CO-EVOLVING REHEARSAL STATES FOR FEDERATED CLASS-INCREMENTAL LEARNING UNDER ASYNCHRONOUS CLASS EXPOSURE

Anonymous authors
Paper under double-blind review

ABSTRACT

Federated class incremental learning (FCIL) must retain earlier classes while clients continue learning from different class histories. At the same round, a class may be current for one client, historical for another, and unseen by a third. We call this disagreement *asynchronous class exposure* (ACE). Existing methods preserve history through replay, prototypes, feature memories, or parameter constraints, but they often treat stored evidence as fixed while the model representation continues to change. As a result, historical evidence can bypass trainable modules, use incompatible encoder coordinates, or become stale over time. We address this gap with Projected Rehearsal Orchestration (PRO), which rehearses current and remembered evidence before the adapter and rebuilds available class statistics under the aggregated global encoder. When fresh examples are unavailable, PRO-MAX estimates class local feature movement from current calibration data and combines client estimates using support weights to maintain the remaining memories. Across six image, text, and graph benchmarks, PRO and PRO-MAX improve retention under heterogeneous exposure, with PRO-MAX achieving the highest reported mean final accuracy among the evaluated methods on all six streams. We further show that the gains are not explained by feature sampling alone. Instead, effective rehearsal depends on keeping historical evidence trainable, compatible with the current representation, and updated as learning evolves.

1 INTRODUCTION

Federated learning improves a shared model without centralizing raw data (McMahan et al., 2017). In federated class-incremental learning (FCIL), this model must learn new classes while retaining earlier ones under a single classifier and without task identity at inference (Dong et al., 2022; 2024). However, one shared model does not imply one shared learning history. At the same round, a class may be current for one client, historical for another, and globally known but locally unseen by a third. We call this phenomenon as *asynchronous class exposure* (ACE), which can arise even under synchronous communication when clients follow different class sequences (Yoon et al., 2021; Chaudhary et al., 2022; Shenaj et al., 2023). As Figure 1 illustrates, FCIL must therefore keep historical knowledge usable as both client histories and the global representation evolve.

Existing methods preserve history through exemplars, generated inputs, prototypes, feature statistics, or parameter constraints (Rebuffi et al., 2017; Zhang et al., 2023; Babakniya et al., 2023; Tan et al., 2022; Li et al., 2025). Under ACE, however, stored evidence may bypass adaptation, use incompatible encoder coordinates, or become stale as the representation changes. These coupled failures motivate treating FCIL as the joint evolution of a global model and a shared rehearsal state, where historical evidence must remain supported, trainable, and representation-compatible.

To resolve this problem, we propose Projected Rehearsal Orchestration (PRO), which maintains compact class-level feature statistics and combines current and remembered evidence through class-balanced pre-adapter rehearsal. Because the representation changes during training, clients then reconstruct available memories under the aggregated global encoder, forming a READ-OPTIMIZE-WRITE lifecycle that keeps historical evidence trainable and coordinate-compatible. When fresh examples are unavailable, PRO-MAX estimates class-local feature movement from temporary cali-

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

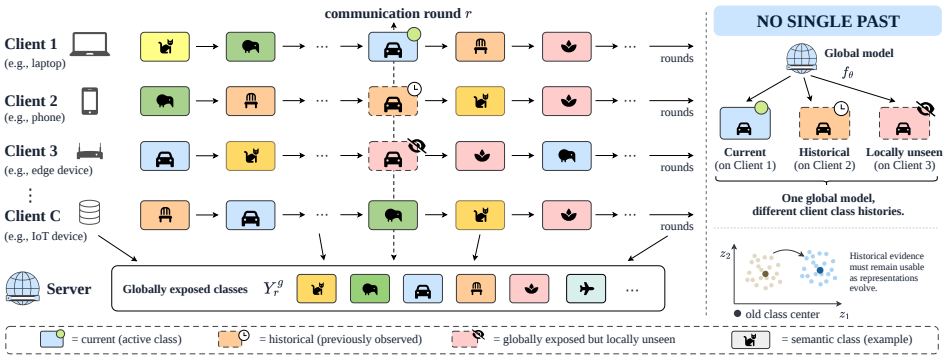


Figure 1: Asynchronous class exposure in federated class-incremental learning. A class can have different client-relative roles, while its stored representation can become incompatible with the evolving encoder.

