SOCIALIZED CONTINUAL LEARNING: RECURSIVE WEIGHT-SPACE EDITING FOR SCALABLE MULTI-EXPERT COLLABORATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Intelligent systems increasingly rely on collaboration among specialized agents to exchange complementary expertise. However, existing collaborative paradigms such as Federated Learning (FL) and Socialized Learning (SL) Yao et al. (2024) typically assume one-shot or fixed-stage interaction, limiting their ability to adapt to open-ended environments where new experts and capabilities emerge continuously. To address this gap, we introduce *Socialized Continual Learning (SCL)*, a continual collaborative learning paradigm that enables previously socialized experts to interact with newly emerging experts while retaining their specialized capabilities. The central challenge is to reconcile newly acquired expertise with knowledge accumulated through previous interactions. To this end, we propose **Recursive Weight Space Editing (RWSE)**, which formulates SCL as a constrained functional alignment problem and recursively integrates new expertise into an evolving shared model. RWSE derives recursive sufficient statistics and an exact closed-form residual editing rule, enabling efficient knowledge integration without revisiting previously incorporated experts. Furthermore, a nonlinear spectrally anisotropic regularization mechanism imposes stronger protection on historically sensitive directions while preserving flexibility for new expertise, thereby mitigating inter-expert interference and maintaining prior specialization. Extensive experiments across diverse continual collaborative learning settings demonstrate the effectiveness of RWSE for continual expert integration.

1 INTRODUCTION

The evolution of intelligent systems increasingly relies on the ability of multiple specialized agents to learn from one another and collectively improve their capabilities in dynamic environments. In many real-world scenarios, individual agents naturally possess complementary expertise, and effective intelligence emerges through interaction, knowledge exchange, and mutual enhancement rather than isolated learning Thompson B. & T. (2022). This motivates *Federated Learning (FL)* McMahan et al. (2017); Huang et al. (2024) and *Socialized Learning (SL)* Yao et al. (2024; 2025); Wang et al. (2025b); Guo et al. (2026), in which multiple experts interact directly to acquire complementary knowledge and improve their collective capabilities.

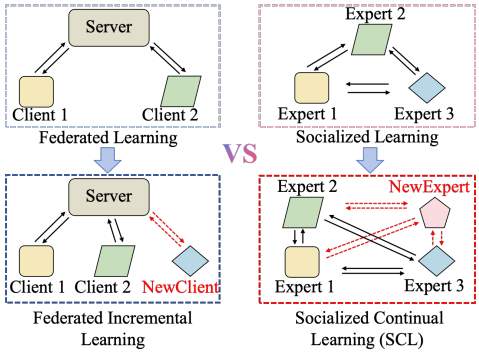
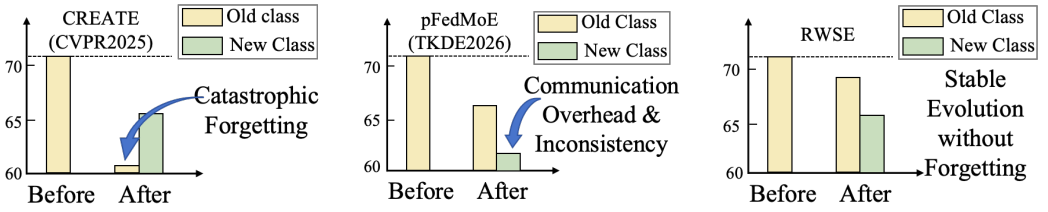


Figure 1: Comparison between our proposed SCL and other learning paradigms.

FL has become a powerful framework for distributed learning, particularly in scenarios with data privacy requirements. As illustrated in Fig. 1, FL enables multiple clients (experts) to collaboratively train server without directly exchanging local data. However, many real-world collaborative systems do not necessarily require privacy-preserving model aggregation, while centralized coordination can itself become undesirable or impractical Zhao et al. (2021); Dong et al. (2024). For example, autonomous vehicles and decentralized multi-agent systems often require agents to retain specialized

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.



(a) Class Incremental Learning (b) Federated Incremental Learning (c) Socialized Continual Learning

Figure 2: Performance comparison of different learning paradigms on CIFAR-100.

capabilities while interacting directly with peers without central coordination. In such settings, the server-based aggregation may hinder individual specialization and compress expertise into a server.

SL has recently emerged as a promising learning paradigm for collaborative intelligence, enabling experts to directly interact and learn from one another without relying on centralized coordination. As illustrated in Fig. 1, each expert shares its specialized knowledge with peers to acquire complementary expertise beyond its original specialization, while preserving and enhancing its own capabilities. SL aims to jointly achieve *versatility*, through acquiring complementary knowledge from peers, and *directionality*, through preserving and enhancing original specialized capabilities. Its decentralized nature allows experts to interact directly and flexibly, facilitating efficient knowledge exchange and mutual enhancement without relying on a central server.

However, both FL and SL are typically formulated as a one-shot or fixed-stage collaborative learning process. This assumption is increasingly inadequate for open-ended environments, where new tasks, capabilities, or specialized experts may emerge continuously. FL has naturally evolved toward *Federated Incremental Learning (FIL)* Li et al. (2025c); Rong et al. (2025), which enables collaborative models to continually acquire new knowledge while preserving previously learned capabilities.

Following the same evolution, we first identify the necessity of extending SL to a continual-learning scenario, which we term *Socialized Continual Learning (SCL)*. As illustrated in Fig. 1, SCL allows previously socialized experts to continue interacting with newly emerging experts, thereby progressively expanding their collective knowledge while retaining their original specialization. This continual interaction, however, introduces a new challenge: **how can experts continuously acquire complementary knowledge from newly emerging experts while preserving the specialized capabilities developed through previous socialized interactions?** On one hand, relying solely on data-driven learning to acquire new knowledge in a space already populated by expert classes can lead to interference and degradation of previously acquired expertise. On the other hand, relying on parameter exchange, where heterogeneity in data and model representations can induce representation collapse and degrade performance. Essentially, the crux of this challenge lies in the structural mismatch between decentralized expertise and a shared parameter space.

Targeting this challenge, we formulate SCL as a constrained functional alignment problem, where each newly arriving expert is recursively integrated with the evolving shared model while preserving historically sensitive directions. To this end, we propose **Recursive Weight Space Editing (RWSE)**, which performs functional-level knowledge alignment and spectrally anisotropic weight-space editing to accommodate new expertise while maintaining prior specialization.

Specifically, RWSE maintains recursive sufficient statistics and derives an exact closed-form residual editing rule, enabling incremental expert integration without revisiting previously incorporated experts. A nonlinear spectral regularization mechanism further provides direction-dependent protection by imposing stronger constraints on historically sensitive parameter directions while retaining flexibility for new knowledge. Together, these components enable continual expert collaboration with effective knowledge expansion and preservation. The contributions are summarized as follows:

- We identify and formulate a new learning paradigm, *Socialized Continual Learning*, which extends conventional SL from one-time knowledge exchange to continual interaction among previously learned and newly emerging experts.
- To support this paradigm, we propose Recursive Weight Space Editing, which formulates continual multi-expert integration as constrained functional alignment with recursive manifold-aware editing for stable and interference-free knowledge accumulation.

- Extensive experiments demonstrate that RWSE consistently outperforms CIL and FIL methods in both expert specialization and generalization capabilities, as shown in Fig. 2.

2 Relate Work

Class Incremental Learning CIL requires a model to sequentially acquire knowledge from a series of new classes while maintaining its performance on previously learned classes. Existing CIL literature can be broadly categorized into three strategic streams:

Replay-based methods Rebuffi et al. (2017); Douillard et al. (2020); Wang et al. (2025c;a); Xu et al. (2026a), which mitigate forgetting by storing a small subset of old data (exemplars) or using generative models to synthesize pseudo-samples from previous distributions. SPARK Wang et al. (2025c) models inter-class relationships with a fuzzy cognitive graph over class prototypes and derives class-level embeddings via feature aggregation. CoLA Wang et al. (2025a) alleviates distribution aliasing in CIL by leveraging two levels supervised contrastive augmentations.

