WHAT TO REPRESENT AND WHAT TO REMEMBER: COMPLEMENTARY REPLAY FOR FEDERATED CLASS-INCREMENTAL LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Exemplar replay-based federated class-incremental learning aims to preserve previously acquired knowledge by replaying selected samples while learning new classes. Existing methods independently score samples in the full feature space and replay those with the highest importance. However, they overlook complementarity within the limited replay set, resulting in redundant selections and insufficient coverage of the historical representation space accumulated across clients and tasks. To address these limitations, this study proposes a federated REpresentation Coverage-Aware repLay Learning framework, termed RECALL, which constructs a compact and complementary replay memory through global critical-feature identification and residual-based exemplar selection. Specifically, it consists of three modules. The global critical feature selection module selects feature coordinates for reconstructing class-balanced current and previous model representations, reducing feature redundancy. The complementary exemplar selection module combines local residual preselection with class-wise global coordination to reduce cross-client redundancy and improve replay coverage. The coverage-weighted feature distillation module constrains representation drift along the selected critical directions, preserving historical knowledge during incremental learning. We also bound full-class reconstruction error for fixed representations in the selected coordinates, accounting for local screening loss and client quotas. Experiments across three benchmark datasets show that RECALL consistently outperforms strong baselines by 1%–8%, validating the effectiveness of complementary and coverage-aware replay.

1 INTRODUCTION

Federated class-incremental learning (FCIL) enables distributed clients to collaboratively learn continually emerging classes without sharing private data while retaining knowledge acquired from previous tasks Dong et al. (2022); Babakniya et al. (2023); Du et al. (2026). This paradigm is particularly suitable for scenarios characterized by distributed data, expanding class spaces, and strict privacy requirements Khademi Nori et al. (2025); Iqbal et al. (2026); Dong et al. (2026). For example, autonomous vehicles across regions must collaboratively adapt their perception models without sharing onboard images Soares & Campos (2024); Yuan et al. (2023). During incremental learning, limited storage prevents clients from retaining all historical samples, while continued optimization for new classes alters existing representations and decision boundaries, causing catastrophic forgetting Li et al. (2024). Data heterogeneity across clients further amplifies local forgetting during model aggregation and weakens the global model's ability to recognize previously learned classes Qi et al. (2026c). Therefore, balancing the acquisition of new classes with the preservation of historical knowledge under limited resources is a central challenge in FCIL.

Existing methods preserve historical knowledge through regularization, generated samples, or stored exemplars. Regularization constrains model updates using earlier parameters or predictions Chen et al. (2025); Yoon et al. (2021); Bakman et al. (2024), while generative replay reconstructs past data distributions without retaining raw samples Zhou et al. (2026); Nguyen et al. (2026); Rong et al. (2025); Tran et al. (2024). Exemplar replay retains real samples for subsequent training Li et al. (2025a); Qi et al. (2026b). However, its limited memory makes the collective information in

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.

the retained set crucial: individually important samples can cover similar directions. At the same time, coordinating candidates in the full representation space communicates feature coordinates that may encode overlapping information. This motivates selecting both the coordinates used to describe historical representations and the samples that span those coordinates.

These two choices address different sources of redundancy. For replay, we likewise use reconstruction quality to assess what a compact selection preserves. The sum of squared ℓ_2 reconstruction errors, equivalently the squared Frobenius norm of the matrix residual, measures the representation energy left outside the retained span. It captures errors in the vectors' directions as well as their lengths, and bounds the output error caused by reconstruction under any fixed linear readout with bounded weights. This motivates a geometric objective for data and feature selection for FCIL.

In this paper, we propose the federated representation coverage-aware replay learning framework, RECALL, which couples feature compression with complementary exemplar selection under a fixed per-task memory budget. Global critical feature selection identifies coordinates that represent the class-balanced current and previous model views. Local screening and global class-wise coordination then select samples that add directions beyond those already retained. Coverage-weighted distillation uses the feature-selection scores to protect these coordinates during subsequent learning. Theoretically, we establish when candidate residuals control full-class reconstruction error, accounting for directions lost during screening and candidates blocked by client quotas.

Extensive experiments on three benchmark datasets include performance comparisons with state-of-the-art (SOTA) methods, ablation studies of key modules, hyperparameter sensitivity analyses, and case studies of replay coverage. The results demonstrate that RECALL consistently improves classification performance while constructing a more compact and representative replay memory, achieving gains of 1%–8% over SOTA methods.

2 Related Work

2.1 Regularization-based methods

Regularization-based methods mitigate catastrophic forgetting by constraining model optimization to reduce interference with previously acquired knowledge. These constraints are typically imposed on model parameters, update directions, or historical predictions Ma et al. (2022); Li et al. (2025b); Gao et al. (2024); Yu et al. (2025). For instance, FedWeIT decomposes model weights into shared and task-adaptive components and selectively transfers relevant knowledge across clients Yoon et al. (2021). FOT restricts aggregated updates to directions orthogonal to the principal subspaces of previous tasks Bakman et al. (2024). FedET employs lightweight enhancers to absorb emerging knowledge and introduces enhancer distillation to balance previously acquired and new knowledge under client heterogeneity Liu et al. (2023). Despite their storage efficiency, these methods remain dependent on the historical knowledge encoded in previous models or constrained subspaces.

2.2 Generative replay methods

Generative replay methods mitigate catastrophic forgetting by reconstructing historical data distributions without retaining real samples. The synthesized samples are replayed with current data or used for knowledge distillation Iqbal et al. (2026); Nguyen et al. (2024); Liang et al. (2024). TARGET trains a server-side generator to recover historical distributions and transfers previous knowledge through exemplar-free distillation Zhang et al. (2023). LANDER introduces label-text embeddings as semantic anchors to guide data-free generation and improve the quality of synthesized samples Tran et al. (2024). CAN leverages client expertise to guide data synthesis and adapts replay to client-specific forgetting, making generative replay more effective under federated heterogeneity Rong et al. (2025). Despite avoiding raw-data storage, they depend heavily on generation quality and incur additional computational costs for training and maintaining generative models.

2.3 Exemplar replay methods

Exemplar replay methods retain a small subset of real historical samples and revisit them alongside current data to preserve previously acquired knowledge Lu et al. (2024); Li et al. (2025a); Qi et al.

(2026a). Since storage is strictly limited, their effectiveness largely depends on how representative samples are selected. Re-Fed coordinates sample caching according to global and local importance, retaining samples that are informative to both local and aggregated models Li et al. (2024). FedCBDR introduces global coordination for class-balanced memory construction and adjusts the training objective to reduce imbalance between historical and emerging classes Qi et al. (2026c). CLIF fuses Lewis weights derived from multiple representation views to improve exemplar importance estimation under client heterogeneity Qi et al. (2026b). However, these methods mainly rank candidates by individual importance in the full representation space without explicitly considering their marginal contributions to the retained set, which may result in redundant replay and insufficient coverage across clients and tasks.

3 PRELIMINARIES

We consider N clients and S sequential tasks. At stage s, client i receives $\mathcal{A}_i^s = \{(x_{i,j}^s, y_{i,j}^s)\}_{j=1}^{m_i^s}$ from a new label set $\mathcal{Y}^s$, with $\mathcal{Y}^s \cap \mathcal{Y}^r = \varnothing$ for $s \neq r$. It retains at most B exemplars from each completed task: $|\mathcal{M}_i^r| \leq B$. Its training data are $\widehat{\mathcal{A}}_i^s = \mathcal{A}_i^s \cup \mathcal{E}_i^{s-1}$, where $\mathcal{E}_i^{s-1} = \bigcup_{r<s} \mathcal{M}_i^r$. Clients train the shared model $F_{\mathbf{w}}$ locally and send updates for server aggregation. The stage-s model predicts all labels in $\mathcal{Y}^{1:s} = \bigcup_{r \leq s} \mathcal{Y}^r$.

Representations are column vectors; data matrices store their transposes as rows. We write $\|\cdot\|_2$ for the Euclidean or matrix spectral norm, $\|\cdot\|_F$ for the Frobenius norm, and $\mathbf{A}^\dagger$ for the Moore-Penrose pseudoinverse. All selection ties use a fixed order of feature indices or client-sample identifiers.

4 METHODOLOGY

RECALL first selects feature coordinates to describe the distributed representations, then selects exemplars in those coordinates, and finally uses their coverage weights to constrain subsequent representation drift. Algorithm 1 gives the training sequence.

4.1 GLOBAL CRITICAL FEATURE SELECTION

Feature selection balances the current representation with the previous model's representation. Let $\mathbf{h}_i^s(x), \mathbf{h}_i^{s-1}(x) \in \mathbb{R}^d$ be the normalized features extracted after stage s by these two models. For each locally present class c, let $n_{i,c}^s = |\{x \in \widehat{\mathcal{A}}_i^s : y(x) = c\}|$ and $\xi_{i,c}^s = (n_{i,c}^s)^{-1/2}$. Classes with no local samples contribute no rows. Define

$$\mathbf{Z}_{i,\mathrm{cur}}^s = \left[\xi_{i,y(x)}^s \mathbf{h}_i^s(x)^\top\right]_{x \in \widehat{\mathcal{A}}_i^s}, \qquad \mathbf{Z}_{i,\mathrm{pre}}^s = \left[\xi_{i,y(x)}^s \mathbf{h}_i^{s-1}(x)^\top\right]_{x \in \widehat{\mathcal{A}}_i^s}. \tag{1}$$

Thus, each present client-class pair has equal total weight within a view. For $\alpha \in [0,1]$, set $\mathbf{W}_{i,\mathrm{cur}}^s = \sqrt{1-\alpha}\mathbf{Z}_{i,\mathrm{cur}}^s$ and $\mathbf{W}_{i,\mathrm{pre}}^s = \sqrt{\alpha}\mathbf{Z}_{i,\mathrm{pre}}^s$. Placing these blocks one above another gives the virtual matrix

$$\mathbf{Z}^s = \left[(\mathbf{W}_{1,\mathrm{cur}}^s)^\top\ (\mathbf{W}_{1,\mathrm{pre}}^s)^\top\ \cdots\ (\mathbf{W}_{N,\mathrm{cur}}^s)^\top\ (\mathbf{W}_{N,\mathrm{pre}}^s)^\top\right]^\top \in \mathbb{R}^{\nu_s \times d}, \quad \nu_s = 2\sum_i |\widehat{\mathcal{A}}_i^s|. \tag{2}$$

Each client contributes its current-view rows followed by its previous-view rows, with the same d feature columns throughout. At $s = 1$, only the current blocks are used, with weight one and $\nu_1 = \sum_i |\mathcal{A}_i^1|$. This matrix is a mathematical representation of the distributed data; the server never assembles it.

