FedRNC: Addressing Spatio-Temporal Label Misalignment in Federated Noisy Class-Incremental Learning

Xingwei Huang[1], Zhaobin Sun[1], Junjie Shi[1], Xin Yang[1], Zengqiang Yan[1]*

[1]School of Electronic Information and Communications, Huazhong University of Science and Technology, Wuhan, China
{completist, zbsun, shijunjie, xinyang2014, z_yan}@hust.edu.cn

Abstract

Federated class-incremental learning (FCIL) aims to incrementally learn new classes across decentralized clients under non-IID data distributions. However, the pervasive challenge of label noise in FCIL has been completely overlooked. In this work, we introduce federated noisy class-incremental learning (FNCIL) and, for the first time, identify a novel form of label noise—**spatio-temporal label misalignment**—where samples from unseen classes are entirely mislabeled as known classes, with their correctly labeled counterparts appearing in latter tasks or other clients. This phenomenon undermines the effectiveness of existing centralized denoising strategies and creates a clear requirement for noise-robust methods in real-world FNCIL scenarios. To tackle this issue, we propose FedRNC, a dual-phase framework that leverages feature-space associations to establish spatio-temporal correspondences between clean global prototypes and noisy cached samples for progressive label correction. Experiments on standard benchmarks demonstrate FedRNC's superiority against existing baselines, along with its plug-and-play capability to upgrade FCIL systems for FNCIL.

Code — https://github.com/ZaC-Hide/FedRNC

Introduction

Federated Learning (FL) (McMahan et al. 2017; Yan et al. 2020; Wu et al. 2023b) has emerged as a promising paradigm for collaboratively training models across distributed clients without sharing raw data. In real-world applications such as healthcare and personalized services (Sun et al. 2024), users often encounter streaming data from disjoint class spaces due to user preferences, regional interests, or temporal trends. This motivates Federated Class-Incremental Learning (FCIL), where each client incrementally observes a non-overlapping subset of classes over time (Wang et al. 2024).

Recent efforts in FCIL primarily focus on mitigating catastrophic forgetting (Ganin et al. 2016) under heterogeneous client data and asynchronous task arrivals, aiming to achieve competitive global performance (Liu et al. 2024; Wang et al. 2024). Typical strategies include server-side knowledge distillation (Douillard et al. 2020), generative replay via GAN-based models (Qi, Zhao, and Li 2023), multi-classifier designs (Yu et al. 2024), and exemplar-based rehearsal buffers that store a subset of past samples (Li et al. 2024). These methods partially preserve prior knowledge but commonly assume clean and reliable data.

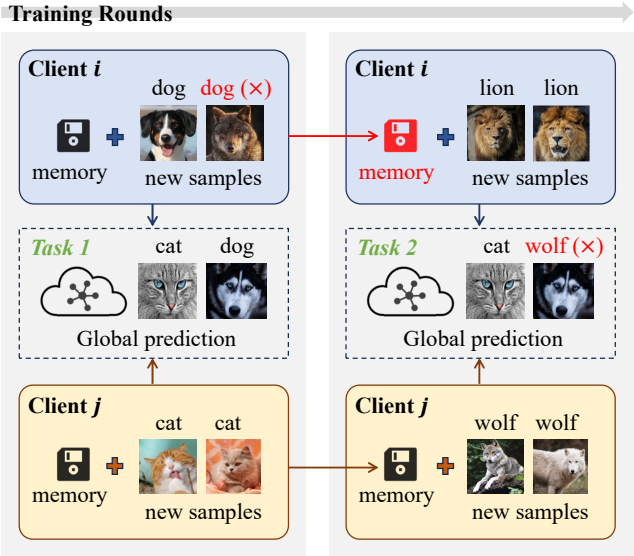

Figure 1: Illustration of **Spatio-Temporal Label Misalignment (STLM)** in FNCIL. Early samples of a new class (e.g., *wolf*) are mislabeled as a known class (e.g., *dog*), causing client i's model to entangle their features. This entanglement affects the global model aggregation, and when client j later learns *wolf*, the mixed *wolf+dog* features lead to persistent misclassification.

However, label noise is pervasive in practical FL deployments due to heterogeneous annotation quality, ambiguous class definitions, and inconsistent human understanding across clients (Xu et al. 2022; Fang and Ye 2022). Existing works on noise-robust FL or continual learning are limited and often rely on over-simplified assumptions. Most treat noise correction as per-task subproblems using local strategies like loss-based filtering or sample reweighting (Han

*Corresponding Author.

et al. 2018; Yu et al. 2019; Chen et al. 2019; Gui, Wang, and Tian 2021). Some further distinguish clean from noisy clients for global reweighting (Jiang et al. 2022; Kim et al. 2022; Zeng et al. 2024). However, these methods generally assume that clean samples for each class are present in the current task or locally accessible, and the noisy samples across different clients are IID or consistent in distribution. Till now, an appropriate modeling framework for characterizing label noise in federated class-incremental learning remains lacking.

In reality, label noise in Federated Noisy Class-Incremental Learning (FNCIL) exhibits more complex spatio-temporal patterns than conventional noise assumptions. When new classes emerge, annotators often mislabel them as known/observed classes due to outdated knowledge, giving rise to a unique type of noise: **true classes of noisy samples are entirely absent from the current task or client**. Unlike conventional label noise, such mislabels may initially appear harmless, as they are treated as hard samples and do not severely distort current decision boundaries. However, once these true classes appear in later tasks or on other clients, such samples—preserved in memory or embedded in model parameters—introduce semantic conflicts during global aggregation, leading to persistent misclassification. As illustrated in Fig. 1, this phenomenon exposes a key limitation of existing denoising methods: **clean samples of all classes are NOT immediately available in FNCIL**. Even worse, relabeling-based methods (Li, Socher, and Hoi 2020; Lai et al. 2024) may further amplify the issue if such mislabels dominate early memory.

To capture this phenomenon, we define a novel and realistic challenge in FNCIL, termed **Spatio-Temporal Label Misalignment (STLM)**. As shown in Fig. 1, noisy instances from a new class (e.g., Class *wolf* mislabeled as Class *dog*) appear early, while clean samples of Class *wolf* only become available in different tasks or other clients. This temporal and spatial diversity in clean class exposure fundamentally limits the effectiveness of conventional denoising methods and calls for a mechanism that can establish cross-task, cross-client consistency without raw data exchange.

