentrate around their class means, and the class prototypes converge to an Equiangular Tight Frame (ETF), yielding a highly symmetric and well-separated latent geometry. This observation has motivated NC-inspired methods that explicitly encourage such structure across a range of tasks [2–4, 20, 21]. However, NC has been studied mainly in multi-class classification, while multi-label settings remain less explored. Li *et al.* [22] reported a generalized NC behavior in balanced multi-label training that features from single-label samples still follow a simplex ETF structure, whereas multi-label features are well approximated as scaled averages of single-label prototypes. MLC-NC [23] exploits NC to improve long-tailed multi-label classification in a centralized setting. To the best of our knowledge, we are the first to investigate multi-label FL through the lens of NC, leveraging it to mitigate interference induced by heterogeneous client distributions and label relationships.

3 Preliminaries

3.1 Problem Formulation

We consider a multi-label, label-skewed FL setting with K clients collaboratively training a shared global model. Each client k holds a private dataset $\mathcal{D}_k = \{(x_i^k, y_i^k)\}_{i=1}^{N_k}$ of size N_k, where the input image $x_i^k \in \mathbb{R}^{H\times W\times 3}$ and the label $y_i^k \in \{0,1\}^C$ is a multi-hot vector over the C classes. Let $\mathcal{C} = \{1, \dots, C\}$ denote the global class index set. Due to label skew, client k only observes a subset $\mathcal{C}_k \subseteq \mathcal{C}$, and even for shared classes, the class-conditional distributions can differ across clients. The goal is to learn a single global model that accurately recognizes all classes in $\mathcal{C}$ without sharing any local client data.

3.2 Neural Collapse (NC)

In this section, we first introduce the structure of the simplex ETF, followed by a discussion of the key properties of NC [1, 22].

Simplex Equiangular Tight Frame (ETF). A simplex ETF matrix $\mathbf{M} = [\mathbf{m}_c]_{c=1}^{C} \in \mathbb{R}^{d\times C}$ composed of C column vectors, each corresponding to a class prototype in $\mathbf{m}_c \in \mathbb{R}^d$. A standard construction is:

$$\mathbf{M} = \sqrt{\frac{C}{C-1}}\,\mathbf{U}\left(\mathbf{I}_C - \frac{1}{C}\mathbf{1}_C\mathbf{1}_C^\top\right), \tag{1}$$

where $\mathbf{U} \in \mathbb{R}^{d\times C}$ denotes a rotation orthogonal matrix ($\mathbf{U}^\top\mathbf{U} = \mathbf{I}_C$). $\mathbf{I}_C$ is the $C \times C$ identity matrix. $\mathbf{1}_C$ is the C-dimensional all-ones vector.

This yields unit-norm columns with equal pairwise inner products:

$$\mathbf{m}_a^\top \mathbf{m}_b = \begin{cases} 1, & a = b, \\ -\dfrac{1}{C-1}, & a \neq b, \end{cases} \qquad a, b \in \mathcal{C}. \tag{2}$$

Hence, all prototypes have equal ℓ_2 norm and identical pairwise angles, forming a centered regular simplex.

Neural Collapse properties. At the end of training, networks exhibit the NC phenomenon. Empirically, the following regularities are observed:

$\mathcal{NC}_1$**: Variability Collapse.** Within-class feature variance collapses, causing features of the same class to concentrate around their class mean.

$\mathcal{NC}_2$**: Convergence to Simplex ETF.** Class means arrange themselves as the vertices of a simplex ETF, forming a maximally symmetric and equidistant configuration.

$\mathcal{NC}_3$**: Self-Duality.** Upon appropriate rescaling, the final-layer classifier weights align with the class means, exhibiting the same simplex ETF geometry.

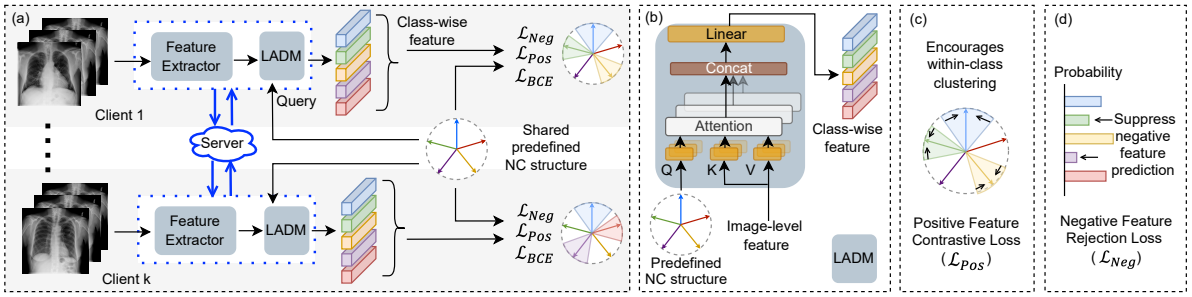


Figure 2: Overview of the proposed FedNCA-ML framework for multi-label label-skewed FL. Subfigure (a) shows the overall architecture, while Subfigures (b)–(d) illustrate the Label-Aware Disentanglement Module (LADM) and the regularization losses. The attention-based LADM extracts label-specific features from image-level features. A predefined ETF matrix acts as both the shared classifier and the source of class-wise query embeddings, ensuring consistent local training across clients. Two regularisation terms are further incorporated to suppress noisy negative features and promote compact intra-class clustering in the latent feature space.

4 Proposed Method

In multi-label, label-skewed FL, each client's data is highly imbalanced and contains missing labels. Clients can also exhibit distinct label co-occurrence patterns, and both label distributions and dependencies may vary widely due to heterogeneous data collection. Together, these factors amplify client drift and hinder stable convergence during local training and aggregation. To address these challenges, we propose **FedNCA-ML**, which leverages an NC-inspired simplex ETF as a shared feature-space reference to encourage balanced learning across classes and provide a consistent decision boundary for aligning local training. An overview is shown in Fig. 2. FedNCA-ML consists of three key components. (1) We employ an attention-based module (LADM; Fig. 2b) to extract class-wise features from the backbone's shared image-level representation, and apply ETF anchoring only to these class-wise features (Fig. 2a). This avoids directly constraining the shared image-level representation, allowing semantic proximity to be naturally preserved in the backbone features. (2) We use a predefined ETF as a globally fixed classifier, providing a consistent alignment target across clients (Fig. 2a). (3) We introduce two additional regularisers (Fig. 2c–d) to further improve clustering in the latent space.

4.1 Label-Aware Disentanglement Module

In principle, a single pooled embedding can encode multiple labels. However, under severe label imbalance, it often entangles class-specific evidence, inducing gradient interference and bias toward majority labels. Motivated by disentanglement-based multi-label learning, we introduce a Label-Aware Disentanglement Module (LADM) to extract class-wise features from backbone representations [24–26]. LADM is further inspired by an analogy to object detection that overlapping bounding boxes indicate that the same region may support multiple instances. Similarly, in multi-label recognition, a single region can provide evidence for multiple semantic categories, particularly when labels co-occur. Accordingly, LADM adopts a DETR-style cross-attention mechanism [27].

Concretely, a set of class queries attends to a grid of image tokens to produce class-specific representations. Unlike DETR, where queries represent instances and are trained with box-level supervision, LADM assigns one fixed query to each class and learns solely from image-level annotations. Each query acts as a soft region selector, aggregating spatial evidence relevant to its class while leveraging contextual information across regions. This yields a set of disentangled, class-aware feature vectors for per-class prediction. In FL, non-IID local data can induce inconsistent class-wise feature extraction across clients, destabilizing aggregation. To encourage global consistency, LADM shares a single query matrix across all clients, anchoring each class to the same query direction. This shared design reduces inter-client conflicts during aggregation, and its effectiveness is validated in the ablation study (Section 5.4, Table 7).

Formally, for client k, each sample $x_i^k \in \mathcal{D}_k$ ($i \in [N_k]$) is encoded by the backbone into a spatial feature map:

$$\mathbf{F}_i^k \in \mathbb{R}^{d \times H' \times W'}, \tag{3}$$

where d denotes the channel dimension and $H' \times W'$ is the feature resolution. The feature map is then flattened into a sequence of $S = H'W'$ tokens:

$$\mathbf{Z}_i^k = \text{reshape}(\mathbf{F}_i^k)^\top \in \mathbb{R}^{S \times d}, \tag{4}$$

and, for notational convenience, we omit the client index in subsequent equations and denote the token sequence as $\mathbf{Z}_i$. To preserve spatial structure, we add a fixed 2D sine–cosine positional embedding [27] to the key and value

projections:

$$\tilde{\mathbf{Z}}_i = \mathbf{Z}_i + \mathbf{P}, \quad \mathbf{P} = \mathrm{PE}_{2\mathrm{D}}(H', W', d), \tag{5}$$

where $\mathrm{PE}_{2\mathrm{D}}(\cdot)$ denotes the fixed sine–cosine positional encoding. This embedding introduces spatial awareness without adding learnable parameters, thereby ensuring a consistent inductive bias across clients.