Given an integer budget $1 \leq d_s \leq d$, the feature reconstruction objective is

$$\min_{\substack{\mathcal{F} \subseteq \{1,\ldots,d\} \\ |\mathcal{F}| \leq d_s}} \mathcal{E}_{\mathrm{feat}}^s(\mathcal{F}), \qquad \mathcal{E}_{\mathrm{feat}}^s(\mathcal{F}) = \min_{\mathbf{A} \in \mathbb{R}^{|\mathcal{F}| \times d}} \|\mathbf{Z}^s - \mathbf{Z}_{:,\mathcal{F}}^s \mathbf{A}\|_F^2. \tag{3}$$

To solve this problem without collecting local representations, RECALL performs federated residual feature selection. Initialize $\mathcal{F}_0^s = \varnothing$ and $\mathbf{R}_{i,v}^{(0)} = \mathbf{W}_{i,v}^s$. At iteration $q = 0, \ldots, d_s - 1$, aggregate

$$\rho_j^{(q)} = \sum_{i,v} \|\mathbf{R}_{i,v}^{(q)}[:,j]\|_2^2, \qquad j_{q+1} \in \arg\max_{j \notin \mathcal{F}_q^s} \rho_j^{(q)}. \tag{4}$$

If the maximum is positive, add j_{q+1} and remove its global projection from every block:

$$\mathbf{u}_{i,v}^{(q)} = \frac{\mathbf{R}_{i,v}^{(q)}[:, j_{q+1}]}{\sqrt{\rho_{j_{q+1}}^{(q)}}}, \qquad \mathbf{a}^{(q)} = \sum_{i,v} (\mathbf{u}_{i,v}^{(q)})^\top \mathbf{R}_{i,v}^{(q)},$$
$$\mathbf{R}_{i,v}^{(q+1)} = \mathbf{R}_{i,v}^{(q)} - \mathbf{u}_{i,v}^{(q)} \mathbf{a}^{(q)}, \quad \mathcal{F}_{q+1}^s = \mathcal{F}_q^s \cup \{j_{q+1}\}. \tag{5}$$

The server aggregates both the scores and projection coefficients. With exact arithmetic and the same tie rule, these are precisely the blocks of centralized residual-norm pivoting; Appendix D proves this implementation equivalence. Selection stops at the budget or a zero residual, yielding $\mathcal{F}^s = \{j_1, \ldots, j_{k_s}\}$ with $k_s \le d_s$. Note that, this is a greedy approximation to Equation 3.

4.2 Complementary Exemplar Selection

The selected coordinates specify what to represent; exemplar selection determines which samples span those coordinates under the memory quotas. For $x \in \mathcal{A}_i^s$, define

$$\mathbf{g}_i^s(x) = \mathrm{Norm}\left(\begin{bmatrix} \sqrt{1-\alpha}\, \mathbf{h}_i^s(x)_{\mathcal{F}^s} \\ \sqrt{\alpha}\, \mathbf{h}_i^{s-1}(x)_{\mathcal{F}^s} \end{bmatrix} \right) \in \mathbb{R}^{p_s}, \qquad p_s = 2k_s, \tag{6}$$

where $\mathrm{Norm}(\mathbf{z}) = \mathbf{z}/\|\mathbf{z}\|_2$ for $\mathbf{z} \neq 0$ and $\mathrm{Norm}(0) = 0$. At $s = 1$, only the current block is used and $p_1 = k_1$. Unlike the vertical stacking in Equation 2, this concatenation assigns one temporal vector to each sample.

Let $\mathcal{A}_{i,c}^s = \{x \in \mathcal{A}_i^s : y(x) = c\}$ and $\mathcal{C}_i^s = \{c : \mathcal{A}_{i,c}^s \neq \varnothing\}$. Client i distributes its budget as evenly as sample availability permits, using integer quotas

$$\sum_{c \in \mathcal{C}_i^s} b_{i,c}^s = \min(B, |\mathcal{A}_i^s|), \qquad 0 \le b_{i,c}^s \le |\mathcal{A}_{i,c}^s|. \tag{7}$$

Unused quotas of small classes are redistributed to classes with remaining samples. Local residual selection retains a candidate set $\mathcal{C}_{i,c}^{s,\mathrm{cand}} \subseteq \mathcal{A}_{i,c}^s$ of size at most $\min(|\mathcal{A}_{i,c}^s|, \lceil \eta b_{i,c}^s \rceil)$, where $\eta \ge 1$.

For global coordination, fix s, c and suppress these indices. Let $\mathcal{C} = \bigcup_i \mathcal{C}_i^{\mathrm{cand}}$ be the uploaded candidates, $\iota(x)$ their originating client, and $b_i = b_{i,c}^s$. For the selected set $\mathcal{M}_t$ after t additions, let $\mathbf{P}_t = \mathbf{Q}_t \mathbf{Q}_t^\top$ project onto the span of its temporal vectors, with $\mathbf{P}_0 = 0$. The eligible set and residual score are

$$\mathcal{U}_t = \{x \in \mathcal{C} \setminus \mathcal{M}_t : |\mathcal{M}_t \cap \mathcal{C}_{\iota(x)}^{\mathrm{cand}}| < b_{\iota(x)}\},$$
$$\Delta_t(x) = \|(\mathbf{I} - \mathbf{P}_t)\mathbf{g}(x)\|_2^2, \qquad x_t \in \arg\max_{x \in \mathcal{U}_t} \Delta_t(x). \tag{8}$$

Here $\mathbf{g}(x) = \mathbf{g}_{\iota(x)}^s(x)$. While $\mathcal{U}_t \neq \varnothing$, add x_t and update the orthonormal basis. Zero-residual additions may fill the remaining quotas but do not expand the span. The same rule constructs local candidate sets using only local class samples and the enlarged budget. Clients transmit candidate vectors, labels, and identifiers; raw samples remain local.

The score in Equation 8 evaluates one eligible candidate, whereas replay should represent the full class, including samples that were never uploaded. We therefore measure E_t, the sum of squared distances from all class-c vectors to the span of the t retained exemplars. A small E_t means that little representation energy remains uncovered. Let $\mathbf{G} \in \mathbb{R}^{n \times p_s}$ store all n stage-s, class-c vectors as rows, and let $\mathbf{H} \in \mathbb{R}^{m \times p_s}$ be its submatrix of m uploaded candidates. Define

$$E_t = \|\mathbf{G}(\mathbf{I} - \mathbf{P}_t)\|_F^2, \quad \mathbf{P}_H = \mathbf{H}^\dagger \mathbf{H},$$
$$\varepsilon_{\mathrm{scr}}^2 = \|\mathbf{G}(\mathbf{I} - \mathbf{P}_H)\|_F^2, \quad \kappa = \|\mathbf{G}\mathbf{H}^\dagger\|_2^2. \tag{9}$$

Here $\mathbf{P}_H$ projects onto the span of all uploaded candidates. Even retaining every candidate leaves error $\varepsilon_{\mathrm{scr}}^2$: this is the information lost during screening that global selection cannot recover. The factor κ converts errors measured on candidates into full-class errors, accounting for directions that carry more energy in the full data. Quotas impose a second restriction. The set $\mathcal{B}_t = \mathcal{C} \setminus (\mathcal{M}_t \cup \mathcal{U}_t)$

contains unselected candidates from clients whose quotas are full. Their residual energy is Ω_t; among the m_t eligible candidates, δ_t is the largest residual:

$$\Omega_t = \sum_{x \in \mathcal{B}_t} \Delta_t(x), \qquad m_t = |\mathcal{U}_t|, \qquad \delta_t = \max_{x \in \mathcal{U}_t} \Delta_t(x), \tag{10}$$

with $\delta_t = 0$ for an empty eligible set. The following theorem shows when these candidate-level quantities certify coverage of the full class and when the next feasible selection must improve it.

Theorem 1. *Fix the representations and candidate sets, and use the exact projectors in Equation 8. If* $\operatorname{rank}(\mathbf{H}) > 0$*, then* $\kappa \geq 1$ *and every feasible selection prefix satisfies*

$$\varepsilon_{\mathrm{scr}}^2 \leq E_t \leq \varepsilon_{\mathrm{scr}}^2 + \kappa(m_t\delta_t + \Omega_t). \tag{11}$$

For $m_t > 0$*, the next residual-maximizing addition satisfies*

$$E_{t+1} \leq E_t - \frac{[E_t - \varepsilon_{\mathrm{scr}}^2 - \kappa\Omega_t]_+}{\kappa m_t}, \qquad [a]_+ = \max(a, 0). \tag{12}$$

Remark 1 The first bound controls how much of the full class the retained memory fails to represent. Good coverage follows when screening loss, the eligible residual bound $m_t\delta_t$, and quota-blocked energy are small, with moderate amplification κ. For $m_t > 0$, the second bound guarantees progress whenever $E_t > \varepsilon_{\mathrm{scr}}^2 + \kappa\Omega_t$. Without blocking, the excess above the screening loss contracts by a factor at most $1 - 1/(\kappa m_t)$ per eligible addition. This identifies when feasible selections improve full-class reconstruction; the guarantee concerns fixed representations, not later training loss. Note that, if $\mathbf{H}$ has rank zero, every selectable vector is zero and $E_t = \varepsilon_{\mathrm{scr}}^2 = \|\mathbf{G}\|_F^2$.

Appendix E gives the proof and shows why neither volume growth nor a larger candidate pool alone guarantees near-optimal coverage.

4.3 Coverage-Weighted Feature Distillation

The feature-pivot scores in Equation 4 determine which coordinates receive stronger protection at the next stage. For the k_s positive pivots selected after stage s, define

$$\omega_q^s = \frac{k_s \rho_{j_q}^{(q-1)}}{\sum_{r=1}^{k_s} \rho_{j_r}^{(r-1)}}, \qquad q = 1, \ldots, k_s. \tag{13}$$

These weights have mean one. During local training at stage $s \geq 2$, use the stored coordinates $\mathcal{F}^{s-1} = \{j_q^{s-1}\}_{q=1}^{k_{s-1}}$ and weights $\boldsymbol{\omega}^{s-1}$. Let $\mathbf{h}_{\mathbf{w}_i}$ be the current local representation, $\mathbf{h}_{\overline{\mathbf{w}}^{s-1}}$ the fixed previous-stage teacher, and $\mathcal{B}_{i,\mathrm{rep}}^s$ the replay portion of the local minibatch. Then

$$\mathcal{L}_{i,\mathrm{cfd}}^s = \frac{1}{|\mathcal{B}_{i,\mathrm{rep}}^s| k_{s-1}} \sum_{x \in \mathcal{B}_{i,\mathrm{rep}}^s} \sum_{q=1}^{k_{s-1}} \omega_q^{s-1} \left([\mathbf{h}_{\mathbf{w}_i}(x)]_{j_q^{s-1}} - [\mathbf{h}_{\overline{\mathbf{w}}^{s-1}}(x)]_{j_q^{s-1}} \right)^2. \tag{14}$$

Set this term to zero at the first stage, for an empty replay minibatch, or when $k_{s-1} = 0$. Let $\mathcal{L}_{i,\mathrm{ce}}^s$ be the mean cross-entropy on the current and replay minibatch over $\mathcal{Y}^{1:s}$. The local objective is

$$\mathcal{L}_i^s(\mathbf{w}_i) = \mathcal{L}_{i,\mathrm{ce}}^s(\mathbf{w}_i) + \lambda_{\mathrm{cfd}} \mathcal{L}_{i,\mathrm{cfd}}^s(\mathbf{w}_i), \qquad \lambda_{\mathrm{cfd}} \geq 0, \tag{15}$$

where $\mathcal{L}_{\mathrm{ce}}$ enables adaptation to the accumulated class space, while $\mathcal{L}_{\mathrm{cfd}}$ selectively constrains representation drift along globally important directions.