bration data and support-weights client estimates to approximately align the remaining memories. Across image, text, and graph streams, PRO and PRO-MAX improve final accuracy and retention under heterogeneous exposure while remaining competitive under aligned histories. Empirical results have shown that these gains are not explained by feature sampling alone, with pre-adapter rehearsal and global-coordinate refresh providing the main lifecycle benefits and class-aware alignment further improving retention when direct reconstruction is unavailable. In summary, our contributions are:

- We formalize ACE and model FCIL as joint evolution of global model and shared rehearsal state.
- We propose PRO to coordinate class-balanced pre-adapter rehearsal and global-coordinate memory refresh, and PRO-MAX to maintain memories when reconstruction is unavailable.
- Empirical results across several benchmarks show that strong retention depends more on how historical evidence is used, refreshed, and maintained than on feature sampling alone.

2 Related Work

Different client learning histories. FCL already considers clients with different task histories, orders, and learning progress (Yoon et al., 2021; Ma et al., 2022; Dong et al., 2022; Shenaj et al., 2023). We build on this setting by using ACE to make each globally exposed class explicitly current, historical, or known only through the federation for each client.

Preserving historical evidence. These different histories make historical knowledge important when a class is absent from current local data. Existing FCIL methods preserve such knowledge through generated samples, distillation, prototypes, feature memories, or class aware replay (Tan et al., 2022; Qi et al., 2025; Rong et al., 2025). We instead focus on whether this evidence remains useful as local learning and aggregation change the representation.

Maintaining historical evidence as representations change. Related methods maintain evolving feature or latent states, such as AF-FCL (Wuerkaixi et al., 2024) and Hybrid Rehearsal (Nori et al., 2025), while SDC (Yu et al., 2020) and LDC (Gomez-Villa et al., 2025) address representation drift in centralized continual learning. We extend this view to FCIL by coordinating both the use and maintenance of historical evidence. PRO rehearses features before the adapter and reconstructs class statistics under the aggregated encoder, while PRO-MAX updates class means when fresh reconstruction is unavailable. We provide the full comparison in Appendix A.

3 Methodology

3.1 Asynchronous Class Exposure

We consider clients $\mathcal{C} = \{1, \ldots, C\}$ and semantic tasks $\mathcal{T} = \{1, \ldots, T\}$. Each task t introduces a disjoint class set $\mathcal{Y}_t$, and inference uses one expanding classifier over the classes exposed to the

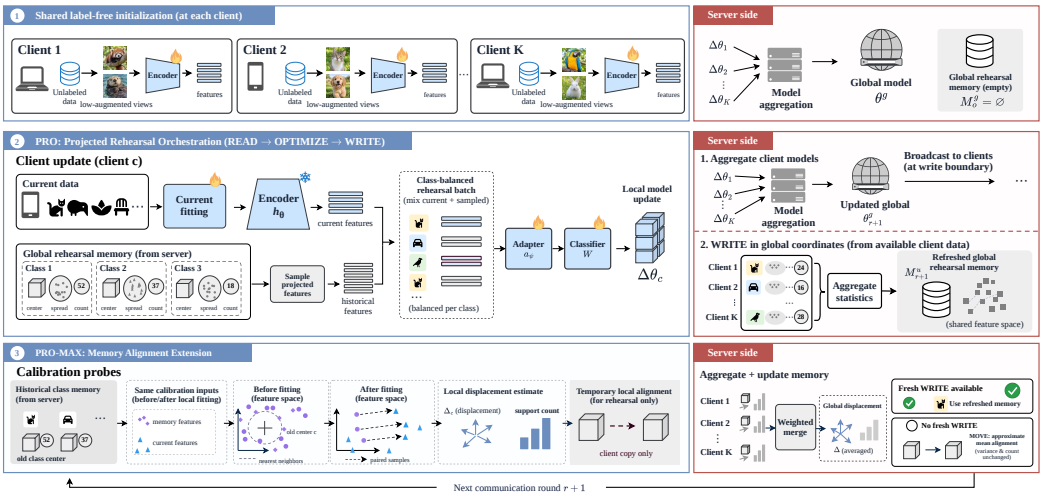


Figure 2: PRO and PRO-MAX training. PRO uses balanced pre-adapter rehearsal and reconstructs statistics under the aggregated encoder. PRO-MAX adds temporary local alignment and confidence-weighted maintenance when fresh WRITE is unavailable. Calibration inputs remain local, and MOVE approximates mean change rather than exact transport.