LADM employs a 4-head cross-attention to obtain class-specific features from the encoded image tokens. We define a shared simplex ETF matrix as

$$\mathbf{M} = [\mathbf{m}_c]_{c=1}^{C} \in \mathbb{R}^{d \times C}, \tag{6}$$

where each column $\mathbf{m}_c$ serves both as a fixed classifier weight and as a class-specific query vector. Given the query $\mathbf{m}_c$ and the encoded image tokens $\tilde{\mathbf{Z}}_i$, LADM computes the class-specific feature using multi-head cross-attention:

$$\mathbf{h}_{ic} \;=\; \mathrm{MultiHeadAttn}\big(\mathbf{m}_c^{\top},\, \tilde{\mathbf{Z}}_i,\, \tilde{\mathbf{Z}}_i\big), \quad c \in \mathcal{C}. \tag{7}$$

The resulting features are stacked to form the class-feature matrix:

$$\mathbf{H}_i \;=\; \begin{bmatrix} \mathbf{h}_{i1}^{\top} \\ \vdots \\ \mathbf{h}_{iC}^{\top} \end{bmatrix} \in \mathbb{R}^{C \times d}. \tag{8}$$

4.2 Neural Collapse-Inspired Feature Alignment

After obtaining class-wise features with LADM, each sample $i \in [N_k]$ yields a feature $\mathbf{h}_{ic} \in \mathbb{R}^d$ for every class $c \in \mathcal{C}$. To prevent client-specific classifier drift, we fix the classifier to a globally shared simplex ETF, which also serves as the LADM query matrix. This shared classifier enhances class separability and mitigates model drift caused by class imbalance and missing labels. Given $\mathbf{h}_{ic}$ and its corresponding prototype $\mathbf{m}_c$, we compute the class logit and apply the sigmoid function to obtain a binary prediction:

$$\hat{y}_{ic} \;=\; \sigma\big(\mathbf{h}_{ic}^{\top}\mathbf{m}_c\big), \tag{9}$$

and train with a binary cross-entropy (BCE) loss:

$$\mathcal{L}_{\mathrm{BCE}} = -\frac{1}{N}\frac{1}{C}\sum_{i=1}^{N}\sum_{c=1}^{C}\Big[y_{ic}\,\log \hat{y}_{ic} + (1-y_{ic})\,\log(1-\hat{y}_{ic})\Big]. \tag{10}$$

Here, $y_{ic} \in \{0, 1\}$ denotes the ground-truth label indicator, and $\sigma(\cdot)$ is the sigmoid function.

4.3 Reducing Noise and Improving Clustering

Each sample produces C class-wise features $\{\mathbf{h}_{ic}\}_{c=1}^{C}$. Let the positive and negative class sets for sample i be defined as $\mathcal{C}_i^{+} = \{\, c \mid y_{ic} = 1 \,\}$ and $\mathcal{C}_i^{-} = \{\, c \mid y_{ic} = 0 \,\}$, respectively. We introduce two regularization terms: one to suppress noise from negative features, and another to promote compact clustering of positive features.

Negative Feature Rejection Loss. While $\mathcal{L}_{\mathrm{BCE}}$ discourages alignment between a negative feature $\mathbf{h}_{ic}$ and its corresponding prototype $\mathbf{m}_c$ when $y_{ic} = 0$, it does not prevent $\mathbf{h}_{ic}$ from spuriously aligning with other class prototypes. To address this, we introduce a penalty on high similarity between each negative feature and all non-self prototypes:

$$\hat{s}_{icr} \;=\; \sigma\big(\mathbf{h}_{ic}^{\top}\mathbf{m}_r\big), \quad c \in \mathcal{C}_i^{-},\; r \in \mathcal{C} \setminus \{c\}, \tag{11}$$

and define the negative feature rejection loss as

$$\mathcal{L}_{\mathrm{Neg}} = -\frac{1}{N}\sum_{i=1}^{N}\frac{1}{|\mathcal{C}_i^{-}|}\sum_{c \in \mathcal{C}_i^{-}}\frac{1}{C-1}\sum_{\substack{r=1 \\ r \neq c}}^{C}\mathbb{I}\big(\hat{s}_{icr} > \tau\big)\,\log\big(1-\hat{s}_{icr}\big) \tag{12}$$

The indicator $\mathbb{I}(\cdot)$ filters out low-similarity pairs, ensuring that only confident negatives contribute to the loss. In our experiments, we set $\tau = 0.3$.

Positive Feature Contrastive Loss. To encourage compact and discriminative class-wise clustering in the latent feature space, we introduce a contrastive loss. This loss drives each positive feature $\mathbf{h}_{ic}$ to be closer to its own prototype than to others through a prototype-based softmax:

$$\mathcal{L}_{\mathrm{Pos}} = -\frac{1}{N}\sum_{i=1}^{N}\frac{1}{|\mathcal{C}_i^{+}|}\sum_{c \in \mathcal{C}_i^{+}}\log\frac{\exp\big(\mathbf{h}_{ic}^{\top}\mathbf{m}_c\big)}{\sum_{r=1}^{C}\exp\big(\mathbf{h}_{ic}^{\top}\mathbf{m}_r\big)}. \tag{13}$$

Algorithm 1 FedNCA-ML

Input: K clients with datasets $\{\mathcal{D}_k\}_{k=1}^{K}$; initial global model w_0; predefined ETF matrix $\mathbf{M}$; learning rate η; local epochs E; communication rounds T.

Server executes:
Initialize $w \leftarrow w_0$
for $t = 0$ to $T-1$ **do** ▷ communication rounds
  for each client $k \in \{1, \dots, K\}$ in parallel **do**
    $w_{t+1}^{k} \leftarrow \text{CLIENTUPDATE}(\mathcal{D}_k, w_t, \mathbf{M})$
  end for
  $w_{t+1} \leftarrow \frac{1}{K}\sum_{k=1}^{K} w_{t+1}^{k}$ ▷ model aggregation
  Broadcast w_{t+1} to all clients
end for

function CLIENTUPDATE($\mathcal{D}_k, w, \mathbf{M}$)
  for $e = 0$ to $E-1$ **do** ▷ local epochs
    for each batch $(x, y) \subset \mathcal{D}_k$ **do**
      $\mathbf{F} \leftarrow \text{FEATUREEXTRACTOR}(w, x)$
      $\mathbf{H} \leftarrow \text{LADM}(\mathbf{F}, \mathbf{M})$ ▷ Eqs. 3, 4, 5, 6, 7, 8
      Compute prediction $\hat{y}_c = \sigma(\mathbf{h}_c^{\top}\mathbf{m}_c)$ ▷ Eqs. 9
      Compute loss $\mathcal{L}_{\text{total}}(w; \mathbf{M}, \mathbf{H}, \hat{\mathbf{y}})$ ▷ Eqs.10,11,12,13,14
      Update $w \leftarrow w - \eta \nabla_w \mathcal{L}_{\text{total}}$
    end for
  end for
  return w
end function

Total Objective. The total training loss is defined as:

$$\mathcal{L}_{\text{total}} \;=\; \mathcal{L}_{\text{BCE}} + \lambda_1 \mathcal{L}_{\text{Neg}} + \lambda_2 \mathcal{L}_{\text{Pos}}. \tag{14}$$

where $\lambda_1, \lambda_2 \geq 0$ are weighting coefficients that balance the contributions of the regularization terms.

4.4 FedNCA-ML

In summary, we propose FedNCA-ML for multi-label, label-skewed FL. Our design focuses on client-side training and introduces an NC-inspired pre-alignment strategy that leverages a shared geometric prior to anchor class-wise representations to consistent directions across clients, mitigating drift caused by heterogeneous local distributions. On the server, aggregation follows the standard FedAvg procedure, averaging client-updated weights in each communication round. The overall pipeline is summarized in Algorithm 1.

5 Experiments

5.1 Dataset and Evaluation Metric

Datasets. We evaluate the proposed method on both general computer vision (CV) and medical imaging benchmarks. For general CV, we use CIFAR-10 [28], PASCAL VOC [29], and MS COCO [30]. Since CIFAR-10 is originally a multi-class dataset, we follow [22] to construct a multi-label variant by composing multiple images into a single composite image and using the union of their labels as the ground truth. For medical imaging, we use DermaMNIST [31] and ChestX-ray14 [32]. DermaMNIST contains 7 skin disease categories in a single-class format and is converted to multi-label using the same strategy as CIFAR-10. ChestX-ray14 is naturally multi-label, comprising 14 thoracic disease categories plus an additional "No Finding" label. Since a significant portion of the dataset, 57% of the training data, is "No Finding" samples (negative cases with all-zero labels), we distribute these samples evenly across all clients. This setup mimics a realistic clinical scenario in which healthy cases are prevalent, while disease cases are relatively rare and unevenly distributed. Detailed information about the datasets and local data distributions under various experimental settings is provided in the Appendix A.