Overall, global critical feature selection reduces the dimensions required to describe the distributed temporal representation space, complementary exemplar selection decreases redundancy within the constrained memory, and coverage-weighted feature distillation preserves the selected directions during incremental optimization. These components jointly connect representation coverage during memory construction with knowledge preservation during model training.

Table 1: Top-1 accuracy (%) of RECALL and competing methods on CIFAR-10, CIFAR-100, and Tiny-ImageNet under different client and task configurations. The 3/5/5 and 5/10/10 settings denote the numbers of tasks for the three datasets, respectively. Data heterogeneity is controlled by $\beta \in \{0.5, 1.0\}$. Results are reported as mean±standard deviation over three random seeds, with the best results highlighted in **bold**.

	CIFAR-10		CIFAR-100		Tiny-ImageNet		CIFAR-10		CIFAR-100		Tiny-ImageNet	
	20 Clients (3/5/5 Task-Splits)						**40 Clients (3/5/5 Task-Splits)**					
	$\beta = 0.5$	$\beta = 1.0$	$\beta = 0.5$	$\beta = 1.0$	$\beta = 0.5$	$\beta = 1.0$	$\beta = 0.5$	$\beta = 1.0$	$\beta = 0.5$	$\beta = 1.0$	$\beta = 0.5$	$\beta = 1.0$
FineTune	$28.47_{\pm 2.2}$	$36.94_{\pm 1.7}$	$18.32_{\pm 2.4}$	$17.22_{\pm 1.4}$	$10.48_{\pm 2.1}$	$10.08_{\pm 1.7}$	$32.64_{\pm 2.1}$	$36.83_{\pm 1.5}$	$17.36_{\pm 1.5}$	$18.57_{\pm 1.7}$	$9.94_{\pm 1.8}$	$10.59_{\pm 1.8}$
FedLwF	$34.72_{\pm 1.7}$	$43.16_{\pm 2.1}$	$24.73_{\pm 1.6}$	$31.49_{\pm 1.6}$	$15.77_{\pm 1.5}$	$16.82_{\pm 1.8}$	$43.71_{\pm 2.2}$	$45.57_{\pm 1.6}$	$25.77_{\pm 1.9}$	$31.26_{\pm 2.0}$	$15.53_{\pm 1.0}$	$15.71_{\pm 1.5}$
TARGET	$30.17_{\pm 1.5}$	$36.89_{\pm 1.7}$	$19.42_{\pm 1.2}$	$24.43_{\pm 1.4}$	$14.39_{\pm 1.3}$	$15.24_{\pm 1.4}$	$32.11_{\pm 1.2}$	$35.21_{\pm 1.4}$	$22.73_{\pm 1.8}$	$25.36_{\pm 1.9}$	$14.94_{\pm 1.8}$	$15.47_{\pm 1.2}$
LANDER	$33.52_{\pm 1.2}$	$40.63_{\pm 1.5}$	$23.32_{\pm 1.2}$	$32.87_{\pm 1.8}$	$16.62_{\pm 0.9}$	$17.84_{\pm 1.3}$	$36.54_{\pm 1.6}$	$40.72_{\pm 2.5}$	$39.38_{\pm 1.6}$	$43.83_{\pm 1.4}$	$15.26_{\pm 1.1}$	$16.53_{\pm 1.5}$
FCIDF	$52.45_{\pm 2.1}$	$53.14_{\pm 1.6}$	$32.88_{\pm 1.9}$	$40.57_{\pm 2.2}$	$15.83_{\pm 1.4}$	$18.22_{\pm 1.4}$	$55.63_{\pm 1.8}$	$59.93_{\pm 1.8}$	$40.08_{\pm 1.6}$	$40.74_{\pm 1.1}$	$14.66_{\pm 1.1}$	$15.92_{\pm 1.6}$
Re-Fed+	$53.19_{\pm 2.1}$	$61.42_{\pm 1.8}$	$43.52_{\pm 1.7}$	$44.83_{\pm 1.8}$	$19.21_{\pm 1.6}$	$19.72_{\pm 1.4}$	$66.62_{\pm 2.1}$	$70.11_{\pm 2.1}$	$44.73_{\pm 1.2}$	$46.82_{\pm 1.8}$	$18.12_{\pm 1.4}$	$18.48_{\pm 1.3}$
FedCBDR	$57.82_{\pm 2.2}$	$65.82_{\pm 2.4}$	$42.86_{\pm 1.7}$	$45.66_{\pm 1.3}$	$18.37_{\pm 1.2}$	$19.49_{\pm 1.5}$	$70.14_{\pm 2.6}$	$73.18_{\pm 1.5}$	$45.92_{\pm 1.5}$	$46.77_{\pm 1.6}$	$19.21_{\pm 1.7}$	$19.69_{\pm 1.3}$
CLIF	$57.77_{\pm 2.1}$	$65.83_{\pm 1.7}$	$43.87_{\pm 1.8}$	$46.19_{\pm 1.8}$	$19.27_{\pm 0.9}$	$19.64_{\pm 1.2}$	$70.31_{\pm 1.9}$	$72.63_{\pm 2.2}$	$46.19_{\pm 0.9}$	$46.92_{\pm 1.8}$	$19.37_{\pm 1.5}$	$20.34_{\pm 1.4}$
RECALL	$\mathbf{63.26}_{\pm 1.4}$	$\mathbf{72.88}_{\pm 1.8}$	$\mathbf{48.49}_{\pm 1.2}$	$\mathbf{49.22}_{\pm 1.2}$	$\mathbf{21.55}_{\pm 1.1}$	$\mathbf{22.08}_{\pm 1.4}$	$\mathbf{77.17}_{\pm 2.1}$	$\mathbf{80.18}_{\pm 1.8}$	$\mathbf{54.12}_{\pm 1.4}$	$\mathbf{54.86}_{\pm 1.6}$	$\mathbf{22.46}_{\pm 1.3}$	$\mathbf{22.85}_{\pm 1.1}$
	20 Clients (5/10/10 Task-Splits)						**40 Clients (5/10/10 Task-Splits)**					
	$\beta = 0.5$	$\beta = 1.0$	$\beta = 0.5$	$\beta = 1.0$	$\beta = 0.5$	$\beta = 1.0$	$\beta = 0.5$	$\beta = 1.0$	$\beta = 0.5$	$\beta = 1.0$	$\beta = 0.5$	$\beta = 1.0$
FineTune	$20.21_{\pm 1.5}$	$23.92_{\pm 1.2}$	$10.98_{\pm 1.1}$	$12.31_{\pm 1.5}$	$7.16_{\pm 0.6}$	$7.42_{\pm 0.8}$	$23.21_{\pm 1.1}$	$25.03_{\pm 1.1}$	$12.84_{\pm 1.5}$	$18.31_{\pm 0.9}$	$6.56_{\pm 0.8}$	$7.43_{\pm 1.1}$
FedLwF	$28.12_{\pm 1.3}$	$29.29_{\pm 1.6}$	$16.23_{\pm 1.1}$	$21.18_{\pm 0.9}$	$12.47_{\pm 1.4}$	$13.61_{\pm 1.2}$	$30.92_{\pm 1.1}$	$32.59_{\pm 1.8}$	$15.44_{\pm 1.6}$	$20.92_{\pm 1.1}$	$14.22_{\pm 1.3}$	$15.24_{\pm 0.8}$
TARGET	$22.19_{\pm 1.4}$	$26.02_{\pm 1.3}$	$13.21_{\pm 0.5}$	$17.73_{\pm 1.1}$	$8.92_{\pm 1.3}$	$9.55_{\pm 0.9}$	$26.58_{\pm 1.1}$	$30.82_{\pm 1.3}$	$13.17_{\pm 1.1}$	$15.88_{\pm 1.1}$	$10.05_{\pm 0.6}$	$11.41_{\pm 0.9}$
LANDER	$24.11_{\pm 0.8}$	$33.14_{\pm 1.2}$	$16.71_{\pm 1.1}$	$21.82_{\pm 1.2}$	$9.11_{\pm 0.6}$	$16.24_{\pm 0.9}$	$27.33_{\pm 0.8}$	$43.12_{\pm 1.1}$	$17.28_{\pm 1.6}$	$21.61_{\pm 0.8}$	$14.73_{\pm 1.1}$	$16.49_{\pm 0.9}$
FCIDF	$47.47_{\pm 1.2}$	$49.33_{\pm 1.6}$	$33.82_{\pm 1.1}$	$36.28_{\pm 0.8}$	$14.57_{\pm 1.2}$	$17.49_{\pm 1.2}$	$52.84_{\pm 1.9}$	$57.52_{\pm 1.7}$	$43.62_{\pm 1.6}$	$44.51_{\pm 1.2}$	$12.89_{\pm 1.1}$	$15.22_{\pm 1.3}$
Re-Fed+	$50.29_{\pm 2.1}$	$59.27_{\pm 1.8}$	$40.23_{\pm 1.4}$	$40.26_{\pm 1.1}$	$15.41_{\pm 1.2}$	$17.21_{\pm 1.2}$	$63.11_{\pm 2.5}$	$68.49_{\pm 1.8}$	$51.38_{\pm 1.9}$	$53.83_{\pm 1.3}$	$16.11_{\pm 0.5}$	$17.29_{\pm 0.7}$
FedCBDR	$50.89_{\pm 1.8}$	$66.32_{\pm 2.1}$	$39.64_{\pm 1.5}$	$40.27_{\pm 1.3}$	$15.11_{\pm 1.1}$	$18.12_{\pm 1.2}$	$65.83_{\pm 2.1}$	$74.83_{\pm 2.0}$	$50.67_{\pm 1.5}$	$52.94_{\pm 1.6}$	$16.11_{\pm 0.8}$	$16.93_{\pm 1.1}$
CLIF	$51.74_{\pm 1.8}$	$62.27_{\pm 1.6}$	$41.34_{\pm 1.5}$	$41.86_{\pm 1.2}$	$16.88_{\pm 1.2}$	$18.01_{\pm 1.3}$	$64.21_{\pm 1.5}$	$70.37_{\pm 1.3}$	$53.69_{\pm 1.3}$	$54.17_{\pm 0.9}$	$18.11_{\pm 1.2}$	$18.35_{\pm 1.4}$
RECALL	$\mathbf{57.60}_{\pm 0.7}$	$\mathbf{71.94}_{\pm 1.6}$	$\mathbf{47.42}_{\pm 1.2}$	$\mathbf{47.80}_{\pm 1.4}$	$\mathbf{18.13}_{\pm 0.8}$	$\mathbf{19.20}_{\pm 1.2}$	$\mathbf{70.66}_{\pm 1.8}$	$\mathbf{76.94}_{\pm 1.2}$	$\mathbf{54.85}_{\pm 1.2}$	$\mathbf{58.93}_{\pm 1.0}$	$\mathbf{19.71}_{\pm 0.9}$	$\mathbf{19.48}_{\pm 1.2}$

5 EXPERIMENTS

5.1 EXPERIMENT SETTINGS

Datasets. Following prior studies Qi et al. (2026c); Li et al. (2024), we evaluate RECALL on three widely used benchmarks,, including CIFAR-10 Krizhevsky et al. (2009), CIFAR-100 Krizhevsky et al. (2009), and Tiny-ImageNet Le & Yang (2015). Client data heterogeneity is simulated using a Dirichlet distribution $Dir(\beta)$, where a smaller β indicates a higher degree of heterogeneity.