Regularization-based approaches Fan et al. (2024b;a); Zang et al. (2026); Li et al. (2025b); Zheng et al. (2025), which employ techniques such as knowledge distillation or parameter-level constraints to penalize significant shifts in the model's representation space during new task adaptation. TaKP Zang et al. (2026) leverages persistent homology to extract stable topological structures and preserves the global feature geometry via a topology alignment mechanism. TagFex Zheng et al. (2025) learns task-agnostic features with an independent model and fuses them with task-specific representations via attention to enhance feature diversity.

Prototype-based methods Gao et al. (2025); Chen et al. (2025); Ni et al. (2025); Xu et al. (2026b); Jiang et al. (2026) establish a prototype for each task and update the prototypes in subsequent phases, classifying samples into the category of the most similar prototype in the inference phase. CREATE Chen et al. (2025) assigns each class a lightweight auto-encoder to learn a compact latent space, ensuring only correct-class samples reconstruct well. Gao et al. (2025) globally normalizes teacher and student logits to resolve the conflict between distillation and cross-entropy learning.

Federated Incremental Learning FIL combines federated learning with continual learning, aiming to enable multiple clients to collaboratively learn from sequentially arriving data streams without sharing raw data. Current research in FIL primarily focuses on two directions:

Distillation-based Regularization Yu et al. (2024b; 2025); Zhang et al. (2025); Nori et al. (2025), which employs global-to-local or peer-to-peer knowledge distillation to maintain consistency across historical and current task representations. pFedMxF Zhang et al. (2025) addresses spatial, temporal and resource heterogeneity via frequency decomposition and an auto-task agnostic classifier. HR Nori et al. (2025) uses a tailored autoencoder with latent exemplars and Lennard-Jones Potential to generate synthetic data.

Architecture-based Isolation McMahan et al. (2017); Yu et al. (2024a); Li et al. (2024); Salami et al. (2025); Yi et al. (2026), which utilizes parameter decomposition or sub-network masking to mitigate inter-task interference in a decentralized manner. pFedDIL Li et al. (2024) uses a parameter-shared auxiliary classifier to measure task correlation, enabling clients to select strategies and transfer knowledge adaptively. pFedMoE Yi et al. (2026) constructs a mixture-of-experts architecture consisting of heterogeneous local experts, shared global proxy experts, and gating networks.

Socialized Learning SL is a paradigm in which individuals acquire knowledge and skills through interaction and collaboration, leveraging collective intelligence to improve learning efficiency and overall system capability Zhao et al. (2021); Wang et al. (2025b).

Existing socialized learning methods primarily focus on information exchange and collaborative learning among multiple agents. A2SL Carpentiero et al. (2024) introduces a dual-adaptation mechanism to handle drifts in hypotheses and likelihood models, while CoSCL Wang et al. (2022) employs a fixed number of narrow sub-networks to learn incremental tasks in parallel. S2L Dong et al. (2024) extends socialized learning to multi-agent reinforcement learning for complex tasks with sparse rewards and sequential structures, and MASC Yao et al. (2024) further exploits interactions among multiple agents to facilitate knowledge sharing. Other studies Yao et al. (2025); Sun et al. (2025) explore cross-task collaboration to coordinate collective capabilities, demonstrat-

ing that effective interaction and knowledge sharing among agents can improve learning efficiency, adaptability, and collective intelligence.

More recent studies have extended socialized learning toward structured, heterogeneous, and complex real-world environments. GeoMoE Cao et al. (2026) leverages Ollivier-Ricci curvature to achieve geometry-guided routing and fusion across Riemannian experts for heterogeneous graph structures, while GHG Li et al. (2025a) introduces Graph Socialized Learning by autonomously selecting interaction partners and accumulating customized graph knowledge through synthetic graphs. AGPC Guo et al. (2026) further explores progressive cross-task collaboration in heterogeneous air-ground perception, enabling coarse-to-fine information exchange across localization, target association, and identity-aware parsing.

3 Problem Formulation

Let $E = \{E_1, E_2, \ldots, t\}$ denotes the T experts, where each expert possesses specialized knowledge acquired from its own data distribution. X is a input space and Y is a label space. The private dataset of the t-th expert is denoted by $\mathcal{D}_t = \{(x_i, y_i)\}_{i=1}^{M_t}$. where M_t is the number of training samples available to t, $x_i \in X$, and $y_i \in Y_t \subseteq Y$.

For example, consider three experts introduced over successive stages. The second stage can therefore be expressed as $(E_1, E_2) \xrightarrow{\text{SCL}} (E_1^{(1)}, E_2^{(1)})$. Initially, two experts E_1 and E_2 satisfy $Y_{E_1} \cap Y_{E_2} = \emptyset$. After the initial socialized learning stage, E_1 and E_2 exchange their complementary expertise and become capable of recognizing the combined class space $Y_{E_1} \cup Y_{E_2}$.

When a new expert E_3 subsequently emerges, the previously learned experts can continue to interact with E_3 to acquire its new expertise, while retaining the knowledge gained from earlier socialized interactions. The second stage can therefore be expressed as $(E_1^{(1)}, E_2^{(1)}, E_3) \xrightarrow{\text{SCL}} (E_1^{(2)}, E_2^{(2)}, E_3^{(2)})$, where the cumulative knowledge space is expanded to $Y_{E_1} \cup Y_{E_2} \cup Y_{E_3}$.

For E_3, the newly learned general classes, corresponding to $Y_{E_1} \cup Y_{E_2}$, should be learned as well as possible while maintaining the performance of the original expert classes (Y_{E_3}) directionally, i.e., aspiring to possess both versatility and directionality.

Versatility Each expert effectively learns a broader range of new general classes, i.e., E_1, $E_3 2$ and E_3 can infer the classes $Y_{E_1} \cup Y_{E_2} \cup Y_{E_3}$ with precision. **Directionality** Each expert effectively learns new general classes while directionally maintaining its superior performance in its respective expert classes, i.e., for $E_3^{(2)}$,

$$\arg \min_{k \in \{E_1^{(2)}, E_2^{(2)}, E_3^{(2)}\}} \mathbb{E}_{x \sim P_{X_{E_3}}} \left[1 - \max_{y \in Y_{E_3}} P(y \mid x; E_k^{(2)}) \right] = E_3^{(2)}. \tag{1}$$

4 Methodology

RWSE is designed for continual expert socialization, where the first expert provides the initial shared model and subsequently arriving experts are incrementally integrated into the existing collaborative state, as shown in Figure 3.

We first consider a single editable linear mapping $W \in \mathbb{R}^{d_{\text{in}} \times d_{\text{out}}}$. Let $\{t\}_{t=1}^{T}$ denote a sequence of experts arriving incrementally. The expert t is represented by a set of functional correspondences (K_t, V_t), where $K_t \in \mathbb{R}^{n_t \times d_{\text{in}}}$ contains the inputs to the editable linear mapping and $V_t \in \mathbb{R}^{n_t \times d_{\text{out}}}$ is the corresponding expert output. In particular,

$$V_t = K_t W_t, \tag{2}$$

where W_t is the corresponding parameter matrix of expert E_t.

4.1 Constrained Functional Alignment

At each collaborative step, the new expert introduces a functional alignment objective

$$\mathcal{L}_{\text{new}}(W) = \lambda \left\| K_t W - V_t \right\|_F^2, \tag{3}$$

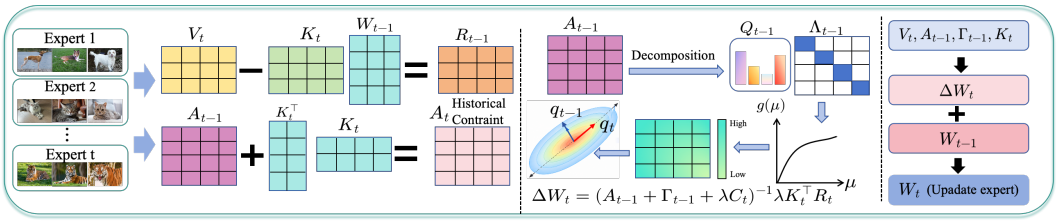


Figure 3: Illustration of our proposed Recursive Weight-Space Editing for Scalable Multi-Expert Collaboration. RWSE enables continual expert integration through recursive historical statistics, residual-based weight editing, and spectrally anisotropic protection.

where $\lambda > 0$ controls the strength of the incoming expertise.

