Inconsistency-Based Federated Active Learning

Chen-Chen Zong , Tong Jin , Sheng-Jun Huang*

College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics
MIIT Key Laboratory of Pattern Analysis and Machine Intelligence, Nanjing, 211106, China
{chencz, tongjin, huangsj}@nuaa.edu.cn

Abstract

Federated learning (FL) enables distributed collaborative learning across local clients while preserving data privacy. However, its practical application in weakly supervised learning (WSL), where only a small subset of data is labeled, remains underexplored. Active learning (AL) is a promising solution for label-limited scenarios, but its adaptation to federated settings presents unique challenges, such as data heterogeneity and noise. In this paper, we propose **I**nconsistency-based **F**ederated **A**ctive **L**earning (IFAL), a novel approach to address these challenges. First, we introduce a data-driven probability formulation that aligns the biases between local and global models in heterogeneous FL settings. Next, to mitigate noise, we propose an inter-model inconsistency criterion that filters out noisy examples and focuses on those with beneficial prediction discrepancies. Additionally, we introduce an intra-model inconsistency criterion to query examples that help refine the model's decision boundaries. By combining these strategies with clustering, IFAL effectively selects a diverse and informative query set. Extensive experiments on benchmark datasets demonstrate that IFAL outperforms state-of-the-art methods.

1 Introduction

Federated learning (FL) is a distributed framework that enables collaborative learning across multiple local clients, where each client contributes to a shared global model on a central server through aggregation, all while ensuring the privacy of local data [Konečnỳ *et al.*, 2016; McMahan *et al.*, 2017; Huang *et al.*, 2021]. Despite extensive research on FL, its practical implementation in weakly supervised learning (WSL) scenarios—where only a subset of examples are labeled and the remainder are unlabeled—remains limited [Fan *et al.*, 2022; Kim *et al.*, 2023]. However, WSL is more aligned with real-world tasks, as acquiring large-scale, fully labeled datasets for each local client is often prohibitively expensive and time-consuming due to the labor-intensive nature of manual labeling [Zong *et al.*, 2025]. Therefore, it is essential to design FL algorithms capable of effectively learning from insufficient labeled data at each local client.

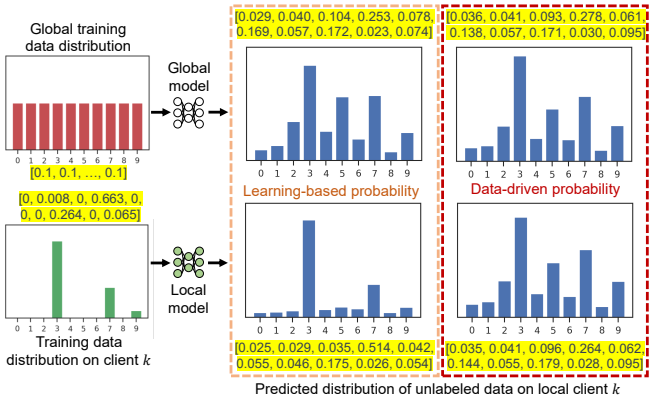

Figure 1: Data heterogeneity scenario in federated learning (CIFAR-10, $\alpha = 0.1$): learning-based and data-driven predicted probability distributions of local and global models on the local unlabeled set.

Active learning (AL) provides an effective solution for label-limited scenarios by iteratively selecting the most informative examples from the unlabeled data pool and querying their labels from the oracle [Settles, 2009; Zong and Huang, 2025]. Existing AL methods typically operate in centralized environments, with two primary example selection criteria: 1) Uncertainty [Li and Sethi, 2006; Balcan *et al.*, 2007; Holub *et al.*, 2008], which selects the most challenging examples based on the model's uncertainty in its predictions; and 2) Diversity [Sener and Savarese, 2017], which queries representative examples to ensure the selected subset approximates the distribution of the original unlabeled data pool.

However, in distributed FL settings, where each client maintains both local and global models and data is isolated from other clients, federated active learning (FAL) introduces unique challenges. For instance, which model is best suited for uncertainty evaluation? Do local representative instances accurately reflect global diversity in data selection? Or are there other sampling criteria that better fit the FL framework?

Recent efforts have been made to address these challenges. For example, LoGo [Kim *et al.*, 2023] uses the local model to obtain the gradient embedding for each example, applies

*Corresponding Author

k-means clustering, and selects the most uncertain example for the global model within each cluster to form the query set. While LoGo considers both uncertainty and diversity and ensures the participation of both models, the authors in [Kim *et al.*, 2023] do not explain why this combination is optimal, and our experiments reveal that its performance is suboptimal. KAFAL [Cao *et al.*, 2023] utilizes prior knowledge of example counts for each class on the client to define a knowledge-specialized probability form, and calculates the Kullback-Leibler (KL) divergence between the global and local models to query a batch of examples with the most inconsistent predictions. Although designing AL query strategies based on inconsistency between local and global models aligns well with FAL's characteristics, relying solely on example counts as a prior tends to overlook temporarily underrepresented or absent classes. Furthermore, considering only the inconsistency between the two models neglects the impact of noisy examples, which leads to suboptimal performance.

In this paper, we begin by considering another key characteristic of FL: data heterogeneity, where the local data distribution may differ significantly from the global distribution. This can lead to substantial prediction discrepancies between local and global models on rare local classes (as shown in Figure 1). Such discrepancies arise from the bias in learning-based probabilities caused by varying training data distributions, which may be irrelevant or even detrimental, potentially undermining the effectiveness of existing AL query strategies. To address this, we propose a data-driven (structural) probability, where the structural probability of each example is determined by its reverse K-nearest neighbors in the representation space, based on a set of example features and pseudo-labels. As illustrated in Figure 1, the structural probability effectively aligns the biases between the models, with prediction inconsistencies reflecting topological structural differences in their representation spaces.

Next, we address the noise problem in the query process, as examples with high prediction inconsistency may not only be informative, hard-to-learn examples but also noisy ones. To distinguish between them, we train an additional model on the client side by distilling knowledge from the global model, and compute the prediction inconsistencies between this model and both the local and global models. On one hand, due to the memory effect of neural networks [Han *et al.*, 2018; Zong *et al.*, 2024], useful knowledge can be transferred from the teacher to the student, while noise information is less likely to be transferred. On the other hand, noise often leads to more random outputs from the model, so for a noisy input, the prediction discrepancies between any pair of the three models are likely to be large. Based on these observations, we propose an inter-model inconsistency criterion to query reliable examples that help narrow the prediction gap between the local and global models.

