FEDERATED COMPLEMENTARY LABEL LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Federated learning enables multiple clients to collaboratively train a global model without sharing their raw data, making it well suited to privacy-sensitive applications such as healthcare and finance. In many such scenarios, however, obtaining exact class labels requires substantial domain expertise, while ruling out unlikely classes is considerably easier. This naturally gives rise to federated complementary-label learning, where each client holds only class-exclusion supervision. A key challenge in this setting is that complementary evidence is inherently sparse. Moreover, data heterogeneity further fragments such evidence across clients, making it difficult for individual clients to obtain sufficiently informative supervision. To address this problem, we propose FEDCUE, a multi-level complementary-evidence learning framework that progressively consolidates fragmented exclusion information from individual samples to local semantic clusters and further to cross-client global clusters. Specifically, FEDCUE associates feature clusters with soft complementary-label distributions, allowing semantically similar samples to share exclusion knowledge locally and enabling compatible complementary evidence to be aggregated across clients without exchanging raw data. We further derive a convergence bound for the resulting time-varying federated objective under standard assumptions. Extensive experiments on synthetic and human-annotated complementary-label benchmarks under various Non-IID settings show that FEDCUE consistently improves federated complementary-label learning. The gains persist across the tested federation scales and participation levels, supporting the proposed multi-level evidence design.

1 INTRODUCTION

Federated learning (FL) (McMahan et al., 2017) enables multiple clients to collaboratively train a shared model without exchanging their raw data. This property makes FL well suited for privacy-sensitive applications such as healthcare, finance, and mobile intelligence (Li et al., 2020a; Kairouz & McMahan, 2021; Wen et al., 2023; Rieke et al., 2020). However, data collected by different clients are often statistically heterogeneous and follow non-independent and identically distributed (Non-IID) distributions (Kou et al., 2026; Li et al., 2020b; Karimireddy et al., 2020; Zhao et al., 2018), making effective collaboration across clients a fundamental challenge (Kou et al., 2026).

However, obtaining ground-truth labels can be difficult and expensive in many real-world applications, especially when annotation requires substantial domain expertise (Litjens et al., 2017). Consider medical imaging as an example. Given a CT image, a general annotator or a junior practitioner may be able to rule out several clearly irrelevant diseases, while determining the exact diagnosis may require an experienced radiologist or even consultation among multiple specialists. In such cases, it is often much easier to obtain exclusion-based supervision, such as "this image does not correspond to disease c," than to obtain its exact class label (Ishida et al., 2017; Yu et al., 2018). Complementary labels naturally describe this type of supervision by specifying classes that a sample does not belong to (Ishida et al., 2017; 2019; Feng et al., 2020). When complementary-labeled data are distributed across different institutions, privacy, security, and regulatory constraints further prevent raw data from being directly centralized (Kairouz & McMahan, 2021; Rieke et al., 2020). This naturally leads to a practical learning scenario in which multiple clients possess only complementary-labeled local data and aim to collaboratively learn a global model through federated learning.

Despite its practical relevance, federated complementary-label learning remains largely underexplored. A complementary label provides only limited exclusion information about an individual

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

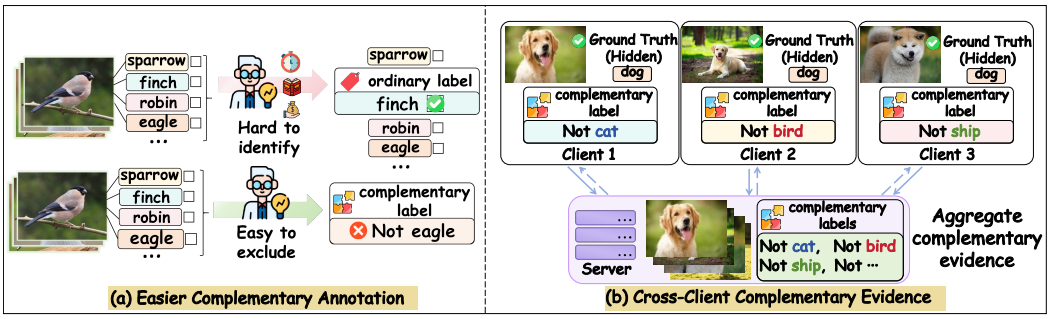

Figure 1: Motivation of federated complementary-label learning. (a) Exact class labels require substantial domain expertise, while excluding unlikely classes is often easier. (b) Semantically similar samples across clients may provide complementary class-exclusion information, whose aggregation yields richer supervision than individual complementary labels.

sample (Ishida et al., 2017; 2019; Feng et al., 2020). Nevertheless, the federated setting also introduces an important opportunity: semantically similar samples distributed across different clients may carry different yet compatible complementary information. As illustrated in Fig. 1, samples from the same underlying semantic cluster may be annotated as "not cat" at one client, "not bird" at another client, and "not ship" at a third client. Each annotation is weak on its own, but their combination provides more informative exclusion supervision.

The key difficulty is therefore not merely the weakness of complementary supervision, but the *fragmentation of complementary evidence* across samples and clients. Under Non-IID data, each client observes only a partial view of the underlying class structure, while semantically related samples may carry different class-exclusion information. Conventional FL mainly exchanges and aggregates model parameters, so such sample-level complementary evidence is only implicitly absorbed into local updates and cannot be explicitly reused across clients. This naturally raises two questions: ❶ *How can fragmented complementary evidence be consolidated within each client?* and ❷ *How can compatible complementary evidence be exploited across clients without exposing individual samples?* Together, these questions call for collaboration beyond model aggregation that explicitly organizes complementary evidence at both local and cross-client levels.

To address these questions, we propose FEDCUE, a multi-level complementary-evidence learning framework for federated complementary-label learning. FEDCUE organizes fragmented complementary evidence through local clustering and cross-client aggregation. Each client groups feature-similar samples and combines their complementary labels into local cluster-level evidence, while only compact cluster statistics are uploaded to the server. The server further aggregates compatible clusters across clients and returns additional complementary supervision to each client. In this way, FEDCUE extends conventional model aggregation with explicit collaboration over complementary evidence, without exchanging raw samples. We further establish a convergence guarantee for FEDCUE while accounting for the round-wise objective changes introduced by reconstructed complementary evidence. Extensive experiments on synthetic and human-annotated complementary-label benchmarks under various Non-IID settings consistently demonstrate the effectiveness of FEDCUE over representative federated baselines. Additional ablation, scalability, and sensitivity analyses further support the proposed multi-level evidence design. Our contributions are summarized as follows:

- *Problem and framework.* To the best of our knowledge, we are the first to systematically study federated learning under exclusively complementary-label supervision. We propose FEDCUE to organize and exploit complementary evidence locally and globally without raw-data sharing.
- *Theory.* We analyze the time-varying objective induced by complementary-evidence reconstruction and derive a convergence bound under standard federated optimization assumptions.
- *Experiments.* We conduct extensive experiments on synthetic and human-annotated benchmarks under various Non-IID settings. The results show that FEDCUE consistently outperforms representative federated baselines. Ablation, scalability, and sensitivity analyses further support the effectiveness and robustness of the proposed multi-level evidence design.

2 Related Work

2.1 Federated Learning

Federated learning (FL) (McMahan et al., 2017) enables multiple clients to collaboratively train a global model while keeping raw data decentralized. Existing studies address heterogeneous client distributions (Karimireddy et al., 2020; Kou et al., 2026; Wu et al., 2026). FedProx (Li et al., 2020b) reduces local-global divergence through proximal regularization, while MOON (Li et al., 2021) uses contrastive learning. Beyond parameter aggregation, FedProto (Tan et al., 2022) exchanges class prototypes and FedMD (Li & Wang, 2019) transfers model predictions. Sharma et al. (2022) study federated learning from positive and negative feedback, treating rejected predictions as complementary labels, but also use positive feedback and a supervised seed model. Existing methods do not explicitly organize complementary evidence across clients.

2.2 Complementary-Label Learning

Complementary-label learning (Ishida et al., 2017) studies weak supervision where annotations specify classes that a sample does not belong to. Existing work mainly focuses on centralized settings, including risk estimators (Ishida et al., 2017; 2019), multi-complementary-label learning (Feng et al., 2020), and consistency- or contrastive-based representation learning (Wang et al., 2021; Jiang et al., 2024). Lin et al. (2023) improve label sharing by propagating complementary labels among nearby instances. These methods assume centralized access to training data, whereas federated complementary evidence is distributed across heterogeneous clients.

3 Problem Setup

We consider federated complementary-label learning with a central server and K clients. Client k holds a private dataset $\mathcal{D}_k = \{(x_i^k, \bar{\mathbf{y}}_i^k)\}_{i=1}^{n_k}$, where $\bar{\mathbf{y}}_i^k \in \mathbb{R}_{\geq 0}^C$ and $\bar{y}_{i,c}^k > 0$ indicates that x_i^k is excluded from class c, with larger values allowing repeated annotations. Ground-truth labels are unavailable during training. For the basic formulation, we assume that all observed complementary labels are valid. We define the normalized complementary-label distribution as $\bar{\mathbf{q}}_i^k = \bar{\mathbf{y}}_i^k / \|\bar{\mathbf{y}}_i^k\|_1$, which covers both single- and multiple-complementary-label settings. Client data are generally Non-IID and may differ in both class and complementary-label distributions.

The model $F_\Theta = g_\phi \circ h_\theta$ consists of an encoder h_θ and a classifier g_ϕ, with $\Theta = (\theta, \phi)$, where h_θ maps inputs to feature representations. For sample x_i^k, the prediction is $\mathbf{p}_i^k(\Theta) = \mathrm{softmax}(g_\phi(h_\theta(x_i^k))) \in \Delta^{C-1}$. At communication round t, the server maintains the global model Θ^t and selects clients $\mathcal{K}_t$ for local training. Our goal is to learn a global classifier from distributed complementary-label data while keeping raw samples and instance-level annotations local.

4 Method

Overview. The key challenge is not the absence of exclusion evidence but its fragmentation: each sample provides limited complementary evidence, while related evidence is distributed within and across clients. FedCUE progressively consolidates this evidence at three levels. It learns from observed complementary labels, constructs local soft complementary distributions from feature clusters, and aggregates matched cluster-level evidence into a global complementary-evidence bank. An overview of FedCUE is shown in Fig. 2.

4.1 Observed Complementary Supervision

For a predicted class distribution $\mathbf{p} \in \Delta^{C-1}$ and a normalized complementary-label distribution $\mathbf{q} \in \Delta^{C-1}$, we define the complementary-label loss as

$$\ell_{\mathrm{cl}}(\mathbf{p}, \mathbf{q}) = -\sum_{c=1}^{C} q_c \log(1 - p_c). \tag{1}$$

A larger value of q_c indicates stronger evidence that class c should be excluded.