To preserve the accumulated knowledge without storing historical raw feature matrices, RWSE maintains a recursive second-order statistic

$$C_t = K_t^\top K_t, \tag{4}$$

and accumulates the historical functional precision as

$$A_t = A_{t-1} + \lambda C_t, (A_1 = \lambda K_1^\top K_1 + \epsilon I, \epsilon > 0) \tag{5}$$

Here, A_{t-1} summarizes the accumulated feature statistics of previously socialized experts. Large eigenvalues of A_{t-1} indicate parameter directions that have been strongly constrained by historical feature distributions. Thus, instead of explicitly replaying historical functional correspondences, RWSE uses A_{t-1} as a compact history-aware metric to control subsequent weight updates.

We therefore formulate the recursive editing step as

$$\begin{aligned} W_t = \arg\min_W \lambda \left\| K_t W - V_t \right\|_F^2 \\ + \operatorname{Tr}\left[(W - W_{t-1})^\top (A_{t-1} + \Gamma_{t-1}) (W - W_{t-1}) \right], \end{aligned} \tag{6}$$

where Γ_{t-1} is a directional protection matrix constructed from the historical precision spectrum, as described in Section 4.3. The first term in Eq. 6 encourages the current shared model to reproduce the new expert's functional behavior, whereas the second term penalizes deviations from the previously socialized state according to the historical importance of different parameter directions.

4.2 Recursive Closed-form Editing

Taking the derivative of Eq. 6 with respect to W and setting it to zero gives

$$\left(A_{t-1} + \Gamma_{t-1} + \lambda K_t^\top K_t\right) W_t = (A_{t-1} + \Gamma_{t-1}) W_{t-1} + \lambda K_t^\top V_t. \tag{7}$$

Define the current functional residual as

$$R_t = V_t - K_t W_{t-1}. \tag{8}$$

$$K_t^\top V_t = K_t^\top K_t W_{t-1} + K_t^\top R_t, \tag{9}$$

Eq. 7 can be rearranged as

$$(A_{t-1} + \Gamma_{t-1} + \lambda C_t) (W_t - W_{t-1}) = \lambda K_t^\top R_t. \tag{10}$$

Therefore, the exact recursive editing rule is

$$W_t = W_{t-1} + (A_{t-1} + \Gamma_{t-1} + \lambda C_t)^{-1} \lambda K_t^\top (V_t - K_t W_{t-1}) \tag{11}$$

Eq 11 provides a direct residual-correction interpretation. $V_t - K_t W_{t-1}$ measures the expertise gap between the incoming expert and the current socialized model, while $K_t^\top (V_t - K_t W_{t-1})$ maps this functional discrepancy back to the weight space. The inverse matrix $(A_{t-1} + \Gamma_{t-1} + \lambda C_t)^{-1}$ then determines the allowable update magnitude according to both historical importance and current feature statistics. Consequently, directions that are strongly constrained by previous experts receive smaller corrections, whereas less constrained directions retain greater plasticity.

4.3 SPECTRALLY ANISOTROPIC HISTORICAL PROTECTION

A conventional isotropic proximal term, $\gamma\|W - W_{t-1}\|_F^2$, assigns the same penalty to all parameter directions. Such a constraint cannot distinguish directions that are important to previously socialized experts from those that remain relatively unconstrained.

To obtain direction-dependent protection, we decompose the historical precision matrix as

$$A_{t-1} = Q_{t-1}\Lambda_{t-1}Q_{t-1}^\top, \tag{12}$$

where $Q_{t-1}^\top Q_{t-1} = I$, $Q_{t-1} = [q_1, \ldots, q_{d_{\text{in}}}]$ and $\Lambda_{t-1} = \text{diag}(\mu_1, \ldots, \mu_{d_{\text{in}}}), \mu_i > 0$. q_i defines a principal parameter direction, where μ_i quantifies the historical constraint strength along that direction, with larger values indicating stronger constraints from previously socialized experts.

Rather than simply rescaling A_{t-1}, RWSE applies a nonlinear spectral transformation:

$$\Gamma_{t-1} = \gamma Q_{t-1} g(\Lambda_{t-1}) Q_{t-1}^\top \tag{13}$$

where $\gamma > 0$ controls the overall protection strength and $g(\cdot)$ is a monotone nonlinear mapping.

$$g(\mu) = \left(\frac{\mu}{\bar{\mu}_{t-1} + \epsilon}\right)^p, \qquad p > 1, \tag{14}$$

$$\bar{\mu}_{t-1} = \frac{1}{d_{\text{in}}} \sum_{i=1}^{d_{\text{in}}} \mu_i. \tag{15}$$

The corresponding anisotropic regularization is

$$\mathcal{L}_{\text{ani}}(W) = \text{Tr}\left[(W - W_{t-1})^\top \Gamma_{t-1}(W - W_{t-1})\right]. \tag{16}$$

Substituting the eigendecomposition into Eq. 16 yields

$$\mathcal{L}_{\text{ani}} = \sum_{i=1}^{d_{\text{in}}} \gamma g(\mu_i) \left\| q_i^\top (W - W_{t-1}) \right\|_2^2. \tag{17}$$

$g(\mu)$ is nonlinear and monotonically increasing, directions with larger historical eigenvalues receive disproportionately larger penalties. Therefore, $\mu_i \uparrow \Longrightarrow \gamma g(\mu_i) \uparrow$, which suppresses updates along historically important directions while retaining greater flexibility in less constrained directions.

RWSE imposes a soft directional constraint rather than an exact orthogonal projection. Historical sensitivity is encoded through direction-dependent curvature, providing a continuous trade-off between knowledge preservation and adaptation. The historical precision matrix is updated as

$$A_t = A_{t-1} + \lambda K_t^\top K_t = \lambda \sum_{i=1}^{t} K_i^\top K_i + \epsilon I. \tag{18}$$

After expert E_i has been integrated, raw feature matrix K_i does not need to be retained. Its contribution is compressed into the second-order statistic $K_i^\top K_i$. At step t, RWSE only requires the current expert's (K_t, V_t), the previous shared model W_{t-1}, and the accumulated historical matrix A_{t-1}. The detailed description of Efficient Recursive Editing can be found in A.5

5 EXPERIMENTS

5.1 EXPERIMENTS SETUP

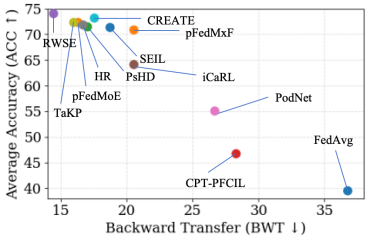
Datasets. We evaluate our method on the widely used CIFAR-100 Krizhevsky & Hinton (2019) and ImageNet-100 Deng et al. (2009), with details provided in Appendix A.2.

Implementation Details. We consider two class-wise data partitioning settings: CIFAR-100-20/5, where 100 classes are evenly distributed across 5 experts (20 classes per expert), and CIFAR-100-10/10, where they are distributed across 10 experts (10 classes per expert).

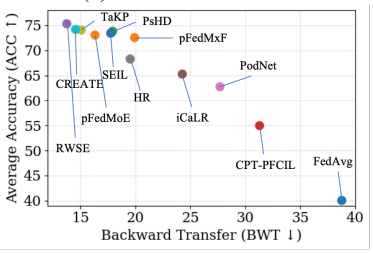
Table 1: Performance comparison on CIFAR-100 and ImageNet-100 datasets (Accuracy %).

