FEDC^2: FROM CLIENT COMPATIBILITY TO CLASS COMPLEMENTARITY FOR FEDERATED LEARNING FROM LABEL PROPORTIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Federated learning from label proportions (Fed-LLP) aims to learn instance-level predictors from distributed clients whose local bags are supervised only by class proportions. Existing LLP methods mainly focus on extracting supervision from aggregate proportions, while federated learning methods typically address client heterogeneity under standard instance-level supervision. However, in Fed-LLP, aggregate supervision is further complicated by substantial client heterogeneity, while the absence of instance-level labels makes such heterogeneity harder to characterize and exploit. Towards this end, we propose **FedC^2**, a hierarchical collaboration framework that organizes cross-client knowledge sharing from client compatibility to class-wise complementarity. **FedC^2** first constructs a target-specific collaboration base from the geometry of clients' observed bag-proportion sets, and then conservatively refines class-wise collaborator contributions according to target-relative coverage and local proportion consistency. We further provide a theoretical analysis characterizing when the resulting hierarchical reweighting tightens class-specific aggregation-error bounds. Extensive experiments across four benchmarks and three levels of client heterogeneity show that **FedC^2** achieves the best accuracy in 10 of 12 settings against representative LLP-based, generic FL, and personalized FL methods. Further analyses support the effectiveness, generality, and transfer relevance of the proposed hierarchical collaboration strategy. A quick visual overview of this work is available at https://fdletf7qxe.github.io.

1 INTRODUCTION

In many real-world applications, instance-level annotations are unavailable or difficult to obtain, while supervision may instead be available in the form of aggregate statistics over groups of instances(Singh et al., 2024; Ma et al., 2025). For example, online platforms may provide feedback at the cohort level(Brahmbhatt et al., 2024; Valverde et al., 2025; Nandy et al., 2022), whereas healthcare and population studies often report group-level prevalence rather than individual annotations(Havaldar et al., 2024; Pacheco et al., 2024). Such settings motivate learning from label proportions (LLP), where each bag of instances is supervised only by its class-proportion vector(Quadrianto et al., 2008; Ma et al., 2026b). The central challenge is to learn an accurate instance-level classifier from such coarse aggregate supervision(Busa-Fekete et al., 2025; Ma et al., 2025).

In practice, proportion-supervised data are often held by multiple institutions or devices, while the underlying instances remain local due to data ownership, governance, privacy, or operational constraints(McMahan et al., 2017; Zhao et al., 2025; Kairouz et al., 2021). As illustrated in Figure 1(a), federated learning from label proportions (Fed-LLP) addresses this setting by allowing each client to learn from local bags with only proportion-level supervision and participate in collaborative model learning without sharing its raw data. Beyond the weak supervision inherited from LLP, this decentralized setting further introduces substantial cross-client heterogeneity(Li et al., 2020; Karimireddy et al., 2020; Hsieh et al., 2020), as clients may exhibit markedly different bag-proportion structures and class compositions. Importantly, such heterogeneity may also create distinct and potentially complementary information across clients (Yan et al., 2024). Recent federated studies highlight the

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.

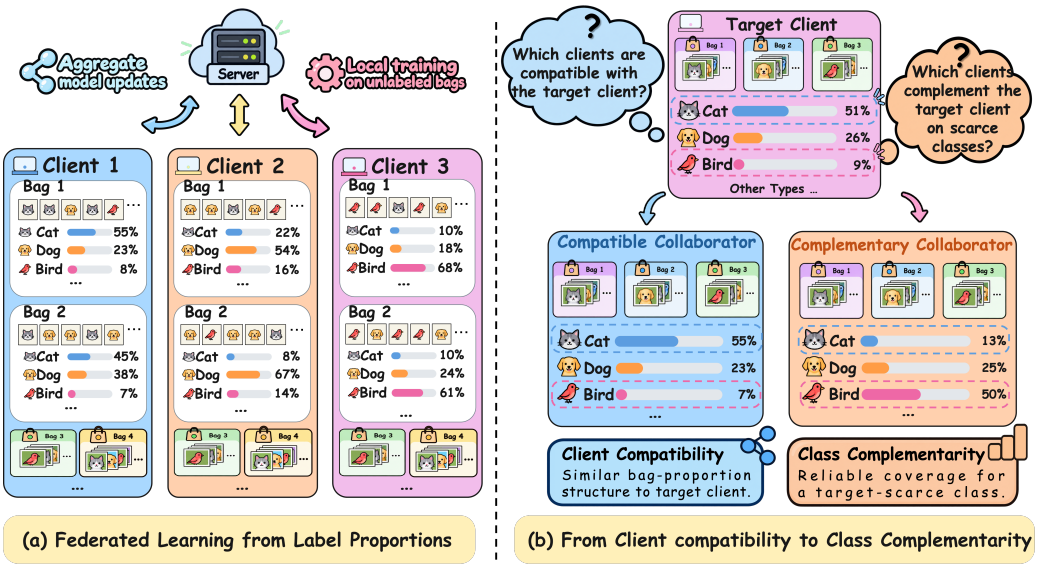


Figure 1: Motivation of FedC^2. (a) Fed-LLP enables collaborative learning from distributed bags using class-proportion supervision. (b) Cross-client collaboration identifies clients compatible with the target and then uses class-wise complementarity to differentiate their contributions by class.

value of cross-client knowledge sharing under heterogeneous local supervision (Kou et al., 2026; Wu et al., 2026; Ye et al., 2023), motivating collaboration in Fed-LLP.

As illustrated in Figure 1(b), effectively organizing cross-client collaboration in Fed-LLP raises two questions. At the client level, such heterogeneity means that not every client is equally appropriate for a given target(Ye et al., 2023; Yan et al., 2024; Huang et al., 2021). This raises the first question: ❶ *Which clients should a given target client collaborate with?* Yet even suitable collaborators need not contribute equally across classes, particularly for classes that are scarce at the target (Xu et al., 2023; Lee & Choi, 2025). This raises a finer-grained question: ❷ *How should the contributions of selected collaborators be differentiated across classes?* Together, these questions suggest that effective collaboration in Fed-LLP should be organized at both the client and class levels.

Building on this perspective, we propose FedC^2, a hierarchical cross-client collaboration framework for Fed-LLP. To establish the collaboration context, FedC^2 characterizes client compatibility from the geometry of client-specific bag-proportion sets and constructs a target-specific aggregation base. Within this context, FedC^2 further models directional class-wise complementarity through target-relative class coverage and collaborator-local reliability. The former captures whether a collaborator provides additional coverage for a class relative to the target, while the latter measures the consistency of its bag-level predictions with local proportion supervision. Because such class-wise evidence is inferred from aggregate rather than instance-level supervision, FedC^2 uses this evidence to conservatively refine the compatibility-guided base. We further provide a theoretical analysis that characterizes the conditions under which compatibility reweighting and class-wise refinement tighten class-specific aggregation-error bounds and relates these bounds to the excess target classifier-only LLP objective. The main contributions of this work are summarized as follows:

- ***Fed-LLP Collaboration Perspective.*** To the best of our knowledge, we are among the first to systematically study cross-client collaboration for federated learning from label proportions under heterogeneous clients. We introduce a coarse-to-fine perspective in which identifying whom a target should collaborate with is followed by differentiating how selected collaborators contribute across classes.

- ***Framework and Theory.*** We instantiate this perspective in FedC^2 through compatibility-guided collaboration and conservative class-wise refinement, and theoretically characterize when the two stages tighten class-specific aggregation-error bounds and their connection to the excess target classifier-only LLP objective under a fixed representation.

- ***Empirical Validation.*** Across four benchmarks and three levels of client heterogeneity, FedC^2 achieves the best accuracy in 10 of 12 settings among representative LLP-based, generic FL, and personalized FL methods. Further analyses support the hierarchical design and the transfer relevance of the proposed class-wise complementarity signal.

2 Related Work

Learning from Label Proportions. Learning from label proportions (LLP) aims to learn instance-level predictors from bags annotated only with class proportions, and has been studied from both algorithmic and theoretical perspectives (Quadrianto et al., 2008; Yu et al., 2014; Li et al., 2024). Recent advances have explored noise-aware learning (Zhang et al., 2022), unbiased risk estimation (Busa-Fekete et al., 2023), latent-label inference (Havaldar et al., 2024), and progressive training strategies (Liu et al., 2025a). More recent approaches further develop proportion-aware objectives and constraints to improve learning from bag-level supervision (Ma et al., 2026b;a). These studies primarily focus on improving learning from bag-level proportions within a single learner, while LLP in federated and heterogeneous environments remains comparatively underexplored.

Federated Learning under Data Heterogeneity. Federated learning (FL) has developed extensive approaches to address statistical heterogeneity, ranging from stabilizing local optimization (Li et al., 2020; Karimireddy et al., 2020; Acar et al., 2021) to adapting collaboration across clients (Ye et al., 2023; Xu et al., 2023; Lee & Choi, 2025; Li et al., 2026). These methods, however, are mainly developed for standard supervised FL, where instance-level labels or class-specific statistics are directly available. In Fed-LLP, clients are supervised only by bag-level class proportions, making the underlying heterogeneity harder to characterize and address.

3 The Proposed Framework: FedC^2

3.1 Problem Formulation

We consider Fed-LLP with C clients and K classes. Client i holds M_i local bags, $\mathcal{D}_i = \{(\mathcal{B}_i^m, \boldsymbol{\pi}_i^m)\}_{m=1}^{M_i}$, where $\mathcal{B}_i^m = \{x_{i,n}^m\}_{n=1}^{N_i^m}$ contains N_i^m instances and $\boldsymbol{\pi}_i^m \in \Delta^{K-1}$ is the corresponding class-proportion vector. Let $N_i = \sum_{m=1}^{M_i} N_i^m$ denote the total number of local instances. At communication round t, client i maintains a model Θ_i^t that predicts $\mathbf{p}_i^t(x) \in \Delta^{K-1}$ for each input x, yielding the bag-level prediction $\bar{\mathbf{p}}_i^{m,t} = \frac{1}{N_i^m} \sum_{n=1}^{N_i^m} \mathbf{p}_i^t(x_{i,n}^m)$.

Training proceeds for T communication rounds. At each round, client i locally updates Θ_i^t on $\mathcal{D}_i$ to obtain $\widehat{\Theta}_i^t$. The server then applies a client-specific collaboration operator $\mathcal{A}_i^t$, which uses the locally updated client models together with the bag proportions, bag sizes, and derived proportion-consistency statistics to construct the next-round model of target client i:

$$\Theta_i^{t+1} = \mathcal{A}_i^t \left(\{\widehat{\Theta}_j^t\}_{j=1}^{C} \right). \tag{1}$$

After the final round, the client-specific models are consolidated into a single global model for evaluation and deployment. Our goal is to learn an accurate global instance-level classifier from heterogeneous proportion-supervised clients without centralizing their underlying instances.

Overview of FedC^2. To address Fed-LLP under client heterogeneity, FedC^2 adopts a coarse-to-fine collaboration strategy. It first exploits the geometry of client-specific bag-proportion sets to identify suitable collaborators for each target client and construct a compatibility-guided collaboration base. FedC^2 then estimates directional class-wise complementarity through target-relative class coverage and collaborator-local reliability, and uses it to refine this collaboration base at the classifier level. The overall framework is illustrated in Figure 2.

3.2 Compatibility-Guided Client Collaboration

Proportion-Space Client Compatibility. We first characterize client-level compatibility using the bag-proportion information directly available in Fed-LLP. Specifically, we represent client i by its

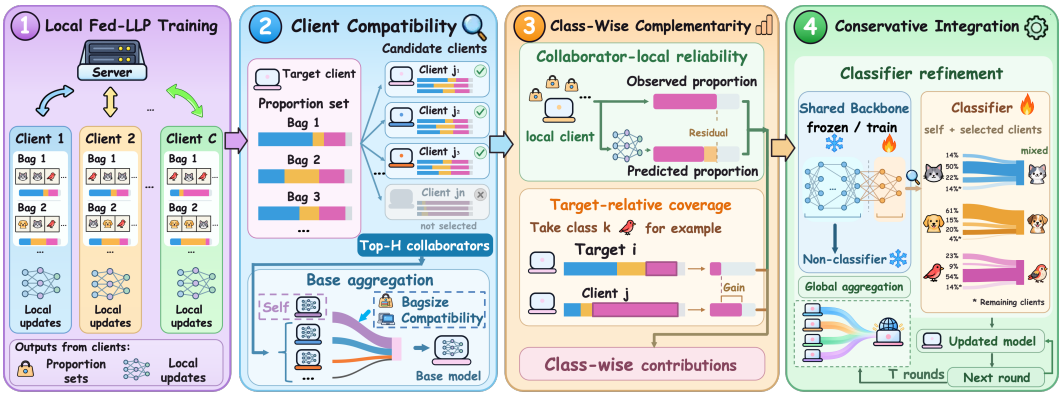


Figure 2: Overview of FedC^2. Client compatibility builds a target-specific collaboration base; class-wise complementarity then conservatively refines classifier aggregation before global consolidation.

observed bag-proportion set $\mathcal{P}_i = \{\boldsymbol{\pi}_i^m\}_{m=1}^{M_i}$. Because clients may contain different numbers of bags and bags are not aligned across clients, we compare their proportion sets using a symmetric nearest-neighbor set discrepancy. For two proportion vectors $\mathbf{p}$ and $\mathbf{q}$, let $\mathrm{JS}(\mathbf{p}, \mathbf{q})$ denote the Jensen–Shannon divergence (Lin, 1991). We define the discrepancy between clients i and j as

$$d_{ij} = \frac{1}{2}\left[\frac{1}{M_i}\sum_{m=1}^{M_i}\min_{1\le n\le M_j}\mathrm{JS}\left(\boldsymbol{\pi}_i^m, \boldsymbol{\pi}_j^n\right) + \frac{1}{M_j}\sum_{n=1}^{M_j}\min_{1\le m\le M_i}\mathrm{JS}\left(\boldsymbol{\pi}_j^n, \boldsymbol{\pi}_i^m\right)\right]. \tag{2}$$

We then convert d_{ij} into a compatibility score $s_{ij} = \exp(-d_{ij}/\sigma)$, where $\sigma > 0$ controls the sensitivity to proportion-space discrepancy and a larger s_{ij} indicates higher compatibility. For each target client i, we rank clients $j \in \{1, \dots, C\} \setminus \{i\}$ by s_{ij} in descending order and define the collaborator set $\mathcal{N}_i$ as the top-H clients, where H denotes the number of selected collaborators.