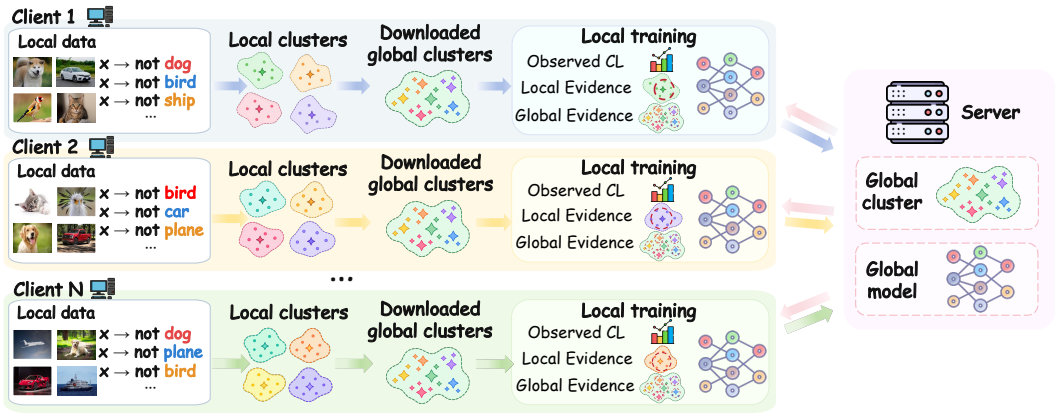

Figure 2: Overview of FEDCUE.

Each client first learns from its observed complementary labels. Given the predicted class probabilities $\mathbf{p}_i^k(\Theta)$ and the normalized complementary-label distribution $\bar{\mathbf{q}}_i^k$, we define the observed complementary-label objective for local training at client k as

$$\mathcal{L}_{\text{obs}}^k(\Theta) = \frac{1}{n_k} \sum_{i=1}^{n_k} \ell_{\text{cl}} \left(\mathbf{p}_i^k(\Theta), \bar{\mathbf{q}}_i^k \right). \tag{2}$$

Minimizing Eq. (2) suppresses classes indicated by the observed complementary labels.

4.2 LOCAL COMPLEMENTARY-EVIDENCE MINING

Complementary labels from semantically similar samples can provide mutually reinforcing exclusion evidence. We therefore construct local semantic clusters and aggregate their complementary labels into soft complementary distributions.

Client k uses the frozen broadcast encoder h_{θ^t} to compute $\widetilde{\mathbf{z}}_i^{k,t} = \text{Norm}(h_{\theta^t}(x_i^k))$, where $\text{Norm}(\mathbf{v}) = \mathbf{v}/\|\mathbf{v}\|_2$. Spherical K-means partitions these features into M_k clusters $\{\mathcal{C}_{k,m}^t\}_{m=1}^{M_k}$. The feature center of cluster m is

$$\mathbf{u}_{k,m}^t = \text{Norm}(|\mathcal{C}_{k,m}^t|^{-1} \sum_{i \in \mathcal{C}_{k,m}^t} \widetilde{\mathbf{z}}_i^{k,t}). \tag{3}$$

We use $d(\mathbf{a}, \mathbf{b}) = \|\mathbf{a} - \mathbf{b}\|_2$ throughout the method. To discard unreliable samples, we retain the neighborhood $\mathcal{N}_{k,m}^t = \{i \in \mathcal{C}_{k,m}^t : d(\widetilde{\mathbf{z}}_i^{k,t}, \mathbf{u}_{k,m}^t) \le r_L\}$, where r_L is the local reliability radius. A retained sample receives the kernel weight $a_{i,m}^{k,t} = \exp[-d(\widetilde{\mathbf{z}}_i^{k,t}, \mathbf{u}_{k,m}^t)^2/\tau_L]$, with temperature $\tau_L > 0$. The soft complementary distribution associated with local cluster m and the corresponding leave-one-out distribution for example i are defined as

$$\begin{cases} \mathbf{q}_{k,m}^{L,t} = \dfrac{\sum_{l \in \mathcal{N}_{k,m}^t} a_{l,m}^{k,t} \bar{\mathbf{q}}_l^k}{\sum_{l \in \mathcal{N}_{k,m}^t} a_{l,m}^{k,t}}, \\ \mathbf{q}_{k,m}^{L,t,-i} = \dfrac{\sum_{l \in \mathcal{N}_{k,m}^t \setminus \{i\}} a_{l,m}^{k,t} \bar{\mathbf{q}}_l^k}{\sum_{l \in \mathcal{N}_{k,m}^t \setminus \{i\}} a_{l,m}^{k,t}}. \end{cases} \tag{4}$$

Let $m(i)$ be the cluster containing example i, and define the eligible set $\mathcal{I}_k^{L,t}$ to contain examples that satisfy the local neighborhood criterion and have at least one other retained neighbor. This ensures that the denominator of $\mathbf{q}_{k,m(i)}^{L,t,-i}$ is nonzero. Using $s_i^{L,t} = a_{i,m(i)}^{k,t}$ to weight each eligible example by its association reliability, we define the local evidence loss as

$$\mathcal{L}_{\text{loc}}^{k,t}(\Theta) = \frac{\sum_{i \in \mathcal{I}_k^{L,t}} s_i^{L,t} \ell_{\text{cl}} \left(\mathbf{p}_i^k(\Theta), \text{sg}(\mathbf{q}_{k,m(i)}^{L,t,-i}) \right)}{\sum_{i \in \mathcal{I}_k^{L,t}} s_i^{L,t} + \epsilon}, \tag{5}$$

where $\mathrm{sg}(\cdot)$ stops gradients through the constructed target and $\epsilon > 0$ prevents numerical division by zero. We set $\mathcal{L}_{\mathrm{loc}}^{k,t} = 0$ when $\mathcal{I}_k^{L,t}$ is empty.

4.3 Federated Complementary-Evidence Aggregation

Local evidence remains limited by the coverage of each individual client. To exploit complementary evidence beyond local observations, client k uploads one cluster-level evidence summary $S_{k,m}^t = (\mathbf{u}_{k,m}^t, \mathbf{q}_{k,m}^{L,t}, n_{k,m}^t)$ for each valid cluster, where $n_{k,m}^t = |\mathcal{N}_{k,m}^t|$ denotes its support size. The server receives only these cluster-level statistics, avoiding the exchange of raw samples, although such communication does not by itself provide a formal privacy guarantee.

The server first applies spherical K-means to the uploaded local cluster representations, obtaining M_G preliminary global centers $\widehat{\mathbf{g}}_j^t$ and their assignments. Let $\mathcal{A}_j^t$ contain the local cluster representations assigned to center j whose distance to $\widehat{\mathbf{g}}_j^t$ does not exceed the global matching radius r_G. For each $(k,m) \in A_j^t$, we define the corresponding aggregation weight as

$$w_{k,m,j}^t = n_{k,m}^t \exp[-d(\mathbf{u}_{k,m}^t, \widehat{\mathbf{g}}_j^t)^2/\tau_G], \tag{6}$$

where $\tau_G > 0$ controls the effect of cluster distance. The global cluster representation is computed by aggregating matched local cluster representations, while the complementary target delivered to client k excludes its own contribution:

$$\begin{cases} \mathbf{g}_j^t = \mathrm{Norm}\left(\dfrac{\sum_{(k',m)\in\mathcal{A}_j^t} w_{k',m,j}^t \mathbf{u}_{k',m}^t}{\sum_{(k',m)\in\mathcal{A}_j^t} w_{k',m,j}^t}\right), \\ \mathbf{q}_{j\to k}^{G,t} = \dfrac{\sum_{(k',m)\in\mathcal{A}_j^t:\, k'\neq k} w_{k',m,j}^t \mathbf{q}_{k',m}^{L,t}}{\sum_{(k',m)\in\mathcal{A}_j^t:\, k'\neq k} w_{k',m,j}^t}. \end{cases} \tag{7}$$

Using the preliminary centers makes the distance weights well-defined before the final global centers are computed. The leave-one-client-out target $\mathbf{q}_{j\to k}^{G,t}$ ensures that the global evidence for client k is contributed only by other clients, and is retained only when the corresponding global cluster contains support from at least one other client.

The server returns the personalized complementary-evidence bank $\mathcal{B}_k^t = \{(\mathbf{g}_j^t, \mathbf{q}_{j\to k}^{G,t})\}$, retaining only valid targets. Client k retrieves

$$\mathcal{N}_{k,i}^{G,t} = \{j : (\mathbf{g}_j^t, \mathbf{q}_{j\to k}^{G,t}) \in \mathcal{B}_k^t,\ d(\widetilde{\mathbf{z}}_i^{k,t}, \mathbf{g}_j^t) \le r_R\}, \tag{8}$$

where r_R is the retrieval radius. For $j \in \mathcal{N}_{k,i}^{G,t}$, the normalized retrieval weight is

$$\omega_{k,i,j}^{G,t} = \mathrm{softmax}_{j\in\mathcal{N}_{k,i}^{G,t}}[-d(\widetilde{\mathbf{z}}_i^{k,t}, \mathbf{g}_j^t)^2/\tau_R],$$

where $\tau_R > 0$ is the retrieval temperature. The global complementary target is computed as

$$\mathbf{q}_{k,i}^{G,t} = \sum_{j\in\mathcal{N}_{k,i}^{G,t}} \omega_{k,i,j}^{G,t} \mathbf{q}_{j\to k}^{G,t}. \tag{9}$$

Let $\mathcal{I}_k^{G,t} = \{i : \mathcal{N}_{k,i}^{G,t} \neq \emptyset\}$ and define

$$s_{k,i}^{G,t} = \max_{j\in\mathcal{N}_{k,i}^{G,t}} \exp[-d(\widetilde{\mathbf{z}}_i^{k,t}, \mathbf{g}_j^t)^2/\tau_R].$$

The global evidence loss is

$$\mathcal{L}_{\mathrm{glo}}^{k,t}(\Theta) = \frac{\sum_{i\in\mathcal{I}_k^{G,t}} s_{k,i}^{G,t} \ell_{\mathrm{cl}}\left(\mathbf{p}_i^k(\Theta), \mathrm{sg}(\mathbf{q}_{k,i}^{G,t})\right)}{\sum_{i\in\mathcal{I}_k^{G,t}} s_{k,i}^{G,t} + \epsilon}. \tag{10}$$

We set this loss to zero when no global cluster passes the retrieval test. The retrieval radius prevents unrelated semantic regions from providing misleading exclusion supervision, while $s_{k,i}^{G,t}$ assigns larger importance to better-matched global evidence.

4.4 Joint Learning and Federated Optimization

To obtain stable semantic clusters, we perform T_w warm-up rounds using only the observed complementary labels. Afterwards, client k optimizes

$$\mathcal{L}_k^t(\Theta) = \mathcal{L}_{\text{obs}}^k(\Theta) + \lambda_L^t \mathcal{L}_{\text{loc}}^{k,t}(\Theta) + \lambda_G^t \mathcal{L}_{\text{glo}}^{k,t}(\Theta), \tag{11}$$

where λ_L and λ_G control the contributions of local and global complementary evidence, respectively, with $\lambda_L^t = \lambda_L \mathbf{1}[t \geq T_w]$ and $\lambda_G^t = \lambda_G \mathbf{1}[t \geq T_w]$.

Each post-warm-up round first exchanges cluster-level evidence using the frozen global encoder h_{θ^t}. Clients then initialize from Θ^t and optimize Eq. (11) for E local epochs, with the constructed complementary targets fixed during local training. The server finally aggregates the resulting local models to obtain the next global model as

$$\Theta^{t+1} = \sum_{k \in \mathcal{K}_t} \frac{n_k}{\sum_{k' \in \mathcal{K}_t} n_{k'}} \Theta_k^{t+1}.$$

This ordering ensures that all exchanged cluster representations are constructed in the same feature space. The complete training procedure is given in Appendix B.

5 Theoretical Analysis

FedCUE induces time-varying objectives because complementary evidence is reconstructed across rounds and the auxiliary-loss weights change after warm-up. Within each round, however, the constructed targets and auxiliary weights remain fixed during local optimization. We analyze the full-participation setting used in our main experiments.