Implementation Details. For fair comparison, all methods use ResNet-18 with a linear classifier. We consider 20 and 40 clients, sample 20% per round, and train for 200 rounds with two local epochs. Data heterogeneity is controlled by $\beta \in \{0.5, 1.0\}$. CIFAR-10 is divided into 3/5 tasks, and CIFAR-100 and Tiny-ImageNet into 5/10 tasks. SGD is used with a learning rate of 0.01 and weight decay of 10^{-5}; batch sizes are 128 and 64 for 20 and 40 clients, respectively. For the 20-client setting, each client retains 60/40, 40/20, and 80/40 exemplars per task for CIFAR-10, CIFAR-100, and Tiny-ImageNet, respectively, and are halved for 40 clients. For RECALL, the feature ratio, candidate multiplier, and distillation weight are tuned over $\{0.2, 0.5, 0.8\}$, $\{1, 2, 4\}$, and $\{0.1, 0.5, 1, 5\}$, respectively. The temporal weight is set 0.5. Other method-specific settings follow the original papers and publicly available implementations. All experiments are evaluated using Top-1 accuracy.

5.2 PERFORMANCE COMPARISON

To examine the effectiveness of RECALL, we benchmark it against eight representative methods: FineTune Tran et al. (2024), FedLwF Li & Hoiem (2017), TARGET Zhang et al. (2023), LANDER Tran et al. (2024), FCIDF Lu et al. (2024), Re-Fed+ Li et al. (2025a), FedCBDR Qi et al. (2026c), and CLIF Qi et al. (2026b). Their performance under different datasets, client scales, task splits, and heterogeneity levels is summarized in Table 1

- RECALL achieves the highest Top-1 accuracy in all settings, outperforming the strongest competing method by 1–8 percentage points, with an average improvement of 4.23 points.
- RECALL maintains consistent advantages under both $\beta = 0.5$ and $\beta = 1.0$, yielding average improvements of 4.25 and 4.21 percentage points, respectively. This demonstrates its robustness to different degrees of non-IID data.
- RECALL remains superior when the number of incremental tasks increases from 3/5/5 to 5/10/10. Its average gains under the two task configurations reach 5.01 and 3.45 percentage points, confirming its ability to preserve historical knowledge over extended learning sequences.
- RECALL consistently performs best with both 20 and 40 clients, even though the per-client replay budget is reduced in the 40-client setting. These results indicate that its complementary replay strategy scales effectively to more distributed and memory-constrained FCIL scenarios.

Table 2: Ablation study of the three components in RECALL with 20 clients under the 5/10/10 task-split setting and different levels of data heterogeneity ($\beta = \{0.5, 1.0\}$).

Variants	GCFS	CES	CWFD	CIFAR-10		CIFAR-100		Tiny-ImageNet	
				$\beta = 0.5$	$\beta = 1.0$	$\beta = 0.5$	$\beta = 1.0$	$\beta = 0.5$	$\beta = 1.0$
FineTune	–	–	–	$20.21_{\pm 1.5}$	$23.92_{\pm 1.2}$	$10.98_{\pm 1.1}$	$12.31_{\pm 1.5}$	$7.16_{\pm 0.6}$	$7.42_{\pm 0.8}$
RECALL w/o GCFS & CWFD	–	✓	–	$50.76_{\pm 1.6}$	$63.28_{\pm 1.4}$	$40.93_{\pm 1.3}$	$41.19_{\pm 1.5}$	$16.38_{\pm 0.9}$	$17.21_{\pm 0.6}$
RECALL w/o GCFS	–	✓	✓	$51.37_{\pm 1.7}$	$64.09_{\pm 1.1}$	$41.11_{\pm 1.4}$	$42.51_{\pm 1.6}$	$16.85_{\pm 0.7}$	$17.53_{\pm 0.6}$
RECALL w/o CES	✓	–	✓	$49.94_{\pm 2.4}$	$56.37_{\pm 1.7}$	$31.37_{\pm 1.4}$	$39.38_{\pm 1.1}$	$14.82_{\pm 0.9}$	$16.37_{\pm 0.6}$
RECALL w/o CWFD	✓	✓	–	$53.78_{\pm 2.2}$	$68.37_{\pm 1.9}$	$43.13_{\pm 1.5}$	$44.51_{\pm 1.5}$	$17.61_{\pm 0.9}$	$17.93_{\pm 1.1}$
RECALL	✓	✓	✓	**57.60**$_{\pm 0.7}$	**71.94**$_{\pm 1.6}$	**47.42**$_{\pm 1.2}$	**47.80**$_{\pm 1.4}$	**18.13**$_{\pm 0.8}$	**19.20**$_{\pm 1.2}$

5.3 Ablation Study

To assess the contribution of each component, we construct four variants of RECALL. Specifically, *w/o GCFS & CWFD* retains only CES to select complementary exemplars in the full feature space and performs replay using the classification loss; *w/o GCFS* performs exemplar selection and feature distillation over all feature dimensions; *w/o CES* replaces complementary exemplar selection with random sampling under the same memory budget; and *w/o CWFD* removes coverage-weighted feature distillation while retaining the selected exemplars. Table 2 reports the results, from which the following observations can be drawn:

- Removing any component consistently degrades accuracy across all datasets and heterogeneity levels, while the complete RECALL achieves the best results in every setting. This confirms the complementary effects of GCFS, CES, and CWFD.
- Replacing CES with class-balanced random selection causes the most substantial performance degradation, particularly on CIFAR-100. This demonstrates that selecting mutually complementary exemplars is essential for adequately covering the historical representation space.
- GCFS improves exemplar selection by eliminating redundant feature dimensions, whereas CWFD further constrains representation drift along the selected critical directions. Their combined use consistently outperforms CES alone, validating the effectiveness of critical-feature-guided selection and preservation.

5.4 Hyperparameter Sensitivity

We examine the sensitivity of RECALL to the feature-selection parameter $\rho \in \{0.2, 0.5, 0.8\}$, candidate multiplier $\eta \in \{1, 2, 4\}$, and distillation weight $\lambda \in \{0.1, 0.5, 1.0, 5.0\}$. For each parameter value, we vary the remaining parameters over their candidate settings. Experiments are conducted on CIFAR-100 with 20 clients, 10 tasks, and $\beta = 0.5$. As shown in Fig. 1, $\rho = 2$ achieves the best performance by balancing information preservation and feature redundancy. Increasing η from 1 to 2 improves accuracy, while further enlargement provides limited benefits, indicating that a moderate candidate pool is sufficient for complementary selection. Performance generally improves with

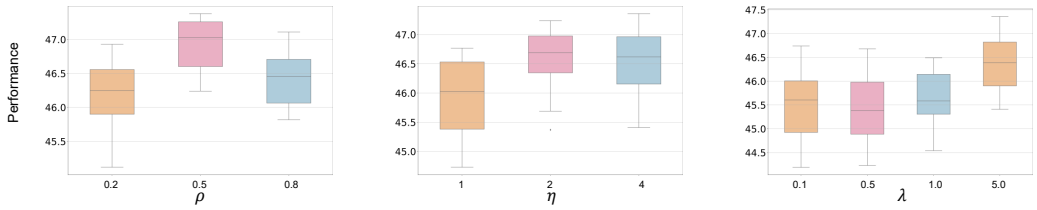

Figure 1: Hyperparameter sensitivity of RECALL to the feature-selection ratio ρ, candidate multiplier η, and distillation weight λ on CIFAR-100 with 20 clients, 10 tasks, and $\beta = 0.5$.

λ, confirming the effectiveness of feature distillation in mitigating representation drift. Overall, RECALL remains stable across all settings and consistently outperforms the strongest baseline.

5.5 Scalability Analysis

To evaluate scalability in large-scale FCIL, we increase the number of clients to 100 while maintaining a participation rate of 20% and keeping the other settings unchanged. Each dataset is divided into five sequential tasks. Each client retains 8, 8, and 16 exemplars per task on CIFAR-10, CIFAR-100, and Tiny-ImageNet, respectively. Table 3 reports the results on three datasets under $\beta \in \{0.5, 1.0\}$. RECALL maintains clear advantages as client data become more fragmented. On CIFAR-10, it outperforms the strongest baseline by 9.04 and 7.36 percentage points under the two heterogeneity levels, respectively. Similar improvements of 4.47 and 4.55 points are observed on Tiny-ImageNet. These results demonstrate that the proposed feature and exemplar coordination mechanisms remain effective with a large number of distributed clients.

Table 3: Top-1 accuracy (%) with 100 clients.

Methods	CIFAR-10		CIFAR-100		Tiny-ImageNet	
	$\beta = 0.5$	$\beta = 1.0$	$\beta = 0.5$	$\beta = 1.0$	$\beta = 0.5$	$\beta = 1.0$
Re-Fed+	$58.94_{\pm 2.1}$	$66.49_{\pm 1.8}$	$40.73_{\pm 1.4}$	$41.91_{\pm 1.5}$	$16.54_{\pm 1.3}$	$17.42_{\pm 1.5}$
FedCBDR	$65.16_{\pm 1.7}$	$72.47_{\pm 1.6}$	$42.57_{\pm 1.1}$	$43.73_{\pm 1.1}$	$18.85_{\pm 1.4}$	$19.63_{\pm 1.2}$
CLIF	$67.71_{\pm 2.4}$	$73.38_{\pm 2.2}$	$43.12_{\pm 1.7}$	$44.81_{\pm 1.1}$	$19.38_{\pm 1.2}$	$19.62_{\pm 1.4}$
RECALL	$\mathbf{76.75}_{\pm 1.6}$	$\mathbf{80.74}_{\pm 1.9}$	$\mathbf{44.35}_{\pm 1.5}$	$\mathbf{45.83}_{\pm 1.3}$	$\mathbf{23.85}_{\pm 1.2}$	$\mathbf{24.18}_{\pm 1.1}$

Table 4: Computational and time costs.