Different from FCIL, FNCIL aims to not only mitigate forgetting but also suppress label noise. For the former goal, exemplar-based methods store fixed past samples but face challenges under label noise: cached mislabeled samples persistently resurface, compounding errors and complicating robust rehearsal. At the same time, their key advantage lies in preserving past data, enabling retrospective correction as new information emerges—unlike exemplar-free methods where noise irreversibly contaminates model weights.

Take the essence and discard the dregs, we propose **FedRNC**, a two-stage federated denoising framework that exploits spatio-temporal feature associations within rehearsal-based FNCIL. Clients maintain bounded buffers of past samples, which amplify noisy exemplar effects but allow for adaptive noise handling. FedRNC first performs local loss-based clustering to separate pseudo-clean and pseudo-noisy samples, curbing noise propagation early on. It then aggregates clean exemplars to form global prototypes and retrospectively refines noisy samples via prototype-guided matching, achieving precise relabeling even for previously unseen classes. This approach preserves privacy, adapts to FCIL dynamics, and effectively mitigates noise. Extensive experiments on multiple benchmarks demonstrate that FedRNC effectively improves robustness under both IID and Non-IID FNCIL scenarios, outperforming state-of-the-art (SOTA) baselines by a large margin. Furthermore, FedRNC can work as a plug-and-play module to seamlessly enhance existing FCIL methods.

The main contributions are as follows:

- We define a new task FNCIL to advance FCIL towards more realistic scenarios.
- We formulate a new label noise type, spatio-temporal label misalignment, in FNCIL and analyze its practical implications.
- We propose **FedRNC**, a two-stage denoising framework to establish spatio-temporal consistency in the feature space.
- We empirically validate the superiority and versatility of FedRNC across multiple benchmarks under diverse noise conditions in FNCIL settings.

Related Work

Federated Class-Incremental Learning FCIL extends class-incremental learning to federated settings, where clients sequentially observe disjoint tasks and collaboratively train a global model without sharing raw data. Classical centralized methods like iCaRL (Rebuffi et al. 2017) combine rehearsal with global aggregation. PODNet (Douillard et al. 2020) introduces feature distillation to preserve representations and has been adapted to FL. To mitigate forgetting under distribution shift, TARGET (Zhang et al. 2023) employs a generative replay strategy via GANs, while FedCBC (Yu et al. 2024) uses class-wise binary classifiers for selective knowledge retention. GLFC (Dong et al. 2022) introduces partial buffering and class-balanced loss, and Re-Fed (Li et al. 2024) prioritizes samples via gradient-based importance scores. Although effective in combating forgetting, these methods assume clean labels and completely ignore potential label noise in FCIL.

Federated Noisy Label Learning Label noise is especially problematic in FL due to data decentralization and client heterogeneity. Local denoising methods like Co-teaching (Han et al. 2018) and DivideMix (Li, Socher, and Hoi 2020) select or relabel instances based on loss statistics, but lack global coordination, limiting their robustness under non-IID conditions. Other approaches exploit global indicators: FedNoRo (Wu et al. 2023a) uses class-decoupled loss for noise-aware client filtering. While these enhance robustness, they struggle with dynamic or latent noise. Replay-based methods such as Self-Purified Replay (Kim et al. 2021) and CNLL (Karim et al. 2022) incorporate denoising into continual learning, yet assume centralized IID data. AF-FCL (Wuerkaixi et al. 2025) explores FCIL with label noise by scoring the task relevance of generated features, effectively identifying unreliable clients. However, it focuses

on client-level filtering rather than modeling temporal noise evolution.

In a short conclusion, FNCIL, especially the spatio-temporal label noise in FNCIL, is under-explored.

FNCIL Formulation

FNCIL can be seen as a variant of FCIL against spatio-temporal label misalignment including the following two settings.

FCIL Setting

We consider a FCIL setting consisting of K clients indexed by $k \in [K]$. Each client k accesses only its local data and learns from a sequence of tasks $\{\mathcal{T}_1, \ldots, \mathcal{T}_n\}$, where $\mathcal{T}_t = \{(x_t^{(i)}, y_t^{(i)})\}_{i=1}^{N_t}$ with N_t samples. The label set for task $\mathcal{T}_t$ is $\mathcal{Y}^t = \{y \mid \exists (x, y) \in \mathcal{T}_t\}$, and the cumulative class set for client k is $\mathcal{Y}_k = \bigcup_{t=1}^{n} \mathcal{Y}^t$. Following common practice, we assume that tasks have disjoint classes: $\mathcal{Y}^t \cap \mathcal{Y}^{t'} = \emptyset$ for $t \neq t'$ and class distributions differ across clients, i.e., $\mathcal{Y}_k \neq \mathcal{Y}_{k'}$ for $k \neq k'$.

To mitigate forgetting, client k maintains a bounded exemplar buffer $\mathcal{B}_k$ storing selected samples from past tasks. When task $\mathcal{T}_k^t$ arrives, training uses both current task data and cached exemplars, forming the local set $\mathcal{T}_{k,\text{local}}^t$. So, the global objective at task t optimizes shared model w^t over all visible classes, written as

$$w^t = \arg\min_{w} \sum_{k=1}^{K} \sum_{(x,y) \in \mathcal{T}_{k,\text{local}}^t} \frac{1}{|\mathcal{T}_{\text{g}}^t|} \mathcal{L}(f_{w_k}(x), y), \qquad (1)$$

where $|\mathcal{T}_{\text{g}}^t| = \sum_{k=1}^{K} |\mathcal{T}_{k,\text{local}}^t|$. After training, $\mathcal{B}_k$ is updated by selecting samples from $\mathcal{T}_k^t$ under a fixed budget $|\mathcal{B}_k| \leq M$. A class-balanced replacement strategy ensures sample representativeness while discarding older exemplars.