Additionally, we propose an intra-model inconsistency criterion, defined as the prediction discrepancy between the learning-based and data-driven probabilities output by the local model for each example. This criterion aims to query the most beneficial examples for improving performance. Finally, we implement Ostu thresholding [Otsu and others, 1975] to select a batch of examples with high scores in both inconsistency metrics and then apply k-means clustering to query a representative subset. We refer to the entire framework as **I**nconsistency-based **F**ederated **A**ctive **L**earning (IFAL).

The main contributions of this paper are as follows:

- To address the challenge posed by data heterogeneity, we propose a data-driven probability formulation based on the topological relationships of examples in the representation space. This formulation effectively mitigates the inductive bias introduced by prior data distributions in different models, enabling a more accurate assessment of the data distributional differences across various model representation spaces.
- We highlight the importance of addressing the noise issue in the AL query process. By introducing an intermediate model and calculating the prediction inconsistencies between this model and both the local and global models, we propose an inter-model inconsistency-based query strategy. This strategy effectively filters out noisy examples while prioritizing those most beneficial for reducing inconsistency between models.
- We propose an intra-model inconsistency-based query strategy by measuring the divergence between the output probability of the classifier head and the proposed structural probability, querying examples that are most beneficial for refining the model's decision boundaries.
- We introduce a hybrid sampling strategy, called IFAL, which first applies Ostu thresholding to decide a subset of examples with high prediction inconsistency and then uses clustering to query a diverse set of examples.
- We conduct extensive experiments on three benchmark datasets, demonstrating that IFAL outperforms current state-of-the-art methods.

2 Related Work

Federated learning (FL) is a distributed machine learning framework that ensures data privacy by training models across multiple clients without sharing their private data. The FedAvg algorithm [McMahan *et al.*, 2017] is a cornerstone of FL, aggregating local model parameters through averaging, which strikes a balance between communication efficiency and computational flexibility. A key challenge in FL is the non-independent and identically distributed (non-IID) nature of data across clients, also known as data heterogeneity, which can severely hinder the performance and convergence of FL. To address this, various approaches have been proposed, including consistency regularization [Li *et al.*, 2020; An *et al.*, 2023], optimized aggregation [Pillutla *et al.*, 2022; Li *et al.*, 2023], and personalized FL methods [Zhang *et al.*, 2023b; Wang *et al.*, 2024]. FL has also been extended to address real-world challenges, such as continual learning [Zhang *et al.*, 2023c], multi-label learning [Liu *et al.*, 2024], semi-supervised learning [Bai *et al.*, 2024], and active learning [Cao *et al.*, 2023; Zhang *et al.*, 2023a], expanding its applicability across diverse domains.

Active learning (AL) is a primary approach for reducing labeling costs by querying the most informative exam-

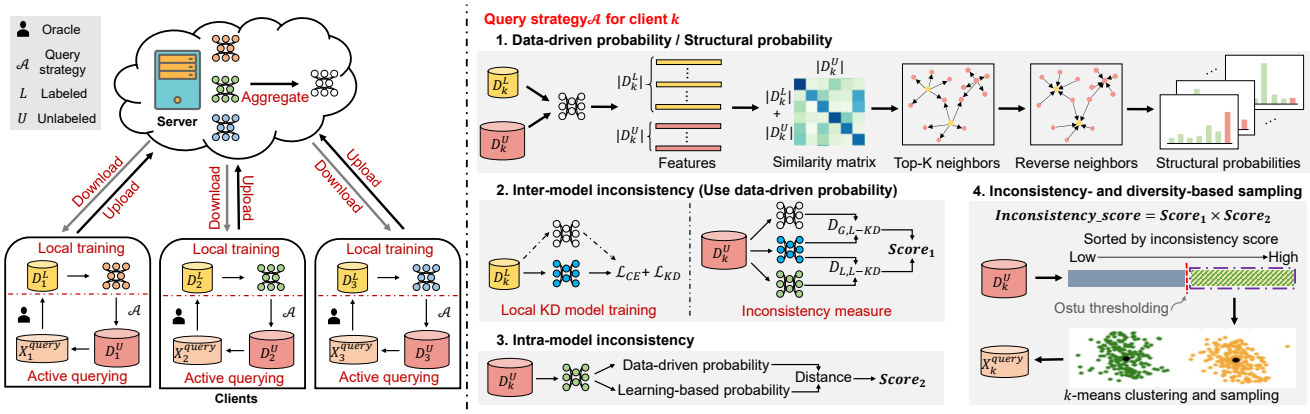

Figure 2: The overview of the proposed IFAL framework. The left part illustrates the basic framework of federated active learning, while the right part demonstrates an example of executing the proposed query strategy on client k.

ples for model training. AL query strategies can be typically categorized into three types: uncertainty-based, diversity-based, and hybrid methods. Uncertainty-based methods prioritize instances for which labeling is least certain. Common uncertainty metrics include entropy-based sampling [Holub *et al.*, 2008], margin-based sampling [Balcan *et al.*, 2007], and least confidence sampling [Li and Sethi, 2006], etc. Diversity-based methods prioritize instances that are most representative or exhibit the greatest feature diversity. Such as clustering-based selection, where instances are selected from distinct clusters [Nguyen and Smeulders, 2004], and core-set selection [Sener and Savarese, 2017], which minimizes the distance between queried instances and the entire dataset. Hybrid methods combine diversity and uncertainty to ensure that annotated data is both representative and uncertainty. These techniques often involve two-stage sampling [Wang *et al.*, 2023; Yuan *et al.*, 2023], leveraging the strengths of both diversity- and uncertainty-based techniques [Ash *et al.*, 2019; Prabhu *et al.*, 2021; Caramalau *et al.*, 2021].