federation without task identity. At communication round r, the selected clients $\mathcal{S}_r$ participate. Client c occupies position $p_{c,r}$ in its task permutation π_c and observes

$$\tau_{c,r} = \pi_c[p_{c,r}], \qquad \mathcal{D}_{c,r} = \{(x_i, y_i)\}_{i=1}^{N_{c,r}}, \qquad y_i \in \widetilde{\mathcal{Y}}_{c,r} \subseteq \mathcal{Y}_{\tau_{c,r}}, \qquad \mathcal{Y}_r^g = \bigcup_{s=0}^{r} \bigcup_{c \in \mathcal{S}_s} \widetilde{\mathcal{Y}}_{c,s}. \quad (1)$$

The observed subset and sample count can differ across clients assigned to the same semantic task. Every evaluated class has at least one supporting client after its first global exposure, but no client is assumed to observe every exposed class.

Let $\mathcal{H}_{c,r} = \bigcup_{s=0}^{r} \widetilde{\mathcal{Y}}_{c,s}$ denote the labeled history of client c. For any globally exposed class $y \in \mathcal{Y}_r^g$, we call it *current* if it appears in the client's data at round r, *historical* if the client has observed it before but not in the current round, and *global only* if it has been exposed elsewhere in the federation but has never been observed by that client. ACE occurs when two participating clients assign different roles to at least one globally exposed class. It describes disagreement between class histories and can therefore occur even when communication and aggregation are synchronous.

The learning target is a single global model over $\mathcal{Y}_r^g$, although each client reveals only its current labeled slice and raw data remain local. Current-only updates provide no direct objective support for locally absent classes, so we model the algorithmic state as $\mathsf{S}_r = (\Theta_r^g, \mathcal{M}_r)$, where $\mathcal{M}_r$ contains constructed class memories with keys $\mathcal{K}_r \subseteq \mathcal{Y}_r^g$. The aim is not to reconstruct unavailable inputs exactly, but to keep historical evidence trainable and coordinate-compatible while Θ_r^g evolves. We expose the relevant training path through the decomposition

$$x \xrightarrow{h_\theta} u \xrightarrow{a_\psi} z \xrightarrow{W} \text{logits}, \qquad \Theta = (\theta, \psi, W), \quad (2)$$

where h_θ, a_ψ, and W are the encoder, adapter, and classifier. PRO rehearses before the adapter and reconstructs fresh memories under the aggregated encoder. PRO-MAX estimates mean movement when no fresh WRITE is available. Figure 2 summarizes the lifecycle, while Appendices B and B.6 detail exposure and rehearsal consistency.

3.2 PRO THROUGH READ, OPTIMIZE, AND WRITE

READ. At round r, the server state contains one tuple for each class with constructed memory,

$$\mathcal{M}_r^u = \{m_{y,r}^u : y \in \mathcal{K}_r\}, \qquad m_{y,r}^u = (\mu_{y,r}^u, \sigma_{y,r}^u, n_{y,r}), \quad (3)$$

where $\mu^u_{y,r}$, $\sigma^u_{y,r}$, and $n_{y,r}$ are the mean, coordinate-wise spread, and construction count. They define

$$Q^u_{y,r} = \mathcal{N}\big(\mu^u_{y,r}, \gamma^2 \operatorname{diag}(\sigma^u_{y,r} \odot \sigma^u_{y,r})\big), \qquad \widetilde{u}_{y,j} = \mu^u_{y,r} + \gamma \sigma^u_{y,r} \odot \epsilon_j, \quad \epsilon_j \sim \mathcal{N}(0, I). \tag{4}$$

Clients can sample historical features without synthesizing inputs. The scalar $\gamma \geq 0$ controls spread, with $\gamma = 0$ giving class-center rehearsal. The diagonal family keeps storage linear in feature dimension and provides a common interface across modalities, but it does not preserve arbitrary correlations or multimodal geometry. All methods share the selected label-free initialization, but $\mathcal{M}^u_0 = \varnothing$ as exposure does not create memory. Under the strict protocol, client c initializes only from

$$\mathcal{U}^{\text{strict}}_c = \{x : (x, y) \in \mathcal{D}^{\text{train}}_{c,0}\}. \tag{5}$$

Full-local-pool initialization is reported separately. Neither initialization creates labeled memory, which arises only through WRITE.