Noise Setting

To simulate realistic label noise under the FCIL setting, we propose a structured noise injection mechanism that reflects the spatio-temporal misalignment of clean samples. Specifically, in each local task $\mathcal{T}_t$ of client k, given its class set $\mathcal{Y}^t$, we randomly divide classes into two disjoint subsets:

- the **visible class set** $\mathcal{Y}_v^t \subset \mathcal{Y}^t$, and
- the **noise-source class set** $\mathcal{Y}_n^t = \mathcal{Y}^t \setminus \mathcal{Y}_v^t$.

For all samples $(x_t^{(i)}, y_t^{(i)})$ with $y_t^{(i)} \in \mathcal{Y}_n^t$, we inject noise by uniformly reassigning their labels to visible classes $y' \in \mathcal{Y}_v^t$, and randomly discard a subset to meet the target noise rate ρ . The resulting task data contains only visible classes $\mathcal{Y}_v^t$, mixed with mislabeled samples originally from $\mathcal{Y}_n^t$. In FCIL, true classes $\mathcal{Y}_n^t$ may only emerge in different tasks or on other clients, creating a spatio-temporal label misalignment that hinders local correction without access to clean exemplars (illustrated in Fig. 2).

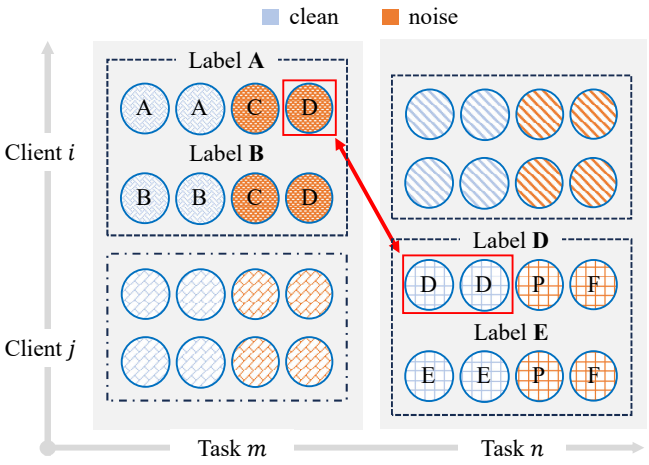

Figure 2: Illustration of noise generation and resulting data composition.

Method

In FNCIL, the primary task of FedRNC is to leverage dynamic replay buffers not only for knowledge retention but also as a platform for retrospective noise correction. Specifically, FedRNC first performs coarse local filtering to suppress dominant noise and then refines residual mislabeled samples through prototype-guided matching across clients.

Stage 1: Local Noise Filtering

Early in training, clean samples align better with the initialized or pre-trained model, yielding faster convergence and lower losses, while noisy samples generate conflicting gradients and higher losses. This results in a bimodal loss distribution, allowing coarse separation via clustering.

Given a task $\mathcal{T}_t = \{(x_i, y_i)\}_{i=1}^{N}$ and local model M_k, after E epochs of training, we record per-sample losses:

$$\ell_i = \mathcal{L}(M_k(x_i), y_i), \qquad (2)$$

where $\mathcal{L}(\cdot, \cdot)$ is the cross-entropy loss. Fitting a two-component Gaussian Mixture Model (GMM) to $\{\ell_i\}$ separates samples into pseudo-clean $\mathcal{T}_\mathcal{C}$ and pseudo-noisy $\mathcal{T}_\mathcal{N} = \mathcal{T}_t \setminus \mathcal{T}_\mathcal{C}$. These are stored in buffers $\mathcal{B}_\mathcal{C}$ and $\mathcal{B}_\mathcal{N}$ respectively, preventing severe noise propagation during replay. To maintain representativeness under fixed capacity, incoming samples are stored with a class-balanced update. When full, the number of retained or discarded samples per class is adjusted to approximate inter-class balance, improving clustering and the following prototype estimation.

Stage 2: Progressive Label correction

Early loss-based filtering offers a coarse separation but can misclassify hard clean samples with high loss or noisy samples close to seen classes. Without prior knowledge of unseen classes in FNCIL, it is impossible to directly recover true labels of such discarded noisy samples. To overcome this, we use the global feature space as supervision. While raw data remains private, aggregated global prototypes from all clients serve as stable semantic anchors (Gao et al. 2024;

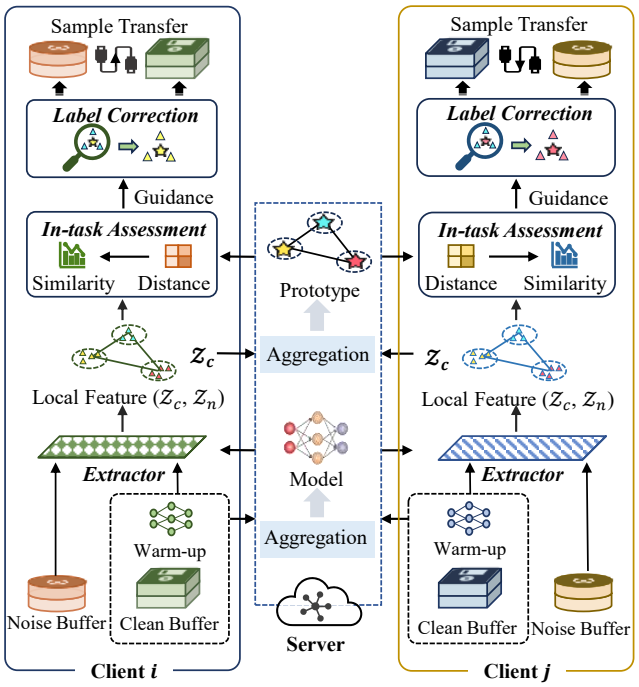

Figure 3: Overview of **Stage 2: Progressive Label correction** in FedRNC.

Wei et al. 2021), enabling clients to match cached features with reliable class prototypes and identify mislabeled samples with previously unseen semantics.

Global Semantic Space Construction. To align semantics across heterogeneous clients, we perform warmup training. Each client fine-tunes its local model on the early-cleaned buffer $\mathcal{B}_\mathcal{C}$ for several epochs. Such models are then aggregated via FedAvg on the server to obtain a global model F that acts as a consistent feature extractor, ensuring aligned class-level embeddings across clients.