Federated active learning (FAL) aims to enhance model performance with minimal labeled data while preserving data privacy by incorporating active querying into the FL framework and selecting the most informative instances on each client for labeling. Early approaches typically apply existing AL strategies directly using either the local model or the global model [Wu *et al.*, 2022; Ahn *et al.*, 2024]. However, due to data heterogeneity, querying examples based on a single model may not be optimal for the global system, thus limiting performance gains. To address this, LoGo [Kim *et al.*, 2023] simultaneously considers local and global inter-class diversity by first clustering unlabeled instances using the local model, and then selecting the most uncertain instances within each cluster based on the global model. KAFAL [Cao *et al.*, 2023] queries the most uncertain examples by measuring the prediction divergence between the global and local models on the same input for the client's specialized classes. However, we find that LoGo's strategy is ineffective, particularly in complex datasets. While KAFAL is a more reasonable approach, it suffers from two main issues: it overlooks temporarily unseen or rare classes on the client side, and it does not account for the interference of noisy examples, leading to suboptimal performance. Therefore, in this paper, we propose a novel inconsistency-based FAL framework that effectively addresses these issues and better aligns with the characteristics of FL.

3 Methodology

3.1 Preliminaries

Notations. Consider the problem of ordinary C-class classification in federated learning (FL). Let f_G be the global model on the central server, and $\{f_{\theta_1}, \ldots, f_{\theta_k}, \ldots, f_{\theta_N}\}$ the set of local models, corresponding one-to-one with the N clients. Assume a total of T communication rounds. In each round t, each client k first downloads the global model f_G^{t-1} from the server to initialize its local model $f_{\theta_k}^t$, and then optimizes $f_{\theta_k}^t$ using its isolated training dataset $\mathcal{D}_k = \{(x_i, y_i)\}_{i=1}^{\mathcal{N}_k}$, where x_i denotes an arbitrary example, y_i is the corresponding ground-truth label belonging to $\{1, \ldots, C\}$, and $\mathcal{N}_k$ is the total number of examples. The updated local models are then uploaded to the server and aggregated to update the global model from f_G^{t-1} to f_G^t.

In federated active learning (FAL), each client k maintains a limited labeled dataset $\mathcal{D}_k^L = \{(x_i, y_i)\}_{i=1}^{\mathcal{N}_k^L}$ for model training and a sufficiently large unlabeled data pool $\mathcal{D}_k^U = \{x_i\}_{i=1}^{\mathcal{N}_k^U}$ for potential labeling, where $\mathcal{N}_i^L$ and $\mathcal{N}_k^U$ denote the total number of instances in $\mathcal{D}_k^L$ and $\mathcal{D}_k^U$, respectively. Let the total number of FAL cycles be M. In each FAL cycle m, a complete FL training is first conducted based on the current labeled datasets $\{\mathcal{D}_1^L|_m, \ldots, \mathcal{D}_N^L|_m\}$, after which each client k queries the b most informative examples (denoted as X_k^{query}) from $\mathcal{D}_k^U|_m$ according to a specified query strategy $\mathcal{A}$, and then sends them to the oracle for labeling, resulting in $\mathcal{D}_k^L|_{m+1} = \mathcal{D}_k^L|_m \cup X_k^{query}$ and $\mathcal{D}_k^U|_{m+1} = \mathcal{D}_k^U|_m \setminus X_k^{query}$.

Overview. The FAL framework is outlined in the left part of Figure 2. In this paper, we aim to design a more effective AL query strategy tailored to the unique characteristics of FAL. First, due to inherent data heterogeneity in FL, significant discrepancies often arise between client and server models, particularly on rare local classes, which are often meaningless and can undermine the effectiveness of existing AL

query strategies. We attribute this to the bias in the learning-based probability induced by varying prior data distributions and propose a data-driven probability formulation to mitigate its effect. Second, we introduce two inconsistency-based sampling criteria: 1) *Inter-model inconsistency*, which uses an intermediate model to compute prediction discrepancies between different model pairs, querying examples that can reduce the prediction gap between local and global models; 2) *Intra-model inconsistency*, which measures the discrepancy between the learning-based and data-driven probability predictions given by the local model, querying examples that can enhance local prediction reliability. Finally, we propose a hybrid sampling criterion that first selects a set of high-information examples with significant prediction inconsistencies and then uses clustering to query a representative batch based on these selections. This approach is referred to as **I**nconsistency-based **F**ederated **A**ctive **L**earning (IFAL), and its overview is presented on the right side of Figure 2.

3.2 Data-Driven Probability

To mitigate the inductive bias in the learning-based predicted probability introduced by varying training distributions, we propose a data-driven (structural) probability, determined by the reverse K-nearest neighbor (rKNN) structure of examples in the representation space. Specifically, in FAL's query phase, for client k, we first extract features from $\mathcal{D}_k^L$ using a specific model and combine them with their true labels to obtain $\mathcal{F}_k^L = \{(z_i, y_i)\}_{i=1}^{\mathcal{N}_k^L}$, and for $\mathcal{D}_k^U$, we extract features and generate pseudo-labels, resulting in $\mathcal{F}_k^U = \{(z_i, \tilde{y}_i)\}_{i=1}^{\mathcal{N}_k^U}$ and $\mathcal{F}_k = \mathcal{F}_k^L \cup \mathcal{F}_k^U = \{(z_i, y_i)\}_{i=1}^{\mathcal{N}_k^L} \cup \{(z_i, \tilde{y}_i)\}_{i=\mathcal{N}_k^L+1}^{\mathcal{N}_k^L+\mathcal{N}_k^U}$. For simplicity, we omit the distinction between y and $\tilde{y}$. Then, we emit K arrows from each instance in $\mathcal{F}_k$ to its K nearest neighbors in $\mathcal{F}_k^U$ based on cosine distance. For any $z_i \in \mathcal{F}_k^U$, its probability given class c can be approximated as:

$$p\left(z_i|c\right) = \frac{\text{\# of Arrows}_{(z_j \in \mathcal{F}_k, c)}}{|Z_k^c|}, \tag{1}$$

where $\text{\# of Arrows}_{(z_j \in \mathcal{F}_k, c)}$ denotes the total number of arrows directed at z_i from features with label $y = c$, and $|Z_k^c|$ represents the total number of features with label $y = c$ in $\mathcal{F}_k$. If z_i is in a region where features with label $y = c$ densely exist, it is likely to receive more arrows, and vice versa.