Let $p_k = n_k/n$, where $n = \sum_{k=1}^K n_k$, and define

$$\begin{aligned} F^t(\Theta) &= \sum_{k=1}^K p_k F_k^t(\Theta), \\ F_k^t(\Theta) &= \mathcal{L}_{\text{obs}}^k(\Theta) + \lambda_L^t \mathcal{L}_{\text{loc}}^{k,t}(\Theta) + \lambda_G^t \mathcal{L}_{\text{glo}}^{k,t}(\Theta). \end{aligned} \tag{12}$$

For the analysis, let H denote the number of local stochastic-gradient updates per communication round. We define the cumulative round-to-round objective variation as

$$V_T = \sum_{t=0}^{T-2} \mathbb{E}\left[\left|F^{t+1}(\Theta^{t+1}) - F^t(\Theta^{t+1})\right|\right], \qquad V_1 = 0. \tag{13}$$

We assume that the round-wise local objectives are uniformly L-smooth, the stochastic gradients are unbiased with bounded variance σ^2, and the weighted gradient heterogeneity is uniformly bounded by ζ^2. The complete assumptions are given in Appendix A.

Theorem 1 *Under Assumptions 1–3, suppose that all clients participate in every round, each client performs H local updates, and $\eta LH \leq 1/8$. After T communication rounds,*

$$\begin{aligned} \frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\left[\|\nabla F^t(\Theta^t)\|^2\right] \leq \frac{4(\Delta_0 + V_T)}{\eta HT} + 4\eta L\sigma^2 \\ + \frac{20}{3}\eta^2 L^2 H^2(\sigma^2 + \zeta^2), \end{aligned} \tag{14}$$

where $\Delta_0 = \mathbb{E}[F^0(\Theta^0)] - F_{\text{inf}}$ and F_{inf} is a common lower bound of the round-wise objectives.

The bound captures the additional variation introduced by the time-varying FedCUE objective. For fixed H and bounded L, σ, ζ, the average squared gradient norm converges to zero when $\eta_T \to 0$, $\eta_T HT \to \infty$, and $V_T/(\eta_T HT) \to 0$.

Remark 1 *Theorem 1 shows that the round-wise reconstruction of complementary evidence does not prevent convergence when the induced objective variation remains controlled. The term V_T explicitly captures the effect of this variation on* FedCUE*'s convergence.*

6 EXPERIMENTS

6.1 EXPERIMENTAL SETUP

Datasets and Experimental Setup. We evaluate FEDCUE on four main benchmarks: CIFAR-10 (Krizhevsky et al., 2009), CIFAR-20 (Krizhevsky et al., 2009), SVHN (Netzer et al., 2011), and the human-annotated CL-CIFAR10 (Wang et al., 2023). For CIFAR-10, CIFAR-20, and SVHN, each training sample is assigned a complementary label uniformly sampled from the non-ground-truth classes. CL-CIFAR10 provides three complementary annotations per image, which we convert into a soft complementary-label distribution while preserving repeated annotations.

We partition the training data over 10 clients using a Dirichlet distribution with $\beta \in \{0.1, 0.5, 0.8\}$, where smaller β indicates stronger heterogeneity. Ground-truth labels are used only for partitioning, not training. All clients participate in every communication round, and all methods are trained for 100 rounds with 5 local epochs per round. We further report results for $\beta = 0.3$, MNIST, and CL-CIFAR20 in Appendix E.1. More implementation details are provided in Appendix C.

Baselines. We compare FEDCUE with two groups of federated baselines. *(i) Standard FL methods.* We include FedAvg (McMahan et al., 2017) and FedProx (Li et al., 2020b). *(ii) Robust FL methods.* We include FedRDN (Yan et al., 2025), FedCDA (Wang et al., 2024), and FedGLOSS-S (Caldarola et al., 2025). FedGLOSS-S denotes our server-side adaptation of FedGloSS with standard SGD clients. All methods use the same data partitions and training setup, isolating gains from complementary-evidence collaboration under the same complementary-label objective.

6.2 MAIN RESULTS

Table 1 and Fig. 3 report the main results. Table 1 summarizes the final test accuracy under three heterogeneity levels, $\beta \in \{0.1, 0.5, 0.8\}$, while Fig. 3 shows the convergence behavior over communication rounds. From these results, we draw the following three findings.

- *Finding 1: Complementary-evidence collaboration provides gains beyond conventional federated optimization.* FEDCUE achieves the best performance in all twelve settings, outperforming the strongest competing baseline by 6.19 percentage points on average and by up to 11.67 points. Notably, this advantage holds not only over FedAvg and FedProx, but also over robust federated methods designed to mitigate client heterogeneity. This suggests that improving model aggregation alone is insufficient in federated complementary-label learning; explicitly sharing and consolidating fragmented complementary evidence provides an additional source of supervision that existing FL methods do not exploit.
- *Finding 2:* FEDCUE *shows particularly large gains when complementary supervision is less informative.* The advantage is particularly pronounced on CIFAR-20, where FEDCUE

Table 1: Comparison with federated baselines under different levels of data heterogeneity. Results denote final test accuracy (%) on synthetic and human-annotated complementary-label settings. Best and second-best results are highlighted in red bold and blue bold, respectively.

Method	CIFAR-10			CIFAR-20			SVHN			CL-CIFAR10		
	β=0.1	β=0.5	β=0.8	β=0.1	β=0.5	β=0.8	β=0.1	β=0.5	β=0.8	β=0.1	β=0.5	β=0.8
Standard FL Methods												
FedAvg	**43.15**	**57.75**	58.27	**18.25**	13.20	13.77	**80.27**	91.09	91.31	**27.59**	35.12	**36.98**
FedProx	42.04	56.39	**58.62**	17.83	**14.84**	13.47	78.38	**91.45**	91.24	26.92	**35.79**	36.50
Robust FL Methods												
FedRDN	40.87	56.16	56.58	15.60	10.52	12.20	76.26	90.36	90.23	27.40	34.38	35.03
FedCDA	38.78	55.86	58.04	17.11	12.58	**13.93**	74.20	90.06	90.56	25.37	35.35	36.92
FedGLOSS-S	42.50	57.03	58.54	18.00	13.63	13.60	78.71	90.91	**91.45**	27.24	35.61	36.04
Proposed Method												
FedCUE	**46.51**	**69.42**	**67.17**	**24.53**	**24.03**	**23.92**	**85.40**	**92.91**	**92.69**	**32.24**	**42.77**	**42.72**

improves the strongest baseline by 6.28–9.99 percentage points across the three heterogeneity levels. With more candidate classes, each complementary label rules out only a smaller fraction of the label space and therefore provides weaker supervision individually. The larger gains in this setting support the main motivation of FEDCUE: aggregating compatible exclusion evidence within and across clients is especially useful when instance-level complementary labels are highly fragmented.

- *Finding 3: The gains persist under realistic complementary annotations and throughout federated training.* On the human-annotated CL-CIFAR10 benchmark, FEDCUE consistently outperforms the strongest baseline by 4.65–6.98 percentage points across all three heterogeneity levels. This shows that the improvement is not tied to the uniform synthetic complementary-label generation process used in the standard benchmarks. Moreover, Fig. 3 shows that the advantage is maintained throughout training rather than emerging only at the final rounds, indicating that the constructed evidence provides useful auxiliary supervision throughout federated optimization.

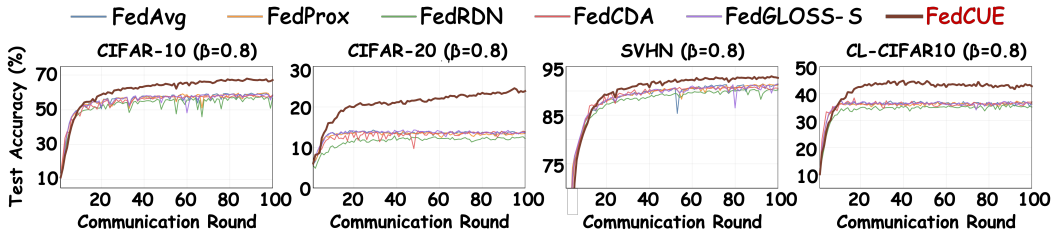

Figure 3: Convergence comparison of FEDCUE and federated baselines across four datasets under $\beta = 0.8$ with 10 clients. Test accuracy is reported over 100 communication rounds.

6.3 ABLATION STUDY

Table 2: Progressive ablation study under moderate data heterogeneity ($\beta = 0.5$).

Config.	CIFAR-10	CIFAR-20	SVHN	CL-CIFAR20
A	23.79	8.45	35.03	6.72
$A + B$	57.75↑33.96	13.20↑4.75	91.09↑56.06	7.64↑0.92
$A + B + C$	68.81↑11.06	23.54↑10.34	**93.26**↑2.17	9.09↑1.45
$\mathbf{A + B + C + D}$	**69.42**↑0.61	**24.03**↑0.49	92.91↓0.35	**9.23**↑0.14

A denotes local complementary-label training, B federated aggregation, C local complementary evidence, and D global complementary evidence. Table 2 reports the moderate $\beta = 0.5$ setting; results under stronger heterogeneity ($\beta = 0.1$) are provided in Appendix E.3.

The ablation results show a consistent progression across the components of FEDCUE. Federated aggregation first substantially improves over purely local training, indicating that complementary supervision confined to a single client remains limited under heterogeneous data. Building local complementary evidence then consistently improves all four datasets, showing that aggregating exclusion information among semantically similar samples effectively reduces instance-level supervision sparsity. Incorporating global evidence further improves three of the four datasets, showing that cross-client complementary evidence can provide additional gains beyond local consolidation, although its benefit is more dataset-dependent. Overall, the results support the central design principle of FEDCUE: local evidence provides the more consistent improvement, while global evidence further exploits compatible complementary information available across clients to enrich the supervision beyond what individual clients can obtain.

6.4 SCALABILITY AND SENSITIVITY

Scalability Analysis. We evaluate FEDCUE under different federation scales and client participation ratios on CIFAR-20 with $\beta = 0.1$.

As shown in Fig. 4, FEDCUE consistently outperforms FedAvg across all settings. Under full participation, the gain remains 6.28–7.16 points as clients increase from 5 to 50, indicating that the

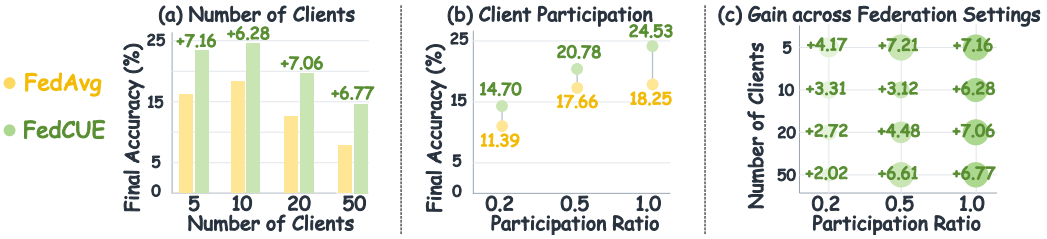

Figure 4: Scalability of FEDCUE under varying client numbers and participation ratios.

overall benefit of FEDCUE is retained as the federation grows. FEDCUE also remains effective under partial participation, despite fewer clients contributing in each round. Across all 12 configurations, the improvement ranges from 2.02 to 7.21 points, showing that the performance gains persist across the tested federation scales and participation ratios.

Parameter Sensitivity. We further examine three key factors on CIFAR-20: evidence weights, cluster granularity, and the unified radius. The results are shown in Fig. 5.