Method	CIFAR-100				ImageNet-100			
	10/10		20/5		10/10		20/5	
	last	avg	last	avg	last	avg	last	avg
iCaRL (CVPR'2017)	48.74	64.15	52.73	65.58	50.54	65.37	60.21	71.36
PodNet (ECCV'2020)	44.41	55.16	48.59	61.81	44.58	62.76	57.62	70.47
PsHD (NeurIPS'2024)	61.38	71.49	65.81	71.67	64.14	73.90	67.88	74.13
CREATE (CVPR'2025)	62.33	73.17	66.85	76.04	64.43	74.28	71.77	77.96
SEIL (PR'2026)	62.74	71.43	59.75	69.51	63.26	73.54	61.76	70.82
TaKP (AAAI'2026)	62.96	72.32	66.96	75.71	63.84	74.08	71.69	78.32
FedAvg (AISTATS'2017)	23.62	39.54	25.35	41.21	23.88	40.13	26.22	43.06
CPT-PFCIL (ICASSP'2025)	29.75	46.79	32.43	44.26	38.87	55.04	41.51	55.85
HR (ICLR'2025)	61.52	71.81	59.55	70.15	55.79	68.37	52.46	65.58
pFedMxF (CVPR'2025)	61.88	70.94	62.37	71.63	60.31	72.55	61.87	72.39
pFedMoE (TKDE'2026)	62.79	72.38	65.44	74.86	62.77	73.16	64.93	75.54
RWSE	**63.85**	**74.13**	**67.79**	**77.34**	**65.28**	**75.31**	**72.96**	**79.52**

Metrics. We compare the average accuracy:$ACC = \frac{1}{N}\sum_{i=1}^{N} A_{N,i}$, where $A_{N,i}$ is the accuracy of the i-th task. Analysis with backward transfer, which presents the degree of forgetting $BWT = \frac{1}{N-1}\sum_{i=1}^{N-1} A_{N,i} - A_{i,i}$, where $A_{i,j}$ represents the test accuracy in j-th task after completely learn i-th task. The forgetting is usually reflected as negative BWT.



(a) CIFAR-100-10/10



(b) ImageNet-100-10/10

Figure 4: Trade-off between ACC and BWT for different methods.

5.2 Experimental Results

Comparative performance. We compare to CIL methods and FIL methods. As shown in Table 1, RWSE consistently achieves the best performance across all datasets and settings, demonstrating its effectiveness for sequential expert integration. It outperforms both rehearsal-based methods (iCaRL, PodNet) and recent strong baselines (CREATE, TaKP), showing that direct functional alignment in weight space is more effective than feature-level regularization. Compared with federated and multi-expert methods (FedAvg, pFedMoE), RWSE avoids performance degradation caused by naive parameter aggregation, benefiting from its structured, statistics-driven weight editing. RWSE achieves better stability–plasticity trade-off by preserving high-precision historical knowledge while efficiently integrating new expertise.

Analysis of Knowledge Preservation and Backward Transfer. Fig. 4 illustrates the trade-off between average accuracy (avg) and backward transfer (BWT) among different methods. Methods located in the upper-left region achieve a better trade-off between accuracy and knowledge retention, as they simultaneously yield higher accuracy and lower forgetting. In contrast, RWSE consistently dominates all competing methods, achieving the highest avg together with the lowest BWT. Compared with all baselines, RWSE establishes a superior balance between knowledge acquisition and forgetting mitigation, indicating that recursive weight-space editing can effectively integrate new knowledge while preserving previously learned representations.

5.3 The Necessity of Versatility and Directionality

We believe that the specific directionality and task generalization of expert models are not mutually exclusive, but should promote each other to achieve more robust continuous collaborative learning.

Table 2: Comparison of detailed accuracy across different classes before and after collaboration.

Method	CIFAR-100						ImageNet-100					
	Before Collaboration			After Collaboration			Before Collaboration			After Collaboration		
	Expert	General	Avg	Expert	General	Avg	Expert	General	Avg	Expert	General	Avg
iCaRL	72.54	0.00	36.27	34.76	49.18	41.97	68.72	0.00	34.36	34.85	48.88	41.87
PODNet	72.27	0.00	36.14	61.63	45.74	53.69	68.43	0.00	34.25	52.33	38.73	45.53
PsHD	72.42	0.00	36.21	51.76	65.34	58.55	68.86	0.00	34.43	43.71	55.62	49.67
CREATE	72.53	0.00	36.27	60.39	66.52	63.46	68.75	0.00	34.38	54.35	62.24	58.30
SEIL	72.58	0.00	36.29	51.03	63.57	57.30	68.59	0.00	34.30	50.17	59.41	53.79
TaKP	72.78	0.00	36.39	59.91	66.88	63.40	68.87	0.00	34.44	54.69	62.87	58.78
FedAvg	70.96	0.00	35.48	60.35	56.77	58.56	67.35	0.00	33.68	57.49	53.76	55.63
CPT-PFCIL	70.85	0.00	35.43	58.43	61.26	59.85	67.13	0.00	33.57	54.37	57.84	56.11
HR	71.12	0.00	35.56	63.77	59.34	61.56	67.31	0.00	33.66	60.55	57.34	58.95
pFedMxF	71.34	0.00	35.67	64.48	59.75	62.12	67.46	0.00	33.73	60.76	57.02	58.89
pFedMoE	71.28	0.00	35.64	66.19	61.25	62.87	67.59	0.00	33.80	60.97	57.58	59.27
RWSE	72.65	0.00	36.33	**69.06**	**65.43**	**67.25**	68.57	0.00	34.29	**64.31**	**61.44**	**62.88**

In this context, generality represents the ability of intelligent agents to capture a wide range of classes of knowledge, while directionality reflects their depth in maintaining specialized performance in expert specific fields. To quantify this characteristic, following the MASC setting Yao et al. (2024), we conducted a deep analysis of different methods by comparing the average accuracy of all classes with the accuracy differences between expert classes and the best performing classes.

We report the performance on the CIFAR-100-10/10 and ImageNet-100-10/10, as shown in Tables 2. **CL forgets original knowledge during growth:** while CIL methods maintain general performance, "expert classes" suffer from rapid degradation. Treating recent tasks as experts fails to resolve this, as catastrophic forgetting prevents these classes from remaining a stable, high-performance set throughout the learning trajectory. **FIL loses sight of the focus among multiple agents:** although FIL shows competitive accuracy, it lacks a prioritized focus. The performance gap between expert and general classes is negligible, suggesting that simplistic aggregation lacks the "directional bias" needed to preserve pre-established expert preferences, leading to a loss of specialized intelligence.

Analysis of versatility and directionality: Versatility reflects the ability of an expert to generalize to broader classes, while directionality measures its capability to preserve the performance of expert classes during knowledge transfer. To evaluate these properties, we analyze the average accuracy over all classes and the performance gap between expert classes and the best-performing general classes. As illustrated in Fig. 5, CIL methods suffer from significant forgetting of expert knowledge. In contrast, FIL methods yield orange bar values close to zero, suggesting insufficient directionality in preserving expert-class performance.

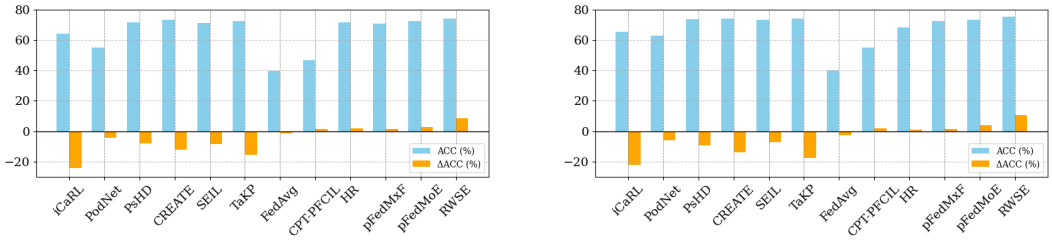
Table 3: Performance comparison on generated datasets under the 10/10 setting (Accuracy %).

Method	CIFAR-100		ImageNet-100	
	last	avg	last	avg
iCaRL	21.07	30.18	22.65	31.67
PodNet	20.34	23.48	20.33	27.69
PsHD	27.49	32.85	28.61	34.72
CREATE	28.53	34.59	29.54	35.12
SEIL	27.87	34.21	28.26	34.91
TaKP	28.54	34.87	28.49	35.45
FedAvg	14.31	21.42	15.14	22.31
CPT-PFCIL	16.43	24.72	17.68	26.85
HR	26.86	31.57	23.96	29.43
pFedMxF	27.44	32.69	26.88	31.56
pFedMoE	28.29	33.95	27.92	33.46
RWSE	**37.97**	**44.29**	**38.74**	**46.52**

5.4 Data-driven Knowledge Distillation

Knowledge distillation involves various methods for extracting knowledge from expert models. In practice, most existing machine learning methods are data-driven, which has led to a direct idea: can task specific knowledge be distilled from data? Inspired by MASC Yao et al. (2024), we utilized the unique data of each expert to obtain the corresponding mean and variance (see supplementary materials for details), thereby generating training data without directly contacting real samples.