Method	CIFAR-10		Tiny-ImageNet	
	Time (s)	GPU Mem.	Time (s)	GPU Mem.
LANDER	27.378	3604	29.748	3987
Re-Fed+	6.352	566	10.966	673
FedCBDR	0.854	682	0.882	736
CLIF (2-views)	2.251	1449	4.033	1546
CLIF (5-views)	3.781	1449	7.711	1546
CLIF (10-views)	6.825	1449	14.207	1546
RECALL	0.735	388	0.938	254

5.6 Computational and Time Costs

This section evaluates the time and peak GPU memory required for replay construction with 20 clients and a ResNet-18 backbone. All runtime and memory measurements are conducted on a single NVIDIA 4090 under the same hardware and software environment. As shown in Table 4, RECALL is the fastest method on CIFAR-10 and remains comparable to FedCBDR on Tiny-ImageNet. It also consumes the least GPU memory on both datasets, benefiting from compact feature-space operations. The lower memory usage on Tiny-ImageNet results from its smaller class-wise candidate groups during replay construction. Overall, RECALL achieves efficient replay construction with limited computational and memory overhead.

5.7 Quantitative Analysis of Replay Complementarity and Coverage

This section evaluates the complementarity of exemplars selected by different replay-based FCIL methods and their coverage of historical representations across clients. CRS measures the effective dimensional diversity of the replay memory, with a higher value indicating less redundancy. Client RCE Mean and Max quantify the average and worst-case residual energy not captured by the replay subspace, respectively. As shown in Fig. 2, RECALL consistently achieves the highest CRS and the lowest Client RCE Mean and Max across both tasks. These results demonstrate that RECALL reduces exemplar redundancy while providing broader and more consistent coverage across clients.

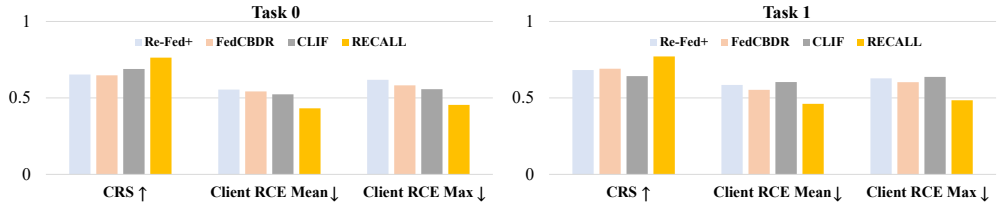

Figure 2: Replay-quality comparison after Tasks 0 and 1 on CIFAR-100 with 20 clients and five incremental tasks. Higher CRS and lower Client RCE Mean and Max indicate better replay complementarity and cross-client coverage.

5.8 Qualitative Analysis of Replay Complementarity and Coverage

This section qualitatively compares the spatial distributions of replay exemplars selected by different methods. Fig. 3 presents t-SNE Van der Maaten & Hinton (2008) visualizations for two representative clients, where circles denote client data representations and squares indicate the selected replay exemplars. CLIF tends to select multiple exemplars from several locally dense regions while leaving some peripheral and disconnected representation branches insufficiently covered. Under a limited replay budget, such local concentration introduces redundant information and reduces the effective coverage of historical knowledge. In contrast, RECALL distributes exemplars more broadly across different classes and representation regions, including both central and boundary structures. Similar patterns can be observed for both clients, indicating that the improvement is not restricted to a particular local distribution. These results show that residual coverage-based selection reduces local redundancy and constructs a more complementary replay memory, consistent with the quantitative CRS and RCE results.

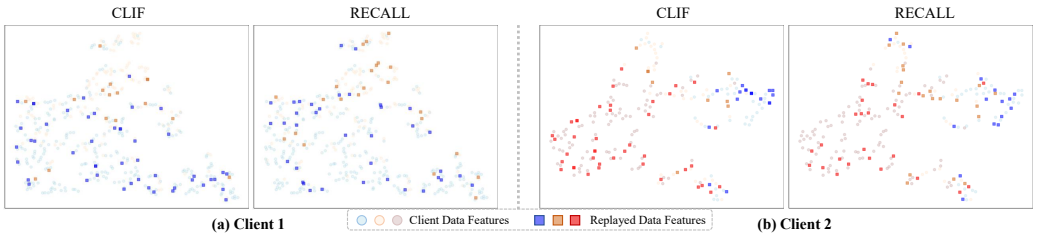

Figure 3: t-SNE visualization of client data representations (circles) and replayed exemplars (squares) selected by CLIF and RECALL for two representative clients. Colors denote different classes. Compared with CLIF, RECALL selects more complementary exemplars that provide broader coverage of the local representation spaces.

6 Conclusions and Future Work

To address replay redundancy and insufficient historical representation coverage under limited memory, this paper proposes RECALL, a complementary replay framework for FCIL. RECALL identifies globally shared critical features, selects exemplars according to their residual coverage contributions, and distills coverage-weighted historical directions during incremental learning. Theoretical analysis establishes the coverage property of the resulting replay memory. Extensive experiments on three datasets under multiple client and task configurations show that RECALL consistently outperforms strong baselines by 1%–8%.

Despite its effectiveness, several directions remain for future exploration. First, the current feature and candidate budgets are fixed across clients and tasks; adapting them to representation drift and data heterogeneity may further improve replay efficiency. Second, the present coverage estimation relies on linear representation subspaces. Extending RECALL to nonlinear or foundation-model representation spaces could improve its applicability to more complex FCIL scenarios.

AI USE STATEMENT

In this work, generative AI tools were used solely for grammar checking and language refinement. They were not used to formulate research ideas, develop methods, design experiments, analyze results, or draw conclusions. All AI-assisted revisions were carefully reviewed by the authors, who take full responsibility for the final content of this work.

REFERENCES

Sara Babakniya, Zalan Fabian, Chaoyang He, Mahdi Soltanolkotabi, and Salman Avestimehr. A data-free approach to mitigate catastrophic forgetting in federated class incremental learning for vision tasks. *Advances in Neural Information Processing Systems*, 36:66408–66425, 2023.

Yavuz Faruk Bakman, Duygu Nur Yaldiz, Yahya H. Ezzeldin, and Salman Avestimehr. Federated orthogonal training: Mitigating global catastrophic forgetting in continual federated learning. In *The Twelfth International Conference on Learning Representations*, 2024.

Leiming Chen, Dehai Zhao, Yongbiao Gao, Jiehan Zhou, and Tan Chee Wei. Fedmtl: Adaptive multi-teacher knowledge distillation for federated continual learning. *Knowledge-Based Systems*, pp. 115160, 2025.

Jiahua Dong, Lixu Wang, Zhen Fang, Gan Sun, Shichao Xu, Xiao Wang, and Qi Zhu. Federated class-incremental learning. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10154–10163. IEEE, 2022.

Jiahua Dong, Wenqi Liang, Yang Cong, Gan Sun, Lixu Wang, Henghui Ding, Yulun Zhang, and Luc Van Gool. Learning from each other: Generalized federated incremental semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(6):6929–6947, 2026.

Jie Du, Wenbing Chen, Peng Liu, Chi-Man Vong, Tianfu Wang, and C. L. Philip Chen. Federated class incremental learning method with high accuracy and extremely low communication cost based on broad learning system. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 56(8):4799–4812, 2026.

Xin Gao, Xin Yang, Hao Yu, Yan Kang, and Tianrui Li. Fedprok: Trustworthy federated class-incremental learning via prototypical feature knowledge transfer. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 4205–4214. IEEE, 2024.

Saeed Iqbal, Muhammad Attique Khan, Ghulam Mustafa, Sarra Ayouni, Abeer Aljohani, and Amir Hussain. Fedcapd: Federated class-incremental learning via capsule distillation and diffusion replay. *Neurocomputing*, pp. 133365, 2026.

Milad Khademi Nori, IL-MIN KIM, and Guanghui Wang. Federated class-incremental learning: A hybrid approach using latent exemplars and data-free techniques to address local and global forgetting. In *International Conference on Learning Representations*, volume 2025, pp. 84398–84411, 2025.

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

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

Yichen Li, Qunwei Li, Haozhao Wang, Ruixuan Li, Wenliang Zhong, and Guannan Zhang. Towards efficient replay in federated incremental learning. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12820–12829. IEEE, 2024.

Yichen Li, Haozhao Wang, Yining Qi, Wei Liu, and Ruixuan Li. Re-fed+: A better replay strategy for federated incremental learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(7):5489–5500, 2025a.

Yichen Li, Yuying Wang, Haozhao Wang, Yining Qi, Tianzhe Xiao, and Ruixuan Li. FedSSI: Rehearsal-free continual federated learning with synergistic synaptic intelligence. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 36151–36169, 2025b.

Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2935–2947, 2017.

Jinglin Liang, Jin Zhong, Hanlin Gu, Zhongqi Lu, Xingxing Tang, Gang Dai, Shuangping Huang, Lixin Fan, and Qiang Yang. Diffusion-driven data replay: A novel approach to combat forgetting in federated class continual learning. In *European Conference on Computer Vision*, pp. 303–319. Springer, 2024.

Chenghao Liu, Xiaoyang Qu, Jianzong Wang, and Jing Xiao. Fedet: A communication-efficient federated class-incremental learning framework based on enhanced transformer. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, pp. 3984–3992, 2023.

Yanyan Lu, Lei Yang, Hao-Rui Chen, Jiannong Cao, Wanyu Lin, and Saiqin Long. Federated class-incremental learning with dynamic feature extractor fusion. *IEEE Transactions on Mobile Computing*, 23(12):12969–12982, 2024.

Yuhang Ma, Zhongle Xie, Jue Wang, Ke Chen, and Lidan Shou. Continual federated learning based on knowledge distillation. In *Ijcai*, pp. 2182–2188, 2022.

Thinh Nguyen, Khoa D Doan, Binh T Nguyen, Danh Le-Phuoc, and Kok-Seng Wong. Overcoming catastrophic forgetting in federated class-incremental learning via federated global twin generator. *arXiv preprint arXiv:2407.11078*, 2024.

Thinh TH Nguyen, Khoa D Doan, Binh T Nguyen, Danh Le-Phuoc, and Kok-Seng Wong. When generator replay degrades: Projected rehearsal orchestration for heterogeneous federated class-incremental learning. *arXiv preprint arXiv:2606.15695*, 2026.

Zhuang Qi, Ying-Peng Tang, Lei Meng, Guoqing Chao, Lei Wu, Han Yu, and Xiangxu Meng. From selection to scheduling: Federated geometry-aware correction makes exemplar replay work better under continual dynamic heterogeneity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17494–17504, June 2026a.