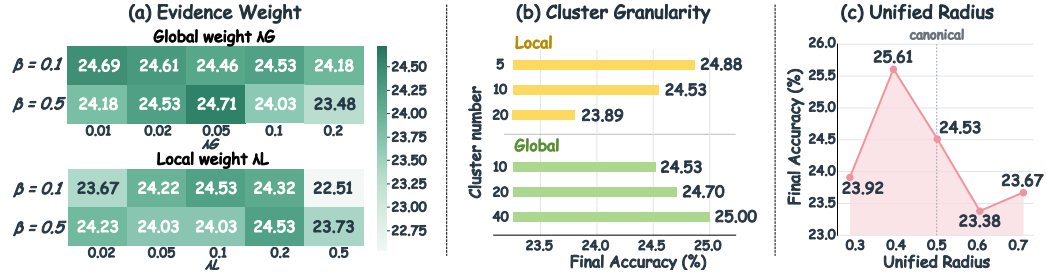

Figure 5: Parameter sensitivity of FEDCUE on CIFAR-20.

Effect of evidence weights. As shown in Fig. 5(a), FEDCUE remains stable over a broad range of λ_L and λ_G. Although the best values vary slightly across heterogeneity levels, the overall trends remain relatively flat. This indicates that the contributions of local and global evidence do not require precise balancing, making FEDCUE robust to the choice of evidence weights.

Effect of cluster granularity. Fig. 5(b) shows that cluster granularity affects local and global evidence differently. Finer local partitioning degrades performance because it fragments limited evidence within each client, while finer global clustering brings gains by reducing the mixing of semantically different cross-client evidence. This suggests that local clustering should preserve sufficient within-group evidence, while global clustering benefits from finer cross-client separation.

Effect of unified radius. As shown in Fig. 5(c), a moderate radius performs best, with $r = 0.4$ achieving an accuracy of 25.61%. A smaller radius rejects potentially useful evidence and reduces evidence coverage, whereas a larger radius admits less compatible matches and weakens evidence reliability. The result therefore reflects a trade-off between evidence coverage and matching quality.

7 CONCLUSION

We study federated complementary-label learning as a problem of fragmented complementary evidence, where sparse exclusion signals are scattered across samples and clients. FEDCUE progressively organizes these signals into local and cross-client evidence, enabling evidence-level collaboration beyond model aggregation without sharing raw data. We further establish a convergence guarantee for the resulting time-varying federated objective. Experiments on synthetic and human-annotated benchmarks under diverse Non-IID settings consistently support the effectiveness, robustness, and scalability of this multi-level design. The gains remain stable across federation scales, participation levels, and parameter choices, suggesting that they arise from better use of distributed complementary evidence rather than narrow tuning. Overall, our results show that federated learning can coordinate not only model parameters, but fragmented supervision itself.

AI USE STATEMENT

Generative AI tools were used solely to aid and polish the writing of this manuscript, including language editing and improvements to grammar, clarity, and presentation. All AI-assisted edits were reviewed by the authors, who take full responsibility for the final content of the paper.

ETHICS STATEMENT

This work uses existing benchmark datasets and does not involve new data collection or human-subject studies. FedCUE keeps raw samples decentralized, although its communication protocol alone does not provide a formal privacy guarantee.

REPRODUCIBILITY STATEMENT

We provide the implementation details, dataset construction procedures, hyperparameter settings, and additional experimental results in the appendix. The assumptions and complete proof of the convergence result are also provided in the appendix.

REFERENCES

Debora Caldarola, Pietro Cagnasso, Barbara Caputo, and Marco Ciccone. Beyond local sharpness: Communication-efficient global sharpness-aware minimization for federated learning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 25187–25197, 2025.

Timothee Cour, Ben Sapp, and Ben Taskar. Learning from partial labels. *The Journal of Machine Learning Research*, 12:1501–1536, 2011.

Lei Feng, Takuo Kaneko, Bo Han, Gang Niu, Bo An, and Masashi Sugiyama. Learning with multiple complementary labels. In *International conference on machine learning*, pp. 3072–3081. PMLR, 2020.

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.

Takashi Ishida, Gang Niu, Weihua Hu, and Masashi Sugiyama. Learning from complementary labels. *Advances in neural information processing systems*, 30, 2017.

Takashi Ishida, Gang Niu, Aditya Menon, and Masashi Sugiyama. Complementary-label learning for arbitrary losses and models. In *International conference on machine learning*, pp. 2971–2980. PMLR, 2019.

Haoran Jiang, Zhihao Sun, and Yingjie Tian. Comco: Complementary supervised contrastive learning for complementary label learning. *Neural Networks*, 169:44–56, 2024.

Peter Kairouz and H Brendan McMahan. Advances and open problems in federated learning. *Foundations and trends in machine learning*, 14(1-2):1–210, 2021.

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*, pp. 5132–5143. PMLR, 2020.

Zhiqiang Kou, Junxiang Wu, Wenke Huang, Wenwen He, Ming-Kun Xie, Changwei Wang, Yuheng Jia, Di Jiang, Yang Liu, Xin Geng, et al. Fedharmony: Harmonizing heterogeneous label correlations in federated multi-label learning. *arXiv preprint arXiv:2604.28024*, 2026.

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

Daliang Li and Junpu Wang. Fedmd: Heterogenous federated learning via model distillation. *arXiv preprint arXiv:1910.03581*, 2019.

Li Li, Yuxi Fan, Mike Tse, and Kuo-Yi Lin. A review of applications in federated learning. *Computers & Industrial Engineering*, 149:106854, 2020a.

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

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, 2020b.

Wei-I Lin, Gang Niu, Hsuan-Tien Lin, and Masashi Sugiyama. Enhancing label sharing efficiency in complementary-label learning with label augmentation. *arXiv preprint arXiv:2305.08344*, 2023.

Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. A survey on deep learning in medical image analysis. *Medical image analysis*, 42: 60–88, 2017.

Jiaqi Lv, Miao Xu, Lei Feng, Gang Niu, Xin Geng, and Masashi Sugiyama. Progressive identification of true labels for partial-label learning. In *international conference on machine learning*, pp. 6500–6510. PMLR, 2020.

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*, pp. 1273–1282. Pmlr, 2017.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Baolin Wu, Andrew Y Ng, et al. Reading digits in natural images with unsupervised feature learning. In *NIPS workshop on deep learning and unsupervised feature learning*, volume 2011, pp. 4. Granada, 2011.

Giorgio Patrini, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, and Lizhen Qu. Making deep neural networks robust to label noise: A loss correction approach. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1944–1952, 2017.

Nicola Rieke, Jonny Hancox, Wenqi Li, Fausto Milletari, Holger R Roth, Shadi Albarqouni, Spyridon Bakas, Mathieu N Galtier, Bennett A Landman, Klaus Maier-Hein, et al. The future of digital health with federated learning. *NPJ digital medicine*, 3(1):119, 2020.

Rahul Sharma, Anil Ramakrishna, Ansel MacLaughlin, Anna Rumshisky, Jimit Majmudar, Clement Chung, Salman Avestimehr, and Rahul Gupta. Federated learning with noisy user feedback. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2726–2739, 2022.

Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33: 596–608, 2020.

Yue Tan, Guodong Long, Lu Liu, Tianyi Zhou, Qinghua Lu, Jing Jiang, and Chengqi Zhang. Fedproto: Federated prototype learning across heterogeneous clients. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pp. 8432–8440, 2022.

Deng-Bao Wang, Lei Feng, and Min-Ling Zhang. Learning from complementary labels via partial-output consistency regularization. In *IJCAI*, pp. 3075–3081, 2021.

Haozhao Wang, Haoran Xu, Yichen Li, Yuan Xu, Ruixuan Li, and Tianwei Zhang. Fedcda: Federated learning with cross-rounds divergence-aware aggregation. In *International Conference on Learning Representations*, volume 2024, pp. 38124–38135, 2024.

Hsiu-Hsuan Wang, Tan-Ha Mai, Nai-Xuan Ye, Wei-I Lin, and Hsuan-Tien Lin. Climage: Human-annotated datasets for complementary-label learning. *arXiv preprint arXiv:2305.08295*, 2023.

Jie Wen, Zhixia Zhang, Yang Lan, Zhihua Cui, Jianghui Cai, and Wensheng Zhang. A survey on federated learning: challenges and applications. *International journal of machine learning and cybernetics*, 14(2):513–535, 2023.

Junxiang Wu, Zhiqiang Kou, Hongwei Zeng, Wenke Huang, Biao Liu, Hanlin Gu, Yuheng Jia, Di Jiang, Yang Liu, and Xin Geng. Trustworthy federated label distribution learning under annotation quality disparity. *arXiv preprint arXiv:2605.04827*, 2026.

Yan Yan and Yuhong Guo. Federated partial label learning with local-adaptive augmentation and regularization. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 16272–16280, 2024.

Yunlu Yan, Huazhu Fu, Yuexiang Li, Jinheng Xie, Jun Ma, Guang Yang, and Lei Zhu. A simple data augmentation for feature distribution skewed federated learning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 25749–25758, 2025.

Xiyu Yu, Tongliang Liu, Mingming Gong, and Dacheng Tao. Learning with biased complementary labels. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 68–83, 2018.

Zhilu Zhang and Mert Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. *Advances in neural information processing systems*, 31, 2018.

Yue Zhao, Meng Li, Liangzhen Lai, Naveen Suda, Damon Civin, and Vikas Chandra. Federated learning with non-iid data. *arXiv preprint arXiv:1806.00582*, 2018.

Zhi-Hua Zhou. A brief introduction to weakly supervised learning. *National science review*, 5(1): 44–53, 2018.

APPENDIX CONTENTS

A Additional Theoretical Analysis 14
A.1 Optimization Setting and Assumptions 14
A.2 Local-Update Drift 15
A.3 One-Round Descent 16
A.4 Proof of Theorem 1 18
B Complete Training Procedure 20
C Detailed Experimental Setup 20
C.1 Datasets and Complementary-Label Supervision 20
C.2 Federated Data Partitioning 21
C.3 Training Configuration and Hyperparameters 21
C.4 Baseline Implementations 22
C.5 Evaluation Protocol 22
D Additional Related Work 22
D.1 Federated Learning Methods 22
D.2 Weakly Supervised Learning 23
D.3 Complementary-Label Learning 23
E Additional Experimental Results 24
E.1 Complete Results across Data Heterogeneity 24
E.2 Additional Convergence Results 27
E.3 Detailed Ablation Results 29
F Further Discussion 29
F.1 Communication Cost 29
F.2 Privacy Discussion 30

A Additional Theoretical Analysis

A.1 Optimization Setting and Assumptions

We analyze a full-participation, fixed-local-step abstraction of FedCUE. All K clients participate in every communication round, and each client performs H stochastic-gradient updates per round. Let

$$p_k = \frac{n_k}{n}, \qquad n = \sum_{k=1}^{K} n_k, \tag{15}$$

so that

$$\sum_{k=1}^{K} p_k = 1. \tag{16}$$

At round t, the complementary targets and loss weights used by FedCUE are fixed throughout local optimization. Conditioned on the round-t evidence construction, the resulting local objective is

$$F_k^t(\Theta) = \mathcal{L}_{\text{obs}}^k(\Theta) + \lambda_L^t \mathcal{L}_{\text{loc}}^{k,t}(\Theta) + \lambda_G^t \mathcal{L}_{\text{glo}}^{k,t}(\Theta), \tag{17}$$

and the corresponding global objective is

$$F^t(\Theta) = \sum_{k=1}^{K} p_k F_k^t(\Theta). \tag{18}$$

During the warm-up stage, $\lambda_L^t = \lambda_G^t = 0$; the inactive auxiliary losses may therefore be taken as zero. After warm-up, the constructed complementary targets remain fixed within each local-optimization stage.