As shown in Table 3, all methods degrade notably without real data, revealing the strong dependence of conventional approaches on high-quality samples. CIL methods remain limited, indicating that feature- or logit-level distillation is insufficient under synthetic distributions, while FIL methods only partially mitigate this issue due to distribution mismatch. In contrast, RWSE consistently out-

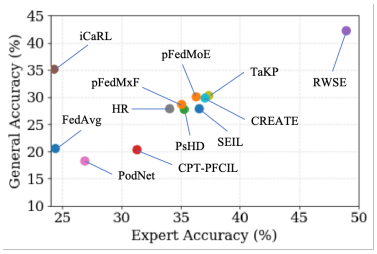


(a) CIFAR-100-10/10 (b) ImageNet-100-10/10

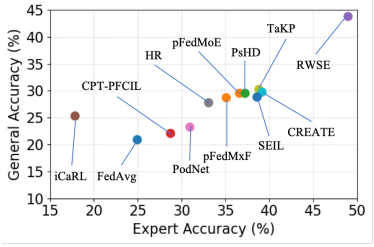
Figure 5: Analysis of versatility and directionality. Blue denote the average ACC across all classes, while orange denote the difference in ACC between expert classes and the best-performing general classes. Blue and orange bars correlate positively with versatility and directionality, respectively.

performs all baselines on both datasets, achieving over 9% improvement in average accuracy. This gain highlights its ability to integrate expert knowledge without relying on real data.

As shown in Fig. 6, we evaluate knowledge distillation with generated data by comparing performance on expert and general classes. Across both CIFAR-100-10/10 and ImageNet-100-10/10, most methods achieve higher accuracy on general classes, indicating a bias of synthetic data toward general semantics and insufficient preservation of expert knowledge. CIL methods exhibit large performance gaps, reflecting severe degradation of expert classes, while FIL methods partially reduce this gap but remain limited by distribution mismatch. In contrast, RWSE achieves the best performance on both expert and general classes, demonstrating a superior balance between directionality and versatility.



(a) Comparison on generated CIFAR-100



(b) Comparison on generated ImageNet-100

Figure 6: Performance comparison after collaboration.

5.5 Ablation Study

To further assess the contribution of each component in RWSE, we conduct an ablation study as reported in Table 4. We evaluate the effects of Recursive Editing (RE) and Spectrally Anisotropic Regularization (SAR). Adding SAR to the RE baseline yields substantial gains in average accuracy, improving performance by 2.14% on CIFAR-100 and 2.51% on ImageNet-100. These results demonstrate that the nonlinear spectral regularization can provide stronger protection for historically sensitive directions than isotropic parameter constraints, while preserving flexibility for incorporating new expertise. The full RWSE framework further achieves the best overall performance, highlighting the complementary roles of recursive functional integration and direction-dependent historical protection in balancing knowledge preservation with continual expertise expansion.

Table 4: Ablation study of the key components.

Dataset	RE	AA	Expert	General	Avg.
CIFAR-100	✓	×	72.57	69.05	70.81
	✓	✓	75.48	72.78	74.13
ImageNet-100	✓	×	73.61	68.85	71.23
	✓	✓	77.16	73.46	75.31

6 Conclusion

We studied socialized continual learning in dynamic environments and identified key limitations of existing approaches, including interference between experts and degradation of specialized knowledge under sequential updates. We proposed Recursive Weight Space Editing, which reframes knowledge integration as constrained functional alignment in weight space. The resulting recursive closed-form update enables incremental knowledge incorporation without storing historical data, while a manifold-aware anisotropic regularization mitigates inter-expert interference by constraining updates along critical directions. This yields a memory-efficient and stable mechanism for continual knowledge accumulation. Empirical results demonstrat its effectiveness in balancing plasticity and stability.

AI USE STATEMENT

Generative AI tools were used solely to assist with language polishing, grammar checking, and improving the clarity, readability, and overall presentation of the manuscript. They were not used to develop the proposed methodology, formulate mathematical claims, design or conduct experiments, analyze results, or draw scientific conclusions. All AI-assisted suggestions and revisions were carefully reviewed and verified by the authors, who take full responsibility for the final content of the manuscript.

REFERENCES

Haifang Cao, Yu Wang, Timing Li, Xinjie Yao, and Pengfei Zhu. Geometric mixture-of-experts with curvature-guided adaptive routing for graph representation learning. *CoRR*, abs/2603.22317, 2026. URL https://doi.org/10.48550/arXiv.2603.22317.

Marco Carpentiero, Virginia Bordignon, Vincenzo Matta, and Ali H. Sayed. Social learning with adaptive models. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2024, Seoul, Republic of Korea, April 14-19, 2024*, pp. 9351–9355. IEEE, 2024. doi: 10.1109/ICASSP48485.2024.10447521.

Huitong Chen, Yu Wang, Yan Fan, Guosong Jiang, and Qinghua Hu. Reducing class-wise confusion for incremental learning with disentangled manifolds. In *CVPR*, pp. 10121–10130. Computer Vision Foundation / IEEE, 2025.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pp. 248–255. IEEE Computer Society, 2009. doi: 10.1109/CVPR.2009.5206848.

Shaokang Dong, Chao Li, Wubing Chen, Hongye Cao, Wenbin Li, and Yang Gao. Multi-agent exploration via self-learning and social learning. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2024, Seoul, Republic of Korea, April 14-19, 2024*, pp. 5055–5059. IEEE, 2024. doi: 10.1109/ICASSP48485.2024.10446068.

Arthur Douillard, Matthieu Cord, Charles Ollion, Thomas Robert, and Eduardo Valle. Podnet: Pooled outputs distillation for small-tasks incremental learning. In *ECCV*, volume 12365 of *Lecture Notes in Computer Science*, pp. 86–102. Springer, 2020. doi: 10.1007/978-3-030-58565-5_6.

Yan Fan, Yu Wang, Pengfei Zhu, Dongyue Chen, and Qinghua Hu. Persistence homology distillation for semi-supervised continual learning. In *NeurIPS*, 2024a.

Yan Fan, Yu Wang, Pengfei Zhu, and Qinghua Hu. Dynamic sub-graph distillation for robust semi-supervised continual learning. In *AAAI*, pp. 11927–11935. AAAI Press, 2024b.

Zijian Gao, Shanhao Han, Xingxing Zhang, Kele Xu, Dulan Zhou, Xinjun Mao, Yong Dou, and Huaimin Wang. Maintaining fairness in logit-based knowledge distillation for class-incremental learning. In *AAAI*, pp. 16763–16771. AAAI Press, 2025.

Zhoupeng Guo, Yunqi Zhu, Zhihe Fan, Xinjie Yao, Ruipu Zhao, Boan Tao, Yiming Sun, Zhen Wang, and Pengfei Zhu. Rethinking air-ground collaboration: A progressive cross-task benchmark and socialized learning framework. *CoRR*, abs/2606.18841, 2026. doi: 10.48550/ARXIV.2606.18841. URL https://doi.org/10.48550/arXiv.2606.18841.

Wenke Huang, Mang Ye, Zekun Shi, Guancheng Wan, He Li, Bo Du, and Qiang Yang. Federated learning for generalization, robustness, fairness: A survey and benchmark. *IEEE Trans. Pattern Anal. Mach. Intell.*, 46(12):9387–9406, 2024. doi: 10.1109/TPAMI.2024.3418862.

Kai Jiang, Xueru Bai, and Feng Zhou. Recurrent network expansion for class incremental learning. *IEEE Trans. Neural Networks Learn. Syst.*, 37(1):122–135, 2026. doi: 10.1109/TNNLS.2025.3601373.

Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. 2019.

Jialu Li, Yu Wang, Pengfei Zhu, Wanyu Lin, Xinjie Yao, and Qinghua Hu. Graphs help graphs: Multi-agent graph socialized learning. In *NeurIPS*, 2025a.

Jiashuo Li, Shaokun Wang, Bo Qian, Yuhang He, Xing Wei, Qiang Wang, and Yihong Gong. Dynamic integration of task-specific adapters for class incremental learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 30545–30555. Computer Vision Foundation / IEEE, 2025b. doi: 10.1109/CVPR52734.2025.02844.