Prototype-Guided Feature Correction. Once the global model F is synchronized and distributed, each client extracts feature representations via

$$z_i = F(x_i). \tag{3}$$

Features from $\mathcal{B}_\mathcal{C}$ and $\mathcal{B}_\mathcal{N}$ form the clean and noisy feature sets $\mathcal{Z}_\mathcal{C}$ and $\mathcal{Z}_\mathcal{N}$, respectively.

In the current task, each client k maintains a cumulative label set $\mathcal{Y}_k^T = \{\mathcal{Y}_k^1, \mathcal{Y}_k^2, \ldots, \mathcal{Y}_k^t\}$, representing all locally observed classes so far. Clean features are grouped by their predicted class, $\mathcal{Z}_{k,\mathcal{C}}^a$ denoted as the set of clean features assigned to class $a \in \mathcal{Y}_k^T$. For each seen class, the local prototype is computed as:

$$\mu_{k,a} = \frac{1}{|\mathcal{Z}_{k,\mathcal{C}}^a|} \sum_{z_i \in \mathcal{Z}_{k,\mathcal{C}}^a} z_i. \tag{4}$$

Then, each client uploads its local prototypes to the server. Defining the global class set as $\mathcal{Y}^T = \bigcup_{k=1}^K \mathcal{Y}_k^T$, for each class $a \in \mathcal{Y}^T$, the server aggregates prototypes into a global prototype:

$$\mu_a = \frac{1}{K_a} \sum_{k=1}^{K_a} \mu_{k,a}, \tag{5}$$

where K_a denotes the number of clients that have the observed/seen class a.

With access to the global prototype set $\mathcal{M} = \{\mu_a \mid a \in \mathcal{Y}^T\}$, each client matches cached samples to these anchors in the feature space. This matching revisits both noisy and previously accepted clean samples, processing buffers $\mathcal{Z}_\mathcal{C}$ and $\mathcal{Z}_\mathcal{N}$ independently via prototype-guided relabeling. This dual approach enables:

- **Refinement of clean samples**: Correcting mislabeled samples missed by initial confidence filtering through global semantic re-evaluation.
- **Reuse of noisy samples**: Salvaging mislabeled yet semantically meaningful instances based on feature-prototype consistency instead of discarding all high-loss samples.

To simplify notation, we use $\mathcal{Z}$ to generically denote either the clean feature set $\mathcal{Z}_\mathcal{C}$ or the noisy feature set $\mathcal{Z}_\mathcal{N}$ in each client, both of which are independently subjected to the same distance-based correction procedure. Upon receiving the global prototype set $\mathcal{M}$ from the server, each client computes distances between cached features and class prototypes. For each class $b \in \mathcal{Y}_k^T$, we compute the Euclidean distance between each feature $z_i \in \mathcal{Z}^b$ and all class prototypes μ_a in $\mathcal{M}$:

$$D_{i,a}^b = \|z_i - \mu_a\|_2, \quad D \in \mathbb{R}^{|\mathcal{Z}^b| \times |\mathcal{M}|}. \tag{6}$$

The number of features in $\mathcal{Z}^b$ nearest to each prototype defines a count vector $n_b \in \mathbb{R}^C$, which is normalized to obtain a similarity vector $s_b \in \mathbb{R}^{|\mathcal{M}|}$, where $s_b^{(a)}$ denotes the similarity between class b and prototype μ_a:

$$s_b^{(a)} = \frac{n_b^{(a)}}{\sum_{j=1}^{|\mathcal{M}|} n_b^{(j)}}. \tag{7}$$

For each task-specific class $b \in \{\mathcal{Y}_k^1, \mathcal{Y}_k^2, \ldots, \mathcal{Y}_k^t\}$, the corresponding similarity vector s_b is masked by excluding entries for current task labels, yielding a truncated vector $s_b \in \mathbb{R}^{|\mathcal{M}| - |\mathcal{Y}_k^t|}$. The final adaptive similarity vector is then computed by taking the element-wise geometric mean across all such vectors:

$$S_{\text{task}}^{(a)} = \left(\prod_{b \in \mathcal{Y}_k^t} s_b^{(a)} \right)^{1/t}, \quad \forall a \notin \mathcal{Y}_k^t. \tag{8}$$

For each global class $a \notin \mathcal{Y}_k^t$, we identify a candidate relabeling set $\mathcal{R}_a^b$ for every local class $b \in \mathcal{Y}_k^t$. Specifically, $\mathcal{R}_a^b$ denotes the subset of features originally assigned to class b that are most similar to prototype μ_a, suggesting a potential relabeling to class a:

$$\mathcal{R}_a^b = \text{Top}_{\lfloor S_{\text{task}}^{(a)} \cdot |\mathcal{Z}^a| \rfloor} \left(\{D_{i,a}^b\} \right), \tag{9}$$

$\text{Top}_k(\cdot)$ denotes selecting the k smallest values in ascending order. If a sample feature appears in multiple sets $\mathcal{R}_a^b$, its relabeling target is determined by the maximum similarity $S_{\text{task}}^{(a)}$, ensuring consistency. The task-level grouping strategy is motivated by the observation that classes within the same task share contextual and temporal coherence. Evaluating them jointly provides a more stable semantic signal than treating each class independently, which is especially beneficial under Non-IID and noisy federated conditions.

Prototype-guided relabeling is independently applied to both feature sets: selected features from $\mathcal{Z}_\mathcal{C}$ are relabeled and kept in $\mathcal{B}_\mathcal{C}$, while those from $\mathcal{Z}_\mathcal{N}$ are moved from $\mathcal{B}_\mathcal{N}$ to $\mathcal{B}_\mathcal{C}$ after relabeling. This progressive refinement adaptively corrects noisy instances over time by leveraging prototype consistency, without sharing raw data. The denoising workflow is illustrated in Fig. 3. The refined buffer $\mathcal{B}_\mathcal{C}$ is then used for local training. Each client trains on this buffer and sends model updates to the server, which aggregates them via averaging and redistributes the global model. This loop enhances label quality and overall performance by tightly coupling denoising with federated optimization.