Client k starts round t from

$$\Theta_k^{t,0} = \Theta^t \tag{19}$$

and performs

$$\Theta_k^{t,h+1} = \Theta_k^{t,h} - \eta g_k^t(\Theta_k^{t,h}), \qquad h = 0, \dots, H-1. \tag{20}$$

The server then performs weighted aggregation:

$$\Theta^{t+1} = \sum_{k=1}^{K} p_k \Theta_k^{t,H}. \tag{21}$$

Let $\mathcal{F}_{t,h}$ denote the optimization history available immediately before local step h of round t, including the round-t complementary targets and loss weights.

Assumption 1 (Smoothness) *For every realized round-t objective and every client k, F_k^t is L-smooth on the region containing the optimization iterates:*

$$\left\|\nabla F_k^t(\Theta_1) - \nabla F_k^t(\Theta_2)\right\| \le L \left\|\Theta_1 - \Theta_2\right\|. \tag{22}$$

Assumption 2 (Unbiased stochastic gradients with bounded variance) *For every client and local step,*

$$\mathbb{E}\left[g_k^t(\Theta_k^{t,h}) \mid \mathcal{F}_{t,h}\right] = \nabla F_k^t(\Theta_k^{t,h}), \tag{23}$$

and

$$\mathbb{E}\left[\left\|g_k^t(\Theta_k^{t,h}) - \nabla F_k^t(\Theta_k^{t,h})\right\|^2 \mid \mathcal{F}_{t,h}\right] \le \sigma^2. \tag{24}$$

Assumption 3 (Bounded weighted gradient heterogeneity) *For every realized round-t objective and every Θ in the optimization region,*

$$\sum_{k=1}^{K} p_k \left\|\nabla F_k^t(\Theta) - \nabla F^t(\Theta)\right\|^2 \le \zeta^2. \tag{25}$$

The complementary-label loss and both evidence losses are non-negative, and $\lambda_L^t, \lambda_G^t \geq 0$. Therefore, the round-wise objectives admit a common finite lower bound, denoted by $F_{\inf}$. For the losses considered here, one may take $F_{\inf} = 0$.

For $t = 0, \ldots, T-2$, define the round-to-round objective variation as

$$v_t = \mathbb{E}\left[\left|F^{t+1}(\Theta^{t+1}) - F^t(\Theta^{t+1})\right|\right], \tag{26}$$

and define the cumulative variation as

$$V_T = \sum_{t=0}^{T-2} v_t, \tag{27}$$

with $V_1 = 0$. This quantity captures round-to-round changes induced by complementary-evidence reconstruction as well as the transition of the auxiliary-loss weights at the end of warm-up.

A.2 Local-Update Drift

We first control the deviation of local iterates from the round-t initialization.

Lemma 1 (Local-update drift) *Under Assumptions 1–3, if $\eta LH \leq 1/4$, then for every $0 \leq h \leq H$,*

$$\sum_{k=1}^{K} p_k \mathbb{E}\left[\left\|\Theta_k^{t,h} - \Theta^t\right\|^2\right] \leq 4\eta^2 h^2 \left(G_t + \sigma^2 + \zeta^2\right), \tag{28}$$

where

$$G_t = \mathbb{E}\left[\left\|\nabla F^t(\Theta^t)\right\|^2\right]. \tag{29}$$

Proof.

Define

$$D_{t,h} = \sum_{k=1}^{K} p_k \mathbb{E}\left[\left\|\Theta_k^{t,h} - \Theta^t\right\|^2\right]. \tag{30}$$

From Eq. (20),

$$\Theta_k^{t,h} - \Theta^t = -\eta \sum_{s=0}^{h-1} g_k^t(\Theta_k^{t,s}). \tag{31}$$

Using

$$\left\|\sum_{s=0}^{h-1} \mathbf{a}_s\right\|^2 \leq h \sum_{s=0}^{h-1} \|\mathbf{a}_s\|^2, \tag{32}$$

we obtain

$$D_{t,h} \leq \eta^2 h \sum_{s=0}^{h-1} \sum_{k=1}^{K} p_k \mathbb{E}\left[\left\|g_k^t(\Theta_k^{t,s})\right\|^2\right]. \tag{33}$$

By Assumption 2,

$$\mathbb{E}\left[\left\|g_k^t(\Theta_k^{t,s})\right\|^2\right] \leq \mathbb{E}\left[\left\|\nabla F_k^t(\Theta_k^{t,s})\right\|^2\right] + \sigma^2. \tag{34}$$

By L-smoothness,

$$\left\|\nabla F_k^t(\Theta_k^{t,s})\right\|^2 \leq 2\left\|\nabla F_k^t(\Theta^t)\right\|^2 + 2L^2\left\|\Theta_k^{t,s} - \Theta^t\right\|^2. \tag{35}$$

Since

$$\nabla F^t(\Theta^t) = \sum_{k=1}^{K} p_k \nabla F_k^t(\Theta^t), \tag{36}$$

the weighted variance decomposition gives

$$\begin{aligned}\sum_{k=1}^{K} p_k \left\|\nabla F_k^t(\Theta^t)\right\|^2 &= \left\|\nabla F^t(\Theta^t)\right\|^2 \\ &\quad + \sum_{k=1}^{K} p_k \left\|\nabla F_k^t(\Theta^t) - \nabla F^t(\Theta^t)\right\|^2 \\ &\le \left\|\nabla F^t(\Theta^t)\right\|^2 + \zeta^2.\end{aligned} \tag{37}$$

Taking expectation and combining Eqs. (33)–(37) yields

$$D_{t,h} \le 2\eta^2 h^2 \left(G_t + \sigma^2 + \zeta^2\right) + 2\eta^2 L^2 h \sum_{s=0}^{h-1} D_{t,s}. \tag{38}$$

We prove Eq. (28) by induction. The case $h = 0$ is immediate. Suppose that, for every $s < h$,

$$D_{t,s} \le 4\eta^2 s^2 \left(G_t + \sigma^2 + \zeta^2\right). \tag{39}$$

Let

$$A_t = G_t + \sigma^2 + \zeta^2. \tag{40}$$

Using

$$\sum_{s=0}^{h-1} s^2 \le \frac{h^3}{3} \tag{41}$$

in Eq. (38),

$$\begin{aligned}D_{t,h} &\le 2\eta^2 h^2 A_t + \frac{8}{3}\eta^4 L^2 h^4 A_t \\ &= \left[2 + \frac{8}{3}(\eta L h)^2\right] \eta^2 h^2 A_t.\end{aligned} \tag{42}$$

Since $\eta L h \le \eta L H \le 1/4$,

$$2 + \frac{8}{3}(\eta L h)^2 \le 2 + \frac{1}{6} < 4. \tag{43}$$

Therefore, Eq. (28) holds for h, completing the induction.

□

A.3 One-Round Descent

We next establish descent of the fixed round-t objective.

Lemma 2 (One-round descent) *Under Assumptions 1–3, if $\eta L H \le 1/8$, then*

$$\begin{aligned}\mathbb{E}\left[F^t(\Theta^{t+1})\right] &\le \mathbb{E}\left[F^t(\Theta^t)\right] - \frac{\eta H}{4} G_t \\ &\quad + L\eta^2 H \sigma^2 + \frac{5}{3} L^2 \eta^3 H^3 (\sigma^2 + \zeta^2).\end{aligned} \tag{44}$$

Proof.

Define

$$\mathbf{a}_t = \nabla F^t(\Theta^t), \tag{45}$$

the weighted stochastic gradient

$$\bar{\mathbf{g}}^{t,h} = \sum_{k=1}^{K} p_k g_k^t(\Theta_k^{t,h}), \tag{46}$$

the corresponding weighted exact local gradient

$$\bar{\nabla}^{t,h} = \sum_{k=1}^{K} p_k \nabla F_k^t(\Theta_k^{t,h}), \tag{47}$$

and the aggregated stochastic-gradient noise

$$\boldsymbol{\xi}^{t,h} = \bar{\mathbf{g}}^{t,h} - \bar{\nabla}^{t,h}. \tag{48}$$

By Assumption 2,

$$\mathbb{E}\left[\boldsymbol{\xi}^{t,h} \mid \mathcal{F}_{t,h}\right] = 0. \tag{49}$$

Moreover, by convexity of the squared norm,

$$\begin{aligned}\mathbb{E}\left[\left\|\boldsymbol{\xi}^{t,h}\right\|^2 \mid \mathcal{F}_{t,h}\right] &\le \sum_{k=1}^{K} p_k \mathbb{E}\left[\left\|g_k^t(\Theta_k^{t,h}) - \nabla F_k^t(\Theta_k^{t,h})\right\|^2 \mid \mathcal{F}_{t,h}\right] \\ &\le \sigma^2.\end{aligned} \tag{50}$$

Since $\{\boldsymbol{\xi}^{t,h}\}_{h=0}^{H-1}$ is a martingale-difference sequence with respect to the local-step filtration,

$$\mathbb{E}\left[\left\|\sum_{h=0}^{H-1} \boldsymbol{\xi}^{t,h}\right\|^2\right] \le H\sigma^2. \tag{51}$$

From Eqs. (20) and (21),

$$\Theta^{t+1} = \Theta^t - \eta \sum_{h=0}^{H-1} \left(\bar{\nabla}^{t,h} + \boldsymbol{\xi}^{t,h}\right). \tag{52}$$

Define the local-gradient drift

$$\mathbf{d}^{t,h} = \bar{\nabla}^{t,h} - \mathbf{a}_t. \tag{53}$$

By Jensen's inequality and Assumption 1,

$$\begin{aligned}\mathbb{E}\left[\left\|\mathbf{d}^{t,h}\right\|^2\right] &\le L^2 \sum_{k=1}^{K} p_k \mathbb{E}\left[\left\|\Theta_k^{t,h} - \Theta^t\right\|^2\right] \\ &= L^2 D_{t,h}.\end{aligned} \tag{54}$$

Hence, by Lemma 1,

$$\begin{aligned}S_t &:= \sum_{h=0}^{H-1} \mathbb{E}\left[\left\|\mathbf{d}^{t,h}\right\|^2\right] \\ &\le 4L^2\eta^2\left(G_t + \sigma^2 + \zeta^2\right) \sum_{h=0}^{H-1} h^2 \\ &\le \frac{4}{3} L^2\eta^2 H^3\left(G_t + \sigma^2 + \zeta^2\right).\end{aligned} \tag{55}$$

Since F^t is a weighted sum of L-smooth objectives, it is also L-smooth. Applying smoothness to Eq. (52) and taking expectation, the linear noise term vanishes by Eq. (49). Using

$$\|\mathbf{x} + \mathbf{y}\|^2 \le 2\|\mathbf{x}\|^2 + 2\|\mathbf{y}\|^2 \tag{56}$$

together with Eq. (51), we obtain

$$\begin{aligned}\mathbb{E}\left[F^t(\Theta^{t+1})\right] \le \mathbb{E}\left[F^t(\Theta^t)\right] &- \eta \sum_{h=0}^{H-1} \mathbb{E}\left[\left\langle \mathbf{a}_t, \bar{\nabla}^{t,h}\right\rangle\right] \\ &+ L\eta^2 \mathbb{E}\left[\left\|\sum_{h=0}^{H-1} \bar{\nabla}^{t,h}\right\|^2\right] + L\eta^2 H\sigma^2.\end{aligned} \tag{57}$$