Yichen Li, Wenchao Xu, Haozhao Wang, Yining Qi, Jingcai Guo, and Ruixuan Li. Personalized federated domain-incremental learning based on adaptive knowledge matching. In Ales Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part XLVI*, Lecture Notes in Computer Science, pp. 127–144. Springer, 2024. doi: 10.1007/978-3-031-72952-2_8.

Yichen Li, Haozhao Wang, Yining Qi, Wei Liu, and Ruixuan Li. Re-fed+: A better replay strategy for federated incremental learning. *IEEE Trans. Pattern Anal. Mach. Intell.*, 47(7):5489–5500, 2025c. doi: 10.1109/TPAMI.2025.3551732.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In Aarti Singh and Xiaojin (Jerry) Zhu (eds.), *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, AISTATS 2017, 20-22 April 2017, Fort Lauderdale, FL, USA*, Proceedings of Machine Learning Research, pp. 1273–1282. PMLR, 2017.

Wenyao Ni, Jiangrong Shen, Qi Xu, and Huajin Tang. ALADE-SNN: adaptive logit alignment in dynamically expandable spiking neural networks for class incremental learning. In *AAAI*, pp. 19712–19720. AAAI Press, 2025. doi: 10.1609/AAAI.V39I18.34171.

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 *ICLR*. OpenReview.net, 2025.

Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H. Lampert. icarl: Incremental classifier and representation learning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5533–5542, 2017. doi: 10.1109/CVPR.2017.587.

Xuankun Rong, Jianshu Zhang, Kun He, and Mang Ye. CAN: leveraging clients as navigators for generative replay in federated continual learning. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, Proceedings of Machine Learning Research. PMLR / OpenReview.net, 2025.

Riccardo Salami, Pietro Buzzega, Matteo Mosconi, Jacopo Bonato, Luigi Sabetta, and Simone Calderara. Closed-form merging of parameter-efficient modules for federated continual learning. In *ICLR*. OpenReview.net, 2025.

Yiming Sun, Xin Li, Pengfei Zhu, Qinghua Hu, Dongwei Ren, Huiying Xu, and Xinzhong Zhu. Task-gated multi-expert collaboration network for degraded multi-modal image fusion. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, Proceedings of Machine Learning Research. PMLR / OpenReview.net, 2025.

Sumers T. Thompson B., Van Opheusden B. and Griffiths T. Complex cognitive algorithms preserved by selective social learning in experimental populations. In *Science*, volume 376, 2022.

Liyuan Wang, Xingxing Zhang, Qian Li, Jun Zhu, and Yi Zhong. Coscl: Cooperation of small continual learners is stronger than a big one. In Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (eds.), *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXVI*, Lecture Notes in Computer Science, pp. 254–271. Springer, 2022. doi: 10.1007/978-3-031-19809-0_15.

Xi Wang, Xu Yang, Kun Wei, Yanan Gu, and Cheng Deng. Class incremental learning via contrastive complementary augmentation. *IEEE Trans. Image Process.*, 34:3663–3673, 2025a.

Xiaofei Wang, Yunfeng Zhao, Chao Qiu, Qinghua Hu, and Victor C. M. Leung. Socialized learning: A survey of the paradigm shift for edge intelligence in networked systems. *IEEE Commun. Surv. Tutorials*, 27(3):2085–2128, 2025b. doi: 10.1109/COMST.2024.3482978.

Yu Wang, Jiabo Xie, Junyan Zheng, Bingxu Lu, Yanxian Bi, and Qinghua Hu. SPARK: simple and parameter-free knowledge embedding with fuzzy cognitive maps for class incremental learning. *IEEE Trans. Fuzzy Syst.*, 33(8):2513–2525, 2025c.

Zhengjin Xu, Xuyang Li, Xiaobin Chang, Wei-Shi Zheng, and Ruixuan Wang. Slowly expanding neural network for class incremental learning. *Pattern Recognit.*, 171:112213, 2026a. doi: 10.1016/J.PATCOG.2025.112213.

Zhengjin Xu, Xuyang Li, Xiaobin Chang, Wei-Shi Zheng, and Ruixuan Wang. Slowly expanding neural network for class incremental learning. *Pattern Recognit.*, 171:112213, 2026b. doi: 10.1016/J.PATCOG.2025.112213.

Xinjie Yao, Yu Wang, Pengfei Zhu, Wanyu Lin, Jialu Li, Weihao Li, and Qinghua Hu. Socialized learning: Making each other better through multi-agent collaboration. In *ICML*, Proceedings of Machine Learning Research, pp. 56927–56945. PMLR / OpenReview.net, 2024.

Xinjie Yao, Yu Wang, Pengfei Zhu, Wanyu Lin, Ruipu Zhao, Zhoupeng Guo, Weihao Li, and Qinghua Hu. Socialized coevolution: Advancing a better world through cross-task collaboration. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, Proceedings of Machine Learning Research. PMLR / OpenReview.net, 2025.

Liping Yi, Han Yu, Gang Wang, Xiaoguang Liu, and Qinghua Hu. pfedmoe: Data-level personalization with mixture of experts in model-heterogeneous personalized federated learning. *IEEE Trans. Knowl. Data Eng.*, 38(3):1905–1918, 2026. doi: 10.1109/TKDE.2026.3656194.

Hao Yu, Xin Yang, Xin Gao, Yihui Feng, Hao Wang, Yan Kang, and Tianrui Li. Overcoming spatial-temporal catastrophic forgetting for federated class-incremental learning. In *Proceedings of the 32nd ACM International Conference on Multimedia, MM 2024, Melbourne, VIC, Australia, 28 October 2024 - 1 November 2024*, pp. 5280–5288. ACM, 2024a.

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 *CVPR*, pp. 4874–4883. Computer Vision Foundation / IEEE, 2025. doi: 10.1109/CVPR52734.2025.00459.

Lang Yu, Lina Ge, Guanghui Wang, Jianghao Yin, Qin Chen, Jie Zhou, and Liang He. Pi-fed: Continual federated learning with parameter-level importance aggregation. *IEEE Internet Things J.*, 11(22):37187–37199, 2024b.

Han Zang, Yongfeng Dong, Linhao Li, Liang Yang, and Yu Wang. Topology-aware knowledge preservation for class-incremental learning. In Sven Koenig, Chad Jenkins, and Matthew E. Taylor (eds.), *AAAI*, pp. 28059–28067. AAAI Press, 2026. doi: 10.1609/AAAI.V40I33.40031.

Yifei Zhang, Hao Zhu, Alysa Ziying Tan, Dianzhi Yu, Longtao Huang, and Han Yu. pfedmxf: Personalized federated class-incremental learning with mixture of frequency aggregation. In *CVPR*, pp. 30640–30650. Computer Vision Foundation / IEEE, 2025.

Yunfeng Zhao, Zhicheng Liu, Chao Qiu, Xiuhua Li, Xiaofei Wang, and Qinghua Hu. Socialized learning for smart cities: Cognitive paradigm, methodology, and solution. *IEEE Wirel. Commun.*, 28(5):200–208, 2021. doi: 10.1109/MWC.111.2100023.

Bowen Zheng, Da-Wei Zhou, Han-Jia Ye, and De-Chuan Zhan. Task-agnostic guided feature expansion for class-incremental learning. In *CVPR*, pp. 10099–10109. Computer Vision Foundation / IEEE, 2025.

A Appendix

A.1 Implementation Details

In this section, we provide a detailed description of the datasets in the main paper. We employ ResNet-18 as the backbone network and optimize the model using stochastic gradient descent (SGD) with a momentum of 0.9 and a weight decay of 2e-4. Each incremental session is trained for 80 epochs with a batch size of 64. All experiments are conducted on NVIDIA RTX 4090 GPU equipped with 24GB memory. The learning rate is set to 0.001, λ is set to 1, and γ is set to 0.8. All results are reported in means over 3 trials.

A.2 Datasets

In this subsection, we provide an introduction to the datasets used in the main paper. We evaluate the performance on CIFAR-100 and ImageNet-100. CIFAR-100 is a classification dataset with 60,000 32 × 32 RGB images from 100 classes. Each class contains 500 training images and 100 testing images. ImageNet-100 is composed of 100 classes with 1,300 images per class for training and 500 images per class for validation. ImageNet-100 resembles real-world scenes with a higher resolution of 256 × 256.