The READ set for client c contains classes supported either by current observations or by an existing shared memory,

$$\mathcal{A}_{c,r} = \widetilde{\mathcal{Y}}_{c,r} \cup \mathcal{K}_r, \qquad q^u_{c,y,r} = \begin{cases} P^u_{c,y,r}, & y \in \widetilde{\mathcal{Y}}_{c,r} \setminus \mathcal{K}_r, \\ Q^u_{y,r}, & y \in \mathcal{K}_r \setminus \widetilde{\mathcal{Y}}_{c,r}, \\ \lambda_{c,y,r} P^u_{c,y,r} + (1 - \lambda_{c,y,r}) Q^u_{y,r}, & y \in \widetilde{\mathcal{Y}}_{c,r} \cap \mathcal{K}_r. \end{cases} \tag{6}$$

For an overlap class, the mixing coefficient follows the within-class mini-batch counts,

$$\lambda_{c,y,r} = \frac{B^{\text{real}}_{c,y,r}}{B^{\text{real}}_{c,y,r} + B^{\text{mem}}_{c,y,r}}. \tag{7}$$

READ cannot create evidence for a class that is neither observed nor remembered. For overlap classes, the count-based mixture prevents either source from being silently discarded and makes the available evidence explicit before optimization. The mixture remains local to the current client and does not claim that its empirical class distribution matches the federation-wide history.

OPTIMIZE. Current evidence must be learned before it supports WRITE, but unrestricted local fitting can distort features around a narrow label subset. Each client initializes from Θ^g_r and minimizes

$$\mathcal{L}^{\text{fit}}_{c,r} = \mathcal{L}^{\text{sup}}_{c,r} + \lambda_{\text{aux}} \mathcal{L}^{\text{aux}}_{c,r} + \lambda_{\text{prox}} \|\Pi_r(\Theta) - \Pi_r(\Theta^g_r)\|^2_2, \tag{8}$$

where Π_r selects the trainable parameters. Supervision uses current labels. While the encoder is plastic, an optional auxiliary term compares normalized projections of two current-input views. It is neither replay nor a teacher signal. The proximal term restrains only trainable parameters around the broadcast model. After the configured plasticity rounds, the encoder is frozen and $\lambda_{\text{aux}} = 0$, leaving the adapter and classifier trainable.

The client then optimizes the balanced rehearsal surrogate using cross-entropy (CE)

$$\widetilde{F}_{c,r}(v) = \frac{1}{|\mathcal{A}_{c,r}|} \sum_{y \in \mathcal{A}_{c,r}} \mathbb{E}_{u \sim q^u_{c,y,r}} \operatorname{CE}(W^\top a_\psi(u), y), \qquad v = (\psi, W), \tag{9}$$

or, equivalently, samples $y \sim \operatorname{Unif}(\mathcal{A}_{c,r})$ and $u \sim q^u_{c,y,r}$ to minimize

$$\mathcal{L}^{\text{PRO}}_{c,r}(v) = \mathbb{E}_{\substack{y \sim \operatorname{Unif}(\mathcal{A}_{c,r}) \\ u \sim q^u_{c,y,r}}} \operatorname{CE}(W^\top a_\psi(u), y). \tag{10}$$

Uniform class sampling prevents current frequencies from defining rehearsal. Historical features also enter before the adapter, so their loss reaches both a_ψ and W. Detached post-adapter replay can update the classifier but cannot constrain a_ψ. This distinction is structural rather than a guarantee against forgetting because the gradient can still be small or the memory incomplete. Appendix B.6 gives the corresponding gradients and formal scope.