For the inner-product term,

$$\begin{aligned}\sum_{h=0}^{H-1}\mathbb{E}\left[\langle \mathbf{a}_t, \bar{\boldsymbol{\nabla}}^{t,h}\rangle\right] &= HG_t + \sum_{h=0}^{H-1}\mathbb{E}\left[\langle \mathbf{a}_t, \mathbf{d}^{t,h}\rangle\right] \\ &\geq \frac{3H}{4}G_t - S_t,\end{aligned} \tag{58}$$

where we used

$$\langle \mathbf{a}, \mathbf{d}\rangle \geq -\frac{1}{4}\|\mathbf{a}\|^2 - \|\mathbf{d}\|^2. \tag{59}$$

For the second-order term,

$$\begin{aligned}\mathbb{E}\left[\left\|\sum_{h=0}^{H-1}\bar{\boldsymbol{\nabla}}^{t,h}\right\|^2\right] &= \mathbb{E}\left[\left\|H\mathbf{a}_t + \sum_{h=0}^{H-1}\mathbf{d}^{t,h}\right\|^2\right] \\ &\leq 2H^2G_t + 2HS_t.\end{aligned} \tag{60}$$

Substituting Eqs. (58) and (60) into Eq. (57) gives

$$\begin{aligned}\mathbb{E}\left[F^t(\Theta^{t+1})\right] \leq \mathbb{E}\left[F^t(\Theta^t)\right] &- \left(\frac{3}{4}\eta H - 2L\eta^2H^2\right)G_t \\ &+ \left(\eta + 2L\eta^2H\right)S_t + L\eta^2H\sigma^2.\end{aligned} \tag{61}$$

Since $\eta LH \leq 1/8$,

$$\frac{3}{4}\eta H - 2L\eta^2H^2 \geq \frac{1}{2}\eta H, \tag{62}$$

and

$$\eta + 2L\eta^2H \leq \frac{5}{4}\eta. \tag{63}$$

Using Eq. (55),

$$\frac{5}{4}\eta S_t \leq \frac{5}{3}L^2\eta^3H^3\left(G_t + \sigma^2 + \zeta^2\right). \tag{64}$$

Furthermore,

$$\begin{aligned}\frac{5}{3}L^2\eta^3H^3G_t &= \frac{5}{3}(\eta LH)^2\eta HG_t \\ &\leq \frac{5}{192}\eta HG_t < \frac{1}{4}\eta HG_t.\end{aligned} \tag{65}$$

Absorbing this term into the descent coefficient yields Eq. (44).

□

A.4 Proof of Theorem 1

For every $t = 0, \ldots, T-2$, the definition of v_t gives

$$\mathbb{E}\left[F^{t+1}(\Theta^{t+1})\right] \leq \mathbb{E}\left[F^t(\Theta^{t+1})\right] + v_t. \tag{66}$$

Combining Eq. (66) with Lemma 2 yields

$$\begin{aligned}\frac{\eta H}{4}G_t \leq \mathbb{E}\left[F^t(\Theta^t) - F^{t+1}(\Theta^{t+1})\right] &+ v_t \\ &+ L\eta^2H\sigma^2 + \frac{5}{3}L^2\eta^3H^3(\sigma^2 + \zeta^2).\end{aligned} \tag{67}$$

For the final round $t = T-1$, no subsequent objective is needed. Lemma 2 directly gives

$$\begin{aligned}\frac{\eta H}{4}G_{T-1} \leq \mathbb{E}\left[F^{T-1}(\Theta^{T-1}) - F^{T-1}(\Theta^T)\right] &\\ + L\eta^2H\sigma^2 + \frac{5}{3}L^2\eta^3H^3(\sigma^2 + \zeta^2).&\end{aligned} \tag{68}$$

Summing Eq. (67) over $t = 0, \dots, T-2$ and adding Eq. (68), the objective terms telescope:

$$\frac{\eta H}{4} \sum_{t=0}^{T-1} G_t \le \mathbb{E}\left[F^0(\Theta^0) - F^{T-1}(\Theta^T)\right] + V_T + TL\eta^2 H\sigma^2 + \frac{5}{3} TL^2\eta^3 H^3(\sigma^2 + \zeta^2). \tag{69}$$

Since

$$F^{T-1}(\Theta^T) \ge F_{\inf}, \tag{70}$$

we obtain

$$\frac{\eta H}{4} \sum_{t=0}^{T-1} G_t \le \Delta_0 + V_T + TL\eta^2 H\sigma^2 + \frac{5}{3} TL^2\eta^3 H^3(\sigma^2 + \zeta^2), \tag{71}$$

where

$$\Delta_0 = \mathbb{E}\left[F^0(\Theta^0)\right] - F_{\inf}. \tag{72}$$

Dividing Eq. (71) by $\eta HT/4$ gives

$$\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\left[\left\|\nabla F^t(\Theta^t)\right\|^2\right] \le \frac{4(\Delta_0 + V_T)}{\eta HT} + 4\eta L\sigma^2 + \frac{20}{3}\eta^2 L^2 H^2(\sigma^2 + \zeta^2), \tag{73}$$

which proves Theorem 1.

Scope of the analysis. The theorem considers full client participation and a common local-step budget H. Full participation matches the setting used for the main experiments. The fixed-H local-SGD formulation is a step-level abstraction of the epoch-based local training used in implementation. Partial participation or client-dependent local-step counts introduce additional sampling or local-work heterogeneity and require corresponding extensions of the analysis.

B Complete Training Procedure

The complete training procedure of FedCUE is summarized in Algorithm 1. It consists of a warm-up stage followed by multi-level complementary-evidence learning. In each post-warm-up round, clients first construct local evidence using the frozen global encoder, the server aggregates cross-client evidence, and clients then perform local optimization with the constructed complementary targets fixed.

Algorithm 1 Training Procedure of FedCUE

Require: Client datasets $\{\mathcal{D}_k\}_{k=1}^{K}$; communication rounds T; warm-up rounds T_w; local epochs E; weights λ_L, λ_G; initial global model Θ^0
Ensure: Final global model Θ^T

Stage 1: Warm-up with observed complementary labels
for $t = 0, \ldots, T_w - 1$ **do**
  Server selects participating clients $\mathcal{K}_t$
  for all $k \in \mathcal{K}_t$ **in parallel do**
    $\Theta_k^{t+1} \leftarrow \textsc{LocalTrain}(\Theta^t, \mathcal{D}_k, \mathcal{L}_{\text{obs}}^k, E)$
  $\Theta^{t+1} \leftarrow \sum_{k \in \mathcal{K}_t} \frac{n_k}{\sum_{k' \in \mathcal{K}_t} n_{k'}} \Theta_k^{t+1}$
Stage 2: Multi-level complementary-evidence learning
for $t = T_w, \ldots, T - 1$ **do**
  Server selects participating clients $\mathcal{K}_t$
  Server broadcasts Θ^t
  Client-side local evidence construction
  for all $k \in \mathcal{K}_t$ **in parallel do**
    $\mathcal{S}_k^t \leftarrow \textsc{ConstructLocalEvidence}(h_{\theta^t}, \mathcal{D}_k)$
  Server-side cross-client evidence aggregation
  $\{\mathcal{B}_k^t\}_{k \in \mathcal{K}_t} \leftarrow \textsc{ConstructGlobalEvidence}(\{\mathcal{S}_k^t\}_{k \in \mathcal{K}_t})$
  Client-side model optimization
  for all $k \in \mathcal{K}_t$ **in parallel do**
    $\Theta_k^{t+1} \leftarrow \textsc{LocalTrain}(\Theta^t, \mathcal{D}_k, \mathcal{L}_{\text{obs}}^k + \lambda_L \mathcal{L}_{\text{loc}}^{k,t} + \lambda_G \mathcal{L}_{\text{glo}}^{k,t}, E)$
  Server-side model aggregation
  $\Theta^{t+1} \leftarrow \sum_{k \in \mathcal{K}_t} \frac{n_k}{\sum_{k' \in \mathcal{K}_t} n_{k'}} \Theta_k^{t+1}$
return Θ^T

C Detailed Experimental Setup

This section provides additional details on the datasets, federated data partitioning, training configurations, baseline settings, and evaluation protocols used in our experiments.

C.1 Datasets and Complementary-Label Supervision

Our experiments cover CIFAR-10, CIFAR-20, SVHN, MNIST, CL-CIFAR10, and CL-CIFAR20. CIFAR-10, CIFAR-20, SVHN, and MNIST are evaluated with synthetically generated complementary labels, whereas CL-CIFAR10 and CL-CIFAR20 provide human-annotated complementary labels. Table 3 summarizes the basic statistics and complementary-label supervision of all datasets.

For the synthetic complementary-label datasets, we generate one complementary label for each training sample by uniformly sampling from all classes other than its ground-truth class. Formally, for a sample with ground-truth label $y_i \in \mathcal{Y}$, the complementary label $\bar{y}_i$ is sampled as

$$\bar{y}_i \sim \text{Uniform}\big(\mathcal{Y} \setminus \{y_i\}\big). \tag{74}$$

The ground-truth labels are retained only for federated data partitioning and test evaluation, while complementary labels are used as the training supervision.

Table 3: Summary of the datasets used in our experiments.

Dataset	#Classes	Image Size	Train	Test	#CLs	CL Type
CIFAR-10	10	$3 \times 32 \times 32$	50,000	10,000	1	Synthetic
CIFAR-20	20	$3 \times 32 \times 32$	50,000	10,000	1	Synthetic
SVHN	10	$3 \times 32 \times 32$	73,257	26,032	1	Synthetic
MNIST	10	$1 \times 28 \times 28$	60,000	10,000	1	Synthetic
CL-CIFAR10	10	$3 \times 32 \times 32$	50,000	10,000	3	Human
CL-CIFAR20	20	$3 \times 32 \times 32$	50,000	10,000	3	Human

For CL-CIFAR10 and CL-CIFAR20, each training image is associated with three human-annotated complementary labels. We preserve repeated annotations when constructing the training target. Specifically, let $m_{i,c}$ denote the number of annotations indicating that sample i does not belong to class c. We construct the normalized complementary-label distribution as

$$\bar{q}_{i,c} = \frac{m_{i,c}}{\sum_{c'=1}^{C} m_{i,c'}}. \tag{75}$$

Thus, repeated annotations assign larger complementary probability to the corresponding class rather than being collapsed into a single binary indicator. We use the original human annotations without additional filtering and thus retain any annotation noise present in the released datasets.

C.2 Federated Data Partitioning

Unless otherwise specified, we use $K = 10$ clients. To simulate Non-IID data, training samples are partitioned across clients according to their ground-truth classes using a Dirichlet distribution with concentration parameter β. Smaller β indicates stronger client-level class heterogeneity. We consider

$$\beta \in \{0.1, 0.3, 0.5, 0.8\}. \tag{76}$$

All clients participate in every communication round unless otherwise stated. For the scalability analysis, we further vary the number of clients as

$$K \in \{5, 10, 20, 50\}, \tag{77}$$

and the client participation ratio as

$$\rho \in \{0.2, 0.5, 1.0\}. \tag{78}$$

C.3 Training Configuration and Hyperparameters

Unless otherwise specified, all main experiments are conducted for 100 communication rounds with 5 local epochs per round. We use a batch size of 64 and a learning rate of 0.01. All reported main results use a CNN backbone.