Experiments

Dataset and Evaluation Protocol

We evaluate our method on three widely-used benchmarks—EMNIST (Cohen et al. 2017), CIFAR-100 (Krizhevsky, Hinton et al. 2009), and Tiny-ImageNet (Le and Yang 2015)—covering low- to high-resolution scenarios, with each dataset split into training and test sets (8:2). Specifically, a global class pool is first constructed using all available categories. Based on this pool, client-specific task streams are generated by randomly sampling disjoint class subsets for each task to simulate realistic federated environments with both inter-client and intra-client label distribution heterogeneity. All clients are assigned tasks in a non-overlapping and non-IID manner.

Data Partition and Noise Generation

To simulate realistic FNCIL scenarios, two types of data partitioning protocols are adopted to reflect different levels of data and noise heterogeneity, including:

- **IID-Noise Setting**: All classes are of the same data amounts, and every task on each client follows an uniform noise level. It helps isolate the effect of temporal misalignment under controlled conditions.
- **Non-IID-Noise Setting**: Sample sizes vary across classes, and each client's local task exhibits a different level of label noise. We adopt the Dirichlet distribution $\text{Dir}(\alpha)$ to control the degree of label distribution skew across clients.

More implementation details are presented in the supplemental materials.

Comparison Methods

A series of representative baselines are selected for comparison, including standard FL methods like FedAvg (McMahan et al. 2017), FedProx (Li et al. 2020), FL combined with CIL methods like iCaRL (Rebuffi et al. 2017), EWC (Kirkpatrick et al. 2017), PODNet (Douillard et al. 2020), and FL combined with noisy class-incremental learning (NCIL) methods like SPR (Kim et al. 2021), S6 (Li et al. 2023), CNLL (Karim et al. 2022), and recent FCIL methods like TARGET (Zhang et al. 2023), FedCBC (Yu et al. 2024), AF-FCL (Wuerkaixi et al. 2025) (an exemplar-free method specifically considers robust learning under noisy clients), GLFC (Dong et al. 2022), and Re-Fed (Li et al. 2024)(exemplar-based). All methods are evaluated under the same task streams and client partitioning. For a fair comparison, exemplar-based methods are allowed the same buffer size as FedRNC.

In addition, we include an *oracle-clean* variant of FedRNC, where all noisy samples are assumed to be correctly relabeled to estimate the potential upper bound.

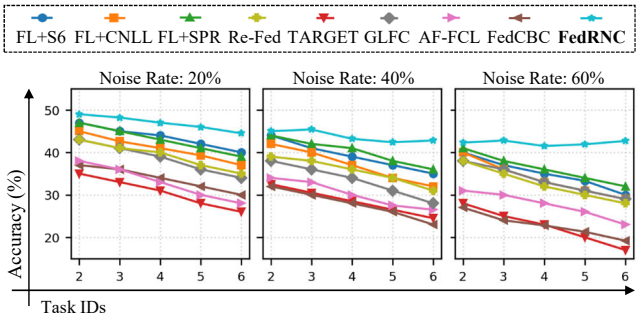

Figure 4: Test accuracy curves across tasks under different noise rates ($\rho = 20\%, 40\%, 60\%$) on CIFAR100.

IID-Noise Evaluation

As summarized in Tab. 1, all methods degrade as noise increases. Exemplar-free methods yield lower absolute accuracy but show greater noise robustness, while exemplar-based ones drop more sharply due to retaining corrupted samples. Although existing denoising strategies alleviate some noise, their lack of temporal and spatial context limits effectiveness in FNCIL. In contrast, FedRNC consistently achieves the best performance across noise levels and datasets. Its advantage becomes especially pronounced at high noise rates, owing to its effective recovery of mislabeled samples via prototype-guided relabeling—demonstrating strong robustness under spatio-temporal misalignment.

To investigate the temporal dynamics during continual learning, we plot task-wise test accuracy under varying noise levels in Fig. 4. As expected, all methods show a downward trend as tasks progress, due to memory limitations and increasing semantic confusion from spatio-temporal noise. Nevertheless, FedRNC consistently outperforms others, particularly under high noise. Thanks to its retrospective relabeling mechanism, which incrementally corrects mislabeled samples with richer contextual cues, FedRNC exhibits a much slower performance degradation over time.

Category	Methods	EMNIST			CIFAR100			Tiny-ImageNet		
		20%	40%	60%	20%	40%	60%	20%	40%	60%
FL+CIL	FedAvg	26.2±1.2	23.8±1.3	22.3±1.3	15.7±0.8	14.6±1.1	12.9±1.0	15.4±1.1	13.7±1.3	12.1±1.4
	FedProx	23.5±1.1	21.8±1.1	20.1±1.2	13.6±1.0	12.0±1.2	10.9±1.4	13.4±0.9	12.0±1.2	10.7±1.4
	FL+iCaRL	53.0±0.9	44.5±1.2	40.3±1.5	35.0±1.0	32.6±1.4	29.1±1.6	32.5±1.1	29.8±1.5	26.7±1.8
	FL+EWC	32.0±1.0	29.9±1.1	27.6±1.3	19.1±0.9	17.7±1.2	16.5±1.3	20.5±1.0	18.1±1.3	16.6±1.5
	FL+PODNet	37.1±0.9	35.5±1.1	33.4±1.4	22.9±0.8	21.3±1.1	19.8±1.4	19.4±0.9	17.9±1.3	16.2±1.6
FL+NCIL	FL+CNLL	56.9±1.0	51.1±1.2	45.7±1.3	37.8±0.8	32.5±1.3	28.4±1.5	35.3±1.0	31.6±1.4	29.3±1.6
	FL+S6	60.7±1.0	56.4±1.1	51.5±1.2	41.2±0.9	38.0±1.2	36.3±1.5	37.8±1.1	36.5±1.4	34.0±1.7
	FL+SPR	65.4±0.8	63.2±1.1	61.9±1.2	41.3±0.7	39.8±1.0	37.5±1.3	37.0±0.9	36.3±1.1	34.7±1.5
FCIL	TARGET	41.5±0.9	38.3±1.1	35.2±1.4	32.5±1.0	29.4±1.3	27.0±1.5	31.7±1.1	28.2±1.4	26.8±1.7
	FedCBC	43.3±1.0	37.3±1.3	33.9±1.6	34.2±1.0	29.8±1.4	25.4±1.6	33.4±1.1	28.4±1.5	25.5±1.5
	AF-FCL	44.8±0.8	42.6±1.2	39.8±1.5	33.8±0.9	32.0±1.2	29.2±1.4	30.8±0.6	28.1±1.4	26.6±1.2
	GLFC	64.6±0.8	58.8±1.1	53.7±1.3	44.3±0.9	39.5±1.2	36.1±1.5	38.1±1.0	33.4±1.1	29.7±1.8
	Re-Fed	65.0±0.7	58.5±1.0	52.5±1.4	**46.7±0.8**	41.5±1.3	36.6±1.6	37.2±0.9	32.0±1.4	29.4±1.7
FNCIL	*Clean*	*74.2±0.5*	*75.8±0.6*	*76.4±0.5*	*51.1±0.6*	*52.3±0.7*	*52.8±0.5*	*45.1±0.5*	*45.9±0.6*	*46.2±0.5*
	FedRNC	**71.3±0.7**	**69.0±0.7**	**71.2±0.5**	45.7±0.7	**44.2±1.0**	**44.5±1.1**	**40.4±1.3**	**38.7±1.3**	**39.9±1.2**