We employ two versions of dataset partitioning. Taking CIFAR-100 as an example: first method involved dividing the dataset evenly among 5 experts (abbreviated as expert-class, general-class-1, general-class-2, general-class-3, and general-class-4), with each expert handling 20 classes. This setup is denoted as CIFAR-100-20/5; second method involved splitting the dataset equally among 10 experts (abbreviated as expert-class, general-class-1, general-class-2, general-class-3, general-class-4, general-class-5, general-class-6, general-class-7, general-class-8, and general-class-9), with each expert handling 10 classes. This setup is represented as CIFAR-100-10/10. Taking one expert as an example, facing the 100 classes in CIFAR100-20/5, the 100 classes are paired off into groups of five, and the groups of classes that this expert is good at are named as expert-class, while the groups of classes that this expert is not good at are named as general-class-1. We list the details of the datasets in Table 5.

A.3 Data-driven Knowledge Distillation

We posit that the feature distribution for expert classes follows Gaussian distribution, and aim to enhance this Gaussian resemblance across distributions. To achieve this, we initially apply Tukey's Ladder of Powers transformation to the features of the target task, as detailed in . This transformation, belonging to a family of power transformations, is designed to reduce distribution skewness and render them more Gaussian-like. It operates as follows:

$$x = \begin{cases} \tilde{x}^{\lambda}, & \text{if } \lambda \neq 0 \\ \log(\tilde{x}), & \text{if } \lambda = 0 \end{cases} \tag{19}$$

where $\tilde{x}$ is a feature vector from an expert class, and λ is a hyper-parameter that adjusts the correction of the distribution. Setting λ to 1 recovers the original feature. Decreasing λ reduces the positive skew of the distribution, while increasing λ enhances it.

The mean of the feature vector from an expert class i is calculated as:

$$\mu_i = \frac{\sum_{j=1}^{n_i} x_j}{n_i}, \tag{20}$$

where x_j is a feature vector of the j-th sample from the expert class i and n_i is the total number of samples in class i. Given the multidimensional nature of the feature vector x_j, we utilize covariance to more accurately represent the variance between any two elements within the feature vector. The covariance matrix Σ_i for features from class i is calculated in the following manner:

$$\Sigma_i = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (x_j - \mu_i)(x_j - \mu_i)^T, \tag{21}$$

Table 5: Details of constructed datasets.

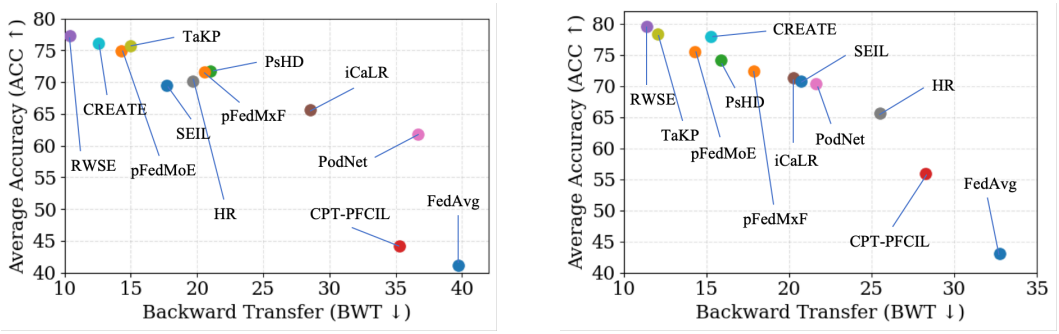
Datasets	Classes	Training	Testing	Detailed classes
CIFAR100-20/5	expert-class	10,000	2,000	apple, aquarium_fish, baby, bear, beaver, bed, bee, beetle, bicycle, bottle, bowl, boy, bridge, bus, butterfly, camel, can, castle, caterpillar, cattle
	general-class-1	10,000	2,000	chair, chimpanzee, clock, cloud, cockroach, couch, crab, crocodile, cup, dinosaur, dolphin, elephant, flatfish, forest, fox, girl, hamster, house, kangaroo, keyboard
	general-class-2	10,000	2,000	lamp, lawn_mower, leopard, lion, lizard, lobster, man, maple_tree, motorcycle, mountain, mouse, mushroom, oak_tree, orange, orchid, otter, palm_tree, pear, pickup_truck, pine_tree
	general-class-3	10,000	2,000	plain, plate, poppy, porcupine, possum, rabbit, raccoon, ray, road, rocket, rose, sea, seal, shark, shrew, skunk, skyscraper, snail, snake, spider
	general-class-4	10,000	2,000	squirrel, streetcar, sunflower, sweet_pepper, table, tank, telephone, television, tiger, tractor, train, trout, tulip, turtle, wardrobe, whale, willow_tree, wolf, woman, worm
CIFAR100-10/10	expert-class	5,000	1,000	apple, aquarium_fish, baby, bear, beaver, bed, bee, beetle, bicycle, bottle
	general-class-1	5,000	1,000	bowl, boy, bridge, bus, butterfly, camel, can, castle, caterpillar, cattle
	general-class-2	5,000	1,000	chair, chimpanzee, clock, cloud, cockroach, couch, crab, crocodile, cup, dinosaur
	general-class-3	5,000	1,000	dolphin, elephant, flatfish, forest, fox, girl, hamster, house, kangaroo, keyboard
	general-class-4	5,000	1,000	lamp, lawn_mower, leopard, lion, lizard, lobster, man, maple_tree, motorcycle, mountain
	general-class-5	5,000	1,000	mouse, mushroom, oak_tree, orange, orchid, otter, palm_tree, pear, pickup_truck, pine_tree
	general-class-6	5,000	1,000	plain, plate, poppy, porcupine, possum, rabbit, raccoon, ray, road, rocket
	general-class-7	5,000	1,000	rose, sea, seal, shark, shrew, skunk, skyscraper, snail, snake, spider
	general-class-8	5,000	1,000	squirrel, streetcar, sunflower, sweet_pepper, table, tank, telephone, television, tiger, tractor
	general-class-9	5,000	1,000	train, trout, tulip, turtle, wardrobe, whale, willow_tree, wolf, woman, worm

To effectively utilize these transformed features in a target task, we compile a set of calibrated statistics $S_y = \{(\mu_1, \Sigma_1), \ldots, (\mu_K, \Sigma_K)\}$ for each class y. We then generate a series of feature vectors labeled y by sampling from these calibrated Gaussian distributions, thereby aligning more closely with the assumed Gaussian feature distribution of expert classes, which is defined as:

$$\mathcal{D}_y = \{(x, y) \mid x \sim \mathcal{N}(\mu, \Sigma),\ \forall(\mu, \Sigma) \in S^y\}\,. \tag{22}$$

A.4 Full Experimental Results

Analysis of Knowledge Preservation and Backward Transfer Fig. 7 illustrates the trade-off between average accuracy (avg) and backward transfer (BWT) among different methods. Methods located in the upper-left region achieve a better trade-off between accuracy and knowledge retention, as they simultaneously yield higher accuracy and lower forgetting. In contrast, RWSE consistently dominates all competing methods, achieving the highest avg together with the lowest BWT. Com-



(a) CIFAR-100-20/5 (b) ImageNet-100-20/5

Figure 7: Trade-off between ACC and BWT for different methods.

Table 6: Comparison of detailed accuracy across different classes before and after collaboration.