Evaluation Metric. Given our focus on label-skewed data distributions, we report both instance-wise (micro) and class-wise (macro) performance metrics. The macro metric provides a balanced evaluation across classes, mitigating bias toward frequent categories. Following standard practice, we report AUC and F1 scores for CIFAR-10, VOC, COCO, and DermaMNIST, and AUC for ChestX-ray14. Together, these metrics comprehensively evaluate overall performance and class-level behaviour.

5.2 Task Setup and Implementation Details

Task Setup. We simulate an FL system with 10 clients to mimic a potential real-world clinical deployment and adopt full client participation in each round. To model heterogeneity, we generate non-IID client distributions by partitioning data with a Dirichlet prior parameterized by the concentration factor β. To further emulate label

Table 1: Comparisons on multi-label CIFAR-10 [28]. Missing-class scenarios are controlled by the class-presence ratio (γ), and non-IID client distributions are generated with a Dirichlet concentration parameter (β). We report both class-wise (macro) and instance-wise (micro) performance.

Method	$\beta = 0.5, \gamma = 0.5$ ($\leq$ 5 of 10 classes/client)				$\beta = 0.1, \gamma = 0.5$ ($\leq$ 5 of 10 classes/client)			
	macro-AUC	macro-F1	micro-AUC	micro-F1	macro-AUC	macro-F1	micro-AUC	micro-F1
Centralized	92.20 ±0.36	61.30 ±0.94	90.04 ±0.65	62.02 ±0.62	92.20 ±0.36	61.30 ±0.94	90.04 ±0.65	62.02 ±0.62
FedAvg [5]	82.29 ±0.46	39.47 ±1.59	81.48 ±0.78	40.26 ±1.09	78.92 ±0.39	31.17 ±0.51	77.62 ±0.24	35.07 ±0.47
FedCurv [7]	82.53 ±0.30	39.96 ±1.28	82.10 ±0.29	40.34 ±1.04	79.06 ±0.51	31.00 ±1.21	77.46 ±0.47	35.03 ±0.54
FedProx [6]	82.45 ±0.34	39.22 ±0.74	81.77 ±0.48	39.66 ±0.53	78.82 ±0.55	30.74 ±0.47	77.40 ±0.31	35.02 ±0.27
SCAFFOLD [13]	82.51 ±0.44	39.98 ±1.41	82.08 ±0.53	40.26 ±1.26	79.00 ±0.23	31.38 ±0.31	77.72 ±0.15	35.54 ±0.16
SphereFed [15]	83.63 ±1.50	42.58 ±3.20	83.50 ±1.87	43.18 ±2.34	80.62 ±1.46	36.83 ±1.39	78.37 ±0.95	38.18 ±1.40
FedLGT [17]	83.52 ±0.68	43.60 ±1.68	83.36 ±0.71	44.03 ±1.71	80.54 ±0.43	36.30 ±0.82	**80.65** ±0.68	39.24 ±0.71
FedNCA-ML	**87.55** ±0.31	**48.17** ±1.65	**87.00** ±0.31	**48.61** ±1.64	**83.80** ±0.54	**38.09** ±0.62	78.90 ±0.66	**41.60** ±0.67

Table 2: Comparisons on PASCAL VOC [29].

Method	$\beta = 0.05, \gamma = 0.5$ ($\leq$ 10 of 20 classes/client)				$\beta = 0.01, \gamma = 0.5$ ($\leq$ 10 of 20 classes/client)			
	macro-AUC	macro-F1	micro-AUC	micro-F1	macro-AUC	macro-F1	micro-AUC	micro-F1
Centralized	95.48 ±0.11	74.61 ±0.10	96.11 ±0.12	76.94 ±0.09	95.48 ±0.11	74.61 ±0.10	96.11 ±0.12	76.94 ±0.09
FedAvg [5]	93.54 ±0.23	57.57 ±0.76	94.13 ±0.33	64.67 ±0.35	93.31 ±0.11	47.73 ±1.03	93.05 ±0.19	60.46 ±0.43
FedCurv [7]	93.53 ±0.33	57.36 ±0.41	94.10 ±0.20	64.60 ±0.26	93.13 ±0.29	48.54 ±0.68	**93.81** ±0.09	60.49 ±1.16
FedProx [6]	93.55 ±0.20	57.04 ±0.24	94.04 ±0.17	64.43 ±0.29	93.14 ±0.25	49.49 ±0.68	93.80 ±0.17	62.18 ±0.51
SCAFFOLD [13]	93.17 ±0.20	57.44 ±0.33	94.03 ±0.11	64.95 ±0.21	**93.41** ±0.17	49.64 ±0.77	93.60 ±0.21	61.44 ±0.56
SphereFed [15]	83.72 ±1.82	33.49 ±1.15	84.38 ±1.10	38.19 ±1.17	84.25 ±2.44	32.44 ±0.21	85.94 ±3.67	35.18 ±1.30
FedLGT [17]	91.93 ±0.42	62.11 ±0.44	91.75 ±0.42	67.58 ±0.45	91.25 ±0.27	56.53 ±2.42	91.46 ±0.41	63.51 ±1.73
FedNCA-ML	**93.82** ±0.36	**64.28** ±0.05	**94.52** ±0.15	**67.61** ±0.49	93.01 ±0.22	**61.08** ±0.10	93.70 ±0.18	**65.05** ±0.53

Table 3: Comparisons on MS COCO [30].

Method	$\beta = 0.05, \gamma = 0.75$ ($\leq$ 60 of 80 classes/client)			
	macro-AUC	macro-F1	micro-AUC	micro-F1
Centralized	94.29 ±0.13	58.43 ±0.02	95.60 ±0.20	64.40 ±0.50
FedAvg [5]	93.66 ±0.16	53.21 ±0.49	94.74 ±0.12	60.35 ±0.15
FedCurv [7]	94.03 ±0.11	54.34 ±0.52	95.40 ±0.32	60.82 ±0.21
FedProx [6]	93.74 ±0.15	53.65 ±0.27	**94.87** ±0.10	60.58 ±0.15
SCAFFOLD [13]	93.55 ±0.22	53.50 ±0.82	94.68 ±0.12	60.15 ±0.86
SphereFed [15]	84.87 ±1.13	35.26 ±0.48	80.62 ±0.56	50.58 ±0.72
FedLGT [17]	90.90 ±0.25	55.68 ±0.83	92.37 ±0.32	**62.76** ±0.76
FedNCA-ML	**94.10** ±0.18	**56.28** ±0.32	94.71 ±0.20	61.71 ±0.52

skew, we introduce a class-presence ratio γ, which restricts the set of classes observed by each client, simulating missing-class scenarios.

Implementation Details. All experiments adopt ResNet-18 [33] as the feature extractor. We train each global model for 100 communication rounds with one local epoch per round, and select the final checkpoint based on the best validation performance. We use a batch size of 32, an initial learning rate of 1×10^{-4}, and AdamW with weight decay 0.01. Models on CIFAR-10 are trained from scratch, while models on the other datasets are initialized with ImageNet-pretrained weights. Accordingly, we set the negative feature rejection coefficient λ_1 to 1 for CIFAR-10 to provide stronger regularisation during scratch training, and to 0.01 for pretrained models, which typically require milder regularisation. The positive feature contrastive coefficient λ_2 is set to 1 across all experiments. We repeat all experiments three times with different random seeds and report the mean and standard deviation.

5.3 Performance Comparison

To evaluate the effectiveness of the proposed method, we compare FedNCA-ML with state-of-the-art approaches on five datasets under nine label-skewed FL settings. As summarized in Tables 1, 2, 3, 4, and 5, FedNCA-ML achieves the best class-wise performance in most cases, highlighting its ability to deliver balanced and generalizable predictions under heterogeneous FL distributions. Specifically, on multi-label CIFAR-10 (Tables 1), under non-IID Dirichlet settings of $\beta = 0.5$ ($\beta = 0.1$) with a maximum of 5 out of 10 classes per client, FedNCA-ML surpasses the second-best approach by 3.92% (3.18%) in class-wise AUC and 4.57% (1.26%) in class-wise F1 score. On multi-label DermaMNIST Tables 4, under $\beta = 0.5$ ($\beta = 0.1$) with up to 5 of 7 classes per client, it achieves improvements of 0.40% (0.77%) in AUC and 0.46% (4.93%) in F1 score. On VOC (Tables 2), under more challenging settings of $\beta = 0.05$ ($\beta = 0.01$) with up to 10 of 20 classes per client, FedNCA-ML outperforms the second-best approach by 2.17% (4.55%) in class-wise F1 score. On COCO (Tables 3), under $\beta = 0.05$ with up to

Table 4: Comparisons on multi-label DermaMNIST [31].