Compatibility-Guided Collaboration Base. Given the collaborator set $\mathcal{N}_i$, we construct a target-specific collaboration base by jointly accounting for proportion-space compatibility and local data size. For $j \in \{i\} \cup \mathcal{N}_i$, we define

$$\alpha_{ij} = \frac{s_{ij}N_j}{\sum_{\ell\in\{i\}\cup\mathcal{N}_i} s_{i\ell}N_\ell}, \qquad s_{ii} = 1, \tag{3}$$

which yields the compatibility-guided base model

$$\Theta_i^{\mathrm{base},t+1} = \sum_{j\in\{i\}\cup\mathcal{N}_i} \alpha_{ij}\widehat{\Theta}_j^t. \tag{4}$$

Since α_{ij} is shared across classes, the base captures *which clients should contribute* but not *how their contributions should vary by class*, motivating the refinement introduced next.

3.3 Complementarity-Guided Class-Wise Refinement

Target-Relative Class Coverage. We first assess whether a selected collaborator provides additional class-specific coverage relative to the target. For client i, we summarize its class-k coverage from the observed bag proportions by the sample-weighted aggregate proportion

$$r_{i,k} = \frac{\sum_{m=1}^{M_i} N_i^m \pi_{i,k}^m}{N_i}. \tag{5}$$

Given target i and collaborator $j \in \mathcal{N}_i$, we define the target-relative class-k coverage gain as

$$g_{ij,k}^{\mathrm{data}} = \frac{[r_{j,k} - r_{i,k}]_+}{\max(1 - r_{i,k}, \epsilon)}, \tag{6}$$

where $[z]_+ = \max(z, 0)$ and $\epsilon > 0$ ensures numerical stability. The positive part retains only coverage that exceeds the target's current class-k proportion, while $1 - r_{i,k}$ represents the maximum

additional class-k proportion available beyond the target's current coverage. Accordingly, $g_{ij,k}^{\mathrm{data}}$ measures the normalized positive difference between collaborator j's and target i's class-k proportions. This signal is directional and target-relative: it is positive only when $r_{j,k} > r_{i,k}$ and lies in $[0, 1]$. However, coverage alone does not indicate whether the collaborator's current predictions are consistent with its own class-k proportion supervision.

Collaborator-Local Reliability. We therefore assess class-k reliability through consistency with the collaborator's own LLP supervision. Let $\mathcal{R}_j \subseteq \{1, \dots, M_j\}$ denote the local bags used for reliability estimation, and let $\widehat{\mathbf{p}}_j^t(x)$ denote the predictive distribution of the locally updated model $\widehat{\Theta}_j^t$. For each $m \in \mathcal{R}_j$, its predicted class-k bag proportion is

$$\widehat{\bar{p}}_{j,k}^{m,t} = \frac{1}{N_j^m} \sum_{n=1}^{N_j^m} \widehat{p}_{j,k}^t(x_{j,n}^m). \tag{7}$$

We average the absolute class-k bag-proportion errors over $\mathcal{R}_j$, weighting each bag by its size

$$e_{j,k}^t = \frac{\sum_{m \in \mathcal{R}_j} N_j^m \left| \widehat{\bar{p}}_{j,k}^{m,t} - \pi_{j,k}^m \right|}{\sum_{m \in \mathcal{R}_j} N_j^m}, \qquad q_{j,k}^t = 1 - e_{j,k}^t. \tag{8}$$

A larger $q_{j,k}^t$ indicates stronger agreement between the model's bag-level predictions and the observed class-k proportions. Importantly, $q_{j,k}^t$ serves as a proportion-consistency-based reliability proxy rather than an estimate of instance-level class-k accuracy.

Conservative Complementarity Integration. We combine target-relative coverage and collaborator-local reliability to define the directional class-wise complementarity score

$$c_{ij,k}^t = g_{ij,k}^{\mathrm{data}} q_{j,k}^t, \qquad j \in \mathcal{N}_i. \tag{9}$$

A larger $c_{ij,k}^t$ indicates stronger class-k complementarity between collaborator j and target i, with greater consistency with the collaborator's local proportion supervision. To limit the influence of large complementarity scores, we set $\widehat{c}_{ij,k}^t = \min(c_{ij,k}^t, c_{\max})$, where $c_{\max} > 0$ is a predefined cap.

When $\sum_{\ell \in \mathcal{N}_i} \alpha_{i\ell} \widehat{c}_{i\ell,k}^t > 0$, we further differentiate the selected collaborators according to class-k complementarity while retaining the compatibility-guided weighting structure:

$$\beta_{ij,k}^t = \frac{\alpha_{ij} \widehat{c}_{ij,k}^t}{\sum_{\ell \in \mathcal{N}_i} \alpha_{i\ell} \widehat{c}_{i\ell,k}^t}, \qquad j \in \mathcal{N}_i, \qquad \beta_{ii,k}^t = 0. \tag{10}$$

The vector $(\beta_{ij,k}^t)_{j \in \mathcal{N}_i}$ defines a class-specific distribution over the selected collaborators.

Since complementarity is estimated from aggregate proportion supervision, we use this distribution as a refinement rather than directly replacing the compatibility-guided weights:

$$\widetilde{\alpha}_{ij,k}^t = \begin{cases} (1-\eta)\alpha_{ij} + \eta \beta_{ij,k}^t, & \sum_{\ell \in \mathcal{N}_i} \alpha_{i\ell} \widehat{c}_{i\ell,k}^t > 0, \\ \alpha_{ij}, & \text{otherwise}, \end{cases} \tag{11}$$

where $j \in \{i\} \cup \mathcal{N}_i$ and $\eta \in [0, 1]$ controls the class-wise refinement strength.

The refined class-wise weights are applied to the corresponding classifier parameters. For a linear classification head, let $\widehat{W}_{j,k}^t$ and $\widehat{b}_{j,k}^t$ denote the class-k classifier row and bias of $\widehat{\Theta}_j^t$. We obtain

$$W_{i,k}^{t+1} = \sum_{j \in \{i\} \cup \mathcal{N}_i} \widetilde{\alpha}_{ij,k}^t \widehat{W}_{j,k}^t, \qquad b_{i,k}^{t+1} = \sum_{j \in \{i\} \cup \mathcal{N}_i} \widetilde{\alpha}_{ij,k}^t \widehat{b}_{j,k}^t, \tag{12}$$

while the non-classifier parameters remain those of $\Theta_i^{\mathrm{base},t+1}$.

3.4 Model Optimization and Global Consolidation

At each communication round $t = 0, \dots, T-1$, client i initializes from Θ_i^t and minimizes the bag-level LLP objective (Yu et al., 2014; Ardehaly & Culotta, 2017)

$$\mathcal{L}_{\mathrm{LLP}}^i(\Theta) = -\sum_{m=1}^{M_i} \frac{N_i^m}{N_i} \sum_{k=1}^{K} \pi_{i,k}^m \log \bar{p}_{i,k}^m(\Theta). \tag{13}$$

The local update yields $\widehat{\Theta}_i^t$, from which client i computes its class-wise reliability vector $\mathbf{q}_i^t$ as described in Sec. 3.3. Each client provides the server with its bag-proportion information $\{(\boldsymbol{\pi}_i^m, N_i^m)\}_{m=1}^{M_i}$, and at each round sends $\widehat{\Theta}_i^t$ together with $\mathbf{q}_i^t$.

Using these quantities, the server performs the compatibility-guided collaboration in Sec. 3.2 followed by the class-wise refinement in Sec. 3.3, producing the next-round client-specific model Θ_i^{t+1}. After training, the client-specific models at the validation-selected checkpoint $t^\star$ are consolidated into a global evaluation model using sample-size weighting (McMahan et al., 2017):

$$\Theta^{\text{global}} = \sum_{i=1}^{C} \frac{N_i}{\sum_{j=1}^{C} N_j} \Theta_i^{t^\star}. \tag{14}$$

3.5 Theoretical Analysis

We analyze how compatibility-guided reweighting changes class-specific aggregation error, when complementarity refinement further tightens the bound while staying close to the base, and how these bounds relate to the target LLP objective. Full proofs and technical conditions appear in Appendix A.

Proposition 3.1. *Conditioning on the compatibility-guided representation, for target i and class k, let $\widehat{\mathbf{u}}_{j,k}^t$ denote client j's locally updated class-k classifier block and $\mathbf{u}_{i,k}^{\star,t+1}$ the class-k block of a fixed minimizer of the target classifier-only LLP objective. For any normalized non-negative weights $\mathbf{w}$, define the class-specific parameter discrepancy and aggregation-error bound as*

$$\varepsilon_{ij,k}^t = \left\| \widehat{\mathbf{u}}_{j,k}^t - \mathbf{u}_{i,k}^{\star,t+1} \right\|_2^2, \qquad \mathcal{B}_{i,k}^t(\mathbf{w}) = \sum_{j=1}^{C} w_j \varepsilon_{ij,k}^t. \tag{15}$$

Let $\rho_j = N_j / \sum_{\ell=1}^{C} N_\ell$ denote the size-based reference distribution and define

$$h_{ij} = \mathbf{1}\{j \in \{i\} \cup \mathcal{N}_i\} \exp(-d_{ij}/\sigma), \qquad d_{ii} := 0,$$

so that $\alpha_{ij} = \rho_j h_{ij} / \mathbb{E}_{\boldsymbol{\rho}}[h_{ij}]$. For $\sigma > 0$,

$$\mathbb{E}_{\boldsymbol{\alpha}_i}[d_{ij}] \le \mathbb{E}_{\boldsymbol{\rho}}[d_{ij}], \qquad \mathcal{B}_{i,k}^t(\boldsymbol{\alpha}_i) - \mathcal{B}_{i,k}^t(\boldsymbol{\rho}) = \frac{\mathrm{Cov}_{\boldsymbol{\rho}}(h_{ij}, \varepsilon_{ij,k}^t)}{\mathbb{E}_{\boldsymbol{\rho}}[h_{ij}]}. \tag{16}$$

Consequently, $\mathcal{B}_{i,k}^t(\boldsymbol{\alpha}_i) \le \mathcal{B}_{i,k}^t(\boldsymbol{\rho})$ if and only if $\mathrm{Cov}_{\boldsymbol{\rho}}(h_{ij}, \varepsilon_{ij,k}^t) \le 0$.

Theorem 3.2. *For $Z_{i,k}^t := \sum_{j \in \mathcal{N}_i} \alpha_{ij} \widehat{c}_{ij,k}^t > 0$, let $\bar{\alpha}_{ij} = \alpha_{ij}/(1-\alpha_{ii})$ for $j \in \mathcal{N}_i$, and $\mu_{i,k}^t = \mathbb{E}_{\bar{\boldsymbol{\alpha}}_i}[\varepsilon_{ij,k}^t]$. Then*

$$\mathcal{B}_{i,k}^t(\widetilde{\boldsymbol{\alpha}}_{i,k}^t) - \mathcal{B}_{i,k}^t(\boldsymbol{\alpha}_i) = \eta \left[\frac{\mathrm{Cov}_{\bar{\boldsymbol{\alpha}}_i}(\widehat{c}_{ij,k}^t, \varepsilon_{ij,k}^t)}{\mathbb{E}_{\bar{\boldsymbol{\alpha}}_i}[\widehat{c}_{ij,k}^t]} + \alpha_{ii}(\mu_{i,k}^t - \varepsilon_{ii,k}^t) \right]. \tag{17}$$

For $\eta > 0$, the refined aggregation-error bound is no larger than the compatibility-guided bound if and only if the bracketed quantity is non-positive.

Moreover,

$$D_{\mathrm{TV}}\left(\widetilde{\boldsymbol{\alpha}}_{i,k}^t, \boldsymbol{\alpha}_i\right) = \eta D_{\mathrm{TV}}\left(\boldsymbol{\beta}_{i,k}^t, \boldsymbol{\alpha}_i\right) \le \eta, \tag{18}$$

where $D_{\mathrm{TV}}(\mathbf{p}, \mathbf{q}) = \frac{1}{2}\|\mathbf{p} - \mathbf{q}\|_1$. If $Z_{i,k}^t = 0$, $\widetilde{\boldsymbol{\alpha}}_{i,k}^t = \boldsymbol{\alpha}_i$.

Remark 3.1. The two terms in Eq. 17 respectively capture whether complementarity favors lower-discrepancy collaborators and the effect of shifting aggregation mass away from the target, thereby characterizing when refinement tightens the bound. Equation 18 further shows that η bounds the deviation from the compatibility-guided base. Under the respective improvement conditions of Proposition 3.1 and Theorem 3.2,

$$\mathcal{B}_{i,k}^t(\widetilde{\boldsymbol{\alpha}}_{i,k}^t) \le \mathcal{B}_{i,k}^t(\boldsymbol{\alpha}_i) \le \mathcal{B}_{i,k}^t(\boldsymbol{\rho}). \tag{19}$$

Proposition 3.3. *Let $F_i^{t+1}(U)$ denote target i's classifier-only LLP objective under the fixed compatibility-guided representation. For normalized non-negative class-wise aggregation weights $\boldsymbol{\Omega}_i = (\mathbf{w}_{i,1}, \ldots, \mathbf{w}_{i,K})$, let $U_i^{t+1}(\boldsymbol{\Omega}_i)$ denote the resulting aggregated classifier head. Under the reference-minimizer and local-smoothness conditions detailed in Appendix A,*

$$0 \le F_i^{t+1}\left(U_i^{t+1}(\boldsymbol{\Omega}_i)\right) - F_i^{t+1}\left(U_i^{\star,t+1}\right) \le \frac{L_i^{t+1}}{2} \sum_{k=1}^{K} \mathcal{B}_{i,k}^t(\mathbf{w}_{i,k}). \tag{20}$$

Table 1: Comparison with representative baselines under varying client-level label heterogeneity. We report test ACC (%); Mini-ImageNet uses 5-way 5-shot episodic evaluation, while the other datasets use top-1 ACC. Best and second-best results are **bolded** and underlined, respectively.