Under Bayes' theorem, we define z_i's data-driven (structural) class probability distribution as

$$p\left(c|z_i\right) = \frac{p\left(z_i|c\right) p\left(c\right)}{\sum_{v=1}^{C} p\left(z_i|v\right) p\left(v\right)}. \tag{2}$$

Here, the prior $p\left(c\right)$ is the probability of observing class c and can be approximately determined by the example count for each class in $\mathcal{F}_k$:

$$p\left(c\right) = \frac{|Z_k^c|}{\sum_{v=1}^{C} |Z_k^v|}. \tag{3}$$

Based on Equations (1), (2), and (3), we can calculate the structural probability of any $x_i \in \mathcal{D}_k^U$ in class c by:

$$p\left(c|x_i\right) = p\left(c|z_i\right) = \frac{\text{\# of Arrows}_{(z_j \in \mathcal{F}_k, c)}}{\sum_{v=1}^{C} \text{\# of Arrows}_{(z_j \in \mathcal{F}_k, v)}}. \tag{4}$$

3.3 Inter-Model Inconsistency

As such, we can measure the inconsistency by evaluating the discrepancy in the structural probabilities output by the local model and the global model for the same example. In detail, for client k, only f_G is adopted to generate pseudo-labels for all examples in $\mathcal{D}_k^U$. Then, both f_{θ_k} and f_G are employed to extract features, and the corresponding structural probabilities are derived based on Equation (4). For example, for $x_i \in \mathcal{D}_k^U$, the structural probability predicted by f_G is denoted as $p_G\left(c|x_i\right)$, while that predicted by f_{θ_k} is $p_k\left(c|x_i\right)$. Here, pseudo-labels are sampled from a single model, ensuring that the measurement of inconsistency focuses more on the topological structural differences of the examples in the representation spaces of different models, thus alleviating the impact of inductive bias. Additionally, the pseudo-labels generated by the global model are generally more accurate and better aligned with the true label distribution.

However, directly evaluating the prediction inconsistency between the local and global models overlooks the detrimental impact of noisy examples[1] on model predictions, especially in AL scenarios where model performance is often suboptimal. Motivated by two key observations: 1) useful knowledge can be transferred from the teacher model to the student, while noisy information is difficult to transfer; and 2) noise often induces random outputs from the model, leading to significant divergence in predictions across multiple models for the same noisy example, we propose training an intermediate model for each client by distilling knowledge from the global model to facilitate example selection.

Let $f_{\theta_{k-KD}}$ denote the intermediate model, with its and f_G's learning-based softmax probabilities in class c represented as $P_{k-KD}\left(c|x_i, \mathbf{T}\right)$ and $P_G\left(c|x_i, \mathbf{T}\right)$, respectively, where $\mathbf{T}$ controls the softness of the logits. Given $x_i \in \mathcal{D}_k^L$, the knowledge distillation (KD) training loss $\mathcal{L}$ for $f_{\theta_{k-KD}}$ is a linear combination of the cross-entropy (CE) loss $\mathcal{L}_{CE}$ and the Kullback-Leibler (KL) divergence loss $\mathcal{L}_{KL}$:

$$\begin{aligned}\mathcal{L} &= (1-\lambda)\,\mathcal{L}_{CE} + \lambda \mathcal{L}_{KL} \\ &= (1-\lambda)\left(-\textstyle\sum_{v=1}^{C} \boldsymbol{y}_i \log P_{k-KD}\left(v|x_i, 1\right)\right) \\ &+ \lambda\left(\mathbf{T}^2 \textstyle\sum_{v=1}^{C} P_G\left(v|x_i, \mathbf{T}\right) \log \frac{P_G\left(v|x_i, \mathbf{T}\right)}{P_{k-KD}\left(v|x_i, \mathbf{T}\right)}\right),\end{aligned} \tag{5}$$

where λ is a balancing factor, $\boldsymbol{y}_i$ is the one-hot form of y_i, and $P_{k-KD}\left(v|x_i, 1\right)$ can be simplified as $P_{k-KD}\left(v|x_i\right)$.

Then, we calculate the prediction divergences of $f_{\theta_{k-KD}}$ with respect to f_{θ_k} and f_G for any $x_i \in \mathcal{D}_k^U$ and denote as $D_{k,k-KD}\left(x_i\right)$ and $D_{G,k-KD}\left(x_i\right)$, respectively. Here, we choose the Wasserstein distance (WD) instead of the commonly used KL or Jensen–Shannon (JS) divergence, as WD is better at handling cases where two distributions barely overlap and is more robust to sparse distributions, noise, and extreme values. Based on $D_{k,k-KD}\left(x_i\right)$ and $D_{G,k-KD}\left(x_i\right)$, we categorize the examples into four typical types:

- **Small $D_{k,k-KD}\left(x_i\right)$ and small $D_{G,k-KD}\left(x_i\right)$.** The three models provide stable and consistent predictions

[1]Noisy examples here also refer to highly challenging instances that are not yet suitable for model learning at the current stage.

for the same input, indicating the example is already grasped by the models and does not require labeling.

- **Small $D_{k,k-KD}(x_i)$ and large $D_{G,k-KD}(x_i)$.** A large prediction divergence between f_G and $f_{\theta_{k-KD}}$ suggests that the input may be difficult or noisy, while a small divergence between f_{θ_k} and $f_{\theta_{k-KD}}$ indicates a lower likelihood of noise. Thus, this example likely contains valuable information and should be labeled.
- **Large $D_{k,k-KD}(x_i)$ and small $D_{G,k-KD}(x_i)$.** A small prediction divergence between f_G and $f_{\theta_{k-KD}}$ indicates that the knowledge contained in the input has been transferred from f_G to $f_{\theta_{k-KD}}$. However, a large divergence between f_{θ_k} and $f_{\theta_{k-KD}}$ suggests a risk of incorrect knowledge transfer. Therefore, labeling the example is also necessary.
- **Large $D_{k,k-KD}(x_i)$ and large $D_{G,k-KD}(x_i)$.** The three models fail to provide consistent and stable predictions for the same input, suggesting that the example may be noisy or too complex for the models to handle at this stage. Hence, labeling is not recommended.