WRITE. OPTIMIZE returns local updates, which the server aggregates with sample-count weights,

$$\Theta^g_{r+1} = \sum_{c \in \mathcal{S}_r} \omega_{c,r} \Theta_{c,r}, \qquad \omega_{c,r} = \frac{n_{c,r}}{\sum_{c' \in \mathcal{S}_r} n_{c',r}}. \tag{11}$$

Local moments cannot be pooled when encoders define different coordinates. At a write boundary, clients with fresh examples for class y re-encode them under $h_{\theta^g_{r+1}}$ without updates and upload a mean $\widehat{\mu}^{u,g}_{c,y,r}$, second moment $\widehat{s}^{u,g}_{c,y,r}$, and count. With $N^{\text{write}}_{y,r} = \sum_{c\in\mathcal{V}_{y,r}} n_{c,y,r}$, server computes

$$\mu^u_{y,r+1} = \frac{\sum_{c\in\mathcal{V}_{y,r}} n_{c,y,r}\widehat{\mu}^{u,g}_{c,y,r}}{N^{\text{write}}_{y,r}}, \tag{12}$$

$$s^u_{y,r+1} = \frac{\sum_{c\in\mathcal{V}_{y,r}} n_{c,y,r}\widehat{s}^{u,g}_{c,y,r}}{N^{\text{write}}_{y,r}}, \qquad \sigma^u_{y,r+1} = \sqrt{\max\{s^u_{y,r+1} - \mu^u_{y,r+1}\odot\mu^u_{y,r+1}, \varepsilon\mathbf{1}\}}. \tag{13}$$

Second-moment aggregation retains within-client variation and variation between client means. WRITE replaces the class entry because all contributions use the same aggregated encoder. This removes coordinate disagreement caused only by different client encoders, although writer support may still be incomplete or unrepresentative. The bank stores one mean, diagonal spread, and count per class. System cost also includes model and memory broadcasts, moment uploads, and the encoder transmission at write boundaries. Appendix H gives the complete accounting. A class can still lack fresh examples when its memory needs maintenance, which motivates PRO-MAX.

3.3 PRO-MAX WITH MOVE WHEN DIRECT WRITE IS UNAVAILABLE

PRO-MAX estimates how local fitting moves features near unsupported memory. Let $\theta^- = \theta^g_r$ and θ^+ be the fitted encoder. For paired calibration inputs, $u^-_i = h_{\theta^-}(x_i)$, $u^+_i = h_{\theta^+}(x_i)$, and $d_i = u^+_i - u^-_i$. Only conditional mean movement on current support is identifiable,

$$\delta_{c,r}(u) = \mathbb{E}_{X\sim P^x_{c,r}}[h_{\theta^+}(X) - h_{\theta^-}(X) \mid h_{\theta^-}(X) = u]. \tag{14}$$

For an unsupported stored class, the client selects the K nearest calibration features to $\mu^u_{y,r}$, sets an adaptive bandwidth, and assigns kernel weights

$$\tau_{c,y} = \max\left\{\text{median}_{i\in\mathcal{N}_{c,y}}\|u^-_i - \mu^u_{y,r}\|_2, 10^{-6}\right\}, \qquad \alpha_{i,y} = \exp\left(-\frac{\|u^-_i - \mu^u_{y,r}\|_2^2}{2\tau^2_{c,y}}\right). \tag{15}$$

It then forms the local displacement and support score

$$\Delta^u_{c,y} = \frac{\sum_{i\in\mathcal{N}_{c,y}}\alpha_{i,y}d_i}{\kappa_{c,y}+\varepsilon}, \qquad \kappa_{c,y} = \sum_{i\in\mathcal{N}_{c,y}}\alpha_{i,y}. \tag{16}$$

The temporary local rehearsal distribution is consequently

$$Q^{u,\text{loc}}_{c,y,r} = \mathcal{N}\left(\mu^u_{y,r} + \Delta^u_{c,y}, \gamma^2\,\text{diag}(\sigma^u_{y,r}\odot\sigma^u_{y,r})\right). \tag{17}$$

This distribution is temporary and affects only local rehearsal. Clients upload $(\Delta^u_{c,y}, \kappa_{c,y})$ rather than calibration examples and never overwrite server state locally. The score $\kappa_{c,y}$ weights aggregation and controls part of the averaging error, but it is not a calibrated probability and does not certify neighborhood locality or transferability. Equations (55) and (57) and the accompanying analysis quantify its effective sample size and error role.