Method	CIFAR-10			SVHN			Fashion-MNIST			Mini-ImageNet		
β	0.25	0.50	0.75	0.25	0.50	0.75	0.25	0.50	0.75	0.25	0.50	0.75
LLP-based methods												
DLLP	13.96	12.33	11.58	19.59	19.59	19.52	55.75	52.51	62.62	30.24	31.81	32.41
PVC	10.00	10.00	10.00	19.59	19.59	19.59	11.56	10.00	11.62	22.56	24.25	22.52
LLP-DC	10.00	10.00	10.00	19.59	19.59	19.59	10.00	10.00	10.00	25.47	27.34	27.55
FL-based methods												
FedProx	14.83	13.38	11.63	19.59	19.41	18.07	57.30	56.07	60.72	30.36	31.43	32.49
FedWMSAM	10.80	10.00	13.14	19.59	19.58	19.55	17.66	21.19	22.39	31.69	30.88	30.46
FedDyn	12.60	13.28	12.09	18.57	19.59	**24.31**	50.27	48.70	55.29	28.04	27.67	27.40
Personalized FL methods												
Ditto	12.22	10.55	10.85	16.49	16.25	16.55	55.76	52.50	62.66	30.10	31.65	32.14
FEROMA	11.49	11.88	10.98	18.67	19.59	19.52	12.55	9.39	11.54	29.86	30.43	29.55
pFedGraph	11.78	10.02	11.99	11.04	9.68	9.57	11.00	19.68	16.34	**32.54**	32.17	31.35
FedC^2	**21.71**	**17.56**	**13.16**	**21.82**	**19.81**	17.18	**71.05**	**67.56**	**66.24**	30.91	**32.60**	**32.75**

4 Experiments

4.1 Experimental Setup

Datasets and Fed-LLP Construction. We evaluate FedC^2 on four image classification benchmarks: CIFAR-10 (Krizhevsky et al., 2009), SVHN (Netzer et al., 2011), Fashion-MNIST (Xiao et al., 2017), and Mini-ImageNet (Vinyals et al., 2016). We first construct the federated setting by distributing the training data across $C = 50$ clients using a per-class Dirichlet partition (Yurochkin et al., 2019) with concentration $\beta \in \{0.25, 0.50, 0.75\}$, where smaller β induces stronger client-level label heterogeneity. Within each client, we further construct the LLP setting by grouping local samples into bags with target size $B = 32$. The class composition of each bag is generated from a second Dirichlet distribution with $\beta_{\text{bag}} = 0.5$ over the locally available classes.

Baselines. We compare FedC^2 with nine representative baselines from three categories: (1) *LLP-based methods*: DLLP (Ardehaly & Culotta, 2017), PVC (Ma et al., 2026b), and LLP-DC (Ma et al., 2026a); (2) *FL-based methods*: FedProx (Li et al., 2020), FedWMSAM (Li et al., 2025), and FedDyn (Acar et al., 2021); and (3) *personalized FL-based methods*: Ditto (Li et al., 2021), FEROMA (Li et al., 2026), and pFedGraph (Ye et al., 2023).

Implementation Details. We use ResNet-18 for all datasets and train for 50 communication rounds with full client participation and 5 local epochs per round. For FedC^2, we set $H = 25$, $\eta = 0.05$, and $c_{\max} = 0.25$. We report top-1 test ACC of the consolidated global evaluation model on CIFAR-10, SVHN, and Fashion-MNIST; Mini-ImageNet uses 5-way 5-shot episodic query ACC of the consolidated encoder. Further implementation details are provided in Appendix B.

4.2 Performance Comparison

Table 1 compares FedC^2 with representative baselines across three levels of client-level label heterogeneity. We note that: ***Obs.*** ❶ ***FedC^2 achieves strong and consistent overall performance across diverse settings.*** It obtains the best ACC in 10 of the 12 settings, including all heterogeneity levels on CIFAR-10 and Fashion-MNIST. ***Obs.*** ❷ ***FedC^2 consistently outperforms LLP-based methods.*** It surpasses the strongest LLP-based baseline in 11 of 12 settings, indicating that stronger local LLP objectives are insufficient to fully address cross-client heterogeneity. ***Obs.*** ❸ ***FedC^2 consistently outperforms generic FL methods.*** It exceeds the strongest FL-based baseline in 10 of 12 settings, suggesting that leveraging label-proportion structure provides additional benefits beyond generic federated optimization. ***Obs.*** ❹ ***FedC^2 consistently outperforms personalized FL methods.*** It exceeds the strongest personalized FL baseline in 10 of 12 settings, suggesting that client-specific collaboration can be further refined by accounting for class-wise complementarity.

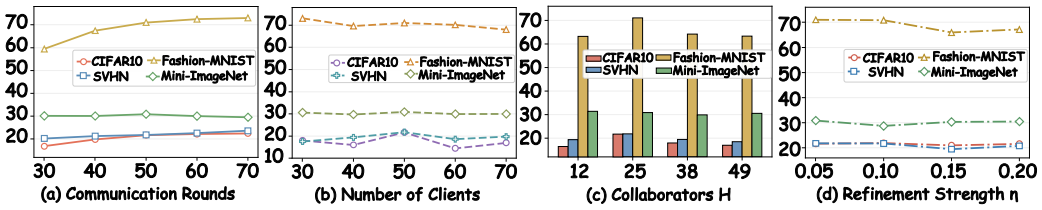


Figure 3: **Sensitivity analysis of FedC^2 under** $\beta = 0.25$ with respect to T, C, H, and η.

4.3 Ablation Studies

We investigate the hierarchical components of FedC^2. Starting from FedAvg (Base), V1 adds compatibility-guided collaboration, while V2 and V3 further augment V1 with coverage-only and reliability-only class-wise refinement, respectively. We find that: ***Obs.*** ❺ ***Compatibility-guided collaboration establishes a stronger aggregation base.*** V1 raises the average ACC from 29.89 to 35.37, providing a strong foundation for subsequent class-wise refinement. ***Obs.*** ❻ ***Coverage and reliability provide complementary signals for class-wise refinement.*** Both V2 and V3 improve over V1, while their joint integration in FedC^2 achieves the highest average ACC of 36.37 and the best performance on three of four datasets.

Table 2: Ablation study of FedC^2 under the $\beta = 0.25$ setting.

Variant	CIFAR-10	SVHN	Fashion-MNIST	Mini-ImageNet	Avg.
Base	13.96	19.59	55.75	30.24	29.89
V1	21.19	19.67	71.23	29.39	35.37
V2	21.58	21.23	**71.80**	30.44	36.26
V3	20.89	20.13	**71.80**	30.33	35.79
FedC^2	**21.71**	**21.82**	71.05	**30.91**	**36.37**

4.4 Mechanism Analysis

Does Complementarity Reflect Actual Transfer? We evaluate whether $c_{ij,k}$ reflects realized class-wise transfer using a leave-one-collaborator-out analysis, with $\Delta\text{ACC}_{ij,k} = \text{ACC}_{\text{full}} - \text{ACC}_{\text{without } j}$. We can see that: ***Obs.*** ❼ ***The complementarity score provides a meaningful ranking signal for realized class-wise transfer.*** As shown in Fig. 4(a), $c_{ij,k}$ is positively correlated with $\Delta\text{ACC}_{ij,k}$ across datasets, with Spearman correlations of 0.524, 0.556, and 0.337 on CIFAR-10, SVHN, and Fashion-MNIST, respectively. These results indicate that collaborators assigned higher complementarity scores tend to provide larger class-wise transfer benefits.

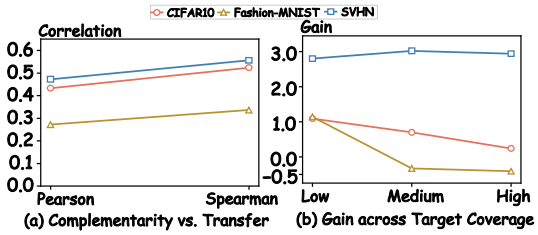


Figure 4: Mechanism analysis of FedC^2.

Which Target Classes Benefit from Refinement? We group target client–class pairs into low-, medium-, and high-coverage subsets and measure the class-wise gain of the full method over the compatibility-guided base. We make the following observation: ***Obs.*** ❽ ***Low-coverage target classes can benefit more from class-wise refinement.*** As shown in Fig. 4(b), the gain on CIFAR-10 decreases from 1.09 to 0.70 and 0.24 percentage points as target coverage increases. Fashion-MNIST shows a similar concentration of benefit in the low-coverage group, whereas SVHN exhibits comparable gains across all three coverage levels.

4.5 Further Analysis

Sensitivity to Training and Federation Scale. We vary communication rounds T and clients C in Fig. 3(a,b). We find out that: ***Obs.*** ❾ ***FedC^2 benefits from sufficient communication while remaining effective across federation scales.*** Increasing T generally improves ACC on CIFAR-10, SVHN, and Fashion-MNIST, after which the gains gradually saturate, while Mini-ImageNet remains comparatively stable. When varying C, the performance exhibits moderate non-monotonic fluctuations but no systematic degradation as the federation scale changes.

Sensitivity to Collaboration Design. We further examine the collaborator-set size H and complementarity refinement strength η, as shown in Fig. 3(c,d). We see that: ***Obs.*** ❿ ***Selective collabo-***

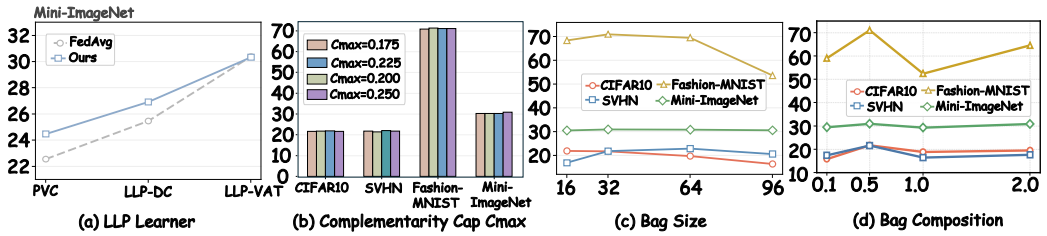


Figure 5: Generalization and robustness analysis of FedC^2 under $\beta = 0.25$ with respect to the local LLP learner, $c_{\max}$, B, and β_{bag}.

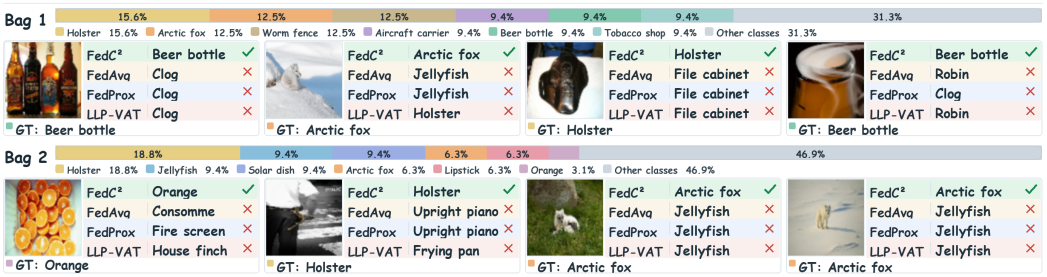


Figure 6: Qualitative prediction examples on Mini-ImageNet training classes.

ration and conservative class-wise refinement lead to more favorable performance. A moderate H generally outperforms either a smaller set or near-global collaboration, suggesting that involving more clients does not necessarily improve collaboration. Relatively small η also provides strong performance, whereas aggressive refinement can reduce ACC on some datasets.

Generalization and Robustness. We examine the LLP learner and complementarity cap in Fig. 5(a,b). The former tests whether collaboration depends on a particular local LLP objective. The latter varies the cap applied to each class-wise complementarity score before refinement. Our results show that: ***Obs.*** ⓫ ***FedC^2 generalizes across LLP learners and remains stable under complementarity caps.*** Integrating FedC^2 consistently improves PVC, LLP-DC, and LLP-VAT on Mini-ImageNet, indicating that its collaboration mechanism is not tied to a particular local LLP objective. Performance also varies mildly across the tested $c_{\max}$ values on all datasets.

Sensitivity to Bag Construction. We examine bag size B and bag-composition concentration β_{bag} in Fig. 5(c,d). We observe that: ***Obs.*** ⓬ ***Bag construction has a noticeable yet dataset-dependent effect, with moderate settings providing a favorable overall operating point.*** Performance is generally stronger at small-to-moderate bag sizes, while overly large bags can degrade CIFAR-10 and Fashion-MNIST. All four datasets achieve their best tested performance at $\beta_{\text{bag}} = 0.5$, despite non-monotonic trends. Overall, moderate bag settings provide a favorable trade-off across datasets.

Qualitative Visualization. Figure 6 shows qualitative predictions on Mini-ImageNet. We find that: ***Obs.*** ⓭ ***FedC^2 more accurately recovers instance classes from bag-level supervision.*** Across the illustrated bags, FedC^2 correctly recognizes more samples, including instances from non-dominant classes, while the baselines more often produce mismatched predictions. This qualitative pattern provides an intuitive complement to the quantitative improvements reported above.

5 Conclusion

In this work, we study federated learning from label proportions (Fed-LLP) under heterogeneous clients and propose FedC^2, a hierarchical collaboration framework that progresses from client compatibility to class-wise complementarity. We provide theoretical analysis to characterize the proposed aggregation mechanism and empirically validate FedC^2 across four benchmarks and varying levels of client heterogeneity. Extensive ablation and further analyses consistently support the effectiveness of the hierarchical design and the relevance of class-wise complementarity. Overall, our results demonstrate the importance of exploiting cross-client structure at multiple granularities when learning instance-level classifiers from federated label-proportion supervision.

AI USE STATEMENT

Generative AI tools were used only to assist with language polishing, grammatical refinement, and the retrieval and organization of publicly available information and references. All technical content, methodology, experiments, analyses, and conclusions were developed and verified by the authors.

ETHICS STATEMENT

This work uses only publicly available benchmark datasets and involves no new data collection or human-subject studies. Although raw instances remain local in the Fed-LLP setting, `FedC`^{2} does not provide formal privacy guarantees; sensitive applications should therefore adopt additional privacy-preserving mechanisms when needed.

REPRODUCIBILITY STATEMENT

The project page is available at https://fdletf7qxe.github.io/, and the implementation is available at https://github.com/fdletf7qxe/FedC2_code. To facilitate reproducibility, we document the dataset splits and evaluation protocols in Appendix B.1, the construction of the Fed-LLP benchmark in Appendix B.2, and the complete training and evaluation protocol, including reliability estimation, checkpoint selection, and global model construction, in Appendix B.3. Baseline implementations and hyperparameters are detailed in Appendix B.4, while the protocols and complete numerical results for the additional mechanism and sensitivity analyses are reported in Appendix D.2.

REFERENCES

Durmus Alp Emre Acar, Yue Zhao, Ramon Matas Navarro, Matthew Mattina, Paul N Whatmough, and Venkatesh Saligrama. Federated learning based on dynamic regularization. *arXiv preprint arXiv:2111.04263*, 2021.

Ehsan Mohammady Ardehaly and Aron Culotta. Co-training for demographic classification using deep learning from label proportions. In *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*, pp. 1017–1024. Ieee, 2017.

Anand Brahmbhatt, Mohith Pokala, Rishi Saket, and Aravindan Raghuveer. Llp-bench: A large scale tabular benchmark for learning from label proportions. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pp. 4374–4381, 2024.

Róbert Busa-Fekete, Heejin Choi, Travis Dick, Claudio Gentile, and Andres Munoz Medina. Easy learning from label proportions. *Advances in Neural Information Processing Systems*, 36:14957–14968, 2023.

Robert Busa-Fekete, Travis Dick, Claudio Gentile, Haim Kaplan, Tomer Koren, and Uri Stemmer. Nearly optimal sample complexity for learning with label proportions. *arXiv preprint arXiv:2505.05355*, 2025.