Method	$\beta = 0.5, \gamma = 0.71$ ($\leq$ 5 of 7 classes/client)				$\beta = 0.1, \gamma = 0.71$ ($\leq$ 5 of 7 classes/client)			
	macro-AUC	macro-F1	micro-AUC	micro-F1	macro-AUC	macro-F1	micro-AUC	micro-F1
Centralized	91.83 ±0.40	64.38 ±1.01	94.48 ±0.25	74.12 ±0.70	91.83 ±0.40	64.38 ±1.01	94.48 ±0.25	74.12 ±0.70
FedAvg [5]	89.72 ±0.29	54.88 ±0.97	92.51 ±0.36	**68.29** ±0.45	83.88 ±1.21	42.79 ±2.03	87.23 ±0.86	62.81 ±0.83
FedCurv [7]	89.48 ±0.18	54.92 ±1.11	92.26 ±0.36	67.97 ±0.92	85.53 ±1.37	43.45 ±1.20	88.87 ±0.50	**64.37** ±0.26
FedProx [6]	89.69 ±0.31	55.78 ±0.94	92.06 ±0.46	68.01 ±1.46	84.45 ±1.34	43.54 ±1.35	88.28 ±1.07	63.25 ±0.63
SCAFFOLD [13]	89.72 ±0.49	55.85 ±0.68	92.45 ±0.32	68.09 ±0.59	84.59 ±1.03	43.20 ±1.00	88.91 ±0.90	63.82 ±1.16
SphereFed [15]	85.09 ±0.87	43.41 ±1.73	89.65 ±0.55	65.24 ±0.58	81.78 ±1.14	40.90 ±1.79	86.59 ±1.87	54.94 ±1.12
FedLGT [17]	87.63 ±0.71	55.82 ±1.34	91.19 ±0.51	67.93 ±0.45	84.91 ±0.79	45.61 ±1.16	89.42 ±0.65	61.15 ±2.43
FedNCA-ML	**90.12** ±0.51	**56.31** ±0.65	**92.85** ±0.48	68.20 ±0.38	**86.30** ±0.85	**50.54** ±1.37	**89.74** ±0.98	63.76 ±1.31

Table 5: Comparisons on ChestX-ray14 [32] with $\leq$ 7 of 14 disease classes/client.

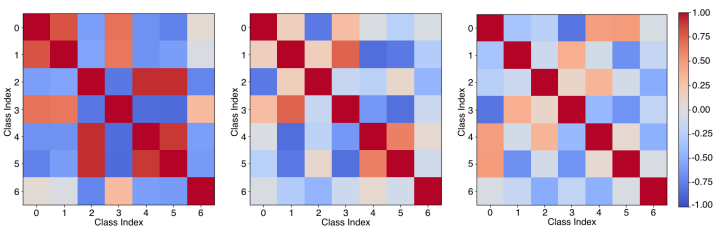
Method	$\beta = 0.5, \gamma = 0.5$		$\beta = 0.1, \gamma = 0.5$	
	macro-AUC	micro-AUC	macro-AUC	micro-AUC
Centralized	71.66 ±0.12	79.54 ±0.36	71.66 ±0.12	79.54 ±0.36
FedAvg [5]	69.02 ±0.24	**73.18** ±0.15	69.05 ±0.78	**77.90** ±0.20
FedCurv [7]	68.72 ±1.19	72.15 ±0.17	69.34 ±0.48	77.36 ±0.27
FedProx [6]	69.02 ±0.21	72.04 ±0.59	69.59 ±0.23	77.43 ±0.24
SCAFFOLD [13]	69.42 ±0.39	72.49 ±0.38	67.45 ±0.54	**77.90** ±0.69
SphereFed [15]	58.35 ±0.57	69.51 ±0.56	61.96 ±0.16	73.59 ±1.02
FedLGT [17]	69.86 ±0.76	72.27 ±1.07	70.16 ±0.37	77.67 ±0.57
FedNCA-ML	**70.55** ±0.15	71.28 ±1.34	**71.28** ±0.15	77.86 ±0.45

60 of 80 classes per client, it yields 0.60% improvement in class-wise F1 score.

On ChestX-ray14 (Tables 5), under non-IID Dirichlet settings of $\beta = 0.5$ and $\beta = 0.1$ with up to 7 of 14 disease classes per client, FedNCA-ML improves class-wise AUC by 0.69% and 1.21%, respectively. However, it achieves slightly lower overall AUC than some other methods. ChestX-ray14 is highly imbalanced with 57% of training and 38% of testing samples labeled as "No Finding". This severe imbalance leads many methods to overpredict the majority class, as reflected in a large gap between class-wise and overall AUC, indicating bias toward majority classes and degraded performance on minority (disease) classes. In medical diagnosis, false positives are generally more tolerable than false negatives, especially for rare diseases. The higher class-wise AUC and the smaller gap between overall and class-wise AUC achieved by FedNCA-ML suggest more balanced predictions and better recognition of minority disease classes.

5.4 Ablation Study

FedNCA-ML Component Analysis. We conduct an ablation study on the multi-label DermaMNIST dataset, which exhibits severe class imbalance (the majority class "melanoma" accounts for 61.97% of the training set and 60.25% of the test set) and pronounced inter-/intra-class variability due to diverse dermatological conditions. We simulate heterogeneous clients with $\beta = 0.1$ and $\gamma = 0.71$, which further amplifies class imbalance and label-relation heterogeneity. Results are reported in Table 6. As shown in the table, using only the predefined ETF classifier degrades performance, because image-level features entangle multiple semantic cues and directly clustering such representations can misguide optimization. In contrast, adding LADM for class-specific feature extraction and applying ETF anchoring to these class-specific features improves class-wise AUC by 0.43% and class-wise F1 by 7.26%. Notably, the lowest per-class F1 increases from 1.15% to 30.27% with an improvement of 29.12%. Finally, introducing the regularisation terms further strengthens discriminative learning and yields more balanced class-wise performance, bringing additional gains of 3.31% in class-wise AUC and 3.43% in class-wise F1.



(a) FC classifier **(b)** LADM + fixed ETF classifier **(c)** LADM + ETF + Dual regularization

Figure 4: Pairwise cosine similarity of class-wise average feature prototypes. Incorporating LADM, ETF-based alignment, and structure-preserving regularization lowers inter-class similarity, reflecting enhanced separability and discrimination.

LADM Analysis. We also investigate the attention mechanism in the LADM module. As shown in Table 7, fixed, well-designed class-wise queries consistently outperform learnable queries across most metrics. This result is intuitive under label-skewed FL, where clients exhibit distinct local distributions. When both the queries and the classifier are learnable, each client can overfit to its local data and label relationships, increasing cross-client inconsistency and degrading the aggregated global model. In

(a) Learnable FC classifier **(b)** LADM + fixed ETF classifier **(c)** LADM + fixed ETF + Dual regularization

Figure 3: t-SNE visualisation of test data feature embeddings on the multi-label DermaMNIST experiment with $\beta = 0.1$ and $\gamma = 0.71$. Each colour represents a class. Observing from subfigure (a), without class-wise feature extraction (LADM), the model relies on undesired information, such as the number of labels per sample, for clustering.

Table 6: Ablation study of the proposed method on multi-label DermaMNIST with $\beta = 0.1$ and $\gamma = 0.71$. We further report the F1 score for each class. The blue shading indicates the class prevalence, ranging from the majority class (60.25%) to the rarest class (5.90%). Dataset details are provided in Appendix A.

ETF Clf	LADM	$\mathcal{L}_{\text{Neg}}$	$\mathcal{L}_{\text{Pos}}$	macro-AUC	macro-F1	micro-AUC	micro-F1	F1 score for each class						
				83.95	40.69	86.68	63.03	44.51	10.28	**34.88**	52.23	1.15	82.28	60.63
✓				83.61	33.35	87.26	61.48	28.09	15.87	2.31	48.00	0.00	81.36	57.80
✓	✓			84.38	47.95	86.70	60.49	**48.16**	26.11	23.83	49.06	30.27	78.88	**79.32**
✓	✓	✓		84.73	45.06	89.71	62.89	48.11	6.72	20.81	50.16	32.14	82.54	74.90
✓	✓		✓	85.20	49.84	89.03	61.06	46.70	**39.83**	33.57	49.38	28.51	79.77	71.11
✓	✓	✓	✓	**87.69**	**51.38**	**90.36**	**63.26**	45.65	35.35	28.41	**53.45**	**36.01**	**83.00**	77.82

Table 7: Ablation study of the class-wise feature extraction block (LADM) on the multi-label DermaMNIST dataset with $\beta = 0.1$ and $\gamma = 0.71$.

query type	query init	macro-AUC	macro-F1	micro-AUC	micro-F1
learnable	random	82.94	47.94	86.33	57.37
learnable	ETF	**85.39**	46.92	84.94	53.37
fixed	ETF	84.38	**47.95**	**86.70**	**60.49**