Zhuang Qi, Ying-Peng Tang, Lei Meng, Xiaoxiao Li, Han Yu, and Xiangxu Meng. Cross-view lewis weight fusion empowering exemplar replay for federated class-incremental learning. In *Forty-third International Conference on Machine Learning*, 2026b.

Zhuang Qi, Ying-Peng Tang, Lei Meng, Han Yu, Xiaoxiao Li, and Xiangxu Meng. Class-wise balancing data replay for federated class-incremental learning. *Advances in Neural Information Processing Systems*, 38:145309–145334, 2026c.

Xuankun Rong, Jianshu Zhang, Kun He, and Mang Ye. CAN: Leveraging clients as navigators for generative replay in federated continual learning. In *Forty-second International Conference on Machine Learning*, 2025.

Ramprasaath R Selvaraju, Michael Cogswell, and et al. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, pp. 618–626, 2017.

Elton F de S Soares and Carlos Alberto V Campos. Towards continual federated learning of monocular depth for autonomous vehicles. In *2024 IEEE 100th Vehicular Technology Conference (VTC2024-Fall)*, pp. 1–5. IEEE, 2024.

Minh-Tuan Tran, Trung Le, Xuan-May Le, Mehrtash Harandi, and Dinh Phung. Text-enhanced data-free approach for federated class-incremental learning. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23870–23880. IEEE, 2024.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

Jaehong Yoon, Wonyong Jeong, Giwoong Lee, Eunho Yang, and Sung Ju Hwang. Federated continual learning with weighted inter-client transfer. In *International conference on machine learning*, pp. 12073–12086. PMLR, 2021.

Hao Yu, Xin Yang, Le Zhang, Hanlin Gu, Tianrui Li, Lixin Fan, and Qiang Yang. Handling spatial-temporal data heterogeneity for federated continual learning via tail anchor. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4874–4883. IEEE, 2025.

Liangqi Yuan, Yunsheng Ma, Lu Su, and Ziran Wang. Peer-to-peer federated continual learning for naturalistic driving action recognition. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 5250–5259. IEEE, 2023.

Jie Zhang, Chen Chen, Weiming Zhuang, and Lingjuan Lyu. Target: Federated class-continual learning via exemplar-free distillation. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4759–4770. IEEE, 2023.

Yurui Zhou, Jia Hu, Geyong Min, and M Jamal Deen. Regularization-enhanced federated class-incremental learning with generative replay for smart edge. *ACM Transactions on Autonomous and Adaptive Systems*, 2026.

APPENDIX

A CONTRIBUTIONS

This study made the following three contributions.

- We reveal that individually important exemplars can still form a redundant replay set and formulate replay construction as a coupled feature–sample coverage problem. Based on this insight, we propose `RECALL` to preserve diverse historical representations under a limited memory budget.
- We develop a complementary exemplar selection module that combines local residual preselection with global class-wise coordination. By evaluating each candidate according to its additional coverage beyond the retained memory, this module addresses set-level redundancy and improves balanced historical coverage across clients and tasks.
- We derive a coverage certificate and a per-selection progress bound for locally screened candidates under client quotas. The analysis separates screening loss, candidate-to-data amplification, and quota-blocked residual energy, with counterexamples showing why residual or volume growth alone does not guarantee near-optimal coverage.

B ALGORITHM

This section presents the pseudocode of the proposed algorithm, as shown in Alg. 1.

C COMMUNICATION COST ANALYSIS

We analyze the communication overhead introduced by `RECALL` beyond standard federated model aggregation. Let P denote the number of model parameters, R the number of communication rounds per stage, and N_p the number of participating clients per round. For dense parameter uploads and downloads, the common model-communication cost is

$$\mathcal{C}^s_{\text{model}} = \Theta\left(RN_pP\right) \tag{16}$$

floating-point values per stage. We therefore focus on the additional communication required for replay construction, which is performed once after each incremental stage.

Global critical feature selection. Let d be the original representation dimension and d_s the number of selected feature dimensions. At each feature-selection step, every client uploads a residual-energy vector and the corresponding local projection coefficients, both of which contain at most d values. The server aggregates these statistics and broadcasts the selected feature index and aggregated projection coefficients. Repeating this procedure for d_s selections introduces

$$\mathcal{C}^s_{\text{GCFS}} = \mathcal{O}\left(Ndd_s\right) \tag{17}$$

additional communicated values. These residual statistics are computed locally, and the complete representation matrices are never transmitted.

Complementary exemplar selection. Let $b^s_{i,c}$ denote the replay quota of class c at client i. The number of candidates uploaded by client i is

$$K^s_i = \sum_{c\in\mathcal{C}^s_i} |\mathcal{C}^{s,\text{cand}}_{i,c}| \le \sum_{c\in\mathcal{C}^s_i} \min\left(|\mathcal{A}^s_{i,c}|, \lceil \eta b^s_{i,c}\rceil\right), \tag{18}$$

where $\eta \ge 1$ is the candidate multiplier. Because $\sum_c b^s_{i,c} \le B$, we have

$$K^s_i \le \eta B + \min(B, |\mathcal{C}^s_i|) \le 2\eta B. \tag{19}$$

Only classes with $b^s_{i,c} > 0$ contribute a ceiling-rounding term, and there are at most B such classes. Each candidate is represented by a temporal vector of at most $2d_s$ values, corresponding to the

Algorithm 1 RECALL for Federated Class-Incremental Learning

Require: Number of clients N; number of stages S; communication rounds R; local datasets $\{\mathcal{A}_i^s\}$; per-task memory budget B; feature budget d_s; temporal coefficient α; candidate multiplier η; distillation coefficient λ_{cfd}.
Ensure: Global model $\mathbf{w}^S$ and local replay memories $\{\mathcal{E}_i^{S-1}\}_{i=1}^N$.

Initialize global model $\mathbf{w}^0$ and $\mathcal{E}_i^0 \leftarrow \varnothing$ for each client i.
for learning stage $s = 1$ to S **do**
  Initialize $\mathbf{w}^{s,0}$ from $\mathbf{w}^{s-1}$, expanding the classifier for the new classes.
  for communication round $r = 1$ to R **do**
    Sample a set of participating clients $\mathcal{C}_r$.
    for all $i \in \mathcal{C}_r$ **in parallel do**
      Initialize the local model $\mathbf{w}_i \leftarrow \mathbf{w}^{s,r-1}$.
      Construct $\widehat{\mathcal{A}}_i^s \leftarrow \mathcal{A}_i^s \cup \mathcal{E}_i^{s-1}$.
      if $s = 1$ **then**
        Update $\mathbf{w}_i$ on $\widehat{\mathcal{A}}_i^s$ using $\mathcal{L}_{i,\text{ce}}^s$.
      else
        Update $\mathbf{w}_i$ using $\mathcal{L}_{i,\text{ce}}^s + \lambda_{\text{cfd}}\mathcal{L}_{i,\text{cfd}}^s$ based on $\mathcal{F}^{s-1}$ and $\boldsymbol{\omega}^{s-1}$.
      end if
      Send the updated parameters $\mathbf{w}_i$ to the server.
    end for
    Aggregate $\{\mathbf{w}_i\}_{i\in\mathcal{C}_r}$ to obtain $\mathbf{w}^{s,r}$.
  end for
  Set $\mathbf{w}^s \leftarrow \mathbf{w}^{s,R}$.
  if $s < S$ **then**
    for all client $i = 1$ to N **in parallel do**
      Extract current and previous representations from $\widehat{\mathcal{A}}_i^s$.
      Construct class-balanced temporal blocks $\{\mathbf{W}_{i,v}^s\}$ as in Equation 2.
    end for
    $\mathcal{F}^s, \boldsymbol{\omega}^s \leftarrow$ GLOBALCRITICALFEATURESELECTION$(\{\mathbf{W}_{i,v}^s\}, d_s)$.
    for all client i and class c **in parallel do**
      Construct key temporal representations $\{\mathbf{g}_i^s(x) : x \in \mathcal{A}_{i,c}^s\}$ using $\mathcal{F}^s$.
      Allocate a class-wise budget $b_{i,c}^s$.
      $\mathcal{C}_{i,c}^{s,\text{cand}} \leftarrow$ LOCALRESIDUALSELECTION$(\mathcal{A}_{i,c}^s, \lceil \eta b_{i,c}^s \rceil)$.
    end for
    for all observed class c **do**
      $\{\mathcal{M}_{i,c}^s\}_{i=1}^N \leftarrow$ GLOBALRESIDUALCOORDINATION$(\{\mathcal{C}_{i,c}^{s,\text{cand}}\}_{i=1}^N, \{b_{i,c}^s\}_{i=1}^N)$.
    end for
    for all client $i = 1$ to N **do**
      Update $\mathcal{M}_i^s \leftarrow \bigcup_c \mathcal{M}_{i,c}^s$ and $\mathcal{E}_i^s \leftarrow \mathcal{E}_i^{s-1} \cup \mathcal{M}_i^s$.
    end for
  end if
  Preserve $\overline{\mathbf{w}}^s \leftarrow \mathbf{w}^s$ as the teacher for stage $s+1$.
end for

current and previous representations restricted to the selected dimensions. Therefore, the candidate-upload cost is

$$\mathcal{C}_{\text{CES}}^s = \mathcal{O}\left(2d_s \sum_{i=1}^N K_i^s\right). \tag{20}$$

After global coordination, the server returns at most B selected identifiers to each client, resulting in an additional $\mathcal{O}(NB)$ integer values. Raw samples remain on their originating clients throughout this process.

Ignoring lower-order metadata terms, the total additional communication of RECALL per stage is

$$\mathcal{C}_{\text{extra}}^s = \mathcal{O}\left(Ndd_s + N\eta Bd_s + NB\right). \tag{21}$$

Relative to the common model-communication cost, the additional communication ratio is approximately

$$\frac{\mathcal{C}^s_{\text{extra}}}{\mathcal{C}^s_{\text{model}}} = \mathcal{O}\left(\frac{Nd_s(d+\eta B)}{RN_pP}\right). \tag{22}$$

Because replay construction is conducted only once per stage, whereas model parameters are exchanged in every communication round, this overhead is amortized over the entire stage.

FineTune and FedLwF introduce no replay-construction communication and therefore provide a lower bound on additional communication. Coordinated replay baselines such as Re-Fed+, Fed-CBDR, and CLIF also exchange method-specific sample scores, candidate information, or multi-view statistics. In contrast, RECALL communicates residual statistics and a bounded candidate set whose size depends on ηB, rather than transmitting all local representations. Generative replay methods such as TARGET and LANDER may additionally exchange or maintain generator-related parameters, whose communication cost depends on the generator architecture.