Table 1: Comparison under increasing noise rates ($\rho = 20\%, 40\%, 60\%$) using the IID-Noise setting. The best results at each noise level are marked in bold.

Non-IID-Noise Evaluation

Under the non-IID-noise setting, both class distributions and noise levels vary across clients according to Dirichlet distributions with concentration parameters as $\alpha = 5$ and $\alpha = 10$ to simulate different degrees of quantity heterogeneity. The noise configurations ρ include $(20\%, 40\%)$ and $(40\%, 60\%)$. Note that higher heterogeneity results in more severe spatio-temporal misalignment, making accurate relabeling more challenging. Consequently, as summarized in Tab. 2, all methods suffer performance degradation as the level of heterogeneity increases. Among them, methods relying solely on local clustering or task-level loss statistics exhibit greater drops, reflecting their sensitivity to non-IID class distributions and imbalanced noise. Comparatively, FedRNC is of better robustness, benefiting from its global prototypes and spatio-temporal matching.

Methods	$\alpha = 5$		$\alpha = 10$	
	(20%,40%)	(40%,60%)	(20%,40%)	(40%,60%)
FedAvg	11.7±1.1	10.8±1.3	13.8±1.2	11.4±1.0
FedProx	11.5±1.3	10.2±1.1	12.5±1.4	11.3±0.9
FL+iCaRL	24.3±1.8	19.3±1.7	24.6±1.6	19.5±1.9
FL+EWC	16.4±1.0	14.1±1.2	17.9±0.9	16.3±1.1
FL+PODNet	18.3±1.2	15.4±1.3	19.4±1.1	16.6±0.8
FL+CNLL	26.1±1.5	23.3±1.6	30.5±1.5	28.2±1.8
FL+S6	28.5±1.6	25.8±1.9	31.3±1.6	29.8±1.4
FL+SPR	28.8±1.3	26.9±1.5	33.4±1.8	31.1±1.7
TARGET	24.4±1.8	20.7±1.3	25.5±1.7	22.1±1.5
FedCBC	26.1±1.9	21.2±1.7	27.6±1.8	22.2±1.2
AF-FCL	24.4±1.6	22.9±1.8	27.8±1.3	24.4±1.4
GLFC	29.9±1.5	26.6±1.6	31.7±1.7	28.8±1.3
Re-Fed	29.5±1.8	25.1±1.4	32.1±1.9	28.5±1.7
Clean	*38.0±0.5*	*39.5±0.6*	*40.9±0.6*	*43.1±0.5*
FedRNC	**33.6±1.4**	**33.2±1.4**	**35.1±1.1**	**35.7±1.3**

Table 2: Comparison under non-IID data (α = 5, 10) with different noise intensities on Tiny-ImageNet.

Ablation Study on Plug-and-Play Properties

To assess the generalizability of FedRNC's denoising strategy, we integrate it as a plug-and-play module into several representative exemplar-based FCIL methods, including FL+iCaRL, GLFC, and Re-Fed. For simplicity, the denoising stage is directly applied prior to their regular local training, without altering their exemplar management or update pipelines. As summarized in Tab. 3, when label noise is present, injecting our denoising process leads to substantial performance gains across all baselines, validating its effectiveness as a general noise-robust pre-processing step.

We further analyze its effectiveness from the perspectives of both stability and plasticity (Yang et al. 2024) across all baselines under the $\rho = 40\%$ noise setting as illustrated in Fig. 5. Here, stability reflects the model's ability to retain knowledge from previous tasks, and plasticity measures its capacity to learn new classes effectively. Consistent performance improvements across all baselines validate our method as a lightweight yet powerful enhancement for replay-based FCIL algorithms, offering robustness in noisy environments.

Ablation Study on Dual-Stage Denoising

To evaluate the necessity of FedRNC's two-stage denoising design, we conduct an ablation study comparing FedRNC with two simplified variants including **w/o PF** which removes prototype-based relabeling, and **w/o LF** which skips initial loss filtering, and three representative baselines (i.e., SPR, CNLL, and S6) that lack spatio-temporal modeling.

As shown in Fig. 6, loss filtering alone suffers from low

Methods	ρ		
	20%	40%	60%
iCaRL	53.0±0.9	44.5±1.2	40.3±1.5
+ours	58.4±0.5	55.7±0.8	56.3±0.7
GLFC	64.6±0.8	58.8±1.1	53.7±1.3
+ours	67.7±0.9	65.0±0.6	65.8±0.4
Re-Fed	65.0±0.7	58.5±1.0	52.5±1.4
+ours	70.9±0.6	67.4±0.5	67.1±0.9
FedRNC	**71.3±0.7**	**69.0±0.7**	**71.2±0.5**

Table 3: Plug-and-play ablation studies of FedRNC onto exemplar-based FCIL baselines evaluated on EMNIST under varying noise levels ($\rho = 20\%, 40\%, 60\%$).