The server confidence-aggregates the uploaded vectors,

$$\Delta^{u,g}_y = \frac{\sum_{c\in\mathcal{S}_r}\kappa_{c,y}\Delta^u_{c,y}}{\sum_{c\in\mathcal{S}_r}\kappa_{c,y}+\varepsilon}. \tag{18}$$

The aggregated vector remains an estimate of the unavailable historical displacement under actual global transition. Client fitting can differ from that transition, and pointwise local estimate need not equal class-average movement. Equations (58) to (60) define target and separate local estimation error from client-to-global mismatch. Fresh reconstruction then has priority over approximation, so

$$m^u_{y,r+1} = \begin{cases}\text{WRITE}_{y,r}, & y\in\mathcal{W}_r,\\ (\mu^u_{y,r}+\Delta^{u,g}_y, \sigma^u_{y,r}, n_{y,r}), & y\in\mathcal{K}_r\setminus\mathcal{W}_r.\end{cases} \tag{19}$$

MOVE changes only the mean and cannot recover covariance or lost encoder information, as discussed following Equation (28). Distant neighborhoods and mismatched transitions introduce the errors bounded in Equations (57) and (60), motivating the controls in Figure 3 and Table 3.

4 EXPERIMENTS

We organize our experiments around three questions, including **(i)** how heterogeneous exposure affects retention, **(ii)** which parts of the rehearsal lifecycle matter, and **(iii)** how far resulting gains extend. To realize this, we first compare aligned and heterogeneous client histories, then examine rehearsal placement, memory refresh, and alignment, before testing replay fidelity, cross-modal performance, sensitivity, and computational cost.

4.1 PROTOCOL

Datasets. We evaluate six benchmarks as follows: **CIFAR-100** (Krizhevsky, 2009) and **TinyImageNet** (Wu et al., 2017) for images, **THUCNews-10** (Guo et al., 2016) and **CLINC150** (Larson et al., 2019) for text, and **Cora** (Sen et al., 2008) and **ogbn-arxiv** (Hu et al., 2020) for node classification. CIFAR-100 is our main controlled setting, while the remaining datasets test larger class spaces and whether the same post-encoder memory interface can be reused across modalities. We use ResNet-18, TextCNN, and two-layer GAT encoders for image, text, and graph data, respectively (He et al., 2016; Kim, 2014; Veličković et al., 2018). The task splits and schedules are reported in Table 7.

Exposure settings. We compare an aligned setting with full task support, full label coverage, and a shared task order against a heterogeneous setting with 40% task support, 50% label coverage, and 60% task-order perturbation. We also vary task support and task order separately to examine how each stream change affects retention. The reported sampler settings are in Table 8.

Baselines. We compare with standard federated methods **FedAvg**, **FedProx**, and **FedProto** (McMahan et al., 2017; Li et al., 2020; Tan et al., 2022); replay-based methods **TARGET**, **MFCL**, and **Hybrid Rehearsal** (Zhang et al., 2023; Babakniya et al., 2023; Nori et al., 2025); and continual or heterogeneous federated methods **FedSpace**, **FedSSI**, and **FedCBDR** (Shenaj et al., 2023; Li et al., 2025; Qi et al., 2025). Text experiments additionally include **FedSeIT** (Chaudhary et al., 2022), while graph experiments include **POWER** and **MOTION** (Zhu et al., 2025; Wan et al., 2025). Their evaluated roles and replay-object budgets are described in Appendix D.5 and Table 9.

Metrics. We report **final average accuracy (FAA)**, **current-task accuracy (CTA)**, and **average forgetting (AF)** to measure final performance, learning of current classes, and retention of earlier classes, respectively. Considering all three helps distinguish genuine continual learning from simply reducing forgetting by learning new classes poorly. Results are five-seed means and standard deviations where available, while mean-only controls remain mean-only.

Table 1: CIFAR-100 retention under aligned and heterogeneous streams. Five-seed means with standard deviations. Higher FAA and lower AF are preferred. Best values are bold, runners-up underlined. CTA is in Table 10.

	Aligned		Heterogeneous	
Method	FAA	AF	FAA	AF
FedAvg	10.84 ± 0.36	79.86 ± 0.81	7.86 ± 0.42	82.74 ± 1.36
TARGET	43.18 ± 0.96	30.66 ± 1.28	27.86 ± 1.74	57.84 ± 2.18
MFCL	44.02 ± 1.05	29.12 ± 1.22	29.42 ± 1.68	54.63 ± 2.04
HR	45.34 ± 0.92	26.38 ± 1.13	32.18 ± 1.52	49.76 ± 1.87
FedCBDR	$\underline{45.76 \pm 0.86}$	$\underline{25.84 \pm 1.05}$	38.96 ± 1.18	39.72 ± 1.48
PRO	43.86 ± 0.82	27.36 ± 1.08	$\underline{43.28 \pm 0.94}$	$\underline{31.54 \pm 1.18}$
PRO-MAX	$\mathbf{46.72 \pm 0.78}$	$\mathbf{23.64 \pm 0.97}$	$\mathbf{46.12 \pm 0.88}$	$\mathbf{26.18 \pm 1.07}$