Based on the above analysis, we propose to calculate the inter-model inconsistency score for any $x_i \in \mathcal{D}_k^U$ by:

$$I_{inter}(x_i) = (D_{k,k-KD}(x_i) + D_{G,k-KD}(x_i)) \cdot \max\left\{\frac{D_{k,k-KD}(x_i)}{D_{G,k-KD}(x_i)}, \frac{D_{G,k-KD}(x_i)}{D_{k,k-KD}(x_i)}\right\}. \quad (6)$$

This equation prioritizes querying examples with one high and one low prediction divergence score, followed by those with both high scores, and finally those with both low scores.

3.4 Intra-Model Inconsistency

The inter-model inconsistency criterion prioritizes querying the most valuable examples for reducing the prediction gap between the local and global models. To query the examples most beneficial for enhancing the local model's performance, here we introduce an intra-model inconsistency criterion.

Since each instance has both a learning-based probability and a data-driven structural probability, we directly define the intra-model inconsistency score as the divergence between the two different prediction probabilities. In detail, for any $x_i \in \mathcal{D}_k^U$ in client k, given $P_k(c|x_i)$ and $p_k(c|x_i)$ for any class c, the intra-model inconsistency score is computed using the WD and denoted as $D_{k,k}(x_i)$ or $I_{intra}(x_i)$. Notably, in this part, $p_k(c|x_i)$ is obtained using the pseudo-labels provided by the local model. Therefore, this inconsistency quantifies the divergence between the learning-based decision boundary and the "abstract" data-driven decision boundary. The queried examples are more likely to be located in regions where the predictions of the two boundaries are inconsistent.

3.5 Inconsistency- and Diversity-Based Sampling

Since $I_{inter}(\cdot)$ and $I_{intra}(\cdot)$ are non-negative for any input, with higher values indicating greater prediction divergence, we directly calculate the total inconsistency score of any $x_i \in \mathcal{D}_k^U$ in client k as

$$I(x_i) = I_{inter}(x_i) \cdot I_{intra}(x_i). \quad (7)$$

Then, we can obtain the set of inconsistency scores for all unlabeled examples on client k, denoted as $\mathcal{I}_k^U = \{(I(x_i))\}_{i=1}^{\mathcal{N}_k^U}$.

To reduce information redundancy and maintain diversity in X_k^{query}, we first select a set of examples with high total inconsistency scores to form a candidate pool $\mathcal{C}_k^{query}$. While we can simply sample the top $\eta\%$ of examples from $\mathcal{D}_k^U$ in descending order of $\mathcal{I}_k^U$, we use automatically Otsu thresholding to decide the threshold τ, thus eliminating the need for an additional hyper-parameter. The candidate pool $\mathcal{C}_k^{query}$ is then formed as

$$\mathcal{C}_k^{query} = \left\{(x_i, z_i) \mid I(x_i) > \tau, 1 \le i \le \mathcal{N}_k^U\right\}, \quad (8)$$

where z_i is the feature of x_i extracted from f_{θ_k}.

Finally, we perform k-means clustering on $\mathcal{C}_k^{query}$ in the representation space to obtain b centroids and select the example closest to each centroid as the target to form X_k^{query}.

4 Experiments

4.1 Implementation Details

Datasets. The experiments are conducted on three benchmark datasets: CIFAR-10 [Krizhevsky *et al.*, 2009], CIFAR-100 [Krizhevsky *et al.*, 2009], and Tiny-Imagenet [Le and Yang, 2015]. CIFAR-10 and CIFAR-100 each contain 60k color images of size 32×32, divided into 50k training images and 10k test images, with 10 and 100 classes, respectively. Tiny-Imagenet is a subset of Imagenet [Deng *et al.*, 2009], consisting of 200 classes, with 500 training images and 50 validation images per class. For the main experiment part, the training data is distributed across $N = 10$ clients following a Dirichlet distribution with parameter $\alpha = 0.1$ to simulate a non-independent and identically distributed (non-IID) data distribution. For the ablation study part, we vary α to $[0.5, 1]$, and adjust N to $[5, 20]$. Visualizations of the different data distributions are shown in the supplementary file.

Implementation Details. The main experiments are conducted using the standard federated learning (FL) framework FedAvg [McMahan *et al.*, 2017], with FedProx [Li *et al.*, 2020] and SCAFFOLD [Karimireddy *et al.*, 2020] additionally examined in the ablation study. The total number of communication rounds T is set to 100, with 5 local update epochs per round. The federated active learning (FAL) process involves 6 cycles for CIFAR-10/100 and 3 cycles for Tiny-Imagenet. Initially, 5% of the examples are randomly selected to form $\mathcal{D}_k^L$ for each client k. In each subsequent AL round, 5% of the examples are queried. A 4-layer CNN is used as the base model, trained with the SGD optimizer (momentum 0.9, weight decay 1e-5, batch size 64). The learning rate (lr) is set to 0.01 and reduced by a factor of 10 after $T > 75$. The hyper-parameter K in reverse K-nearest neighbor (rKNN) is generally set to 250. The local distillation model is trained similarly for 5×100 epochs, with early stopping applied to reduce training time, and the lr is reduced after the (5×75)-th epoch. For the parameters in Equation (5), λ is 0.9, and $\mathbf{T}$ is 4, which are common settings in knowledge distillation tasks. We repeat all experiments three times on GeForce RTX 3090 GPUs and record the average results for three random seeds.

Baselines. We select nine AL query strategies for comparison, which can be further categorized into five groups:

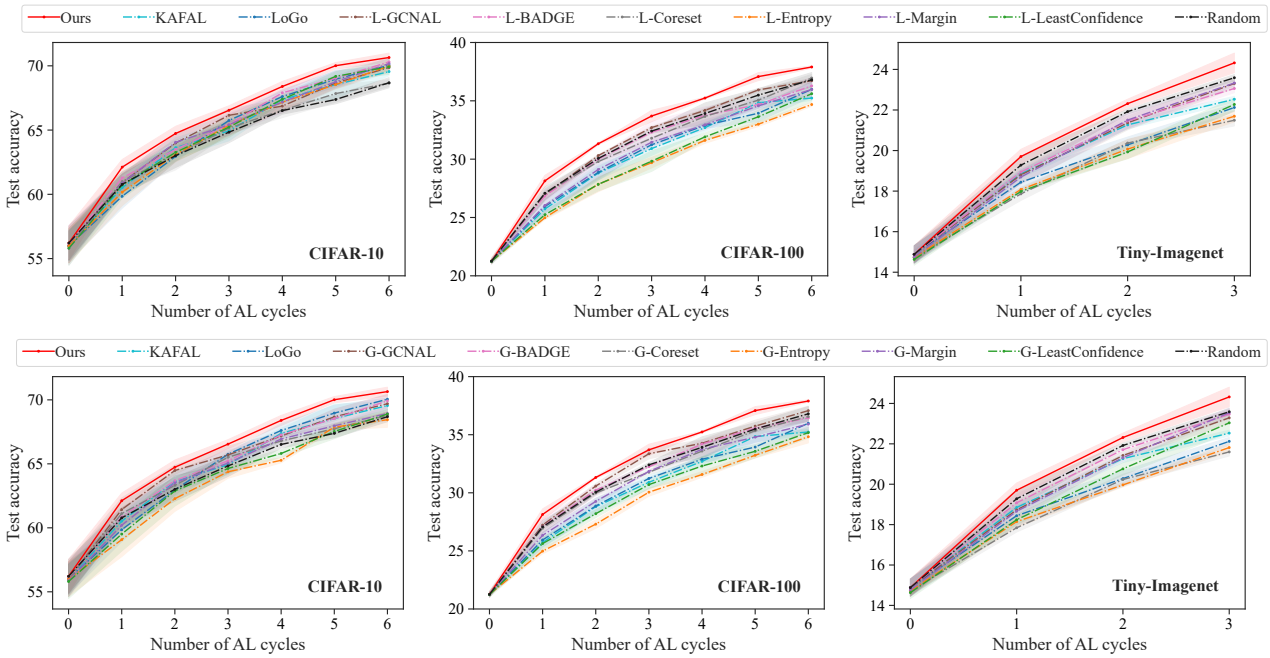

Figure 3: Test accuracy comparison on CIFAR-10 (left), CIFAR-100 (middle), and Tiny-ImageNet (right) with $\alpha = 0.1$. Results are repeated with three random seeds.

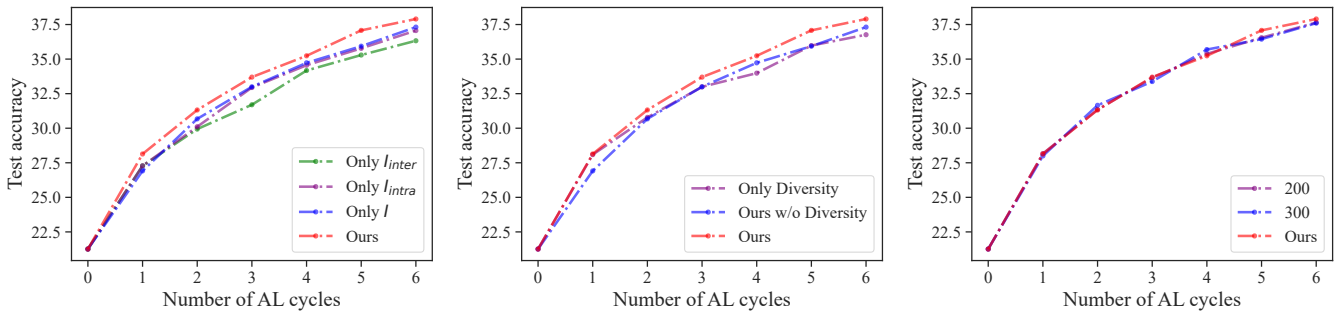

Figure 4: Ablation study on CIFAR-100 with $\alpha = 0.1$. Results are repeated with three random seeds.

(1) Random, which randomly selects examples from the unlabeled data pool for labeling. (2) Traditional uncertainty-based strategies, including entropy-based sampling (Entropy) [Holub *et al.*, 2008], margin-based sampling (Margin) [Balcan *et al.*, 2007], and least confidence sampling (LeastConfidence) [Li and Sethi, 2006]. (3) Traditional diversity-based strategy, Coreset [Sener and Savarese, 2017]. (4) Traditional hybrid strategies, including BADGE [Ash *et al.*, 2019] and GCNAL [Caramalau *et al.*, 2021]. (5) Latest FAL strategies, KAFAL [Cao *et al.*, 2023] and LoGo [Kim *et al.*, 2023]. For traditional strategies, we evaluate their performance separately using the local and global models, denoted by the prefixes "L-" and "G-", such as L-Entropy and G-Entropy.

4.2 Performance Comparison

Figure 3 displays the test accuracy of various methods on CIFAR-10, CIFAR-100, and Tiny-Imagenet. The specific numerical results are provided in the supplementary file.

As the number of AL cycles increases, all methods generally show improvement with the gradual addition of labeled instances. However, our method achieves the highest final test accuracy across all datasets, and in most AL cycles, the curve of our method consistently lies above those of the other methods, demonstrating its superiority. Additionally, we make the following observations. 1) As the difficulty of the dataset increases, all methods, except ours, gradually degrade in performance, with some even falling below random sampling. 2) For traditional strategies, the relative advantages of the local and global models used for executing AL queries vary across datasets. Generally, for uncertainty-based strategies and the hybrid strategy BADGE, the local model performs better on CIFAR-10, while the global model outperforms on other datasets. For the diversity-based strategy Coreset, the global model is superior on CIFAR-10, while the local model has a slight advantage on other datasets. Only for the hybrid strategy GCNAL, the global model consistently outperforms