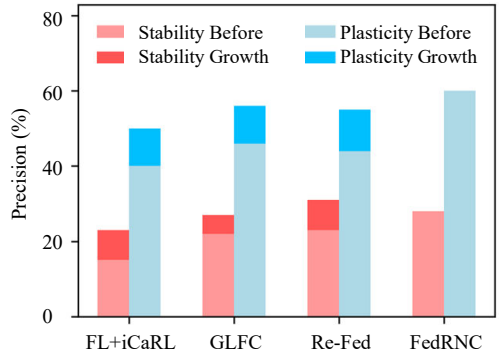

Figure 5: Stability and plasticity gain under noisy settings ($\rho = 40\%$), where each bar shows the improvement introduced by FedRNC's denoising module (color-coded within each bar).

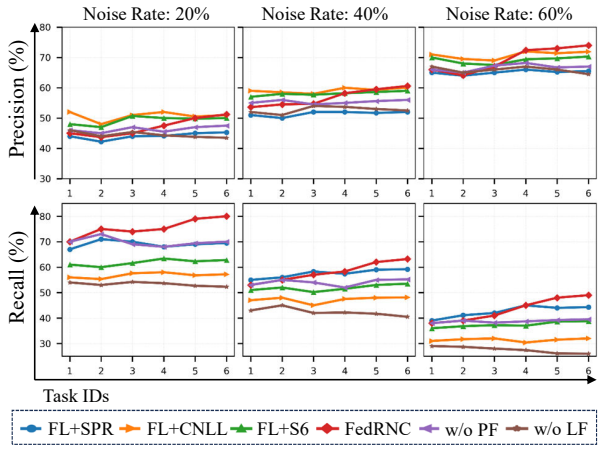

Figure 6: Cumulative precision and recall of filtered clean samples over tasks.

precision and recall due to hard samples and noisy signals, while prototype matching struggles with recall under heavy noise as corrupted features distort class centers. By combining both, FedRNC consistently achieves better precision and recall. Its advantage—especially in recall—grows over time by leveraging increasingly reliable temporal knowledge to recover mislabeled yet valuable samples.

Methods	IID		Non-IID	
	BACC (↑)	Forget (↓)	BACC (↑)	Forget (↓)
FedRNC	**40.4±1.3**	**5.8±0.2**	**35.1±1.1**	**6.3±0.3**
w/o LF	32.6±2.2	10.7±1.0	29.4±2.3	12.1±1.2
w/o PF	34.8±1.6	9.2±0.6	31.5±1.8	10.3±0.9
w/o PF-TS	36.1±1.4	8.5±0.5	33.7±1.6	9.4±0.7
w/o PF-NR	36.8±1.2	8.1±0.4	34.5±1.2	9.0±0.6
w/o PF-LC	35.7±1.5	8.9±0.5	32.8±1.7	9.7±0.7

Table 4: Component-wise ablation studies of FedRNC on Tiny-ImageNet (IID noise: $\rho = 40\%$, Non-IID noise: $\rho = (20\%,40\%)$, $\alpha = 10$).

Ablation Study on Component Effectiveness

Component-wise ablation studies are summarized in Tab. 4. Removing task-wise similarity (**w/o PF-TS**) or noisy sample reuse (**w/o PF-NR**) degrades performance, showing the importance of contextual matching and progressive correction. Simply excluding noisy samples (**w/o PF-LC**) without label correction is also suboptimal, confirming the utility of constructive sample recovery. Such results confirm that: (1) the first-stage filtering is critical for reliable prototype computation; (2) matching-based screening is more effective when supported by prior denoising; and (3) effective reuse and relabeling of corrected samples amplifies the robustness of FedRNC against label noise.

Conclusion

This paper presents a practical yet underexplored challenge in FNCIL: spatio-temporal label misalignment (STLM), where noisy labels stem from delayed class emergence across clients and tasks. This noise type disrupts generalization in a persistent, accumulative manner and invalidates conventional relabeling-based denoising. To address this, we propose **FedRNC**, a two-stage replay-based framework that first filters noise via local loss clustering and then relabels using global prototype-guided correction based on spatio-temporal consistency. Extensive experiments on three benchmarks under various noise settings validate FedRNC's robustness over existing federated, continual, and denoising baselines. We believe STLM and FedRNC open promising directions for building noise-robust, memory-efficient FNCIL systems that better reflect real-world scenarios.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 62202179 and in part by the National Key Research and Development Program 2024YFE0217700.

References

Chen, P.; Liao, B. B.; Chen, G.; and Zhang, S. 2019. Understanding and utilizing deep neural networks trained with noisy labels. In *International conference on machine learning*, 1062–1070. PMLR.

Cohen, G.; Afshar, S.; Tapson, J.; and Van Schaik, A. 2017. EMNIST: Extending MNIST to handwritten letters. In *2017 international joint conference on neural networks (IJCNN)*, 2921–2926. IEEE.

Dong, J.; Wang, L.; Fang, Z.; Sun, G.; Xu, S.; Wang, X.; and Zhu, Q. 2022. Federated class-incremental learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10164–10173.

Douillard, A.; Cord, M.; Ollion, C.; Robert, T.; and Valle, E. 2020. Podnet: Pooled outputs distillation for small-tasks incremental learning. In *European Conference on Computer Vision*, 86–102. Springer.

Fang, X.; and Ye, M. 2022. Robust federated learning with noisy and heterogeneous clients. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10072–10081.

Ganin, Y.; Ustinova, E.; Ajakan, H.; Germain, P.; Larochelle, H.; Laviolette, F.; March, M.; and Lempitsky, V. 2016. Domain-adversarial training of neural networks. *Journal of machine learning research*, 17(59): 1–35.

Gao, X.; Yang, X.; Yu, H.; Kang, Y.; and Li, T. 2024. Fedprok: Trustworthy federated class-incremental learning via prototypical feature knowledge transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4205–4214.

Gui, X.-J.; Wang, W.; and Tian, Z.-H. 2021. Towards understanding deep learning from noisy labels with small-loss criterion. *arXiv preprint arXiv:2106.09291*.