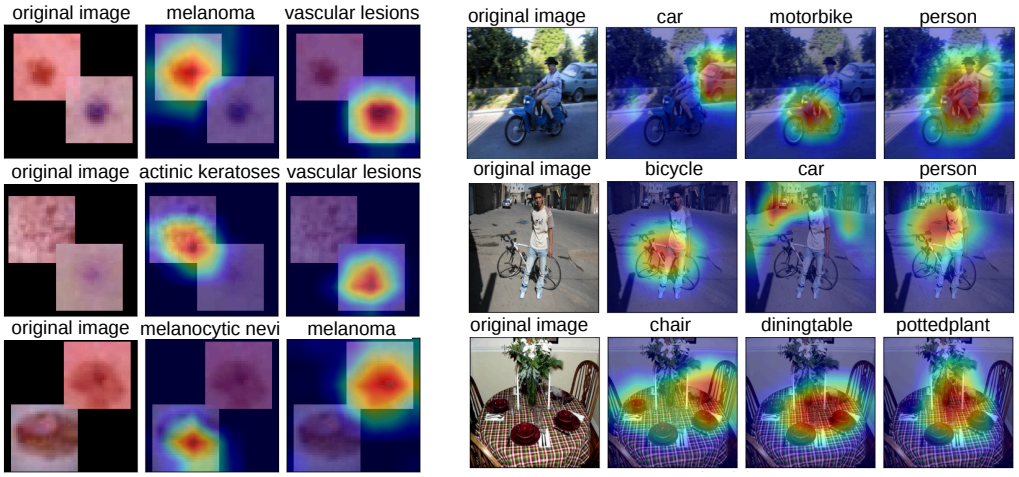


Figure 5: Examples of Grad-CAM visualizations on the multi-label DermaMNIST and VOC datasets. Each subfigure shows an input image alongside class-specific Grad-CAM maps for all corresponding ground-truth labels. From the visualizations, FedNCA-ML captures class-specific evidence for each target class from the shared global image-level features. Redder regions indicate stronger model responses for the corresponding class.

contrast, predefined and well-separated query embeddings provide a stable shared reference, encouraging more consistent optimization and improving robustness across non-IID clients.

Visualization. To further analyse the model's behaviour, we present t-SNE visualisations of the test data in the latent feature space (Figure 3) and the pairwise cosine similarity between class-wise average features (Figure 4), obtained under different model architectures and training strategies on the multi-label DermaMNIST dataset. As shown in Figure 3a, when a conventional learnable fully connected (FC) classifier is used, the resulting feature representations exhibit poor clustering. Notably, the model appears to rely on undesired information, grouping features by the number of labels per sample rather than solely by semantic content. By incorporating LADM (Figure 3b, 4b), which extracts single-class features, and further adding a predefined ETF classifier to regulate feature distribution across clients, the model learns to cluster features based on meaningful semantic attributes. This results in improved clustering quality, as evidenced by the substantially reduced pairwise cosine similarity between

class-wise average features, indicating enhanced inter-class separability and stronger discriminative capability. Finally, incorporating additional regularization terms during training yields even more compact and semantically coherent feature clusters, with further reductions in prototype similarity (Figure 3c, 4c). In addition, Figure 5 presents Grad-CAM visualizations, demonstrating that FedNCA-ML consistently focuses on semantically relevant regions for each target class. This indicates that the model can localize class-specific evidence in the image, supporting accurate class-wise prediction.

6 Conclusion

This paper tackles the challenging problem of multi-label label-skewed FL. This task is complicated by three intertwined factors: severe label imbalance, multi-label co-occurrence bias, and cross-client inconsistency in both label distributions and label relationships. To address these issues, we propose FedNCA-ML, a pre-learning FL framework that structures and optimizes the latent feature space with a Neural Collapse–inspired geometry, promoting cross-client representation consistency under non-IID data. FedNCA-ML integrates a class-wise feature extraction module with a predefined ETF as a shared geometric reference, inducing NC-style clustering in multi-label settings and guiding clients toward a coherent optimization objective. We further introduce two regularization losses to suppress noisy signals and encourage compact, well-separated class-wise clustering in the latent space. Experiments on five datasets under nine different FL settings demonstrate the effectiveness and robustness of the proposed method.

Acknowledgments

This work was supported by the UKRI grant EP/X040186/1 (Turing AI Fellowship). This work was also partly supported by the InnoHK-funded Hong Kong Centre for Cerebrocardiovascular Health Engineering (COCHE) Project 2.1 (Cardiovascular risks in early life and fetal echocardiography).

References

[1] 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.

[2] 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.

[3] Yibo Yang, Haobo Yuan, Xiangtai Li, Zhouchen Lin, Philip Torr, and Dacheng Tao. Neural collapse inspired feature-classifier alignment for few-shot class incremental learning. *arXiv preprint arXiv:2302.03004*, 2023.

[4] Jintong Gao, He Zhao, Dan dan Guo, and Hongyuan Zha. Distribution alignment optimization through neural collapse for long-tailed classification. In *Forty-first International Conference on Machine Learning*, 2024.

[5] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017.

[6] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. *Proceedings of Machine learning and systems*, 2:429–450, 2020.

[7] Neta Shoham, Tomer Avidor, Aviv Keren, Nadav Israel, Daniel Benditkis, Liron Mor-Yosef, and Itai Zeitak. Overcoming forgetting in federated learning on non-iid data. *arXiv preprint arXiv:1910.07796*, 2019.

[8] Lin Zhang, Yong Luo, Yan Bai, Bo Du, and Ling-Yu Duan. Federated learning for non-iid data via unified feature learning and optimization objective alignment. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4420–4428, 2021.

[9] Kuangpu Guo, Yuhe Ding, Jian Liang, Zilei Wang, Ran He, and Tieniu Tan. Exploring vacant classes in label-skewed federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 16960–16968, 2025.

[10] Xin-Chun Li and De-Chuan Zhan. Fedrs: Federated learning with restricted softmax for label distribution non-iid data. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, pages 995–1005, 2021.

[11] Qinbin Li, Bingsheng He, and Dawn Song. Model-contrastive federated learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10713–10722, 2021.

[12] 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.

[13] Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for federated learning. In *International conference on machine learning*, pages 5132–5143. PMLR, 2020.

[14] Zebang Shen, Juan Cervino, Hamed Hassani, and Alejandro Ribeiro. An agnostic approach to federated learning with class imbalance. In *International Conference on Learning Representations*, 2021.

[15] Xin Dong, Sai Qian Zhang, Ang Li, and HT Kung. Spherefed: Hyperspherical federated learning. In *European Conference on Computer Vision*, pages 165–184. Springer, 2022.

[16] Zhaobin Sun, Nannan Wu, Junjie Shi, Li Yu, Kwang-Ting Cheng, and Zengqiang Yan. Fedmlp: Federated multi-label medical image classification under task heterogeneity. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 394–404. Springer, 2024.

[17] I-Jieh Liu, Ci-Siang Lin, Fu-En Yang, and Yu-Chiang Frank Wang. Language-guided transformer for federated multi-label classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 13882–13890, 2024.

[18] Jack Lanchantin, Tianlu Wang, Vicente Ordonez, and Yanjun Qi. General multi-label image classification with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16478–16488, 2021.

[19] 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.

[20] Kun Wei, Zhe Xu, and Cheng Deng. Compress to one point: Neural collapse for pre-trained model-based class-incremental learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 21465–21473, 2025.

[21] Xuantong Liu, Jianfeng Zhang, Tianyang Hu, He Cao, Yuan Yao, and Lujia Pan. Inducing neural collapse in deep long-tailed learning. In *International conference on artificial intelligence and statistics*, pages 11534–11544. PMLR, 2023.

[22] Pengyu Li, Yutong Wang, Xiao Li, and Qing Qu. Neural collapse in multi-label learning with pick-all-label loss. 2023.

[23] Zijian Tao, Shao-Yuan Li, Wenhai Wan, Jinpeng Zheng, Jia-Yao Chen, Yuchen Li, Sheng-Jun Huang, and Songcan Chen. Mlc-nc: Long-tailed multi-label image classification through the lens of neural collapse. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 20850–20858, 2025.

[24] Shilong Liu, Lei Zhang, Xiao Yang, Hang Su, and Jun Zhu. Query2label: A simple transformer way to multi-label classification. *arXiv preprint arXiv:2107.10834*, 2021.

[25] Tal Ridnik, Gilad Sharir, Avi Ben-Cohen, Emanuel Ben-Baruch, and Asaf Noy. Ml-decoder: Scalable and versatile classification head. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 32–41, 2023.

[26] Pengyu Xu, Lin Xiao, Bing Liu, Sijin Lu, Liping Jing, and Jian Yu. Label-specific feature augmentation for long-tailed multi-label text classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 10602–10610, 2023.

[27] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020.