4.2 EMPIRICAL RESULTS

Retention under heterogeneous exposure. We first ask whether methods that retain earlier classes under aligned histories remain stable when client histories differ. As shown in Table 1, HR and FedCBDR reach 45.34 and 45.76 FAA in the aligned setting, both above PRO at 43.86, confirming that the comparison includes strong replay baselines. Under heterogeneous exposure, however, TARGET, MFCL, and HR lose 13.16–15.32 FAA points and FedCBDR falls from 45.76 to 38.96, while PRO

and PRO-MAX drop by only 0.58 and 0.60 points and finish at 43.28 and 46.12. AF follows the same pattern, showing that the proposed lifecycle is less sensitive to the evaluated exposure changes rather than benefiting only from an easier aligned setting.

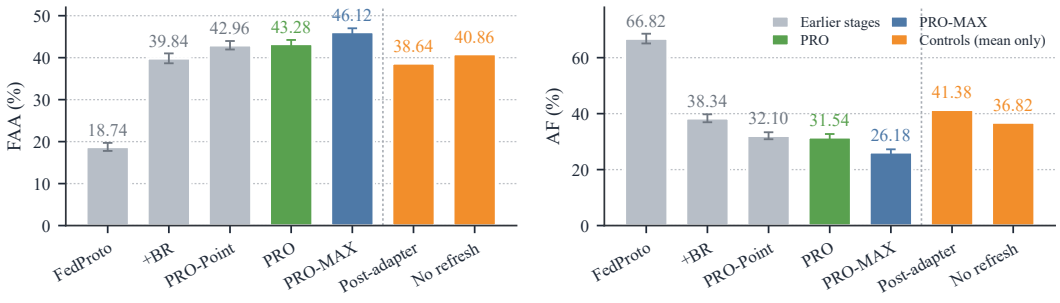
Table 2 shows that the same pattern persists when we change the stream more gradually. Reducing task support from 80% to 20% lowers FAA by 9.42 points for FedCBDR but only 4.40 and 4.44 points for PRO and PRO-MAX. Increasing task-order perturbation from 0% to 90% gives drops of 8.02, 2.92, and 2.66 points, respectively. These are separate stream interventions rather than direct measurements of ACE, but together they show that the proposed variants remain less sensitive across the tested conditions. We next examine which parts of the rehearsal lifecycle account for this stability.

Table 2: CIFAR-100 FAA endpoint sweeps. Support and order are separate interventions. At 0% order perturbation, support and label coverage remain sparse. Best values are bold, runners-up underlined. Full sweeps are in Tables 26 and 27.

	Task support		Order perturbation	
Method	80%	20%	0%	90%
FedCBDR	43.84	34.42	43.20	35.18
PRO	<u>44.26</u>	<u>39.86</u>	<u>43.76</u>	<u>40.84</u>
PRO-MAX	**47.12**	**42.68**	**46.62**	**43.96**

What makes historical rehearsal useful? Figure 3 examines whether the gains can be explained by sampling features around stored class centers alone. Starting from balanced pre-adapter point rehearsal, global refresh raises FAA from 39.84 to 42.96 and reduces AF from 38.34 to 32.10, whereas adding feature spread changes FAA by only 0.32 points and AF by 0.56. The larger change in this sequence therefore comes from keeping stored evidence compatible with the model that later consumes it rather than from feature sampling alone.

The targeted controls in Figure 3 support the same conclusion from the training side. Post-adapter replay reaches 38.64 FAA and 41.38 AF, while removing global refresh gives 40.86 and 36.82, compared with 43.28 and 31.54 for PRO. Pre-adapter rehearsal lets historical evidence affect the trainable adapter, while refresh reconstructs that evidence under the current global encoder. These controls are not a full factorial decomposition, but together they show that useful rehearsal depends on both where historical evidence enters training and how it is reconstructed. This leaves one important case: classes for which fresh examples are no longer available.



(a) FAA. (b) AF.