Haizhou Du, Yiran Xiang, Yiwen Cai, Xiufeng Liu, Zonghan Wu, Huan Huo, and Guodong Long. Fedfree: Breaking knowledge-sharing barriers through layer-wise alignment in heterogeneous federated learning. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 82256–82283. Curran Associates, Inc., 2025. doi: 10.52202/085713-2753.

Dongyang Fan, Bettina Messmer, Nikita Doikov, and Martin Jaggi. On-device collaborative language modeling via a mixture of generalists and specialists. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 15833–15861. PMLR, 13–19 Jul 2025.

Shreyas Havaldar, Navodita Sharma, Shubhi Sareen, Karthikeyan Shanmugam, and Aravindan Raghuveer. Learning from label proportions: Bootstrapping supervised learners via belief propagation. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 20596–20619, 2024.

Kevin Hsieh, Amar Phanishayee, Onur Mutlu, and Phillip Gibbons. The non-iid data quagmire of decentralized machine learning. In *International Conference on Machine Learning*, pp. 4387–4398. PMLR, 2020.

Yutao Huang, Lingyang Chu, Zirui Zhou, Lanjun Wang, Jiangchuan Liu, Jian Pei, and Yong Zhang. Personalized cross-silo federated learning on non-iid data. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 7865–7873, 2021.

Peter Kairouz, H. Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista A. Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, Rafael G. L. D'Oliveira, Hubert Eichner, Salim El Rouayheb, David Evans, Josh Gardner, Zachary Garrett, Adrià Gascón, Badih Ghazi, Phillip B. Gibbons, Marco Gruteser, Zaïd Harchaoui, Chaoyang He, Lie He, Zhouyuan Huo, Ben Hutchinson, Justin Hsu, Martin Jaggi, Tara Javidi, Gauri Joshi, Mikhail Khodak, Jakub Konečný, Aleksandra Korolova, Farinaz Koushanfar, Sanmi Koyejo, Tancrède Lepoint, Yang Liu, Prateek Mittal, Mehryar Mohri, Richard Nock, Ayfer Özgür, Rasmus Pagh, Hang Qi, Daniel Ramage, Ramesh Raskar, Mariana Raykova, Dawn Song, Weikang Song, Sebastian U. Stich, Ziteng Sun, Ananda Theertha Suresh, Florian Tramèr, Praneeth Vepakomma, Jianyu Wang, Li Xiong, Zheng Xu, Qiang Yang, Felix X. Yu, Han Yu, and Sen Zhao. Advances and open problems in federated learning. *Found. Trends Mach. Learn.*, 14(1-2):1–210, 2021. doi: 10.1561/2200000083.

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.

Gyuejeong Lee and Daeyoung Choi. Class-wise federated averaging for efficient personalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1773–1782, October 2025.

Gene Li, Lin Chen, Adel Javanmard, and Vahab Mirrokni. Optimistic rates for learning from label proportions. In *The Thirty Seventh Annual Conference on Learning Theory*, pp. 3437–3474. PMLR, 2024.

Mohan Li, Dario Fenoglio, Martin Gjoreski, and Marc Langheinrich. Federated learning with profile mapping under distribution shifts and drifts. *arXiv preprint arXiv:2602.07671*, 2026.

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.

Tian Li, Shengyuan Hu, Ahmad Beirami, and Virginia Smith. Ditto: Fair and robust federated learning through personalization. In *International conference on machine learning*, pp. 6357–6368. PMLR, 2021.

Tianle Li, Yongzhi Huang, Linshan Jiang, Chang Liu, Qipeng Xie, Wenfeng Du, Lu Wang, and Kaishun Wu. Fedwmsam: Fast and flat federated learning via weighted momentum and sharpness-aware minimization. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 4425–4462. Curran Associates, Inc., 2025. doi: 10.52202/085713-0157.

Jianhua Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, 37(1):145–151, 1991.

Jiabin Liu, Bo Wang, Yuping Zhang, Huadong Wang, Biao Li, Xin Shen, and Gang Kou. Progressive training for learning from label proportions. *IEEE Transactions on Neural Networks and Learning Systems*, 36(11):19951–19963, 2025a.

Qiqi Liu, Jiaqiang Li, Yuchen Liu, Yaochu Jin, Lingjuan Lyu, Xiaohu Wu, and Han Yu. Personalized federated learning under local supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4069–4079, October 2025b.

Tianhao Ma, Han Chen, Juncheng Hu, Yungang Zhu, and Ximing Li. Forming auxiliary high-confident instance-level loss to promote learning from label proportions. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20592–20601. IEEE, 2025.

Tianhao Ma, Ximing Li, Changchun Li, and Renchu Guan. Learning from label proportions with dual-proportion constraints. *arXiv preprint arXiv:2603.21153*, 2026a.

Tianhao Ma, Wei Wang, Ximing Li, Gang Niu, and Masashi Sugiyama. Learning from label proportions via proportional value classification. In *The Fourteenth International Conference on Learning Representations*, 2026b.

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.

Jay Nandy, Rishi Saket, Prateek Jain, Jatin Chauhan, Balaraman Ravindran, and Aravindan Raghuveer. Domain-agnostic contrastive representations for learning from label proportions. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pp. 1542–1551, 2022.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y. Ng. Reading digits in natural images with unsupervised feature learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning 2011*, 2011.

Carolina Pacheco, Florence Yellin, René Vidal, and Benjamin Haeffele. Vertex proportion loss for multi-class cell detection from label proportions. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 366–376. Springer, 2024.

Anik Pramanik, Murat Kantarcioglu, Vincent Oria, and Shantanu Sharma. Feddag: Clustered federated learning via global data and gradient integration for heterogeneous environments. In *The Fourteenth International Conference on Learning Representations*, 2026.

Novi Quadrianto, Alex J Smola, Tiberio S Caetano, and Quoc V Le. Estimating labels from label proportions. In *Proceedings of the 25th International Conference on Machine learning*, pp. 776–783, 2008.

Minhyuk Seo, Taeheon Kim, Hankook Lee, Jonghyun Choi, and Tinne Tuytelaars. Co-lora: Collaborative model personalization on heterogeneous multi-modal clients. In *The Fourteenth International Conference on Learning Representations (ICLR)*, 2026.

Xinyi Shang, Yang Lu, Yiu-ming Cheung, and Hanzi Wang. Fedic: Federated learning on non-iid and long-tailed data via calibrated distillation. In *2022 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6. IEEE, 2022.

Sagalpreet Singh, Navodita Sharma, Shreyas Havaldar, Rishi Saket, and Aravindan Raghuveer. Learning from label proportions and covariate-shifted instances. *arXiv preprint arXiv:2411.12334*, 2024.

Mahendra Singh Thapa and Rui Li. Harnessing heterogeneous statistical strength for personalized federated learning via hierarchical Bayesian inference. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 59285–59303. PMLR, 13–19 Jul 2025.

Jonathan Valverde, Tiansheng Yao, Xiang Li, Yuan Gao, Yin Zhang, Andrew Evdokimov, Adam Kraft, Samuel Ieong, Jerry Zhang, Ed H Chi, et al. Cross-batch aggregation for streaming learning from label proportions in industrial-scale recommendation systems. In *Proceedings of the Nineteenth ACM Conference on Recommender Systems*, pp. 935–939, 2025.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, koray kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

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.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Jian Xu, Xinyi Tong, and Shao-Lun Huang. Personalized federated learning with feature alignment and classifier collaboration. *arXiv preprint arXiv:2306.11867*, 2023.

Kunda Yan, Sen Cui, Abudukelimu Wuerkaixi, Jingfeng Zhang, Bo Han, Gang Niu, Masashi Sugiyama, and Changshui Zhang. Balancing similarity and complementarity for federated learning. *arXiv preprint arXiv:2405.09892*, 2024.

Rui Ye, Zhenyang Ni, Fangzhao Wu, Siheng Chen, and Yanfeng Wang. Personalized federated learning with inferred collaboration graphs. In *International conference on machine learning*, pp. 39801–39817. PMLR, 2023.

Felix X Yu, Krzysztof Choromanski, Sanjiv Kumar, Tony Jebara, and Shih-Fu Chang. On learning from label proportions. *arXiv preprint arXiv:1402.5902*, 2014.

Xin Yu, Zelin He, Ying Sun, Lingzhou Xue, and Runze Li. Understanding the statistical accuracy-communication trade-off in personalized federated learning with minimax guarantees. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 73009–73050. PMLR, 13–19 Jul 2025.

Mikhail Yurochkin, Mayank Agarwal, Soumya Ghosh, Kristjan Greenewald, Nghia Hoang, and Yasaman Khazaeni. Bayesian nonparametric federated learning of neural networks. In *International conference on machine learning*, pp. 7252–7261. PMLR, 2019.

Chenyu Zhang and Navid Azizan. Personalized collaborative learning with affinity-based variance reduction. In *The Fourteenth International Conference on Learning Representations*, 2026.

Jianxin Zhang, Yutong Wang, and Clay Scott. Learning from label proportions by learning with label noise. *Advances in Neural Information Processing Systems*, 35:26933–26942, 2022.

Michael Zhang, Karan Sapra, Sanja Fidler, Serena Yeung, and Jose M Alvarez. Personalized federated learning with first order model optimization. *arXiv preprint arXiv:2012.08565*, 2020.

Joshua Zhao, Saurabh Bagchi, Salman Avestimehr, Kevin Chan, Somali Chaterji, Dimitris Dimitriadis, Jiacheng Li, Ninghui Li, Arash Nourian, and Holger Roth. The federation strikes back: A survey of federated learning privacy attacks, defenses, applications, and policy landscape. *ACM Computing Surveys*, 57(9):1–37, 2025.

Hao Zheng, Shiyu Song, Zhigang Hu, Meiguang Zheng, Liu Yang, Aikun Xu, Rongchang Zhao, Ruizhi Pu, Ruiyi Fang, and Boyu Wang. Intra-class unbiased prototype aggregation and classifier collaboration for personalized federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 28795–28803, 2026.

APPENDIX CONTENTS

A Additional Theoretical Analysis **15**

A.1 Conditional Classifier Setup and Auxiliary Lemmas 15

A.2 Proof of Proposition 3.1 16

A.3 Proof of Theorem 3.2 17

A.4 Smoothness-Based Connection to the Target LLP Objective 20

B Additional Experimental Details **21**

B.1 Dataset Details 21

B.2 Fed-LLP Benchmark Construction 22

B.3 Training and Implementation Details 22

B.4 Baseline Implementations 23

C Additional Related Work **24**

C.1 Learning from Label Proportions 24

C.2 Federated Learning under Heterogeneity and Collaboration 24

D Additional Experimental Results **25**

D.1 Additional Qualitative Results 25

D.2 Numerical Results for Main-Text Analyses 30

D.3 Communication Dynamics. 31

A Additional Theoretical Analysis

A.1 Conditional Classifier Setup and Auxiliary Lemmas

We first formalize the conditional classifier analysis used in Sec. 3.5. Consider target client i at communication round t. After compatibility-guided aggregation, let $\phi_i^{\mathrm{base},t+1}$ denote the non-classifier parameters of $\Theta_i^{\mathrm{base},t+1}$. Throughout the analysis, we condition on $\phi_i^{\mathrm{base},t+1}$ and study only the effect of class-specific classifier reweighting.

Let $U = (\mathbf{u}_1, \ldots, \mathbf{u}_K)$ denote the classifier head, where $\mathbf{u}_k$ concatenates the classifier weight and bias associated with class k. All classifier-parameter quantities below are evaluated in the common parameterization used by federated optimization and server-side parameter aggregation. Conditioned on $\phi_i^{\mathrm{base},t+1}$, we define the classifier-only target LLP objective as

$$F_i^{t+1}(U) = -\sum_{m=1}^{M_i} \frac{N_i^m}{N_i} \sum_{k=1}^{K} \pi_{i,k}^m \log \bar{p}_{i,k}^m \left(\phi_i^{\mathrm{base},t+1}, U\right), \tag{21}$$

where $\bar{p}_{i,k}^m(\phi, U)$ denotes the predicted class-k proportion of bag $\mathcal{B}_i^m$ under non-classifier parameters ϕ and classifier head U.

Assumption A.1 (Reference Minimizer). Conditioned on $\phi_i^{\mathrm{base},t+1}$, the classifier-only objective in Eq. 21 admits a finite minimizer $U_i^{\star,t+1}$ in the unconstrained classifier parameter space. When multiple minimizers exist, we fix one reference minimizer for the given target i and round t and use it throughout the corresponding analysis. This is an explicit attainment assumption for the conditional classifier objective and is not implied merely by finiteness of the local dataset.

Let $\mathbf{u}_{i,k}^{\star,t+1}$ denote the class-k parameter block of $U_i^{\star,t+1}$. For locally updated client j, let $\widehat{\mathbf{u}}_{j,k}^t$ denote its class-k classifier parameter block, and define

$$\varepsilon_{ij,k}^t = \left\| \widehat{\mathbf{u}}_{j,k}^t - \mathbf{u}_{i,k}^{\star,t+1} \right\|_2^2. \tag{22}$$

This quantity measures classifier-parameter discrepancy relative to the fixed target reference solution in the parameterization used by the optimization; it should not be interpreted as instance-level classification error.

Lemma A.1 (Class-Specific Aggregation Upper Bound)**.** *Let* $\mathbf{w} = (w_1, \ldots, w_C)$ *satisfy* $w_j \geq 0$ *and* $\sum_{j=1}^C w_j = 1$*, and define the aggregated class-*k *classifier block as*

$$\mathbf{u}_{i,k}^{t+1}(\mathbf{w}) = \sum_{j=1}^{C} w_j \widehat{\mathbf{u}}_{j,k}^t. \tag{23}$$

Then

$$\left\| \mathbf{u}_{i,k}^{t+1}(\mathbf{w}) - \mathbf{u}_{i,k}^{\star,t+1} \right\|_2^2 \leq \mathcal{B}_{i,k}^t(\mathbf{w}), \tag{24}$$

where

$$\mathcal{B}_{i,k}^t(\mathbf{w}) = \sum_{j=1}^{C} w_j \varepsilon_{ij,k}^t. \tag{25}$$

Proof. Using $\sum_j w_j = 1$, we have

$$\begin{aligned} &\left\| \mathbf{u}_{i,k}^{t+1}(\mathbf{w}) - \mathbf{u}_{i,k}^{\star,t+1} \right\|_2^2 \\ &= \left\| \sum_{j=1}^{C} w_j \left(\widehat{\mathbf{u}}_{j,k}^t - \mathbf{u}_{i,k}^{\star,t+1} \right) \right\|_2^2 \\ &\leq \sum_{j=1}^{C} w_j \left\| \widehat{\mathbf{u}}_{j,k}^t - \mathbf{u}_{i,k}^{\star,t+1} \right\|_2^2 \\ &= \mathcal{B}_{i,k}^t(\mathbf{w}), \end{aligned} \tag{26}$$

where the inequality follows from Jensen's inequality for the squared Euclidean norm. □

For later use, we next establish a generic identity for score-guided reweighting. For a probability distribution $\mathbf{p} = (p_j)_{j\in\mathcal{I}}$ over a finite index set $\mathcal{I}$, we write $\mathbb{E}_{\mathbf{p}}[a_j] = \sum_{j\in\mathcal{I}} p_j a_j$ and use $\mathrm{Cov}_{\mathbf{p}}(\cdot,\cdot)$ for the corresponding weighted covariance.