[28] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[29] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88:303–338, 2010.

[30] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[31] Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Scientific Data*, 10(1):41, 2023.

[32] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M. Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3462–3471, 2017.

[33] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Appendix

A Dataset

In this section, we introduce the datasets used in our experiments. To evaluate both the effectiveness and real-world applicability of the proposed method, we conduct experiments on datasets from both general computer vision (CV) as well as medical imaging domains. Specifically, we use CIFAR-10 [28], PASCAL VOC [29], and MS COCO [30] as general CV benchmarks, and DermaMNIST [31] and ChestX-ray14 [32] as medical imaging datasets. To simulate non-IID federated learning (FL) settings, we partition the data using a Dirichlet distribution, with the concentration parameter β controlling the degree of heterogeneity. To further model missing-class scenarios, we constrain the number of classes available to each client using the class presence ratio γ, which specifies the proportion of total classes present locally. Dataset-specific settings are detailed below.

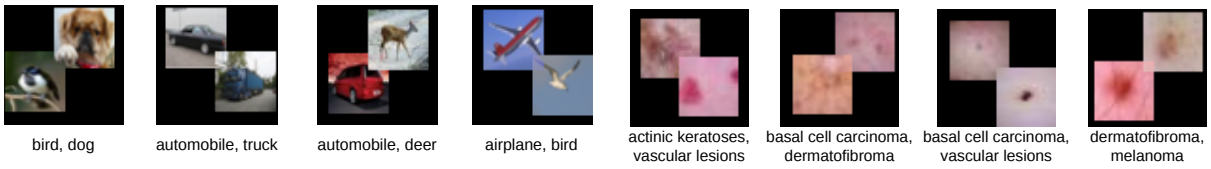
CIFAR-10 [28] is a widely used benchmark dataset in CV. It comprises 10 classes: *airplane*, *automobile*, *bird*, *cat*, *deer*, *dog*, *frog*, *horse*, *ship*, and *truck*. To adapt CIFAR-10 to the multi-label setting, following [22], we construct composite samples by combining multiple original images into a single image, with the associated class labels forming the multi-label ground truth. Specifically, we retain a portion of the original single-label images and augment the dataset by generating an equal number of synthetic samples for every possible pairwise class combination. Each composite sample is created by randomly selecting two images from different classes and merging them into a single image. Examples of the resulting multi-label composite samples are shown in Figure 6a. We evaluate the proposed method on the resulting multi-label CIFAR-10 dataset under two FL settings: $\beta = 0.5,\ \gamma = 0.5$ and $\beta = 0.1,\ \gamma = 0.5$. The corresponding class-wise data distributions across clients are illustrated in Figure 7.

DermaMNIST [31] is a skin lesion classification dataset. It consists of 7 diagnostic categories: *actinic keratoses*, *basal cell carcinoma*, *benign keratosis-like lesions*, *dermatofibroma*, *melanocytic nevi*, *melanoma*, and *vascular lesions*. To adapt DermaMNIST to the multi-label setting, we adopt a strategy similar to that used for multi-label CIFAR-10. Given the long-tailed nature of the original dataset, we preserve its intrinsic class distribution by retaining all original single-label samples. We then augment the dataset by generating an equal number of synthetic samples for each possible pairwise label combination. Examples of the resulting composite multi-label samples are shown in Figure 6b. The final dataset consists of the complete set of original samples together with the newly generated multi-label samples. We evaluate the proposed method on this multi-label DermaMNIST dataset under two FL settings: $\beta = 0.5, \gamma = 0.71$ and $\beta = 0.1, \gamma = 0.71$. The corresponding class-wise data distributions across clients are illustrated in Figure 8.

PASCAL VOC [29] is a widely used benchmark dataset in CV. It contains approximately 11,500 images, each annotated with one or more object categories from a predefined set of 20 classes. These categories include: *aeroplane*, *bicycle*, *bird*, *boat*, *bottle*, *bus*, *car*, *cat*, *chair*, *cow*, *dining table*, *dog*, *horse*, *motorbike*, *person*, *potted plant*, *sheep*, *sofa*, *train*, and *TV monitor*. This dataset presents a challenging multi-label classification task due to substantial class imbalance, high intra-class variability, and frequent inter-class co-occurrence. For instance, the *person* class commonly appears alongside many other object categories. Under label-skewed FL settings, these challenges are further amplified by non-IID client distributions and inconsistent label co-occurrence patterns across clients. We evaluate the proposed method on PASCAL VOC under two FL settings: $\beta = 0.05, \gamma = 0.5$ and $\beta = 0.01, \gamma = 0.5$. The resulting class-wise data distributions across clients are shown in Figure 9.

MS COCO [30] is another widely used benchmark dataset in CV. It contains a large-scale collection of natural images, each annotated with one or more object categories selected from a predefined set of 80 classes. Compared with PASCAL VOC, MS COCO poses an even more challenging multi-label classification task due to its larger label space, more complex visual scenes, and denser object co-occurrence patterns. In particular, many images contain multiple objects with diverse scales, occlusions, and cluttered backgrounds, leading to substantial class imbalance, high intra-class variability, and frequent inter-class co-occurrence. These categories include *person*, *bicycle*, *car*, *motorcycle*, *airplane*, *bus*, *train*, *truck*, *boat*, *traffic light*, *fire hydrant*, *stop sign*, *parking meter*, *bench*, *bird*, *cat*, *dog*, *horse*, *sheep*, *cow*, *elephant*, *bear*, *zebra*, *giraffe*, *backpack*, *umbrella*, *handbag*, *tie*, *suitcase*, *frisbee*, *skis*, *snowboard*, *sports ball*, *kite*, *baseball bat*, *baseball glove*, *skateboard*, *surfboard*, *tennis racket*, *bottle*, *wine glass*, *cup*, *fork*, *knife*, *spoon*, *bowl*, *banana*, *apple*, *sandwich*, *orange*, *broccoli*, *carrot*, *hot dog*, *pizza*, *donut*, *cake*, *chair*, *couch*, *potted plant*, *bed*, *dining table*, *toilet*, *tv*, *laptop*, *mouse*, *remote*, *keyboard*, *cell phone*, *microwave*, *oven*, *toaster*, *sink*, *refrigerator*, *book*, *clock*, *vase*, *scissors*, *teddy bear*, *hair drier*, and *toothbrush*, covering a broad range of everyday object categories. In the label-skewed FL setting, these difficulties are further amplified by non-IID client distributions and inconsistent co-occurrence patterns across clients. We conduct experiments under the FL setting $\beta = 0.05, \gamma = 0.75$, and visualize the resulting class-wise client distributions in Figure 11.

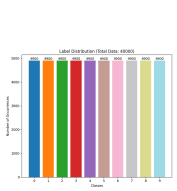
ChestX-ray14 [32] is a large-scale medical imaging dataset widely used for automated thoracic disease classification. It contains 112,120 frontal-view chest X-ray images collected from 30,805 unique patients. Each image is



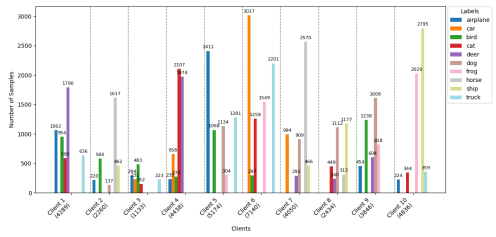
(a) Examples for the multi-label CIFAR-10 dataset.

(b) Examples for the multi-label DermaMNIST dataset.

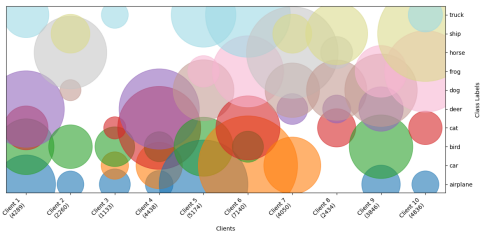
Figure 6: Examples for the multi-label CIFAR-10 and DermaMNIST datasets. Each subfigure displays a composite image along with its corresponding set of labels.



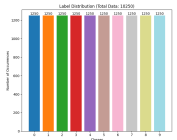
(a) Overall training data distribution.



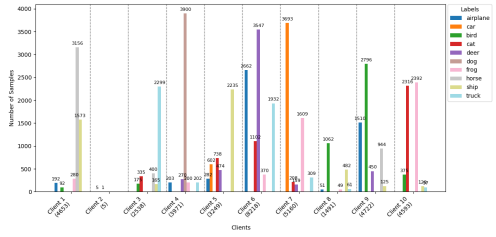
(b) Per-client training data distribution with $\beta = 0.5$ and $\gamma = 0.5$.