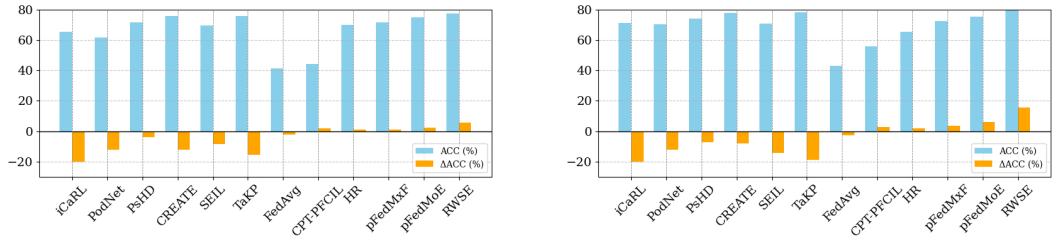
Method	CIFAR-100						ImageNet-100					
	Before Collaboration			After Collaboration			Before Collaboration			After Collaboration		
	Expert	General	Avg	Expert	General	Avg	Expert	General	Avg	Expert	General	Avg
iCaRL	66.25	0.00	33.13	31.48	46.45	38.97	63.41	0.00	31.71	28.49	40.16	34.33
PODNet	66.34	0.00	33.17	52.36	41.92	47.14	63.35	0.00	31.68	51.45	45.82	48.64
PsHD	66.39	0.00	33.20	54.48	43.87	49.16	63.47	0.00	31.74	52.76	46.21	49.49
CREATE	66.52	0.00	33.26	48.19	59.35	53.27	63.66	0.00	31.83	48.19	54.87	51.53
SEIL	66.41	0.00	33.21	43.64	54.88	49.26	63.58	0.00	31.79	47.65	53.47	50.56
TaKP	66.55	0.00	33.28	48.73	59.34	54.04	63.61	0.00	31.81	48.92	55.41	52.17
FedAvg	65.42	0.00	32.71	54.00	50.05	52.03	62.33	0.00	31.17	50.63	40.02	45.33
CPT-PFCIL	64.96	0.00	32.48	49.73	55.85	54.29	62.56	0.00	31.28	40.36	51.65	46.51
HR	65.36	0.00	32.68	57.37	54.12	55.75	62.45	0.00	31.23	54.78	49.57	52.18
pFedMxF	65.69	0.00	32.85	58.44	52.71	55.58	62.62	0.00	31.31	55.06	48.73	51.90
pFedMoE	65.61	0.00	32.81	59.19	54.03	56.61	62.58	0.00	31.29	55.46	50.84	53.15
RWSE	66.54	0.00	33.27	**61.22**	**56.79**	**59.01**	63.45	0.00	31.73	**59.73**	**54.18**	**56.96**

pared with all baselines, RWSE establishes a superior balance between knowledge acquisition and forgetting mitigation, indicating that recursive weight-space editing can effectively integrate new knowledge while preserving previously learned representations.

The Necessity of Versatility and Directionality We report the performance on the CIFAR-100-20/5 and ImageNet-100-20/5, as shown in Tables 6. **CL forgets original knowledge during growth:** while CIL methods maintain general performance, "expert classes" suffer from rapid degradation. Treating recent tasks as experts fails to resolve this, as catastrophic forgetting prevents these classes from remaining a stable, high-performance set throughout the learning trajectory. **FIL loses sight of the focus among multiple agents:** although FIL shows competitive accuracy, it lacks a prioritized focus. The performance gap between expert and general classes is negligible, suggesting that simplistic aggregation lacks the "directional bias" needed to preserve pre-established expert preferences, leading to a loss of specialized intelligence.

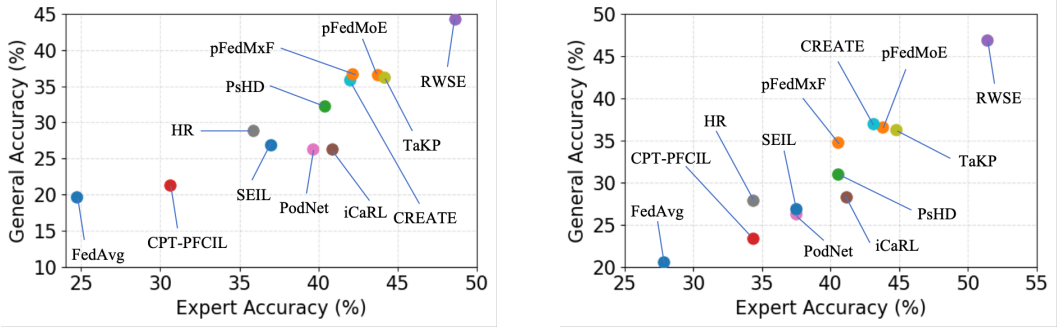
Analysis of versatility and directionality Versatility reflects the ability of an agent to generalize to broader classes, while directionality measures its capability to preserve the performance of expert classes during knowledge transfer. To evaluate these properties, we analyze the average accuracy over all classes and the performance gap between expert classes and the best-performing general classes. As illustrated in Fig. 8, most CIL methods suffer from significant forgetting of expert knowledge, as indicated by the negative values in the orange bars. In contrast, most FIL methods yield orange bar values close to zero, suggesting insufficient directionality in preserving expert-class performance.

Data-driven Knowledge Distillation As shown in Table 7, all methods degrade notably without real data, revealing the strong dependence of conventional approaches on high-quality samples. CIL methods remain limited, indicating that feature- or logit-level distillation is insufficient under



(a) CIFAR-100-20/5 (b) ImageNet-100-20/5

Figure 8: Analysis of versatility and directionality. Blue bars denote the average accuracy across all classes, while orange bars denote the difference in accuracy between expert classes and the best-performing general classes. Blue and orange bars correlate positively with versatility and directionality, respectively.



(a) Comparison on generated CIFAR-100 (b) Comparison on generated ImageNet-100

Figure 9: Performance comparison after collaboration.

synthetic distributions, while FIL methods only partially mitigate this issue due to distribution mismatch. In contrast, RWSE consistently outperforms all baselines on both datasets, achieving over 9% improvement in average accuracy. This gain highlights its ability to integrate expert knowledge without relying on real data. This advantage stems from the fact that RWSE operates directly in the weight space, enabling structured knowledge editing and reducing dependency on data quality.

Table 7: Performance comparison on generated datasets under the 20/5 setting (Accuracy %).

Method	CIFAR-100		ImageNet-100	
	last	avg	last	avg
iCaRL	24.31	33.58	25.94	34.73
PodNet	22.67	32.95	23.52	31.88
PsHD	30.13	36.32	30.49	35.75
CREATE	31.27	38.94	33.87	40.06
SEIL	26.14	31.93	27.44	32.17
TaKP	32.69	40.23	34.25	40.54
FedAvg	15.76	22.18	16.05	24.23
CPT-PFCIL	18.57	25.98	21.31	28.83
HR	26.45	32.37	24.18	31.09
pFedMxF	31.33	39.45	30.14	37.65
pFedMoE	32.56	40.17	30.63	38.88
RWSE	**38.58**	**46.47**	**40.35**	**49.14**

As shown in fig. 9, we evaluate knowledge distillation with generated data by comparing performance on expert and general classes. Across both CIFAR-100-20/5 and ImageNet-100-20/5, most methods achieve higher accuracy on general classes, indicating a bias of synthetic data toward general semantics and insufficient preservation of expert knowledge. CIL methods exhibit large performance gaps, reflecting severe degradation of expert classes, while FIL methods partially reduce this gap but remain limited by distribution mismatch. In contrast, RWSE achieves the best performance on both expert and general classes, demonstrating a superior balance between directionality and versatility.

A.5 Efficient Recursive Editing

The direct computation of $\left(A_{t-1} + \Gamma_{t-1} + \lambda K_t^\top K_t\right)^{-1}$ requires inversion in the d_{in}-dimensional feature space. Define

$$H_{t-1} = A_{t-1} + \Gamma_{t-1}. \tag{23}$$

Then Eq. 11 involves $(H_{t-1} + \lambda K_t^\top K_t)^{-1}$.

Applying the Woodbury matrix identity yields

$$(H_{t-1} + \lambda K_t^\top K_t)^{-1} = H_{t-1}^{-1} - H_{t-1}^{-1} K_t^\top \left(\lambda^{-1} I + K_t H_{t-1}^{-1} K_t^\top\right)^{-1} K_t H_{t-1}^{-1}. \tag{24}$$

The large $d_{\text{in}} \times d_{\text{in}}$ inversion is therefore transformed into an $n_t \times n_t$ system. When $n_t \ll d_{\text{in}}$, this substantially reduces the cost of the sample-dependent inversion.

Moreover, because Γ_{t-1} is constructed from the eigendecomposition of A_{t-1}, we have

$$H_{t-1} = Q_{t-1} \left[\Lambda_{t-1} + \gamma g(\Lambda_{t-1})\right] Q_{t-1}^\top, \tag{25}$$

so that

$$H_{t-1}^{-1} = Q_{t-1} \left[\Lambda_{t-1} + \gamma g(\Lambda_{t-1})\right]^{-1} Q_{t-1}^\top. \tag{26}$$

Thus, once the historical eigensystem is available, the inverse associated with the anisotropic anchor can be evaluated directly in the spectral domain.