Lemma A.2 (Score-Tilting Identity). *Let $\mathbf{p} = (p_j)_{j\in\mathcal{I}}$ be a probability distribution over a finite index set $\mathcal{I}$, and let $a_j \geq 0$ satisfy $\mathbb{E}_{\mathbf{p}}[a_j] > 0$. Define the score-tilted distribution $\mathbf{p}^{(a)}$ by*

$$p_j^{(a)} = \frac{p_j a_j}{\mathbb{E}_{\mathbf{p}}[a_j]}, \qquad j \in \mathcal{I}. \tag{27}$$

Then, for any scalar quantity e_j,

$$\mathbb{E}_{\mathbf{p}^{(a)}}[e_j] - \mathbb{E}_{\mathbf{p}}[e_j] = \frac{\mathrm{Cov}_{\mathbf{p}}(a_j, e_j)}{\mathbb{E}_{\mathbf{p}}[a_j]}. \tag{28}$$

Proof. By the definition of $\mathbf{p}^{(a)}$,

$$\begin{aligned}
\mathbb{E}_{\mathbf{p}^{(a)}}[e_j] - \mathbb{E}_{\mathbf{p}}[e_j] &= \frac{\mathbb{E}_{\mathbf{p}}[a_j e_j]}{\mathbb{E}_{\mathbf{p}}[a_j]} - \mathbb{E}_{\mathbf{p}}[e_j] \\
&= \frac{\mathbb{E}_{\mathbf{p}}[a_j e_j] - \mathbb{E}_{\mathbf{p}}[a_j]\mathbb{E}_{\mathbf{p}}[e_j]}{\mathbb{E}_{\mathbf{p}}[a_j]} \\
&= \frac{\mathrm{Cov}_{\mathbf{p}}(a_j, e_j)}{\mathbb{E}_{\mathbf{p}}[a_j]}.
\end{aligned} \tag{29}$$

□

The next elementary covariance property will be used to formalize the monotonicity induced by compatibility-based collaborator selection.

Lemma A.3 (Oppositely Ordered Covariance). *Let $\mathbf{p} = (p_j)_{j\in\mathcal{I}}$ be a probability distribution over a finite index set $\mathcal{I}$. If two scalar sequences $\{a_j\}_{j\in\mathcal{I}}$ and $\{b_j\}_{j\in\mathcal{I}}$ satisfy*

$$(a_j - a_\ell)(b_j - b_\ell) \leq 0 \qquad \text{for all } j, \ell \in \mathcal{I}, \tag{30}$$

then

$$\mathrm{Cov}_{\mathbf{p}}(a_j, b_j) \leq 0. \tag{31}$$

Proof. The weighted covariance admits the pairwise representation

$$\mathrm{Cov}_{\mathbf{p}}(a_j, b_j) = \frac{1}{2}\sum_{j\in\mathcal{I}}\sum_{\ell\in\mathcal{I}} p_j p_\ell (a_j - a_\ell)(b_j - b_\ell). \tag{32}$$

By Eq. 30, every pairwise product is non-positive, while $p_j p_\ell \geq 0$. Therefore, $\mathrm{Cov}_{\mathbf{p}}(a_j, b_j) \leq 0$. □

A.2 Proof of Proposition 3.1

Recall the size-based reference distribution $\rho_j = N_j / \sum_{\ell=1}^{C} N_\ell$ and define

$$h_{ij} = \mathbf{1}\{j \in \{i\} \cup \mathcal{N}_i\} \exp(-d_{ij}/\sigma). \tag{33}$$

Since $d_{ii} = 0$, we have $h_{ii} = 1$. Moreover, $\rho_i > 0$ because $N_i > 0$, and hence

$$\mathbb{E}_{\boldsymbol{\rho}}[h_{ij}] \geq \rho_i h_{ii} = \rho_i > 0. \tag{34}$$

Thus, Lemma A.2 is applicable on $\mathcal{I} = \{1, \ldots, C\}$ with $\mathbf{p} = \boldsymbol{\rho}$ and $a_j = h_{ij}$.

We first verify that the compatibility-guided weights are exactly the corresponding score-tilted distribution. Using the definition of $\boldsymbol{\rho}$,

$$\begin{aligned}
\frac{\rho_j h_{ij}}{\mathbb{E}_{\boldsymbol{\rho}}[h_{ij}]} &= \frac{N_j h_{ij}}{\sum_{\ell=1}^{C} N_\ell h_{i\ell}} \\
&= \frac{N_j \exp(-d_{ij}/\sigma)}{\sum_{\ell\in\{i\}\cup\mathcal{N}_i} N_\ell \exp(-d_{i\ell}/\sigma)}
\end{aligned} \tag{35}$$

for $j \in \{i\} \cup \mathcal{N}_i$, and the value is zero otherwise. Since $s_{ij} = \exp(-d_{ij}/\sigma)$ and $s_{ii} = 1$, this recovers the compatibility-guided weights in Sec. 3.2. Therefore,

$$\alpha_{ij} = \frac{\rho_j h_{ij}}{\mathbb{E}_{\boldsymbol{\rho}}[h_{ij}]}. \tag{36}$$

Thus, $\boldsymbol{\alpha}_i$ is precisely the h_{ij}-tilted distribution obtained from $\boldsymbol{\rho}$ according to Lemma A.2.

We next characterize the effect of this reweighting on proportion-space discrepancy. For non-self clients j and ℓ, the top-H construction and the strict monotonic decrease of $\exp(-d/\sigma)$ in d for $\sigma > 0$ imply

$$d_{ij} < d_{i\ell} \quad \Longrightarrow \quad h_{ij} \geq h_{i\ell}. \tag{37}$$

Indeed, if the higher-discrepancy client is selected, the lower-discrepancy client must also be selected; if both are selected, the exponential weighting preserves the reverse ordering. The self index satisfies the same ordering because $d_{ii} = 0$, $h_{ii} = 1$, $d_{ij} \geq 0$, and $h_{ij} \leq 1$ for all j. When two discrepancies are equal, their difference is zero regardless of tie breaking. Consequently,

$$(h_{ij} - h_{i\ell})\,(d_{ij} - d_{i\ell}) \leq 0 \qquad \text{for all } j, \ell \in \{1, \ldots, C\}. \tag{38}$$

Lemma A.3 therefore gives

$$\mathrm{Cov}_{\boldsymbol{\rho}}\,(h_{ij}, d_{ij}) \leq 0. \tag{39}$$

Applying Lemma A.2 with $e_j = d_{ij}$ and using Eq. 36, we obtain

$$\begin{aligned} \mathbb{E}_{\boldsymbol{\alpha}_i}[d_{ij}] - \mathbb{E}_{\boldsymbol{\rho}}[d_{ij}] &= \frac{\mathrm{Cov}_{\boldsymbol{\rho}}(h_{ij}, d_{ij})}{\mathbb{E}_{\boldsymbol{\rho}}[h_{ij}]} \\ &\leq 0. \end{aligned} \tag{40}$$

Thus, compatibility-guided reweighting yields a weighted average proportion-space discrepancy no larger than that under the size-based reference distribution.

For the aggregation-error bound,

$$\mathcal{B}^t_{i,k}(\boldsymbol{\alpha}_i) = \mathbb{E}_{\boldsymbol{\alpha}_i}\left[\varepsilon^t_{ij,k}\right], \qquad \mathcal{B}^t_{i,k}(\boldsymbol{\rho}) = \mathbb{E}_{\boldsymbol{\rho}}\left[\varepsilon^t_{ij,k}\right]. \tag{41}$$

Applying Lemma A.2 again, now with $e_j = \varepsilon^t_{ij,k}$, yields

$$\mathcal{B}^t_{i,k}(\boldsymbol{\alpha}_i) - \mathcal{B}^t_{i,k}(\boldsymbol{\rho}) = \frac{\mathrm{Cov}_{\boldsymbol{\rho}}\left(h_{ij}, \varepsilon^t_{ij,k}\right)}{\mathbb{E}_{\boldsymbol{\rho}}[h_{ij}]}. \tag{42}$$

Since $\mathbb{E}_{\boldsymbol{\rho}}[h_{ij}] > 0$, Eq. 42 implies

$$\mathcal{B}^t_{i,k}(\boldsymbol{\alpha}_i) \leq \mathcal{B}^t_{i,k}(\boldsymbol{\rho}) \quad \Longleftrightarrow \quad \mathrm{Cov}_{\boldsymbol{\rho}}\left(h_{ij}, \varepsilon^t_{ij,k}\right) \leq 0. \tag{43}$$

This proves Proposition 3.1. □

A Sufficient Condition for Compatibility Improvement. A sufficient condition for Eq. 43 is that

$$(h_{ij} - h_{i\ell})(\varepsilon^t_{ij,k} - \varepsilon^t_{i\ell,k}) \leq 0 \qquad \text{for all } j, \ell. \tag{44}$$

That is, higher effective compatibility does not correspond to larger classifier-parameter discrepancy. Lemma A.3 then gives $\mathrm{Cov}_{\boldsymbol{\rho}}(h_{ij}, \varepsilon^t_{ij,k}) \leq 0$, and hence $\mathcal{B}^t_{i,k}(\boldsymbol{\alpha}_i) \leq \mathcal{B}^t_{i,k}(\boldsymbol{\rho})$. This condition is sufficient but not necessary.

A.3 Proof of Theorem 3.2

We first consider the case $Z^t_{i,k} > 0$. Recall that

$$Z^t_{i,k} = \sum_{j \in \mathcal{N}_i} \alpha_{ij} \widehat{c}^t_{ij,k}. \tag{45}$$

Since every summand is non-negative, $Z^t_{i,k} > 0$ implies that $\alpha_{ij}\widehat{c}^t_{ij,k} > 0$ for at least one $j \in \mathcal{N}_i$. Hence,

$$1 - \alpha_{ii} = \sum_{j \in \mathcal{N}_i} \alpha_{ij} > 0, \tag{46}$$

and the normalized base distribution over the selected collaborators,

$$\bar{\alpha}_{ij} = \frac{\alpha_{ij}}{1-\alpha_{ii}}, \qquad j \in \mathcal{N}_i, \tag{47}$$

is well defined.

Using Eq. 47, we can rewrite the normalization term as

$$\begin{aligned} Z_{i,k}^t &= (1-\alpha_{ii}) \sum_{j \in \mathcal{N}_i} \bar{\alpha}_{ij} \widehat{c}_{ij,k}^t \\ &= (1-\alpha_{ii}) \mathbb{E}_{\bar{\boldsymbol{\alpha}}_i} \left[\widehat{c}_{ij,k}^t \right]. \end{aligned} \tag{48}$$

Since both $Z_{i,k}^t > 0$ and $1-\alpha_{ii} > 0$,

$$\mathbb{E}_{\bar{\boldsymbol{\alpha}}_i} \left[\widehat{c}_{ij,k}^t \right] > 0. \tag{49}$$

For $j \in \mathcal{N}_i$, the complementarity-weighted distribution therefore satisfies

$$\begin{aligned} \beta_{ij,k}^t &= \frac{\alpha_{ij} \widehat{c}_{ij,k}^t}{Z_{i,k}^t} \\ &= \frac{\bar{\alpha}_{ij} \widehat{c}_{ij,k}^t}{\mathbb{E}_{\bar{\boldsymbol{\alpha}}_i} [\widehat{c}_{ij,k}^t]}. \end{aligned} \tag{50}$$

Thus, on $\mathcal{N}_i$, $\boldsymbol{\beta}_{i,k}^t$ is exactly the score-tilted distribution obtained from $\bar{\boldsymbol{\alpha}}_i$ using $\widehat{c}_{ij,k}^t$ as the tilting score. When viewed as a distribution over all C clients, we set $\beta_{ii,k}^t = 0$ and assign zero mass outside $\mathcal{N}_i$.

Recall that

$$\mu_{i,k}^t = \mathbb{E}_{\bar{\boldsymbol{\alpha}}_i} \left[\varepsilon_{ij,k}^t \right]. \tag{51}$$

Applying Lemma A.2 on $\mathcal{I} = \mathcal{N}_i$ to Eq. 50, with $\mathbf{p} = \bar{\boldsymbol{\alpha}}_i$, $a_j = \widehat{c}_{ij,k}^t$, and $e_j = \varepsilon_{ij,k}^t$, yields

$$\begin{aligned} \mathcal{B}_{i,k}^t(\boldsymbol{\beta}_{i,k}^t) &= \mathbb{E}_{\boldsymbol{\beta}_{i,k}^t} \left[\varepsilon_{ij,k}^t \right] \\ &= \mu_{i,k}^t + \frac{\mathrm{Cov}_{\bar{\boldsymbol{\alpha}}_i} \left(\widehat{c}_{ij,k}^t, \varepsilon_{ij,k}^t \right)}{\mathbb{E}_{\bar{\boldsymbol{\alpha}}_i} [\widehat{c}_{ij,k}^t]}. \end{aligned} \tag{52}$$

On the other hand, the compatibility-guided bound decomposes into the target contribution and the selected-collaborator contribution:

$$\begin{aligned} \mathcal{B}_{i,k}^t(\boldsymbol{\alpha}_i) &= \alpha_{ii} \varepsilon_{ii,k}^t + \sum_{j \in \mathcal{N}_i} \alpha_{ij} \varepsilon_{ij,k}^t \\ &= \alpha_{ii} \varepsilon_{ii,k}^t + (1-\alpha_{ii}) \mu_{i,k}^t. \end{aligned} \tag{53}$$

Subtracting Eq. 53 from Eq. 52 gives

$$\begin{aligned} &\mathcal{B}_{i,k}^t(\boldsymbol{\beta}_{i,k}^t) - \mathcal{B}_{i,k}^t(\boldsymbol{\alpha}_i) \\ &= \frac{\mathrm{Cov}_{\bar{\boldsymbol{\alpha}}_i} \left(\widehat{c}_{ij,k}^t, \varepsilon_{ij,k}^t \right)}{\mathbb{E}_{\bar{\boldsymbol{\alpha}}_i} [\widehat{c}_{ij,k}^t]} + \alpha_{ii} \left(\mu_{i,k}^t - \varepsilon_{ii,k}^t \right). \end{aligned} \tag{54}$$

The two terms in Eq. 54 have distinct roles. The covariance term captures the score–discrepancy alignment among the selected collaborators, whereas $\alpha_{ii}(\mu_{i,k}^t - \varepsilon_{ii,k}^t)$ captures the effect of shifting aggregation mass away from the target. In particular, when $\varepsilon_{ii,k}^t < \mu_{i,k}^t$, the target has lower discrepancy than the selected collaborators on average, making the displacement term positive. When $\varepsilon_{ii,k}^t \geq \mu_{i,k}^t$, the displacement term is non-positive.