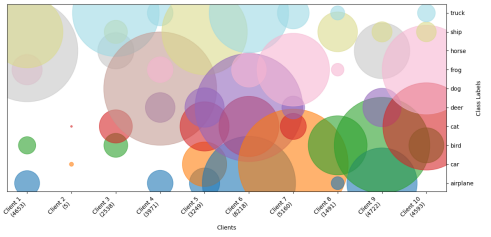
(c) Per-client training data bubble chart under $\beta = 0.5$ and $\gamma = 0.5$.



(d) Testing data distribution.



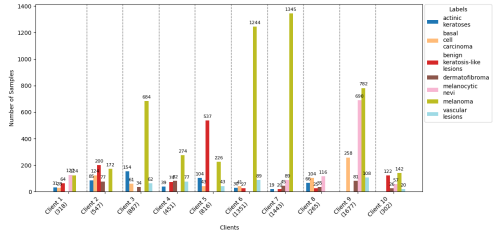
(e) Per-client training data distribution under $\beta = 0.1$ and $\gamma = 0.5$.



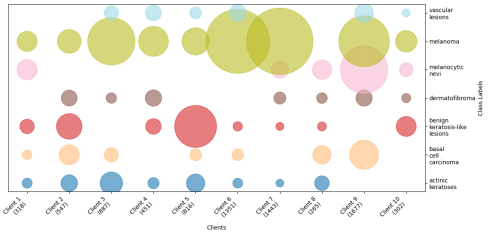
(f) Per-client training data bubble chart under $\beta = 0.1$ and $\gamma = 0.5$.

Figure 7: Distribution of data across local clients in the CIFAR-10 [28] experiments. The class presence ratio (γ) is set to 0.5 ($\leq$ 5 of 10 classes per client). Non-IID client distributions are simulated using the Dirichlet factor (β).

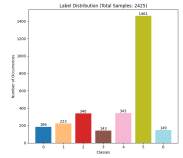
(a) Overall training data distribution.



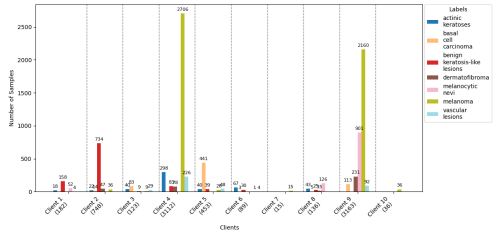
(b) Per-client training data distribution under $\beta = 0.5$ and $\gamma = 0.71$.



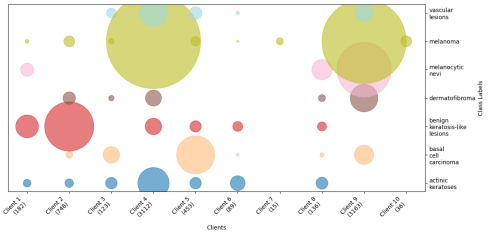
(c) Per-client training data bubble chart under $\beta = 0.5$ and $\gamma = 0.71$.



(d) Testing data distribution.

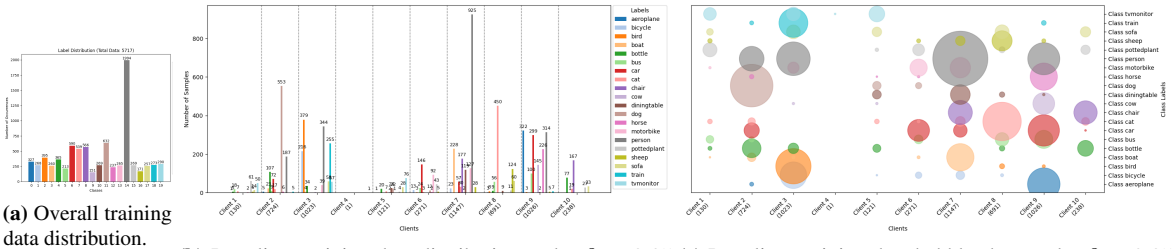


(e) Per-client training data distribution under $\beta = 0.1$ and $\gamma = 0.71$.



(f) Per-client training data bubble chart under $\beta = 0.1$ and $\gamma = 0.71$.

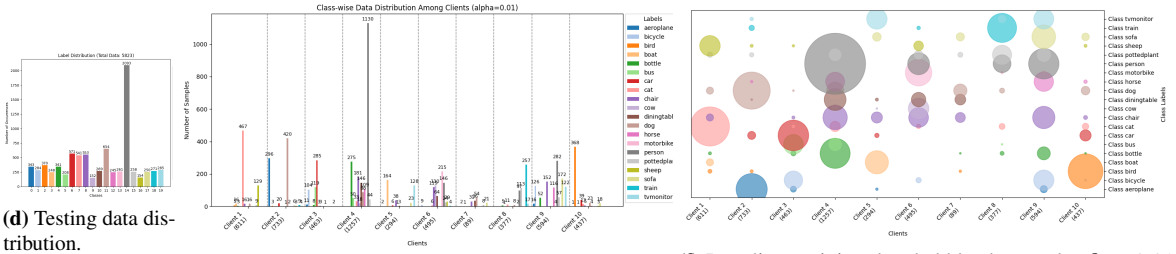
Figure 8: Distribution of data across local clients in the DermaMNIST [31] experiments. The class presence ratio (γ) is set to 0.71 ($\leq$ 5 of 7 classes per client). Non-IID client distributions are simulated using the Dirichlet factor (β).



(a) Overall training data distribution.

(b) Per-client training data distribution under $\beta = 0.05$ and $\gamma = 0.5$.

(c) Per-client training data bubble chart under $\beta = 0.05$ and $\gamma = 0.5$.

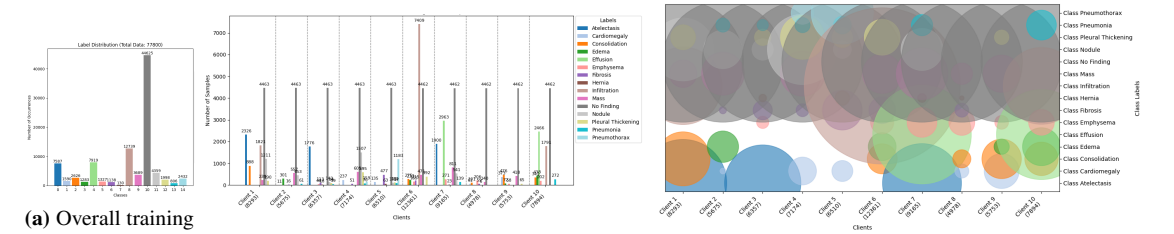


(d) Testing data distribution.

(e) Per-client training data distribution under $\beta = 0.01$ and $\gamma = 0.5$.

(f) Per-client training data bubble chart under $\beta = 0.01$ and $\gamma = 0.5$.

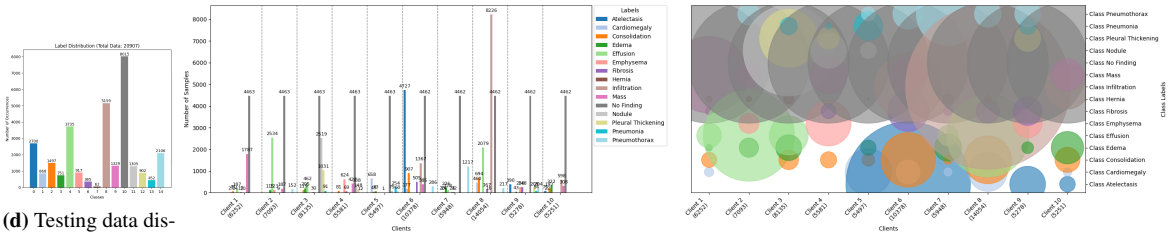
Figure 9: Distribution of data across local clients in the PASCAL VOC [29] experiments. The class presence ratio (γ) is set to 0.5 ($\leq$ 10 of 20 classes per client). Non-IID client distributions are simulated using the Dirichlet factor (β).



(a) Overall training data distribution.

(b) Per-client training data distribution under $\beta = 0.5$ and $\gamma = 0.5$.

(c) Per-client training data bubble chart under $\beta = 0.5$ and $\gamma = 0.5$.

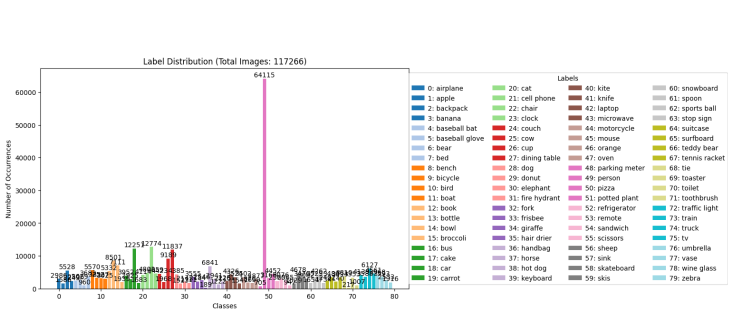


(d) Testing data distribution.