For FedCUE, the local and global complementary-evidence losses are weighted by

$$\lambda_L = 0.1, \qquad \lambda_G = 0.1. \tag{79}$$

Evidence-based training is activated after a warm-up period of 20 communication rounds.

The default configuration uses 10 local clusters and 10 global clusters. The three evidence radii are set to

$$r_L = r_G = r_R = 0.5, \tag{80}$$

where r_L, r_G, and r_R denote the local evidence radius, server-side global matching radius, and client-side retrieval radius, respectively. The corresponding temperature parameters are set to

$$\tau_L = \tau_G = \tau_R = 0.5. \tag{81}$$

Table 4 summarizes the default experimental configuration.

Table 4: Default experimental configuration used for the main FEDCUE experiments.

Hyperparameter	Value
Number of clients K	10
Communication rounds	100
Local epochs	5
Batch size	64
Learning rate	0.01
Client participation	1.0
Random seed	2026
Backbone	CNN
Warm-up rounds	20
Local evidence weight λ_L	0.1
Global evidence weight λ_G	0.1
Number of local clusters	10
Number of global clusters	10
Local evidence radius r_L	0.5
Global matching radius r_G	0.5
Retrieval radius r_R	0.5
Local kernel temperature τ_L	0.5
Global matching temperature τ_G	0.5
Retrieval temperature τ_R	0.5

C.4 BASELINE IMPLEMENTATIONS

We compare FEDCUE with two standard federated learning methods, FedAvg and FedProx, and three robust federated learning methods, FedRDN, FedCDA, and FedGLOSS-S. All compared methods use the same CNN backbone and are evaluated under the same client data partitions and the same number of communication rounds.

To adapt the baselines to complementary-label supervision, we replace the standard supervised classification objective with the observed complementary-label loss while retaining each method's original federated optimization mechanism. FedAvg performs standard weighted model aggregation, whereas FedProx additionally applies proximal regularization during local training. FedRDN and FedCDA retain their respective method-specific components. For FedGLOSS-S, we use the server-side adaptation described in the main text with standard SGD clients.

C.5 EVALUATION PROTOCOL

We report the test accuracy of the global model after the final communication round rather than selecting the best intermediate checkpoint. Unless otherwise specified, all reported results correspond to the model obtained after 100 communication rounds. Ground-truth class labels are used only for test evaluation and federated data partitioning, and are never used for local training.

D ADDITIONAL RELATED WORK

D.1 FEDERATED LEARNING METHODS

Federated learning (FL) (McMahan et al., 2017) aims to learn a shared model across multiple clients while keeping local data decentralized. FedAvg (McMahan et al., 2017) establishes the standard federated optimization paradigm through local client updates followed by weighted server-side model aggregation.

Client heterogeneity remains a major challenge in FL. FedProx (Li et al., 2020b) introduces proximal regularization to reduce local-global model divergence. MOON (Li et al., 2021) further improves federated optimization through model-level contrastive learning that aligns local representations with global knowledge.

Beyond model aggregation, recent studies explore exchanging higher-level knowledge among clients. FedProto (Tan et al., 2022) communicates class prototypes to improve representation consistency across clients, while FedMD (Li & Wang, 2019) transfers knowledge through predictions on a shared public dataset. Federated learning has also been studied with weak user feedback. Sharma et al. (2022) combine positive and negative feedback, treating rejected model predictions as complementary labels during federated training. Their setting differs from ours by relying on both feedback types and a centrally trained seed model.

Despite these advances, existing federated methods are not specifically designed to organize complementary evidence within and across heterogeneous clients. FedCUE instead focuses on complementary-only supervision and explicitly consolidates fragmented exclusion information at both local and cross-client levels.

D.2 Weakly Supervised Learning

Weakly supervised learning trains predictive models from incomplete, inexact, or noisy supervision, reducing reliance on expensive manual annotations (Zhou, 2018). Existing studies can be broadly categorized into noisy-label learning, partial-label learning, semi-supervised learning, and complementary-label learning.

Noisy-label learning considers settings where observed labels may be incorrect and improves robustness through sample selection (Han et al., 2018), loss correction (Patrini et al., 2017), or robust regularization (Zhang & Sabuncu, 2018). Partial-label learning associates each sample with a candidate label set containing the unknown ground-truth class and aims to resolve this ambiguity during training (Cour et al., 2011; Lv et al., 2020). Recent work further extends this setting to federated partial-label learning, including FedPLL-LAAR (Yan & Guo, 2024). Semi-supervised learning instead combines limited labeled data with abundant unlabeled data, commonly using consistency regularization and pseudo-labeling (Sohn et al., 2020).

Complementary-label learning differs from these settings by providing class-exclusion supervision, where annotations specify classes that a sample does not belong to rather than candidate classes that may contain the true label. This distinction motivates methods that directly exploit exclusion information. We further discuss such methods and their relation to federated learning below.

D.3 Complementary-Label Learning

Complementary-label learning (CLL) (Ishida et al., 2017) studies weakly supervised classification where each training sample is associated with a complementary label indicating a class it does not belong to. Compared with conventional supervision, complementary labels can be easier to obtain because annotators may identify implausible categories without determining the exact class. CLL has therefore attracted attention as an alternative when precise annotation is expensive.

Early studies focus on theoretically unbiased learning objectives from complementary labels. An unbiased risk estimator is introduced for complementary-label learning (Ishida et al., 2017), while subsequent work investigates risk-estimation variance and non-negative risk formulations for improved training stability (Ishida et al., 2019). Several studies extend CLL to multiple complementary labels (Feng et al., 2020), while more recent approaches exploit semantic relationships among samples through consistency or contrastive representation learning (Wang et al., 2021; Jiang et al., 2024). Lin et al. (2023) further propose complementary-label augmentation, which explicitly propagates complementary labels among nearby instances to improve label sharing in centralized CLL.

This local propagation idea is related to FedCUE's local evidence construction, but the learning settings and collaboration mechanisms differ. The centralized CLL methods discussed above assume that training samples and complementary labels are jointly accessible, whereas FedCUE considers evidence distributed across heterogeneous clients. FedCUE organizes exclusion information locally and further aggregates compatible cluster-level evidence across clients without sharing raw data, extending evidence organization from within-client samples to cross-client collaboration.

E Additional Experimental Results

E.1 Complete Results across Data Heterogeneity

To complement the main results, we provide complete comparisons under four levels of data heterogeneity, $\beta \in \{0.1, 0.3, 0.5, 0.8\}$. In addition to the settings reported in the main text, we include the intermediate heterogeneity level $\beta = 0.3$ and additional results on MNIST and CL-CIFAR20. Tables 5– 10 report the final test accuracy of all compared methods on the six datasets.

Table 5: CIFAR-10 test accuracy (%) under different levels of data heterogeneity. Best and second-best results in each column are highlighted in red bold and blue bold, respectively.

Method	CIFAR-10			
	β=0.1	β=0.3	β=0.5	β=0.8
Standard FL Methods				
FedAvg	**43.15**	**55.38**	**57.75**	58.27
FedProx	42.04	54.87	56.39	**58.62**
Robust FL Methods				
FedRDN	40.87	54.88	56.16	56.58
FedCDA	38.78	44.80	55.86	58.04
FedGLOSS-S	42.50	54.92	57.03	58.54
Proposed Method				
FedCUE	**46.51**	**62.39**	**69.42**	**67.17**

CIFAR-10. As shown in Table 5, FedCUE achieves the highest accuracy across all four heterogeneity levels. The improvement over the strongest baseline ranges from 3.36 to 11.67 percentage points, with particularly large gains at $\beta = 0.3$ and $\beta = 0.5$.

Table 6: CIFAR-20 test accuracy (%) under different levels of data heterogeneity. Best and second-best results in each column are highlighted in red bold and blue bold, respectively.

Method	CIFAR-20			
	β=0.1	β=0.3	β=0.5	β=0.8
Standard FL Methods				
FedAvg	**18.25**	**16.69**	13.20	13.77
FedProx	17.83	15.76	**14.84**	13.47
Robust FL Methods				
FedRDN	15.60	13.59	10.52	12.20
FedCDA	17.11	15.18	12.58	**13.93**
FedGLOSS-S	18.00	15.38	13.63	13.60
Proposed Method				
FedCUE	**24.53**	**24.46**	**24.03**	**23.92**

CIFAR-20. Table 6 shows that FEDCUE consistently outperforms all competing methods across the four heterogeneity settings, with gains of 6.28–9.99 percentage points over the strongest baseline. The larger margins are consistent with the weaker supervision provided by each complementary label when the number of classes increases.

Table 7: SVHN test accuracy (%) under different levels of data heterogeneity. Best and second-best results in each column are highlighted in red bold and blue bold, respectively.

Method	SVHN			
	β=0.1	β=0.3	β=0.5	β=0.8
Standard FL Methods				
FedAvg	**80.27**	89.36	91.09	91.31
FedProx	78.38	90.08	**91.45**	91.24
Robust FL Methods				
FedRDN	76.26	89.37	90.36	90.23
FedCDA	74.20	88.02	90.06	90.56
FedGLOSS-S	78.71	**90.17**	90.91	**91.45**
Proposed Method				
FedCUE	**85.40**	**91.08**	**92.91**	**92.69**

SVHN. As reported in Table 7, FEDCUE achieves the highest accuracy under all four heterogeneity levels. The improvement is particularly pronounced under strong heterogeneity ($\beta = 0.1$), while remaining positive across the other settings.

Table 8: MNIST test accuracy (%) under different levels of data heterogeneity. Best and second-best results in each column are highlighted in red bold and blue bold, respectively.

Method	MNIST			
	β=0.1	β=0.3	β=0.5	β=0.8
Standard FL Methods				
FedAvg	81.75	99.17	99.03	99.26
FedProx	78.79	**99.25**	99.07	99.25
Robust FL Methods				
FedRDN	81.01	99.18	**99.11**	**99.29**
FedCDA	**96.21**	98.99	98.89	98.88
FedGLOSS-S	82.67	99.06	98.99	98.68
Proposed Method				
FedCUE	**86.14**	**99.38**	**99.44**	**99.51**

MNIST. Table 8 shows that FEDCUE achieves the best performance for $\beta \geq 0.3$, where all methods reach high accuracy. Under the strongest heterogeneity setting ($\beta = 0.1$), FedCDA performs better than FEDCUE.

Table 9: CL-CIFAR10 test accuracy (%) under different levels of data heterogeneity. Best and second-best results in each column are highlighted in red bold and blue bold, respectively.

Method	CL-CIFAR10			
	β=0.1	β=0.3	β=0.5	β=0.8
Standard FL Methods				
FedAvg	**27.59**	**35.17**	35.12	**36.98**
FedProx	26.92	34.47	**35.79**	36.50
Robust FL Methods				
FedRDN	27.40	34.62	34.38	35.03
FedCDA	25.37	31.78	35.35	36.92
FedGLOSS-S	27.24	34.99	35.61	36.04
Proposed Method				
FedCUE	**32.24**	**40.23**	**42.77**	**42.72**

CL-CIFAR10. On the human-annotated CL-CIFAR10 benchmark, Table 9 shows that FEDCUE consistently achieves the highest accuracy across all four heterogeneity levels, outperforming the strongest baseline by 4.65–6.98 percentage points. This confirms that the improvement extends beyond synthetically generated complementary labels.