For $Z_{i,k}^t > 0$ and $\eta \in [0,1]$, the refined aggregation weights satisfy

$$\widetilde{\boldsymbol{\alpha}}_{i,k}^t = (1-\eta)\boldsymbol{\alpha}_i + \eta \boldsymbol{\beta}_{i,k}^t. \tag{55}$$

Since $\mathcal{B}_{i,k}^t$ is linear in the aggregation weights,

$$\begin{aligned}
&\mathcal{B}_{i,k}^t(\widetilde{\boldsymbol{\alpha}}_{i,k}^t) - \mathcal{B}_{i,k}^t(\boldsymbol{\alpha}_i) \\
&= \eta \left[\mathcal{B}_{i,k}^t(\boldsymbol{\beta}_{i,k}^t) - \mathcal{B}_{i,k}^t(\boldsymbol{\alpha}_i)\right] \\
&= \eta \left[\frac{\mathrm{Cov}_{\bar{\boldsymbol{\alpha}}_i}\left(\widehat{c}_{ij,k}^t, \varepsilon_{ij,k}^t\right)}{\mathbb{E}_{\bar{\boldsymbol{\alpha}}_i}[\widehat{c}_{ij,k}^t]} + \alpha_{ii}\left(\mu_{i,k}^t - \varepsilon_{ii,k}^t\right)\right].
\end{aligned} \tag{56}$$

This establishes the decomposition in Theorem 3.2.

For $\eta > 0$, the sign of the bound change in Eq. 56 is determined exactly by the bracketed quantity. Therefore,

$$\mathcal{B}_{i,k}^t(\widetilde{\boldsymbol{\alpha}}_{i,k}^t) \le \mathcal{B}_{i,k}^t(\boldsymbol{\alpha}_i) \tag{57}$$

if and only if

$$-\frac{\mathrm{Cov}_{\bar{\boldsymbol{\alpha}}_i}\left(\widehat{c}_{ij,k}^t, \varepsilon_{ij,k}^t\right)}{\mathbb{E}_{\bar{\boldsymbol{\alpha}}_i}[\widehat{c}_{ij,k}^t]} \ge \alpha_{ii}\left(\mu_{i,k}^t - \varepsilon_{ii,k}^t\right). \tag{58}$$

For $\eta = 0$, the refined and base distributions coincide, and the aggregation-error bounds are equal. For any $\eta > 0$, the interpolation parameter scales the magnitude of the refinement-induced change without altering its sign.

It remains to establish the bounded-refinement property. Here, both $\boldsymbol{\alpha}_i$ and $\boldsymbol{\beta}_{i,k}^t$ are viewed as distributions over the full client set, using the zero-extension convention above for $\boldsymbol{\beta}_{i,k}^t$. From Eq. 55,

$$\widetilde{\boldsymbol{\alpha}}_{i,k}^t - \boldsymbol{\alpha}_i = \eta\left(\boldsymbol{\beta}_{i,k}^t - \boldsymbol{\alpha}_i\right). \tag{59}$$

Hence,

$$\begin{aligned}
D_{\mathrm{TV}}\left(\widetilde{\boldsymbol{\alpha}}_{i,k}^t, \boldsymbol{\alpha}_i\right) &= \frac{1}{2}\left\|\widetilde{\boldsymbol{\alpha}}_{i,k}^t - \boldsymbol{\alpha}_i\right\|_1 \\
&= \eta D_{\mathrm{TV}}\left(\boldsymbol{\beta}_{i,k}^t, \boldsymbol{\alpha}_i\right) \\
&\le \eta,
\end{aligned} \tag{60}$$

where the last inequality follows because the total variation distance between two probability distributions is at most one. Thus, η directly bounds the total-variation departure of the refined aggregation distribution from the compatibility-guided base.

Finally, when $Z_{i,k}^t = 0$, the refinement rule directly gives

$$\widetilde{\boldsymbol{\alpha}}_{i,k}^t = \boldsymbol{\alpha}_i, \tag{61}$$

so the compatibility-guided weights are recovered exactly and the corresponding total variation distance is zero. This completes the proof of Theorem 3.2. □

Complementarity Tilting Property. Equation 50 further gives

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{\beta}_{i,k}^t}\left[\widehat{c}_{ij,k}^t\right] - \mathbb{E}_{\bar{\boldsymbol{\alpha}}_i}\left[\widehat{c}_{ij,k}^t\right] \\
&= \frac{\mathrm{Var}_{\bar{\boldsymbol{\alpha}}_i}\left(\widehat{c}_{ij,k}^t\right)}{\mathbb{E}_{\bar{\boldsymbol{\alpha}}_i}[\widehat{c}_{ij,k}^t]} \ge 0.
\end{aligned} \tag{62}$$

Thus, independently of its relation to classifier-parameter discrepancy, the complementarity tilt shifts the normalized collaborator distribution toward larger complementarity scores. Equality holds when the complementarity scores are constant on the support of $\bar{\boldsymbol{\alpha}}_i$. In this case, $\boldsymbol{\beta}_{i,k}^t = \bar{\boldsymbol{\alpha}}_i$, which removes score-based differentiation among collaborators but does not in general recover $\boldsymbol{\alpha}_i$, since the latter also assigns mass to the target itself. Because a common positive rescaling of all active complementarity scores cancels under normalization, this property concerns their relative pattern rather than their absolute magnitude.

A.4 Smoothness-Based Connection to the Target LLP Objective

We finally connect the class-specific aggregation-error bound to the target LLP objective defined in Eq. 21. For each class k, let $\mathbf{w}_{i,k} = (w_{i1,k}, \dots, w_{iC,k})$ be a non-negative normalized aggregation distribution, and define

$$\mathbf{u}_{i,k}^{t+1}(\mathbf{w}_{i,k}) = \sum_{j=1}^{C} w_{ij,k} \widehat{\mathbf{u}}_{j,k}^{t}. \tag{63}$$

Collecting the K class-specific blocks gives the aggregated classifier head

$$U_i^{t+1}(\mathbf{\Omega}_i) = \left(\mathbf{u}_{i,1}^{t+1}(\mathbf{w}_{i,1}), \dots, \mathbf{u}_{i,K}^{t+1}(\mathbf{w}_{i,K}) \right), \tag{64}$$

where

$$\mathbf{\Omega}_i = (\mathbf{w}_{i,1}, \dots, \mathbf{w}_{i,K}). \tag{65}$$

We identify the classifier head with the concatenation of its class-specific parameter blocks, so that

$$\|U - U'\|_2^2 = \sum_{k=1}^{K} \|\mathbf{u}_k - \mathbf{u}'_k\|_2^2. \tag{66}$$

To specify the region relevant to the analysis, define

$$\mathcal{U}_{i,k}^{t} = \operatorname{conv}\left(\left\{ \mathbf{u}_{i,k}^{\star,t+1}, \widehat{\mathbf{u}}_{1,k}^{t}, \dots, \widehat{\mathbf{u}}_{C,k}^{t} \right\} \right), \tag{67}$$

and

$$\mathcal{U}_i^t = \mathcal{U}_{i,1}^t \times \cdots \times \mathcal{U}_{i,K}^t. \tag{68}$$

The region $\mathcal{U}_i^t$ is compact and convex and contains both the reference minimizer and every classifier head induced by the class-wise aggregation distributions considered above.

Assumption A.2 (Local Classifier Smoothness). Conditioned on the fixed compatibility-guided representation $\phi_i^{\text{base},t+1}$, the classifier-only target LLP objective $F_i^{t+1}(U)$ is differentiable and L_i^{t+1}-smooth on $\mathcal{U}_i^t$.

This is a local regularity condition on the compact convex region generated blockwise by the reference solution and the locally updated classifier blocks, rather than a global smoothness assumption on the full model.

Proposition A.4 (Connection to the Target LLP Objective)**.** *Under Assumptions A.1 and A.2, for any ordered collection of class-wise aggregation distributions $\mathbf{\Omega}_i$,*

$$\begin{aligned} 0 &\le F_i^{t+1}\left(U_i^{t+1}(\mathbf{\Omega}_i)\right) - F_i^{t+1}\left(U_i^{\star,t+1}\right) \\ &\le \frac{L_i^{t+1}}{2} \sum_{k=1}^{K} \mathcal{B}_{i,k}^{t}(\mathbf{w}_{i,k}). \end{aligned} \tag{69}$$

Proof. The first inequality follows because $U_i^{\star,t+1}$ is a minimizer of F_i^{t+1}. Moreover, by Assumption A.1, $U_i^{\star,t+1}$ is a finite unconstrained minimizer of the differentiable classifier-only objective. Hence, the first-order optimality condition gives

$$\nabla F_i^{t+1}\left(U_i^{\star,t+1}\right) = 0. \tag{70}$$

By Assumption A.2 and the standard smoothness inequality,

$$\begin{aligned} &F_i^{t+1}\left(U_i^{t+1}(\mathbf{\Omega}_i)\right) - F_i^{t+1}\left(U_i^{\star,t+1}\right) \\ &\le \left\langle \nabla F_i^{t+1}\left(U_i^{\star,t+1}\right), U_i^{t+1}(\mathbf{\Omega}_i) - U_i^{\star,t+1} \right\rangle \\ &\quad + \frac{L_i^{t+1}}{2} \left\| U_i^{t+1}(\mathbf{\Omega}_i) - U_i^{\star,t+1} \right\|_2^2 \\ &= \frac{L_i^{t+1}}{2} \left\| U_i^{t+1}(\mathbf{\Omega}_i) - U_i^{\star,t+1} \right\|_2^2. \end{aligned} \tag{71}$$

Using the block decomposition in Eq. 66,

$$\begin{aligned}
&\left\| U_i^{t+1}(\mathbf{\Omega}_i) - U_i^{\star,t+1} \right\|_2^2 \\
&= \sum_{k=1}^{K} \left\| \mathbf{u}_{i,k}^{t+1}(\mathbf{w}_{i,k}) - \mathbf{u}_{i,k}^{\star,t+1} \right\|_2^2 \\
&\leq \sum_{k=1}^{K} \mathcal{B}_{i,k}^{t}(\mathbf{w}_{i,k}),
\end{aligned} \tag{72}$$

where the last inequality follows from Lemma A.1. Substituting Eq. 72 into Eq. 71 proves Proposition A.4. □

For convenience, define the resulting smoothness-based excess-objective upper bound as

$$\mathcal{R}_i^{t+1}(\mathbf{\Omega}_i) = \frac{L_i^{t+1}}{2} \sum_{k=1}^{K} \mathcal{B}_{i,k}^{t}(\mathbf{w}_{i,k}). \tag{73}$$

By Proposition A.4, $\mathcal{R}_i^{t+1}(\mathbf{\Omega}_i)$ upper-bounds the excess target classifier-only LLP objective under the fixed representation.

We now consider the three ordered collections of aggregation distributions corresponding to the hierarchy analyzed in the main text:

$$\begin{aligned}
\mathbf{\Omega}_i^{\rho} &= (\boldsymbol{\rho}, \ldots, \boldsymbol{\rho}), \\
\mathbf{\Omega}_i^{\alpha} &= (\boldsymbol{\alpha}_i, \ldots, \boldsymbol{\alpha}_i), \\
\widetilde{\mathbf{\Omega}}_i^{t} &= (\widetilde{\boldsymbol{\alpha}}_{i,1}^{t}, \ldots, \widetilde{\boldsymbol{\alpha}}_{i,K}^{t}).
\end{aligned} \tag{74}$$

A sufficient condition for the corresponding hierarchy of smoothness-based bounds is that Eq. 43 holds for every class k, and, for $\eta > 0$, that Eq. 58 holds for every class satisfying $Z_{i,k}^{t} > 0$. For classes with $Z_{i,k}^{t} = 0$, the refined and compatibility-guided weights coincide by the exact fallback rule. When $\eta = 0$, the refinement stage also coincides with the compatibility-guided base. Under these conditions,

$$\mathcal{R}_i^{t+1}\left(\widetilde{\mathbf{\Omega}}_i^{t}\right) \leq \mathcal{R}_i^{t+1}\left(\mathbf{\Omega}_i^{\alpha}\right) \leq \mathcal{R}_i^{t+1}\left(\mathbf{\Omega}_i^{\rho}\right). \tag{75}$$

Thus, conditioned on the same target representation, the corresponding smoothness-based upper bounds are ordered as in Eq. 75 under the stated sufficient conditions. This orders derived upper bounds rather than the realized target LLP objective values themselves. Moreover, $\boldsymbol{\rho}$ serves only as a size-based classifier-weighting reference under the same fixed target representation and should not be interpreted as an end-to-end FedAvg model. Accordingly, the analysis does not establish monotonic decrease of the realized training objective across communication rounds or a performance guarantee for the final consolidated global model.

B Additional Experimental Details

B.1 Dataset Details

We evaluate on four image classification benchmarks: CIFAR-10, SVHN, Fashion-MNIST, and Mini-ImageNet. Table 3 summarizes their dataset statistics and train/validation/test splits.

For CIFAR-10, SVHN, and Fashion-MNIST, we reserve a stratified 10% subset of the official training data for validation and keep the official test split for final evaluation. Mini-ImageNet follows the standard class-disjoint $64/16/20$ train/validation/test split. For preprocessing, CIFAR-10 uses random cropping and horizontal flipping during training, Fashion-MNIST images are replicated to three channels, and Mini-ImageNet images are resized to 84×84.

Table 3: Dataset statistics and data splits used in our experiments.

Dataset	Classes	Train	Validation	Test
CIFAR-10	10	45,000	5,000	10,000
SVHN	10	65,931	7,326	26,032
Fashion-MNIST	10	54,000	6,000	10,000
Mini-ImageNet	64 / 16 / 20	38,400	9,600	12,000

Mini-ImageNet Episodic Evaluation. Mini-ImageNet uses mutually disjoint sets of 64 training, 16 validation, and 20 test classes. The model is trained only on the 64 training classes. For validation and test, each episode samples five classes from the corresponding split, with five labeled support images and 15 query images per class. The consolidated global model is used only as a feature extractor: for each class, we average its five support embeddings to construct an episode-specific prototype, and assign each query to its nearest prototype under squared Euclidean distance. The encoder is kept fixed during evaluation, with no gradient updates or fine-tuning. The learned 64-way training head remains in the checkpoint but is bypassed when predicting validation and test queries. Episode accuracy is computed over the query set and averaged across episodes.