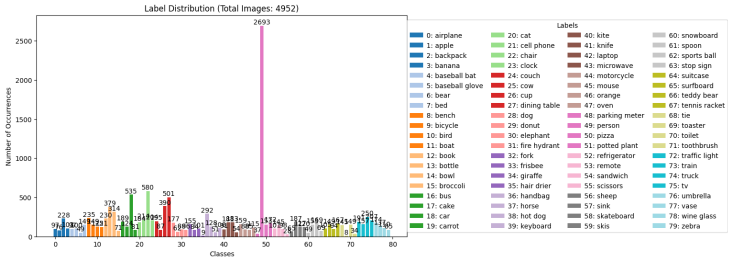
(e) Per-client training data distribution under $\beta = 0.1$ and $\gamma = 0.5$.

(f) Per-client training data bubble chart under $\beta = 0.1$ and $\gamma = 0.5$.

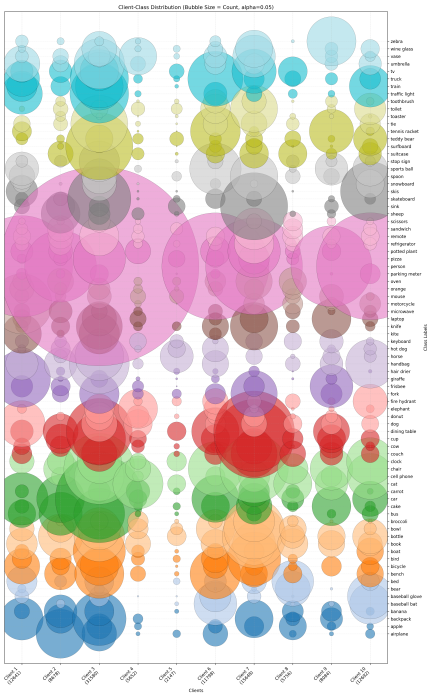
Figure 10: Distribution of data across local clients in the ChestX-ray14 [32] experiments. The ChestX-ray14 dataset contains 14 thoracic disease categories and an additional "No Finding" label. Since a large portion of the dataset (57% of the training data) is "No Finding" samples (i.e., negative cases with all-zero labels), we distribute these samples evenly across all clients to reflect a realistic clinical scenario in which healthy cases are prevalent. The class presence ratio (γ) is set to 0.5 ($\leq$ 7 of 14 disease classes per client). Non-IID client distributions are simulated using the Dirichlet factor (β).



(a) Overall training data distribution.



(b) Testing data distribution.



(c) Per-client training data distribution under $\beta = 0.05$ and $\gamma = 0.75$.

Figure 11: Distribution of data across local clients in the COCO [30] experiments. The class presence ratio (γ) is set to 0.75 ($\leq$ 60 of 80 classes per client). Non-IID client distributions are simulated using the Dirichlet factor (β).

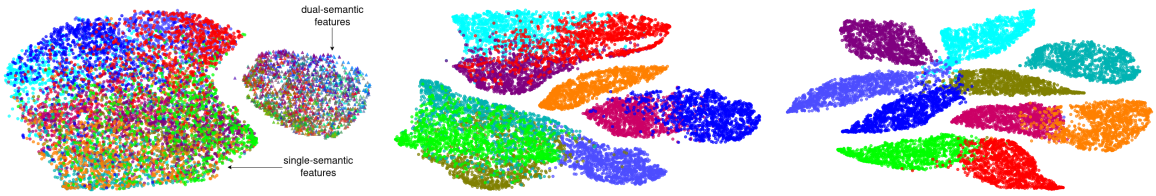
annotated with one or more disease labels, making the dataset naturally suited to multi-label classification. These labels are extracted from the corresponding radiology reports using natural language processing techniques. The dataset includes 14 disease categories: *Atelectasis*, *Cardiomegaly*, *Consolidation*, *Edema*, *Effusion*, *Emphysema*, *Fibrosis*, *Hernia*, *Infiltration*, *Mass*, *Nodule*, *Pleural Thickening*, *Pneumonia*, and *Pneumothorax*. In addition, a *No Finding* label is used to indicate negative samples in which none of the 14 diseases are present. Due to its large scale, real-world variability, and inherent label noise, ChestX-ray14 provides a challenging benchmark for developing and evaluating deep learning models in medical image analysis, particularly in multi-label and imbalanced scenarios. We conduct experiments on this dataset under two label-skewed FL configurations: $\beta = 0.5$, $\gamma = 0.5$ and $\beta = 0.1$, $\gamma = 0.5$. Since a significant portion of the dataset (57% of the training data) is *No Finding* samples, we distribute these samples evenly across all clients in both settings. This setup mimics realistic clinical scenarios where healthy cases are common, while disease cases are relatively rare and unequally distributed across institutions. Figure 10 illustrates the resulting class-wise data distributions across clients.

B Experiment (Additional)

B.1 Ablation Study (Additional)

To complement the results reported in the main manuscript, we present an additional ablation study to evaluate the contribution of each component in our proposed method. The experiments are conducted on the multi-label CIFAR-10 dataset [28], where the client data distribution is configured using a Dirichlet concentration parameter of $\beta = 0.5$ and a class-presence ratio of $\gamma = 0.5$ (i.e., each client has access to at most 5 out of 10 classes). The experimental results are summarized in Tables 8 and 9, and Figures 12 and 13. As shown in Table 8, integrating LADM for class-specific feature extraction, along with the use of a predefined ETF classifier to encourage class-wise clustering alignment across clients in the latent feature space, leads to performance improvements of 3.38% in class-wise AUC and 3.99% in class-wise F1 score. Additional gains are achieved by introducing regularization terms that enhance the model's discriminative capacity, resulting in a further class-wise increase of 1.88% in class-wise AUC and 4.58% in class-wise F1 score. Regarding LADM specifically, the results in Table 9 demonstrate that using fixed, well-designed class-wise queries is generally more effective than learnable queries across most evaluation metrics.

To further analyse model behaviour, we visualize the latent feature distributions of the test set using t-SNE under different architectural and training configurations. As shown in Figure 12, incorporating LADM for feature disentanglement, along with a predefined ETF classifier to regulate feature distribution across clients,



(a) Learnable FC classifier **(b)** LADM + fixed ETF classifier **(c)** LADM + fixed ETF + Dual regularization

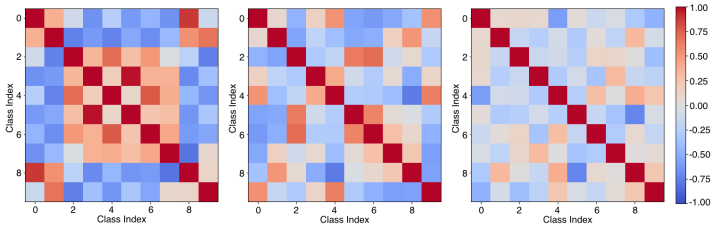
Figure 12: t-SNE visualisation of test data feature embeddings on the multi-label CIFAR-10 experiment with $\beta = 0.5$, $\gamma = 0.5$. Each colour represents a class. Observing from subfigure (a), without feature disentanglement (LADM), the model appears to rely on undesired information, such as the number of labels per sample, for clustering.

Table 8: Ablation study of the proposed method on the multi-label CIFAR-10 dataset [28]. β is set to 0.5 and γ is set to 0.5.

ETF Clf	LADM	$\mathcal{L}_{\text{Neg}}$	$\mathcal{L}_{\text{Pos}}$	macro-AUC	macro-F1	micro-AUC	micro-F1
				82.46	40.65	81.83	41.24
✓				83.44	15.29	81.75	16.77
✓	✓			85.84	44.64	85.10	45.45
✓	✓	✓		87.72	48.22	86.16	48.93
✓	✓		✓	86.78	44.46	86.08	45.28
✓	✓	✓	✓	**87.72**	**49.22**	**87.13**	**49.60**

Table 9: Ablation study of the class-wise feature extraction block - LADM on the multi-label CIFAR-10 dataset.

query type	query init	macro-AUC	macro-F1	micro-AUC	micro-F1
learnable	random	84.15	38.10	83.70	39.30
learnable	ETF	**86.26**	43.61	**85.41**	43.87
fixed	ETF	85.84	**44.64**	85.10	**45.45**



(a) FC classifier **(b)** LADM + fixed ETF classifier **(c)** LADM + ETF + Dual regularization

Figure 13: Pair-wise cosine similarity between test data class-wise average feature prototypes on the multi-label CIFAR-10 dataset.

enables the model to focus on semantic content rather than irrelevant factors such as the number of labels present in each sample. Furthermore, the addition of regularization terms during training leads to more compact and semantically coherent clusters, while also reducing the similarity among class-wise average prototypes. As illustrated in Figure 13, the pairwise cosine similarity between class-wise average features decreases with the inclusion of LADM and the ETF classifier, and is further reduced by the added regularization terms, indicating enhanced inter-class separability and stronger discriminative capability.