Han, B.; Yao, Q.; Yu, X.; Niu, G.; Xu, M.; Hu, W.; Tsang, I.; and Sugiyama, M. 2018. Co-teaching: Robust training of deep neural networks with extremely noisy labels. *Advances in neural information processing systems*, 31.

Jiang, X.; Sun, S.; Wang, Y.; and Liu, M. 2022. Towards federated learning against noisy labels via local self-regularization. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, 862–873.

Karim, N.; Khalid, U.; Esmaeili, A.; and Rahnavard, N. 2022. Cnll: A semi-supervised approach for continual noisy label learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3878–3888.

Kim, C. D.; Jeong, J.; Moon, S.; and Kim, G. 2021. Continual learning on noisy data streams via self-purified replay. In *Proceedings of the IEEE/CVF international conference on computer vision*, 537–547.

Kim, S.; Shin, W.; Jang, S.; Song, H.; and Yun, S.-Y. 2022. FedRN: Exploiting k-reliable neighbors towards robust federated learning. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, 972–981.

Kirkpatrick, J.; Pascanu, R.; Rabinowitz, N.; Veness, J.; Desjardins, G.; Rusu, A. A.; Milan, K.; Quan, J.; Ramalho, T.; Grabska-Barwinska, A.; et al. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13): 3521–3526.

Krizhevsky, A.; Hinton, G.; et al. 2009. Learning multiple layers of features from tiny images.(2009).

Lai, L.; Ohn-Bar, E.; Arora, S.; and Yi, J. S. K. 2024. Uncertainty-guided never-ending learning to drive. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15088–15098.

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

Li, G.; Wang, P.; Luo, Q.; Liu, Y.; and Ke, W. 2023. Online noisy continual relation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 13059–13066.

Li, J.; Socher, R.; and Hoi, S. C. 2020. Dividemix: Learning with noisy labels as semi-supervised learning. *arXiv preprint arXiv:2002.07394*.

Li, T.; Sahu, A. K.; Zaheer, M.; Sanjabi, M.; Talwalkar, A.; and Smith, V. 2020. Federated optimization in heterogeneous networks. *Proceedings of Machine learning and systems*, 2: 429–450.

Li, Y.; Li, Q.; Wang, H.; Li, R.; Zhong, W.; and Zhang, G. 2024. Towards efficient replay in federated incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12820–12829.

Liu, Y.; Kang, Y.; Zou, T.; Pu, Y.; He, Y.; Ye, X.; Ouyang, Y.; Zhang, Y.-Q.; and Yang, Q. 2024. Vertical federated learning: Concepts, advances, and challenges. *IEEE transactions on knowledge and data engineering*, 36(7): 3615–3634.

McMahan, B.; Moore, E.; Ramage, D.; Hampson, S.; and y Arcas, B. A. 2017. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, 1273–1282. PMLR.

Qi, D.; Zhao, H.; and Li, S. 2023. Better generative replay for continual federated learning. *arXiv preprint arXiv:2302.13001*.

Rebuffi, S.-A.; Kolesnikov, A.; Sperl, G.; and Lampert, C. H. 2017. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2001–2010.

Sun, Z.; Wu, N.; Shi, J.; Yu, L.; Cheng, K.-T.; and Yan, Z. 2024. Fedmlp: Federated multi-label medical image classification under task heterogeneity. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 394–404. Springer.

Wang, L.; Zhang, X.; Su, H.; and Zhu, J. 2024. A comprehensive survey of continual learning: Theory, method and application. *IEEE transactions on pattern analysis and machine intelligence*, 46(8): 5362–5383.

Wei, J.; Zhu, Z.; Cheng, H.; Liu, T.; Niu, G.; and Liu, Y. 2021. Learning with noisy labels revisited: A study using real-world human annotations. *arXiv preprint arXiv:2110.12088*.

Wu, N.; Yu, L.; Jiang, X.; Cheng, K.-T.; and Yan, Z. 2023a. FedNoRo: Towards noise-robust federated learning by addressing class imbalance and label noise heterogeneity. *arXiv preprint arXiv:2305.05230*.

Wu, N.; Yu, L.; Yang, X.; Cheng, K.-T.; and Yan, Z. 2023b. Fediic: Towards robust federated learning for class-imbalanced medical image classification. In *International*

Conference on Medical Image Computing and Computer-Assisted Intervention, 692–702. Springer.

Wuerkaixi, A.; Cui, S.; Zhang, J.; Yan, K.; Han, B.; Niu, G.; Fang, L.; Zhang, C.; and Sugiyama, M. 2025. Accurate forgetting for heterogeneous federated continual learning. *arXiv preprint arXiv:2502.14205*.

Xu, J.; Chen, Z.; Quek, T. Q.; and Chong, K. F. E. 2022. Fedcorr: Multi-stage federated learning for label noise correction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10184–10193.

Yan, Z.; Wicaksana, J.; Wang, Z.; Yang, X.; and Cheng, K.-T. 2020. Variation-aware federated learning with multi-source decentralized medical image data. *IEEE Journal of Biomedical and Health Informatics*, 25(7): 2615–2628.

Yang, X.; Yu, H.; Gao, X.; Wang, H.; Zhang, J.; and Li, T. 2024. Federated continual learning via knowledge fusion: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 36(8): 3832–3850.

Yu, H.; Yang, X.; Gao, X.; Feng, Y.; Wang, H.; Kang, Y.; and Li, T. 2024. Overcoming spatial-temporal catastrophic forgetting for federated class-incremental learning. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 5280–5288.

Yu, X.; Han, B.; Yao, J.; Niu, G.; Tsang, I.; and Sugiyama, M. 2019. How does disagreement help generalization against label corruption? In *International conference on machine learning*, 7164–7173. PMLR.

Zeng, B.; Yang, X.; Chen, Y.; Shen, Z.; Yu, H.; and Zhang, Y. 2024. FedES: federated early-stopping for hindering memorizing heterogeneous label noise. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, 5416–5424.

Zhang, J.; Chen, C.; Zhuang, W.; and Lyu, L. 2023. Target: Federated class-continual learning via exemplar-free distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4782–4793.