B.2 Fed-LLP Benchmark Construction

Federated Client Partition. For each dataset, we first distribute the training data across $C = 50$ clients. For each class k, its samples are partitioned among the clients according to an C-dimensional Dirichlet distribution with concentration β. We consider $\beta \in \{0.25, 0.50, 0.75\}$, where a smaller β induces stronger client-level label heterogeneity.

LLP Bag Construction. Within each client, we further organize the local samples into LLP bags with target size $B = 32$. For each bag, a class-composition vector is sampled from a Dirichlet distribution with concentration $\beta_{\text{bag}} = 0.5$ over the classes with remaining local samples. The corresponding class counts are sampled subject to the available samples of each class, and the selected samples are removed before constructing the next bag. This process continues until all local samples are assigned, with the terminal bag allowed to contain fewer than B samples. The empirical class proportions of each constructed bag define its label-proportion vector $\boldsymbol{\pi}_i^m$ used for LLP training.

B.3 Training and Implementation Details

Model and Local Training. We use ResNet-18 as the backbone for all four datasets, with a 3×3 stride-1 convolutional stem and no initial max-pooling. All $C = 50$ clients participate in every communication round, and each client performs 5 local epochs before aggregation. Local training uses SGD with momentum 0.9 and weight decay 5×10^{-4}. The learning rate starts from 0.05 and follows a cosine decay schedule over the communication rounds.

The local optimization unit is an LLP bag rather than an individual sample. Each local SGD step processes a mini-batch of up to eight bags. For each bag, we weight the LLP proportion loss by N_i^m/N_i, its fraction of the client's training instances, and sum the weighted losses over the mini-batch without further normalization by the number of bags. The default `FedC`^{2} experiments use the DLLP-KL local solver. Model selection is based exclusively on validation performance, and the selected checkpoint is used for final test evaluation.

Reliability Estimation. For each client j, the reliability subset $\mathcal{R}_j$ consists of the first $\min(16, M_j)$ training bags in the fixed stored bag order. Reliability is computed using only the bag images and their observed class proportions, without access to instance-level labels. Specifically, we average the predicted class probabilities within each bag and compare the resulting bag-level prediction with its observed proportion vector, with bag-size weighting as defined in Sec. 3.3. We recompute $\mathbf{q}_j$ after local training every five communication rounds, starting from round 0, and reuse the most recently cached value between refreshes.

Checkpoint Selection. Model selection is performed exclusively on the validation set. For `FedC`^{2} on CIFAR-10, SVHN, and Fashion-MNIST, we select a common communication round $t^\star$ that maximizes the arithmetic mean of the client-wise validation top-1 accuracies. For Mini-ImageNet, the

Table 4: Default training and method configuration used in the main experiments.

Setting	Value
Backbone	ResNet-18
Number of clients C	50
Communication rounds T	50
Client participation	100%
Local epochs	5
Bags per optimization step	8
Target bag size B	32
Optimizer	SGD
Momentum	0.9
Initial learning rate	0.05
Weight decay	5×10^{-4}
Learning-rate schedule	Cosine
Random seed	0
Collaborator-set size H	25
Complementarity strength η	0.05
Complementarity cap $c_{\max}$	0.25
Compatibility bandwidth σ	Mean pairwise compatibility distance

selection criterion is the mean client 5-way 5-shot validation accuracy over the evaluated rounds. When multiple checkpoints attain the same best validation value, we retain the earliest one. All client checkpoints are loaded from the same selected round, rather than selecting a different best round for each client. The test set is not used for checkpoint selection.

Global Evaluation Model. For `FedC`^{2}, FEROMA, and pFedGraph, after checkpoint selection we construct a single global evaluation model by sample-size-weighted consolidation of the selected client models:

$$\Theta^{\text{eval}} = \sum_{i=1}^{C} \frac{N_i}{\sum_{j=1}^{C} N_j} \Theta_i^{t^\star}, \tag{76}$$

where $t^\star$ denotes the validation-selected checkpoint. This one-time consolidation is used only for global evaluation and is distinct from each method's training-time aggregation or personalization mechanism. Ditto is instead evaluated using the native global model maintained during training.

`FedC`^{2}-Specific Settings. For each target client, H denotes the number of selected collaborators excluding the target itself. We use $H = 25$ in the main experiments. The compatibility-kernel bandwidth σ is set to the mean pairwise compatibility distance across all client pairs, rather than to a fixed constant. All reported experimental partitions yield $\sigma > 0$; the current implementation rejects the degenerate zero-bandwidth case. For class-wise complementarity refinement, we set $\eta = 0.05$ and $c_{\max} = 0.25$. Unless explicitly varied in the sensitivity studies, these settings are kept fixed throughout all experiments.

B.4 Baseline Implementations

For clarity, we use the original method names in the main paper and omit the common FedAvg/DLLP wrappers from the displayed baseline names. For the LLP-based methods, we use standard sample-count-weighted FedAvg as the federated aggregation scheme. For the FL-based and personalized FL methods, we use the basic DLLP objective for local proportion-based training while retaining the method-specific federated optimization or personalization mechanism. Table 5 summarizes the corresponding Fed-LLP implementations and key hyperparameters. All compared methods use the same dataset splits, client partitions, LLP bags, backbone, communication-round budget, and evaluation protocol unless a method-specific component requires otherwise.

Table 5: Implementation of the compared baselines in the Fed-LLP setting. The method names shown in the main paper omit the common FedAvg/DLLP wrappers for readability.

Method	Category	Fed-LLP Implementation	Key Hyperparameters
DLLP	LLP-based	FedAvg with DLLP	–
PVC	LLP-based	FedAvg with PVC	–
LLP-DC	LLP-based	FedAvg with LLP-DC	$\lambda = 0.5$, $\tau_{\text{conf}} = 0.6$
FedProx	FL-based	DLLP with FedProx regularization	$\mu = 0.01$
FedWMSAM	FL-based	DLLP with FedWMSAM optimization	$\rho = 0.01$, $\lambda = 0.01$
FedDyn	FL-based	DLLP with FedDyn optimization	$\alpha = 0.1$
Ditto	Personalized FL	DLLP with Ditto personalization	$\lambda = 1.0$
FEROMA	Personalized FL	DLLP with target-specific FEROMA aggregation	5-round warm-up, 512 descriptor samples
PFedGraph	Personalized FL	DLLP with pFedGraph collaboration	$\lambda = 0.01$, $\alpha_{\text{scale}} = 0.08$

C Additional Related Work

C.1 Learning from Label Proportions

Learning from label proportions (LLP) considers a weakly supervised setting in which individual labels are unavailable and supervision is provided only through the class proportions of groups or bags. Early studies established the basic LLP formulation and investigated how instance-level predictors can be learned from such aggregate supervision (Quadrianto et al., 2008; Yu et al., 2014). With the development of deep learning, subsequent work extended LLP to neural models and explored how bag-level proportion information can be exploited for representation and classifier learning (Ardehaly & Culotta, 2017). These studies laid the foundation for modern LLP by showing that meaningful instance-level prediction can be learned without direct instance annotations, while also motivating more effective learning objectives and inference strategies under proportion-level supervision.

Recent LLP studies have moved beyond direct proportion matching toward stronger mechanisms for extracting instance-level supervision from bag-level proportions. Label-noise-based LLP (Zhang et al., 2022) formulates proportion supervision through noisy instance labels, while EasyLLP (Busa-Fekete et al., 2023) constructs debiased estimates of instance-level losses from aggregate labels. Belief-propagation-based learning (Havaldar et al., 2024) further infers latent instance labels, whereas PT-LLP (Liu et al., 2025a) progressively bridges bag-level proportion fitting and instance-level training. More recently, PVC (Ma et al., 2026b) enhances instance discrimination beyond direct proportion matching, while LLP-DC (Ma et al., 2026a) jointly exploits bag-level proportion constraints and instance-level pseudo-labeling. Together, these advances provide increasingly effective ways to recover fine-grained supervision from aggregate label proportions.

Theoretical studies have also investigated the learnability and generalization properties of LLP under aggregate supervision (Yu et al., 2014; Li et al., 2024), providing complementary understanding of when reliable instance-level prediction can be recovered from label proportions. Despite these advances, existing LLP research is predominantly formulated around a single learner and focuses on how to better exploit bag-level supervision itself. In federated environments, however, heterogeneous clients introduce an additional question of how information should be shared across different local proportion distributions. This motivates us to study hierarchical collaboration in Fed-LLP, from client-level compatibility to target-relative class-wise complementarity.

C.2 Federated Learning under Heterogeneity and Collaboration

Optimization under Statistical Heterogeneity. FedAvg (McMahan et al., 2017) establishes the standard federated optimization paradigm by iteratively aggregating locally trained models, but its performance can deteriorate when client data are strongly non-IID. Subsequent approaches improve robustness to statistical heterogeneity (Seo et al., 2026) through different optimization mechanisms. FedProx (Li et al., 2020) constrains local updates with a proximal regularizer, SCAFFOLD (Karimireddy et al., 2020) corrects client drift using control variates, and FedDyn (Acar et al., 2021) introduces dynamic regularization to better align local and global optimization. More recent methods further improve optimization and generalization under heterogeneous data through techniques such as calibrated distillation and sharpness-aware training (Shang et al., 2022; Li et al., 2025). These methods (Zhang & Azizan, 2026) primarily address how to stabilize federated optimization under

heterogeneous local objectives, whereas our work focuses on how heterogeneous clients should collaborate when supervision is available only through bag-level label proportions.

Personalized and Collaborative Federated Learning. Beyond stabilizing optimization under non-IID data, personalized and collaborative FL further adapts how information is shared across clients. Ditto (Li et al., 2021) jointly maintains global and personalized objectives, while FedFomo (Zhang et al., 2020) learns client-specific combinations of models from other participants. pFedGraph (Ye et al., 2023) further represents inter-client relationships through an inferred collaboration graph, and FEROMA (Li et al., 2026) adapts aggregation according to client distribution profiles. These approaches (Pramanik et al., 2026) demonstrate the benefit of moving beyond uniform global aggregation toward client-dependent collaboration. However, their collaboration is primarily organized at the client or model level (Yu et al., 2025), leaving open how collaborator contributions should be differentiated across classes, particularly when only bag-level proportion supervision is available.

Class-Aware Federated Collaboration. Recent personalized FL methods have further explored collaboration beyond the whole-model level. FedPAC (Xu et al., 2023) combines shared representation learning with client-specific classifier collaboration, while cwFedAvg (Lee & Choi, 2025) constructs class-specific global models according to client class distributions and recombines them for personalization (Liu et al., 2025b). FedIC (Zheng et al., 2026) further groups same-class prototypes into intra-class subspaces and performs classifier collaboration within the resulting groups. These methods demonstrate the benefit of exploiting class-dependent (Du et al., 2025) structures in heterogeneous FL (Thapa & Li, 2025), but rely on standard supervised settings where instance labels, class distributions, or class prototypes can be directly derived. In contrast, Fed-LLP provides only bag-level class proportions (Fan et al., 2025). `FedC`^{2} therefore organizes collaboration hierarchically: it first establishes a compatibility-guided collaboration context and then refines collaborator contributions according to target-relative class complementarity.

D Additional Experimental Results

D.1 Additional Qualitative Results

We provide additional qualitative prediction examples to complement the visualization in the main paper. Figs. 7–10 show representative bags from CIFAR-10, SVHN, Fashion-MNIST, and Mini-ImageNet, together with their bag-level class proportions, sampled instances, and top-1 predictions from `FedC`^{2} and representative baselines. These examples further illustrate how different methods recover instance-level classes from proportion-only supervision under the Fed-LLP setting. For Mini-ImageNet, the qualitative examples are drawn from LLP bags over the training classes and are separate from the class-disjoint episodic evaluation described in Appendix B.1.