The feature budget d_s and candidate multiplier η provide explicit communication–performance trade-offs. Increasing d_s preserves more representation information but linearly increases both feature-selection and candidate-transmission costs. Reducing d_s lowers communication, although an overly small feature subspace may increase the feature approximation error $\mathcal{E}_{\text{feat}}$. Increasing η enlarges the candidate pool and may improve cross-client coverage, but its communication cost also grows linearly with η. As shown in the hyperparameter analysis, moderate values of d_s and η provide a favorable balance between replay quality and communication efficiency.

D Federated Implementation of Residual Pivoting

The equivalence associated with Equation 5 concerns centralized residual-norm pivoting. It requires exact aggregation of both the scores and the projection coefficients.

Let $\mathbf{R}^{(q)}$ place the local residual blocks $\mathbf{R}^{(q)}_{i,v}$ one above another in the order shown in Equation 2. At $q=0$, this equals $\mathbf{Z}^s$ because the temporal weights are included in $\mathbf{W}^s_{i,v}$. If the local and centralized residuals agree at iteration q, then

$$\|\mathbf{R}^{(q)}[:,j]\|_2^2 = \sum_{i,v}\|\mathbf{R}^{(q)}_{i,v}[:,j]\|_2^2 = \rho_j^{(q)}. \tag{23}$$

The fixed tie rule therefore selects the same pivot. For a positive pivot, form $\mathbf{u}^{(q)}$ by stacking the local vectors $\mathbf{u}^{(q)}_{i,v}$ in the same order. This vector has norm one, and

$$(\mathbf{u}^{(q)})^\top \mathbf{R}^{(q)} = \sum_{i,v}(\mathbf{u}^{(q)}_{i,v})^\top \mathbf{R}^{(q)}_{i,v} = \mathbf{a}^{(q)}. \tag{24}$$

Consequently, each update in Equation 5 is a block of $\mathbf{R}^{(q+1)} = (\mathbf{I} - \mathbf{u}^{(q)}(\mathbf{u}^{(q)})^\top)\mathbf{R}^{(q)}$. Induction proves equality of the full pivot sequence and residuals. A zero maximum means that all remaining columns already lie in the selected span, so both procedures stop.

E Coverage Guarantees

E.1 Proof of Theorem 1

The proof uses nested orthogonal projections and a spectral-norm bound, followed by the actual feasible pivot rule.

Proof. Since $\mathcal{M}_t \subseteq \mathcal{C}$, the selected span lies inside the candidate span, so $\mathbf{P}_H\mathbf{P}_t = \mathbf{P}_t\mathbf{P}_H = \mathbf{P}_t$. Hence

$$\begin{aligned}\mathbf{G}(\mathbf{I}-\mathbf{P}_t) &= \mathbf{G}(\mathbf{I}-\mathbf{P}_H) + \mathbf{G}(\mathbf{P}_H-\mathbf{P}_t),\\ E_t &= \varepsilon^2_{\text{scr}} + \|\mathbf{G}(\mathbf{P}_H-\mathbf{P}_t)\|_F^2.\end{aligned} \tag{25}$$

The equality of squared norms holds because the two right-hand projectors have orthogonal ranges. Using $\mathbf{P}_H = \mathbf{H}^\dagger\mathbf{H}$ gives

$$\mathbf{G}(\mathbf{P}_H-\mathbf{P}_t) = \mathbf{G}\mathbf{H}^\dagger\mathbf{H}(\mathbf{I}-\mathbf{P}_t), \qquad E_t - \varepsilon^2_{\text{scr}} \le \kappa\|\mathbf{H}(\mathbf{I}-\mathbf{P}_t)\|_F^2. \tag{26}$$

Selected rows have zero residual. Partitioning the other candidate rows into $\mathcal{U}_t$ and $\mathcal{B}_t$ yields

$$\|\mathbf{H}(\mathbf{I}-\mathbf{P}_t)\|_F^2 = \sum_{x\in\mathcal{U}_t} \Delta_t(x) + \Omega_t \le m_t\delta_t + \Omega_t. \tag{27}$$

This proves Equation 11, including $m_t = 0$.

When $\delta_t > 0$, let $\mathbf{q}_t = (\mathbf{I}-\mathbf{P}_t)\mathbf{g}(x_t)/\sqrt{\delta_t}$. It is a unit vector orthogonal to the selected span, so $\mathbf{P}_{t+1} = \mathbf{P}_t + \mathbf{q}_t\mathbf{q}_t^\top$ and

$$E_t - E_{t+1} = \|\mathbf{G}\mathbf{q}_t\|_2^2 \ge |\mathbf{g}(x_t)^\top \mathbf{q}_t|^2 = \delta_t. \tag{28}$$

The inequality uses the presence of the selected record in $\mathbf{G}$. For $m_t > 0$, Equation 11 implies

$$\delta_t \ge \frac{[E_t - \varepsilon_{\text{scr}}^2 - \kappa\Omega_t]_+}{\kappa m_t}, \tag{29}$$

which proves Equation 12. If $\delta_t = 0$, the span is unchanged and the same bound holds with zero decrement.

Finally, because $\mathbf{H}$ is a row submatrix of $\mathbf{G}$, $\|\mathbf{G}\mathbf{z}\|_2 \ge \|\mathbf{H}\mathbf{z}\|_2$ for every $\mathbf{z}$. Choose any nonzero $\mathbf{u} \in \mathrm{range}(\mathbf{H})$, the column space of $\mathbf{H}$, and put $\mathbf{z} = \mathbf{H}^\dagger\mathbf{u}$. Then $\|\mathbf{G}\mathbf{H}^\dagger\mathbf{u}\|_2 \ge \|\mathbf{u}\|_2$, proving $\kappa \ge 1$ when $\mathrm{rank}(\mathbf{H}) > 0$. □

F Other Results

F.1 Definitions of Replay Complementarity and Coverage Metrics

This section formally defines the metrics used to evaluate replay complementarity and historical representation coverage. Let $\mathbf{G}_{i,c} \in \mathbb{R}^{n_{i,c}\times d}$ denote the historical representation matrix of class c at client i, and let $\mathbf{M}_c \in \mathbb{R}^{m_c\times d}$ contain the replay representations of class c selected across all clients. Writing $\mathbf{1}_a$ for the length-a all-ones vector, the global class mean is $\boldsymbol{\mu}_c = (\sum_i n_{i,c})^{-1}\sum_i \mathbf{G}_{i,c}^\top \mathbf{1}_{n_{i,c}}$ for a class with at least one record. We define

$$\widetilde{\mathbf{G}}_{i,c} = \mathbf{G}_{i,c} - \mathbf{1}_{n_{i,c}}\boldsymbol{\mu}_c^\top, \qquad \widetilde{\mathbf{M}}_c = \mathbf{M}_c - \mathbf{1}_{m_c}\boldsymbol{\mu}_c^\top. \tag{30}$$

Complementarity Rank Score. Let

$$\widetilde{\mathbf{M}}_c = \mathbf{U}_c\boldsymbol{\Sigma}_c\mathbf{V}_c^\top \tag{31}$$

be the singular value decomposition of the centered replay matrix. For a positive integer evaluation rank q, retain classes with $r_c = \min(q, m_c, d) > 0$ and positive leading singular-value energy. Define

$$r_c = \min(q, m_c, d), \qquad p_{c,j} = \frac{\sigma_{c,j}^2}{\sum_{\ell=1}^{r_c}\sigma_{c,\ell}^2}, \tag{32}$$

where $\sigma_{c,j}$ is the j-th singular value. The class-wise CRS is

$$\mathrm{CRS}_c^{(q)} = \frac{\exp\left(-\sum_{j=1}^{r_c} p_{c,j}\log(p_{c,j}+\varepsilon)\right)}{r_c}, \tag{33}$$

where $\varepsilon > 0$ is the numerical stabilizer and the logarithm is natural. Let $\mathcal{C}$ be the set of classes meeting the preceding validity conditions. If $\mathcal{C} \neq \varnothing$, the overall CRS is

$$\mathrm{CRS}^{(q)} = \frac{1}{|\mathcal{C}|}\sum_{c\in\mathcal{C}} \mathrm{CRS}_c^{(q)}. \tag{34}$$

A higher CRS indicates that the replay representations span more balanced directions and contain less redundancy.

Client Residual Coverage Error. Let $\mathbf{V}_{c,q}$ contain the leading

$$q_c = \min\left(q, \operatorname{rank}\left(\widetilde{\mathbf{M}}_c\right)\right) \tag{35}$$

right singular vectors of $\widetilde{\mathbf{M}}_c$. The residual coverage error for class c at client i is defined as

$$\mathrm{RCE}_{i,c}^{(q)} = \frac{\left\|\widetilde{\mathbf{G}}_{i,c}\left(\mathbf{I} - \mathbf{V}_{c,q}\mathbf{V}_{c,q}^{\top}\right)\right\|_F^2}{\left\|\widetilde{\mathbf{G}}_{i,c}\right\|_F^2}. \tag{36}$$

For this metric, let $\mathcal{C}_i$ contain the classes at client i with $\|\widetilde{\mathbf{G}}_{i,c}\|_F^2 > 0$. Client-level RCE averages over these classes:

$$\mathrm{RCE}_i^{(q)} = \frac{1}{|\mathcal{C}_i|}\sum_{c\in\mathcal{C}_i}\mathrm{RCE}_{i,c}^{(q)}. \tag{37}$$

We then report the average and worst-client errors:

$$\mathrm{ClientRCEMean}^{(q)} = \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}}\mathrm{RCE}_i^{(q)}, \tag{38}$$

$$\mathrm{ClientRCEMax}^{(q)} = \max_{i\in\mathcal{I}}\mathrm{RCE}_i^{(q)}, \tag{39}$$

where $\mathcal{I} = \{i : \mathcal{C}_i \neq \varnothing\}$ is assumed nonempty. Lower values indicate better average and worst-client coverage. If a class has no replay vectors or its centered replay matrix has rank zero, use the zero projector, giving RCE equal to one for nonzero-energy data. Zero-energy client–class blocks are excluded. An aggregate with no valid classes or clients is undefined and is not reported. We use $q = 32$ in all evaluations.

CRS measures spectral balance. Client RCE uses the same projection-residual form as E_t, but applies it to centered evaluation representations and a rank-truncated replay span. The theorem concerns the uncentered, selected-coordinate vectors used for selection and does not directly bound these evaluation metrics. Neither metric affects training or exemplar selection.

F.2 Performance Comparison of Different Methods across Incremental Stages

This section compares the performance of different methods throughout the incremental learning process. As shown in Fig. 4, all methods achieve comparable accuracy at the initial stage because no historical knowledge needs to be retained. After new tasks are introduced, their performance decreases substantially due to the increasing number of classes and the interference between newly acquired and previously learned knowledge. On CIFAR-10 and CIFAR-100, the performance of most methods gradually stabilizes or slightly recovers during the later stages, whereas the decline remains more pronounced on the more challenging Tiny-ImageNet dataset. Nevertheless, `RECALL` consistently maintains higher accuracy than the competing replay-based methods at all subsequent stages across the three datasets. Its advantage is particularly evident during the later stages, when the replay memory must simultaneously preserve knowledge from multiple previous tasks. By identifying globally informative feature dimensions and selecting exemplars according to their residual coverage contributions, `RECALL` avoids allocating the limited memory to redundant samples and retains a broader range of historical representations. Consequently, it provides more reliable knowledge preservation as the task sequence progresses. These results demonstrate that the proposed complementary replay strategy effectively alleviates performance degradation and improves long-term knowledge retention throughout federated class-incremental learning.