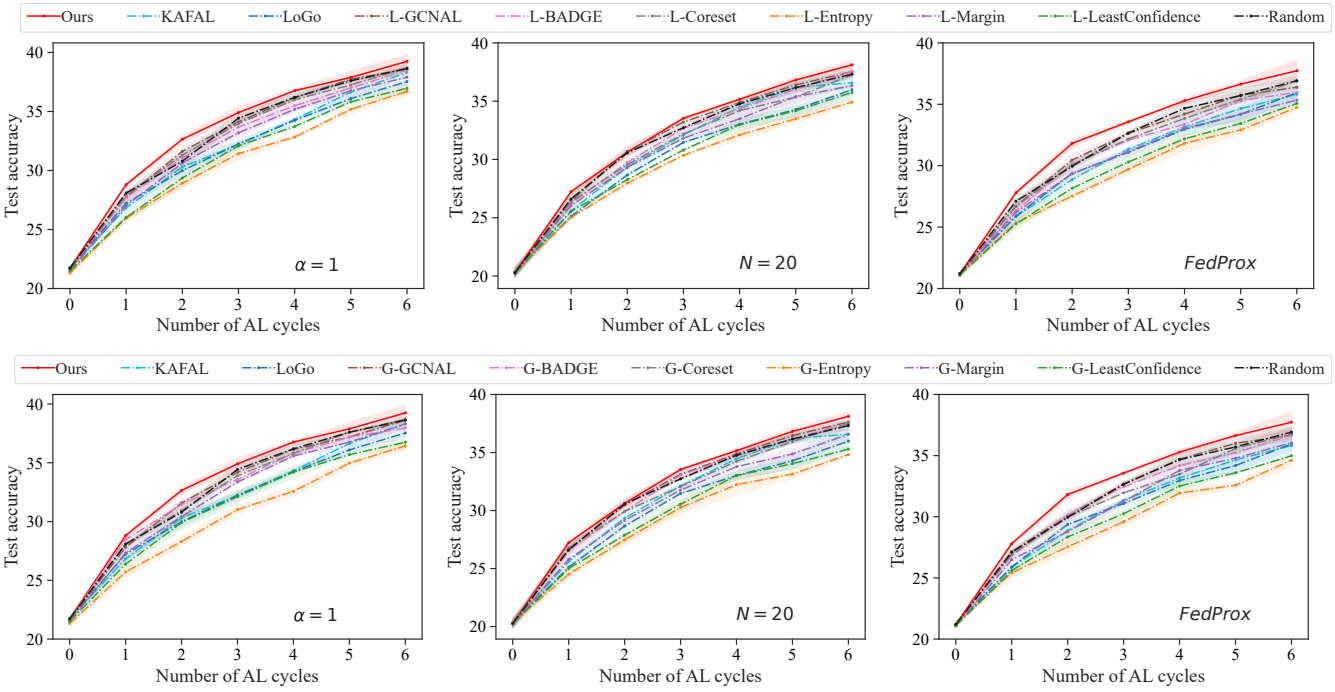

Figure 5: Test accuracy on CIFAR-100 with $\alpha = 1$, $N = 20$, and the FedProx federated learning (FL) framework, respectively. Results are repeated with three random seeds. The results with $\alpha = 0.5$, $N = 5$, and SCAFFOLD FL frameworks are shown in the supplementary file.

the local model. 3) Although KAFAL and LoGo are specifically designed for FAL, their performance does not show a significant advantage over other methods. The notable performance improvement of our method over theirs indirectly validates the rationale behind our approach and the effectiveness of its design.

4.3 Ablation Studies

Effect of each component. To validate the effectiveness of each component in IFAL, we first conduct experiments with three distinct inconsistency query strategies: inter-model inconsistency, intra-model inconsistency, and their combination, as shown in Figure 4 (left). We then present the diversity ablation experiment in Figure 4 (middle), where "Only Diversity" means solely applying k-means clustering to the unlabeled examples, and "Ours w/o Diversity" refers to directly selecting the batch of examples with the highest inconsistency scores. The results show that removing any of the components leads to performance degradation, which corroborate the soundness of our strategy design.

Effect of hyper-parameter K. Figure 4 (right) illustrates the effect of the hyper-parameter K in rKNN on IFAL's performance, with K set to [200, 250, 300]. The results show that varying the value of K within a certain range has little impact on IFAL's performance.

Robustness to varying federal settings. We further conduct experiments to assess the impact of different levels of data heterogeneity (α), client size (N), and FL frameworks on IFAL's performance. The results are shown in Figure 5, where the left plot changes α to 1, the middle plot changes N to 20, and the right plot changes the FL framework to FedProx. Additional results, where α is set to 0.5, N is set to 5, and the FL framework is switched to SCAFFOLD, are provided in the supplementary file. The results demonstrate that IFAL is independent of the specific FL framework adopted and exhibits strong generalization across diverse training setups, highlighting the versatility and robustness of IFAL in real-world FL applications.

5 Conclusion

In this paper, we introduced IFAL, a novel inconsistency-based federated active learning framework designed to address the challenges posed by data heterogeneity and noisy examples in query strategy design for federated learning. By leveraging a data-driven probabilistic formulation, IFAL aligns the biases between local and global models, enabling a more accurate assessment of model prediction inconsistencies by capturing the structural differences in the representation space. IFAL incorporates two inconsistency criteria: inter-model inconsistency and intra-model inconsistency. The former introduces an intermediate model and utilizes the randomness in predictions for noisy examples, effectively filtering out noise and querying examples that are truly valuable for reducing the prediction divergence between local and global models. The latter leverages the prediction divergence between learning-based and data-driven probabilities to identify examples most useful for refining the local model's decision boundary. By combining these strategies with clustering, IFAL forms a diverse and informative query set. Extensive experiments and analyses confirm the superiority of IFAL over state-of-the-art methods across various settings.

Acknowledgements

This work was supported by the Natural Science Foundation of Jiangsu Province of China (BK20222012) and the NSFC (U2441285, 62222605).

References

[Ahn *et al.*, 2024] Jin-Hyun Ahn, Yeeun Ma, Seoyun Park, and Cheolwoo You. Federated active learning (f-al): an efficient annotation strategy for federated learning. *IEEE Access*, 2024.

[An *et al.*, 2023] Xuming An, Li Shen, Han Hu, and Yong Luo. Federated learning with manifold regularization and normalized update reaggregation. *Advances in Neural Information Processing Systems*, 36:55097–55109, 2023.

[Ash *et al.*, 2019] Jordan T Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. Deep batch active learning by diverse, uncertain gradient lower bounds. *arXiv preprint arXiv:1906.03671*, 2019.

[Bai *et al.*, 2024] Sikai Bai, Shuaicheng Li, Weiming Zhuang, Jie Zhang, Kunlin Yang, Jun Hou, Shuai Yi, Shuai Zhang, and Junyu Gao. Combating data imbalances in federated semi-supervised learning with dual regulators. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 10989–10997, 2024.