Figure 7: **Additional qualitative prediction examples on CIFAR-10.**

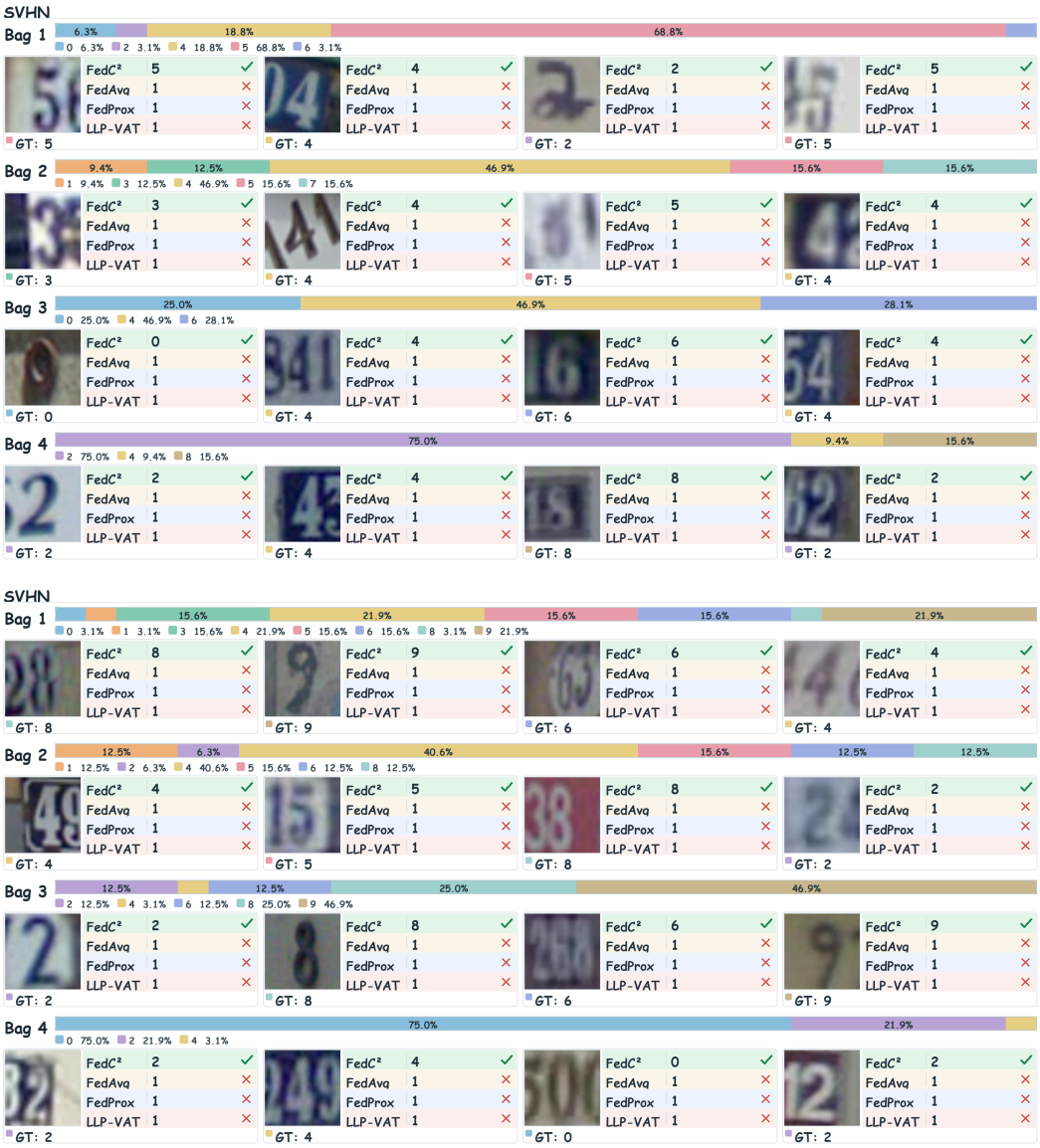


Figure 8: **Additional qualitative prediction examples on SVHN.**

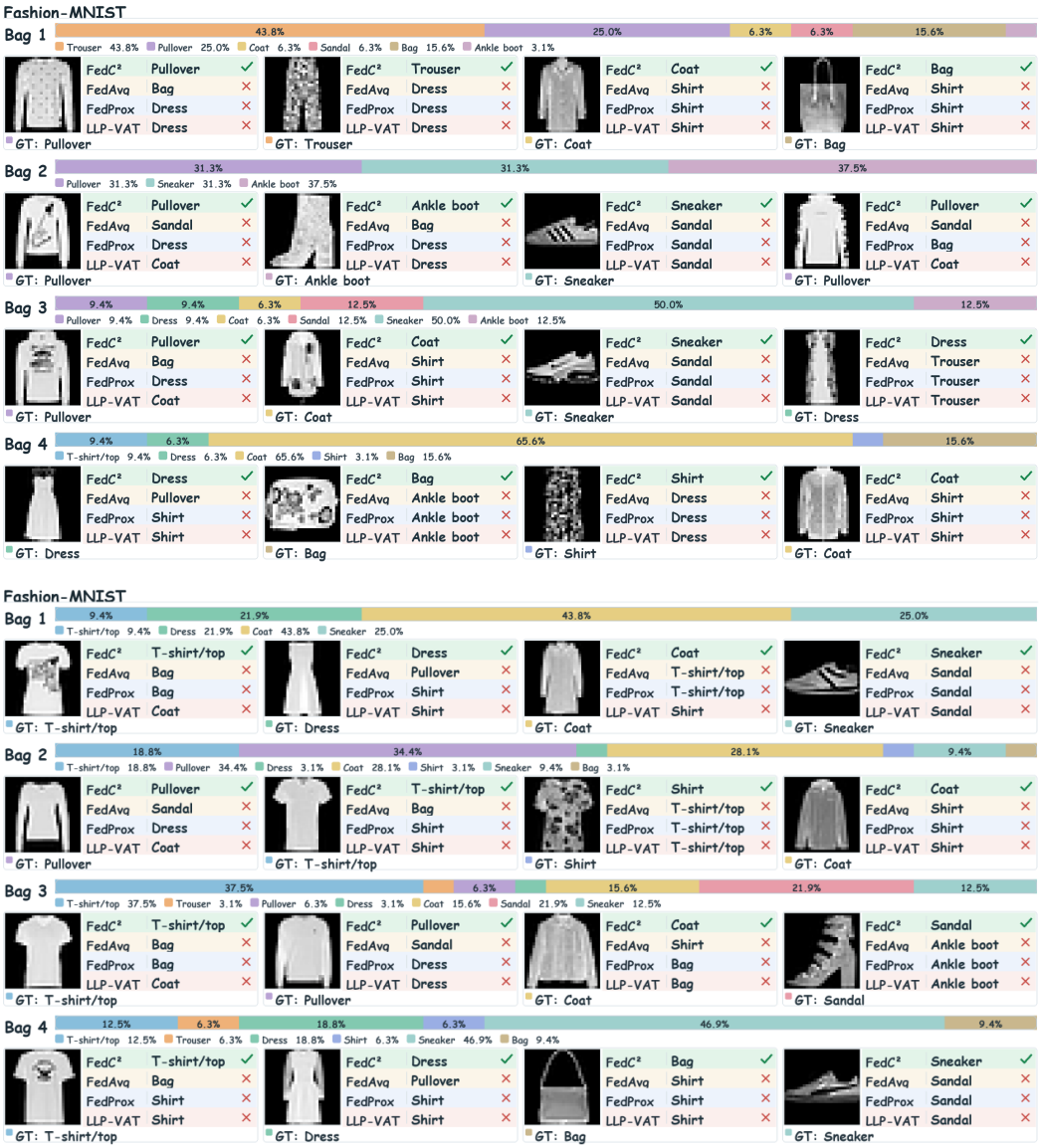


Figure 9: **Additional qualitative prediction examples on Fashion-MNIST.**



Figure 10: **Additional qualitative prediction examples on Mini-ImageNet.**

Table 6: Numerical results for the mechanism analyses in Fig. 4.

Dataset	Transfer Correlation		Gain by Target Coverage		
	Pearson	Spearman	Low	Medium	High
CIFAR-10	0.433	0.524	1.09	0.70	0.24
SVHN	0.473	0.556	2.80	3.02	2.94
Fashion-MNIST	0.272	0.337	1.14	−0.33	−0.41

D.2 Numerical Results for Main-Text Analyses

We provide the complete numerical results underlying the mechanism, sensitivity, and generalization analyses in Figs.3, 4 and 5. Unless otherwise specified, $\beta = 0.25$. We report test ACC (%): top-1 ACC for CIFAR-10, SVHN, and Fashion-MNIST, and 5-way 5-shot episodic ACC for Mini-ImageNet following the protocol in Appendix B.1.

Leave-One-Collaborator-Out Analysis. To examine whether the complementarity score reflects realized class-wise transfer, we perform a leave-one-collaborator-out analysis at the validation-selected checkpoint. For each target client i and class k, we consider the highest-scoring, median-scoring, and, when available, lowest strictly positive-scoring collaborators according to $c_{ij,k}$ within the selected collaborator set $\mathcal{N}_i$. For a selected collaborator j, we construct the counterfactual classifier by setting only its aggregation weight for class k to zero and renormalizing the remaining class-k weights. The target backbone and all other classifier rows are kept fixed, and no retraining is performed. We then define

$$\Delta\mathrm{ACC}_{ij,k} = \mathrm{ACC}_{\mathrm{full},i,k} - \mathrm{ACC}_{\mathrm{without}\ j,i,k}, \tag{77}$$

where both accuracies are evaluated on test samples whose ground-truth label is class k. Duplicate (i, j, k) combinations arising from the collaborator selection procedure are removed before computing the correlation statistics, and confidence intervals are estimated using 10,000 bootstrap resamples with target clients as the resampling unit. As shown in Table 6, $c_{ij,k}$ exhibits positive Pearson and Spearman correlations with $\Delta\mathrm{ACC}_{ij,k}$ across all three datasets, indicating that the proposed complementarity score provides a meaningful ranking signal for realized class-wise transfer.

Target-Coverage Stratification. To examine how the benefit of class-wise refinement depends on the target client's class coverage, we consider all target client–class pairs within each dataset and rank them by their target-side class coverage. The pairs are divided into three approximately equal-frequency groups corresponding to the bottom, middle, and top thirds of the coverage distribution, denoted as low, medium, and high coverage, respectively. For each group, we report the average refinement gain,

$$\mathrm{Gain} = \mathrm{ACC}_{\mathrm{full}} - \mathrm{ACC}_{\mathrm{compat}}, \tag{78}$$

where $\mathrm{ACC}_{\mathrm{compat}}$ denotes the accuracy obtained with compatibility-guided collaboration before class-wise complementarity refinement. As shown in Table 6, CIFAR-10 exhibits a clear decrease in gain from low- to high-coverage groups, while Fashion-MNIST shows a positive gain primarily for low-coverage classes and negative gains for the medium- and high-coverage groups. In contrast, SVHN shows comparable gains across all three coverage levels. These results suggest that low target coverage can indicate greater benefit from class-wise refinement on some datasets, but this pattern is not universal.

Sensitivity Analysis. Table 7 reports the complete numerical results underlying Fig. 3. Increasing the communication budget generally improves performance before saturation, whereas the effects of federation scale are non-monotonic and dataset dependent. For the collaboration design, moderate collaborator-set sizes and relatively conservative refinement strengths generally provide more favorable performance.

Generalization and Sensitivity Analysis. Table 8 reports the complete numerical results underlying Fig. 5. The collaboration mechanism remains effective with different local LLP learners, while its performance is relatively stable across the tested complementarity caps. The effects of bag size and bag composition are more dataset dependent, with moderate settings providing a favorable overall operating point.

Table 7: Complete numerical results for the sensitivity analyses in Fig. 3. Default settings are highlighted in blue.

Setting	Value	CIFAR-10	SVHN	Fashion-MNIST	Mini-ImageNet
(a) Communication Rounds T					
T	30	16.83	20.17	59.50	30.12
T	40	19.77	21.24	67.50	30.07
T	**50**	21.71	21.82	71.05	30.91
T	60	22.15	22.55	72.57	30.01
T	70	22.39	23.61	73.06	29.50
(b) Number of Clients C					
C	30	17.96	17.49	73.14	30.56
C	40	15.95	19.38	69.63	29.71
C	**50**	21.71	21.82	71.05	30.91
C	60	14.40	18.57	70.17	29.90
C	70	16.88	19.73	68.00	29.96
(c) Collaborator-Set Size H					
H	12	16.45	19.35	63.23	31.40
H	**25**	21.71	21.82	71.05	30.91
H	38	17.96	19.46	64.19	29.89
H	49	16.99	18.49	63.33	30.54
(d) Refinement Strength η					
η	**0.05**	21.71	21.82	71.05	30.91
η	0.10	21.95	21.73	70.89	28.74
η	0.15	21.02	19.44	65.97	30.32
η	0.20	21.55	20.76	67.22	30.47

Table 8: Complete numerical results for the generalization and sensitivity analyses in Fig. 5. Default settings are highlighted in blue.

(a) Compatibility with Local LLP Learners			
Local Learner	**FedAvg**	**FedC^2**	**Gain**
PVC	22.56	24.48	+1.92
LLP-DC	25.47	26.93	+1.46
LLP-VAT	30.32	30.39	+0.07

(b) Complementarity Cap $c_{\max}$				
$c_{\max}$	**CIFAR-10**	**SVHN**	**Fashion-MNIST**	**Mini-ImageNet**
0.175	21.70	21.81	70.87	30.31
0.200	21.76	21.36	71.26	30.31
0.225	21.95	21.97	71.07	30.31
0.250	21.71	21.82	71.05	30.91

(c) Bag Size B				
B	**CIFAR-10**	**SVHN**	**Fashion-MNIST**	**Mini-ImageNet**
16	21.88	16.89	68.46	30.48
32	21.71	21.82	71.05	30.91
64	19.72	22.89	69.57	30.82
96	16.43	20.56	53.62	30.56

(d) Bag Composition β_{bag}				
β_{bag}	**CIFAR-10**	**SVHN**	**Fashion-MNIST**	**Mini-ImageNet**
0.1	15.82	17.38	59.09	29.48
0.5	21.71	21.82	71.05	30.91
1.0	18.86	16.41	52.46	29.31
2.0	19.51	17.60	64.59	30.86

D.3 Communication Dynamics.

We further examine the validation trajectory of FedC^2 throughout training on all four datasets. Figure 11 visualizes the complete communication dynamics, while Table 9 reports the corresponding validation ACC at aligned five-round checkpoints. CIFAR-10, SVHN, and Fashion-MNIST im-

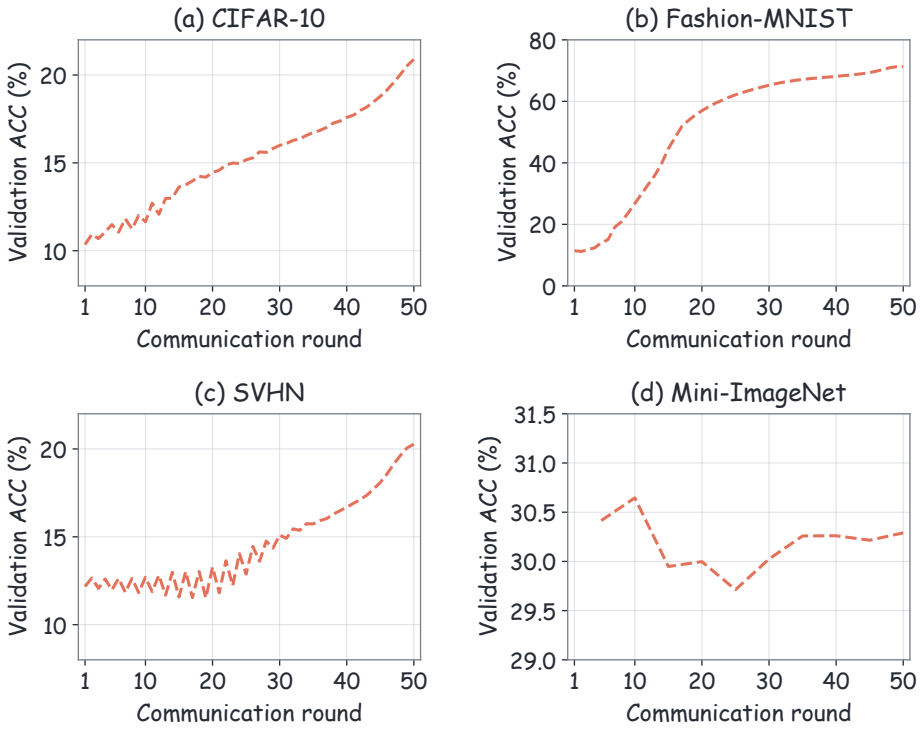


Figure 11: **Validation trajectories of FedC^2 over communication rounds.** CIFAR-10, SVHN, and Fashion-MNIST are evaluated after every communication round, while Mini-ImageNet is evaluated every five rounds.

Table 9: Validation ACC (%) of FedC^2 at aligned communication checkpoints corresponding to Fig. 11.

Round	CIFAR-10	SVHN	Fashion-MNIST	Mini-ImageNet
5	11.49	11.98	14.00	30.42
10	11.65	12.70	26.78	**30.65**
15	13.63	11.58	44.62	29.95
20	14.45	13.28	56.94	30.00
25	15.17	12.88	62.13	29.71
30	16.00	15.10	65.23	30.02
35	16.72	15.73	67.14	30.26
40	17.58	16.68	68.14	30.26
45	18.78	18.09	69.35	30.22
50	**20.89**	**20.27**	**71.33**	30.29

prove substantially over training, whereas Mini-ImageNet remains comparatively stable with mild fluctuations.