F.3 Performance Comparison under Different Client Participation Fractions per Round

This section evaluates the performance of different methods under varying client participation fractions. As shown in Fig. 5, increasing the participation fraction from 0.1 generally leads to improved accuracy, since the global model can incorporate information from a larger number of clients during each communication round. This improvement is particularly evident on CIFAR-10, where

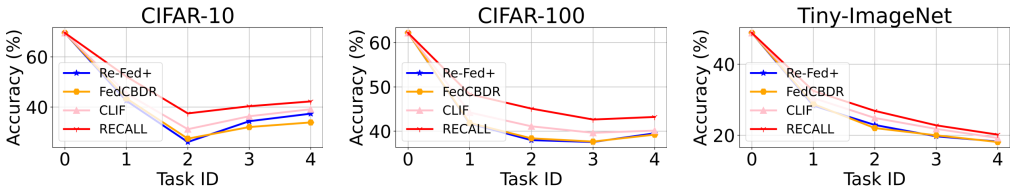

Figure 4: Performance comparison of different methods across incremental stages with $\beta = 0.5$, using 5 incremental tasks on three datasets. The accuracy at each stage is evaluated over all classes observed up to that stage. Higher values indicate better performance.

most methods achieve progressively better performance as the participation fraction increases. On CIFAR-100 and Tiny-ImageNet, the performance gains become relatively moderate at higher participation fractions and exhibit slight fluctuations, indicating that the benefit of involving additional clients varies across datasets. Despite these differences, RECALL consistently achieves the highest accuracy under all participation fractions on the three datasets. Its superiority is maintained from low to high participation levels and becomes particularly evident at moderate and high participation fractions. These results indicate that the effectiveness of RECALL does not depend on a specific client participation setting. By selecting complementary exemplars that cover diverse historical representations, RECALL effectively utilizes the knowledge provided by the participating clients and maintains robust performance under different levels of client availability.

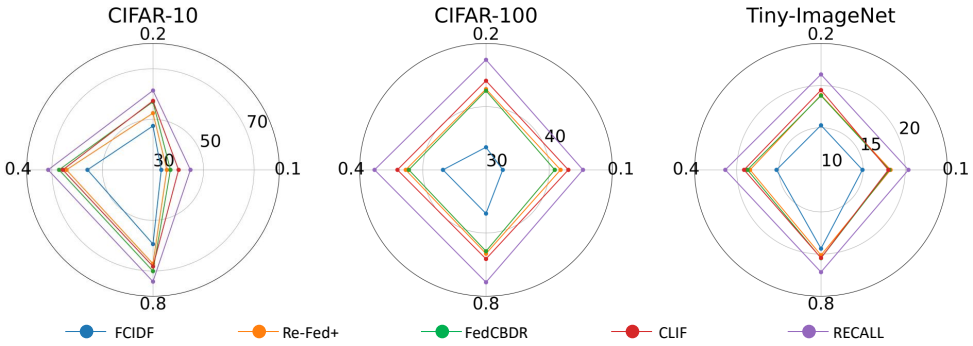

Figure 5: Performance comparison under client participation fractions of 0.1, 0.2, 0.4, and 0.8 per communication round with $\beta = 0.5$, using three incremental tasks on CIFAR-10 and five incremental tasks on CIFAR-100 and Tiny-ImageNet. Larger radial values indicate higher accuracy and better performance.

F.4 Performance Comparison under Different Local Training Epochs per Round

This section evaluates the robustness of different methods to the number of local training epochs per communication round. As shown in Fig. 6, increasing the number of local epochs from 1 to 4 generally improves the performance of all methods across the three datasets. When only one local epoch is performed, local models may not sufficiently learn the newly introduced classes before aggregation. Increasing the number of epochs allows more thorough local optimization and therefore produces more informative model updates. However, further increasing the number of local epochs to 8 leads to noticeable performance degradation on CIFAR-10 and Tiny-ImageNet, while the performance on CIFAR-100 largely saturates. Under heterogeneous client distributions, excessive local training may cause local models to overfit their current data and drift farther from the global optimization objective, thereby intensifying interference with previously acquired knowledge. In comparison, RECALL consistently achieves the best overall performance under different numbers of local epochs across all three datasets. Although its accuracy also decreases under excessive local training on CIFAR-10 and Tiny-ImageNet, it maintains a clear advantage over the competing methods. This is because its complementary replay memory provides broader coverage of historical representations during local

optimization, reducing the bias toward newly arrived data. These results demonstrate that RECALL remains effective under varying local training intensities, with four local epochs generally providing a favorable balance between local optimization and knowledge retention.

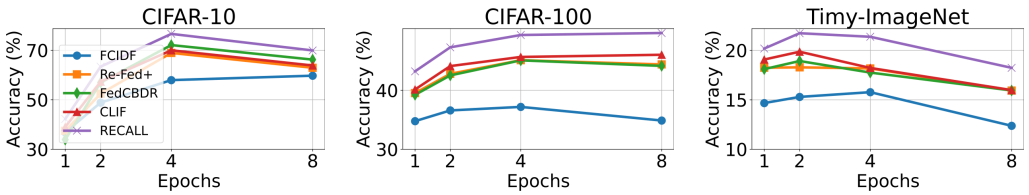

Figure 6: Performance comparison under different numbers of local training epochs per communication round with $\beta = 0.5$, using three incremental tasks on CIFAR-10 and five incremental tasks on CIFAR-100 and Tiny-ImageNet. Higher accuracy indicates better performance.

F.5 Performance Comparison under Different Memory Buffer Sizes

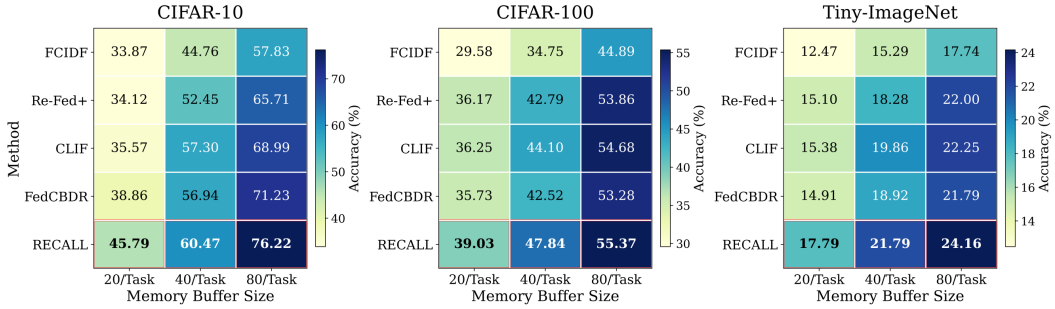

Figure 7: Performance comparison under memory buffer sizes of 20, 40, and 80 exemplars per task with $\beta = 0.5$, using three incremental tasks on CIFAR-10 and five incremental tasks on CIFAR-100 and Tiny-ImageNet. Each cell reports classification accuracy, with darker colors indicating better performance within each dataset.

This section evaluates the performance of different replay-based methods under varying memory buffer sizes. As shown in Fig. 7, increasing the number of stored exemplars per task consistently improves the accuracy of all methods, since a larger memory buffer preserves more historical information during incremental learning. However, when the available memory is severely constrained, the quality and complementarity of the selected exemplars become particularly important. Nevertheless, RECALL achieves the highest accuracy across all buffer sizes on the three datasets. Notably, with only 20 exemplars per task, RECALL obtains accuracies of 45.79%, 39.03%, and 17.79% on CIFAR-10, CIFAR-100, and Tiny-ImageNet, respectively, clearly outperforming the strongest competing methods. Its advantage is also maintained as the buffer size increases to 40 and 80 exemplars per task, indicating that its effectiveness is not limited to a particular memory budget. By prioritizing exemplars with complementary residual coverage contributions, RECALL avoids allocating the limited memory to redundant historical samples and retains a more representative replay set. These results demonstrate that RECALL makes more effective use of available memory and achieves robust performance under both restrictive and relatively sufficient replay budgets.

F.6 Qualitative Analysis of Attention on Previous-Task Samples

This section uses Grad-CAM Selvaraju et al. (2017) to qualitatively examine model attention on samples from previously learned tasks after learning subsequent tasks. Fig. 8 compares Grad-CAM heatmaps for the same previous-task images under three configurations: Finetune, Finetune with CES, and Finetune with both CES and CWFD. In several examples, Finetune produces dispersed or displaced responses, with strong activations appearing outside the main object regions. These patterns suggest that learning new tasks may weaken the model's focus on relevant visual structures in

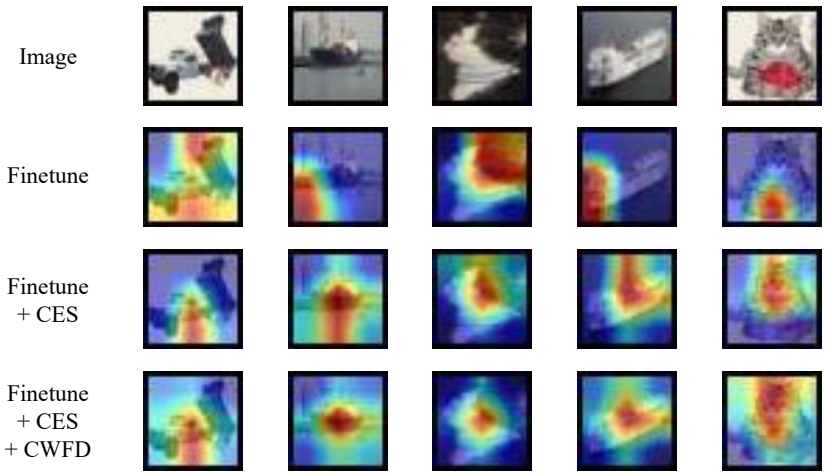

Figure 8: Activation maps for identical input images under Finetune, Finetune with CES, and Finetune with CES and CWFD. The first row shows the original images, and subsequent rows show the corresponding activation maps. Warmer colors indicate stronger responses.

previous-task samples. Incorporating CES shifts the responses toward the main objects, suggesting that complementary replay helps preserve useful historical representations. Adding CWFD further refines the heatmaps in several examples, yielding responses that more closely align with the visible object structures. This observation is consistent with its role in preserving historical representation directions during incremental learning. Although Grad-CAM heatmaps alone do not quantify forgetting, these visualizations provide qualitative support for the complementary roles of CES and CWFD in retaining object-relevant representations for previously learned tasks.