[Balcan *et al.*, 2007] Maria-Florina Balcan, Andrei Broder, and Tong Zhang. Margin based active learning. In *International Conference on Computational Learning Theory*, pages 35–50. Springer, 2007.

[Cao *et al.*, 2023] Yu-Tong Cao, Ye Shi, Baosheng Yu, Jingya Wang, and Dacheng Tao. Knowledge-aware federated active learning with non-iid data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22279–22289, 2023.

[Caramalau *et al.*, 2021] Razvan Caramalau, Binod Bhattarai, and Tae-Kyun Kim. Sequential graph convolutional network for active learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9583–9592, 2021.

[Deng *et al.*, 2009] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[Fan *et al.*, 2022] Chenyou Fan, Junjie Hu, and Jianwei Huang. Private semi-supervised federated learning. In *IJCAI*, pages 2009–2015, 2022.

[Han *et al.*, 2018] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. *Advances in neural information processing systems*, 31, 2018.

[Holub *et al.*, 2008] Alex Holub, Pietro Perona, and Michael C Burl. Entropy-based active learning for object recognition. In *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–8. IEEE, 2008.

[Huang *et al.*, 2021] Sheng-Jun Huang, Chen-Chen Zong, Kun-Peng Ning, and Haibo Ye. Asynchronous active learning with distributed label querying. In *IJCAI*, pages 2570–2576, 2021.

[Karimireddy *et al.*, 2020] 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.

[Kim *et al.*, 2023] SangMook Kim, Sangmin Bae, Hwanjun Song, and Se-Young Yun. Re-thinking federated active learning based on inter-class diversity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3944–3953, 2023.

[Konečnỳ *et al.*, 2016] Jakub Konečnỳ, H Brendan McMahan, Daniel Ramage, and Peter Richtárik. Federated optimization: Distributed machine learning for on-device intelligence. *arXiv preprint arXiv:1610.02527*, 2016.

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

[Le and Yang, 2015] Yann Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015.

[Li and Sethi, 2006] Mingkun Li and Ishwar K Sethi. Confidence-based active learning. *IEEE transactions on pattern analysis and machine intelligence*, 28(8):1251–1261, 2006.

[Li *et al.*, 2020] 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.

[Li *et al.*, 2023] Zexi Li, Tao Lin, Xinyi Shang, and Chao Wu. Revisiting weighted aggregation in federated learning with neural networks. In *International Conference on Machine Learning*, pages 19767–19788. PMLR, 2023.

[Liu *et al.*, 2024] 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.

[McMahan *et al.*, 2017] 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.

[Nguyen and Smeulders, 2004] Hieu T Nguyen and Arnold Smeulders. Active learning using pre-clustering. In *Proceedings of the twenty-first international conference on Machine learning*, page 79, 2004.

[Otsu and others, 1975] Nobuyuki Otsu et al. A threshold selection method from gray-level histograms. *Automatica*, 11(285-296):23–27, 1975.

[Pillutla *et al.*, 2022] Krishna Pillutla, Sham M Kakade, and Zaid Harchaoui. Robust aggregation for federated learning. *IEEE Transactions on Signal Processing*, 70:1142–1154, 2022.

[Prabhu *et al.*, 2021] Viraj Prabhu, Arjun Chandrasekaran, Kate Saenko, and Judy Hoffman. Active domain adaptation via clustering uncertainty-weighted embeddings. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8505–8514, 2021.

[Sener and Savarese, 2017] Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. *arXiv preprint arXiv:1708.00489*, 2017.

[Settles, 2009] Burr Settles. Active learning literature survey. 2009.

[Wang *et al.*, 2023] Fan Wang, Zhongyi Han, Zhiyan Zhang, Rundong He, and Yilong Yin. Mhpl: Minimum happy points learning for active source free domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20008–20018, 2023.

[Wang *et al.*, 2024] Jiaqi Wang, Xingyi Yang, Suhan Cui, Liwei Che, Lingjuan Lyu, Dongkuan DK Xu, and Fenglong Ma. Towards personalized federated learning via heterogeneous model reassembly. *Advances in Neural Information Processing Systems*, 36, 2024.

[Wu *et al.*, 2022] Xing Wu, Jie Pei, Cheng Chen, Yimin Zhu, Jianjia Wang, Quan Qian, Jian Zhang, Qun Sun, and Yike Guo. Federated active learning for multicenter collaborative disease diagnosis. *IEEE transactions on medical imaging*, 42(7):2068–2080, 2022.

[Yuan *et al.*, 2023] Jiakang Yuan, Bo Zhang, Xiangchao Yan, Tao Chen, Botian Shi, Yikang Li, and Yu Qiao. Bi3d: Bi-domain active learning for cross-domain 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15599–15608, 2023.

[Zhang *et al.*, 2023a] Chen Zhang, Yu Xie, Hang Bai, Xiongwei Hu, Bin Yu, and Yuan Gao. Federated active semi-supervised learning with communication efficiency. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2023.

[Zhang *et al.*, 2023b] Jianqing Zhang, Yang Hua, Hao Wang, Tao Song, Zhengui Xue, Ruhui Ma, and Haibing Guan. Fedala: Adaptive local aggregation for personalized federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 11237–11244, 2023.

[Zhang *et al.*, 2023c] Jie Zhang, Chen Chen, Weiming Zhuang, and Lingjuan Lyu. Target: Federated class-continual learning via exemplar-free distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4782–4793, 2023.

[Zong and Huang, 2025] Chen-Chen Zong and Sheng-Jun Huang. Rethinking epistemic and aleatoric uncertainty for active open-set annotation: An energy-based approach. *arXiv preprint arXiv:2502.19691*, 2025.

[Zong *et al.*, 2024] Chen-Chen Zong, Ye-Wen Wang, Ming-Kun Xie, and Sheng-Jun Huang. Dirichlet-based prediction calibration for learning with noisy labels. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17254–17262, 2024.

[Zong *et al.*, 2025] Chen-Chen Zong, Ye-Wen Wang, Kun-Peng Ning, Hai-Bo Ye, and Sheng-Jun Huang. Bidirectional uncertainty-based active learning for open-set annotation. In *European Conference on Computer Vision*, pages 127–143. Springer, 2025.