Table 10: CL-CIFAR20 test accuracy (%) under different levels of data heterogeneity. Best and second-best results in each column are highlighted in red bold and blue bold, respectively.

Method	CL-CIFAR20			
	β=0.1	β=0.3	β=0.5	β=0.8
Standard FL Methods				
FedAvg	7.59	7.40	7.64	7.48
FedProx	7.48	7.34	7.27	7.25
Robust FL Methods				
FedRDN	**7.86**	7.85	7.26	7.66
FedCDA	7.61	**7.97**	7.72	**7.87**
FedGLOSS-S	7.65	7.64	**7.83**	7.75
Proposed Method				
FedCUE	**9.12**	**9.09**	**9.23**	**9.28**

CL-CIFAR20. Similarly, Table 10 shows that FEDCUE consistently outperforms all baselines on CL-CIFAR20, with gains of 1.12–1.41 percentage points over the strongest competing method. The improvement across all four β values further supports its effectiveness under human complementary supervision.

E.2 Additional Convergence Results

To complement the convergence results reported in the main text, we further present the complete training trajectories under $\beta = 0.1$ and $\beta = 0.5$. For each heterogeneity level, we compare FedCUE with all competing methods on CIFAR-10, CIFAR-20, SVHN, and CL-CIFAR10 over 100 communication rounds. Fig. 6 and 7 provide the corresponding training trajectories. Overall, FedCUE reaches competitive accuracy in the early stage of training and maintains its advantage as communication proceeds, indicating that the final improvements are not caused by isolated late-round fluctuations.

CIFAR-10. FedCUE quickly separates from the competing methods after the early communication rounds and maintains higher accuracy throughout most of training under both heterogeneity levels. The advantage is especially clear under $\beta = 0.5$, where the gap gradually increases as training proceeds and remains substantial near convergence. Under the stronger $\beta = 0.1$ setting, the training trajectories exhibit larger fluctuations, but FedCUE still maintains the highest final accuracy.

CIFAR-20. The advantage is particularly clear on CIFAR-20. Under both $\beta = 0.1$ and $\beta = 0.5$, FedCUE forms a clear performance gap over the baselines during the early-to-middle stages of training and preserves this gap through the final rounds. This behavior is consistent with the larger final improvements reported on CIFAR-20, where the weaker information provided by each individual complementary label makes evidence consolidation especially useful.

SVHN. On SVHN, FedCUE also shows stable convergence under both settings. Under strong heterogeneity ($\beta = 0.1$), FedCUE maintains a visible advantage over the competing methods after the initial training stage despite round-to-round fluctuations. When $\beta = 0.5$, all methods reach relatively high accuracy and the performance gap becomes smaller, while FedCUE still converges to the best final performance.

CL-CIFAR10. On the human-annotated CL-CIFAR10 benchmark, FedCUE maintains a clear advantage throughout training under both heterogeneity levels. The separation from the baselines emerges early and remains stable during the later communication rounds, rather than appearing only at the end of training. This further shows that the benefit of complementary-evidence collaboration extends to realistic human annotations and is not specific to synthetically generated complementary labels.

Overall. Across the two heterogeneity levels, the convergence curves show that the advantage of FedCUE is established during training rather than emerging only at the final communication rounds. The improvement is generally more visible on CIFAR-20 and CL-CIFAR10, where complementary supervision is relatively more challenging, while the gap is smaller on SVHN as all methods approach higher accuracy. Similar trends under both $\beta = 0.1$ and $\beta = 0.5$ indicate that the observed gains are not specific to a single degree of client heterogeneity.

Under stronger heterogeneity ($\beta = 0.1$), the trajectories exhibit larger round-to-round variations, as expected from more imbalanced client data. Nevertheless, FedCUE maintains a persistent advantage on most datasets, suggesting that complementary-evidence collaboration remains effective even when local class distributions are highly skewed.

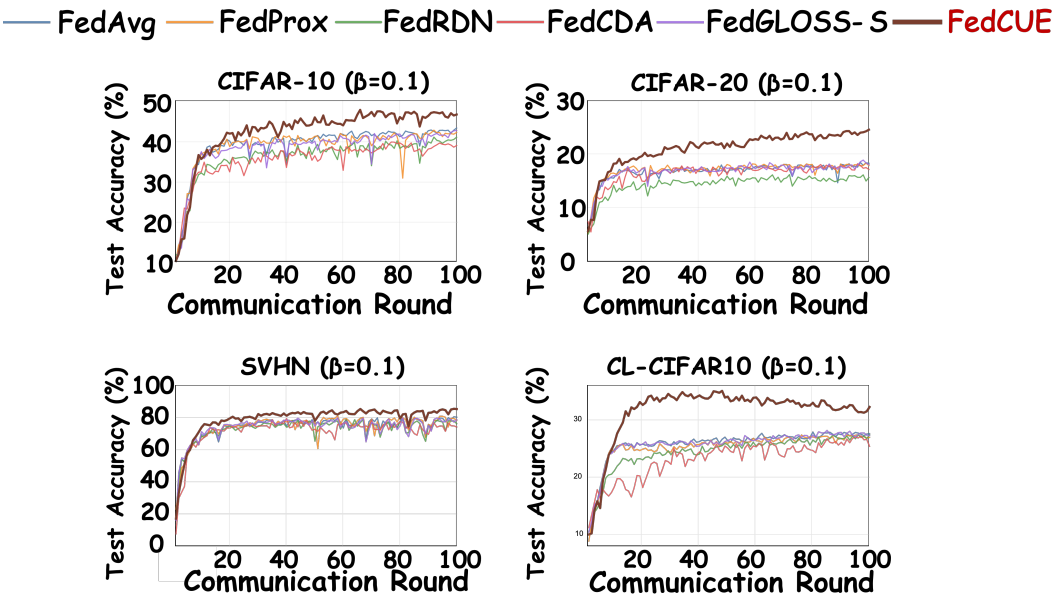

Figure 6: Convergence comparison under strong data heterogeneity ($\beta = 0.1$) on (a) CIFAR-10, (b) CIFAR-20, (c) SVHN, and (d) CL-CIFAR10. Test accuracy is reported over 100 communication rounds.

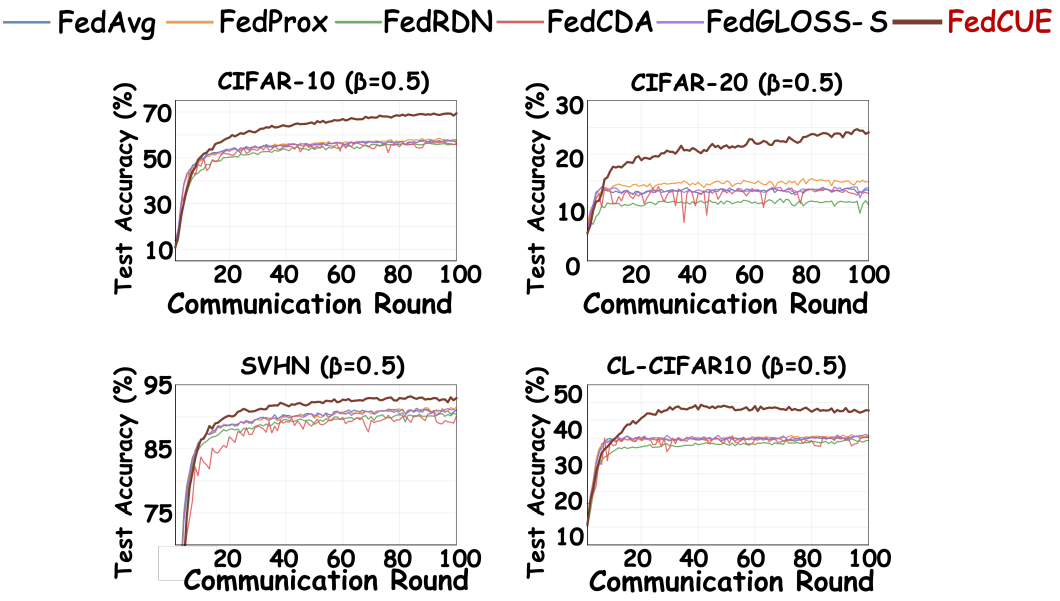

Figure 7: Convergence comparison under moderate data heterogeneity ($\beta = 0.5$) on (a) CIFAR-10, (b) CIFAR-20, (c) SVHN, and (d) CL-CIFAR10. Test accuracy is reported over 100 communication rounds.

E.3 Detailed Ablation Results

Table 11 reports the progressive ablation results under strong data heterogeneity ($\beta = 0.1$), complementing the moderate heterogeneity setting reported in the main text.

Table 11: Progressive ablation study under strong data heterogeneity ($\beta = 0.1$).

Config.	CIFAR-10	CIFAR-20	SVHN	CL-CIFAR20
A	18.16	9.71	18.37	6.53
$A + B$	43.15 ↑24.99	18.25 ↑8.54	80.27 ↑61.90	7.59 ↑1.06
$A + B + C$	48.26 ↑5.11	22.08 ↑3.83	85.13 ↑4.86	8.83 ↑1.24
$A + B + C + D$	46.51 ↓1.75	24.53 ↑2.45	85.40 ↑0.27	9.12 ↑0.29

The overall trend remains consistent with the main-text results. Federated aggregation substantially improves over local-only training, and introducing local complementary evidence further improves performance across all four datasets. Adding global evidence yields additional gains on CIFAR-20, SVHN, and CL-CIFAR20, while slightly reducing accuracy on CIFAR-10. These results suggest that local evidence provides the more consistent improvement, whereas the contribution of global evidence can vary across datasets under strong heterogeneity.

F Further Discussion

F.1 Communication Cost

FedCUE introduces cluster-level evidence communication in addition to standard federated model exchange. Let P denote the number of model parameters, d the feature dimension, C the number of classes, and M_L and M_G the numbers of local and global clusters, respectively.

For each participating client, standard federated training requires transmitting the model parameters between the client and server, corresponding to approximately $2P$ scalar values per communication round. After the warm-up stage, FedCUE additionally uploads at most M_L local evidence summaries. Each summary contains a d-dimensional cluster representation, a C-dimensional complementary-label distribution, and one scalar support size. The additional client-to-server communication is therefore

$$M_L(d + C + 1) \tag{82}$$

scalar values per participating client.

The server returns at most M_G global evidence entries, each consisting of a d-dimensional global cluster representation and a C-dimensional complementary-label distribution. The corresponding server-to-client overhead is

$$M_G(d + C). \tag{83}$$

Thus, the total additional communication introduced by FedCUE is bounded by

$$M_L(d + C + 1) + M_G(d + C) \tag{84}$$

scalars per participating client and post-warm-up round. Under the default configuration, $M_L = M_G = 10$. Importantly, this overhead depends on the number of clusters rather than the number of local training samples and is typically small compared with transmitting the full model parameters.

F.2 PRIVACY DISCUSSION

FEDCUE keeps raw training samples and instance-level complementary annotations on local clients. During complementary-evidence collaboration, clients transmit only aggregated cluster-level statistics, including cluster representations, complementary-label distributions, and support sizes. Consequently, individual samples and their annotations are not directly exchanged with the server or other clients.

However, the use of aggregated statistics does not by itself provide a formal privacy guarantee. As with other federated learning protocols, communicated model parameters and intermediate statistics may still contain information about local data. FEDCUE should therefore be viewed as preserving data decentralization rather than providing formal privacy protection. Developing formal privacy mechanisms for complementary-evidence exchange is beyond the scope of this work.