Figure 3: CIFAR-100 rehearsal construction and use. The first five settings form the sequential comparison with five-seed standard deviations. Targeted controls are reported as means only. Complete values and CTA are in Tables 11 and 12.

Maintaining memories without fresh examples. Table 3 tests whether PRO-MAX benefits from class-aware memory movement rather than simply from running an additional calibration step. Zero transport remains within 0.04 points of PRO in both FAA and AF, whereas full PRO-MAX improves FAA by 2.80 points and reduces AF by 5.32 points relative to that control. Uniform neighbor weighting gives an intermediate result, while shuffled transport lowers center alignment and worsens

retention. The pattern supports both class association and local weighting for approximate memory maintenance. Since MOVE remains an approximation, we next ask whether simply providing more faithful historical features could explain the remaining gain.

Table 3: Transport controls on heterogeneous CIFAR-100. Alignment is cosine similarity to fresh global class centers. Higher alignment and FAA and lower AF are preferred. Best values are bold, runners-up underlined. CTA is in Table 13.

Variant	Alignment	FAA	AF
PRO	0.73 ± 0.03	43.28 ± 0.94	31.54 ± 1.18
Zero transport	0.74 ± 0.03	43.32 ± 0.96	31.50 ± 1.20
Shuffled class transport	0.68 ± 0.04	42.86 ± 1.08	33.06 ± 1.36
Uniform neighbor weights	$\underline{0.81 \pm 0.03}$	$\underline{44.90 \pm 0.95}$	$\underline{28.66 \pm 1.16}$
PRO-MAX	$\mathbf{0.86 \pm 0.02}$	$\mathbf{46.12 \pm 0.88}$	$\mathbf{26.18 \pm 1.07}$

Is better replay alone enough? Table 4 separates replay quality from the learned state by replacing projected replay with fresh historical features while keeping the starting model and memory fixed. Fresh features reduce MMD from 0.08 to 0.03 under normal support and from 0.11 to 0.04 under weak support, but FAA improves by only 0.12 and 0.24 points. The weak condition also remains worse after replacement. Thus, better historical features help, but do not remove the effect of weaker prior learning, showing that retention depends on both replay quality and how historical evidence is learned and maintained.

Table 4: Within-PRO replay-source substitution from matched model and memory states. Lower MMD and AF and higher FAA are preferred. Each condition contains only two replay sources, so rank styling is omitted. CTA and the retained-input reference are in Table 14.

Condition	Replay source	MMD	FAA	AF
Normal	Projected	0.08 ± 0.01	43.28 ± 0.94	31.54 ± 1.18
Normal	Fresh features	0.03 ± 0.01	43.40 ± 0.93	31.24 ± 1.17
Weak	Projected	0.11 ± 0.02	42.22 ± 1.07	33.02 ± 1.32
Weak	Fresh features	0.04 ± 0.01	42.46 ± 1.04	32.50 ± 1.29

Does the same interface work across modalities? Table 5 shows that the same post encoder memory interface remains useful across image, text, and node classification benchmarks. PRO-MAX achieves the highest reported mean FAA on all six datasets, outperforming the strongest evaluated non-PRO comparator by 6.04-8.34 points and reducing AF by 9.44-13.76 points. It also consistently improves over PRO, suggesting that the same rehearsal design can be reused across the evaluated modalities while leaving modality specific processing to the underlying encoder.

Table 5: Six-benchmark heterogeneous-stream results. The comparator has the highest evaluated non-PRO FAA, and AF reduction is a mean difference. Best values are bold, runners-up underlined per dataset. Full results are in Tables 16, 17 and 23 to 25.

		FAA			AF reduction
Dataset	Comparator	Comparator	**PRO**	**PRO-MAX**	(points)
CIFAR-100	FedCBDR	38.96 ± 1.18	$\underline{43.28 \pm 0.94}$	$\mathbf{46.12 \pm 0.88}$	13.54
TinyImageNet	FedCBDR	29.46 ± 1.18	$\underline{33.28 \pm 1.04}$	$\mathbf{37.54 \pm 0.98}$	13.76
THUCNews-10	FedSeIT	73.28 ± 1.18	$\underline{78.42 \pm 1.04}$	$\mathbf{81.62 \pm 0.96}$	9.44
CLINC150	FedSeIT	59.74 ± 1.14	$\underline{63.92 \pm 1.06}$	$\mathbf{67.84 \pm 0.98}$	10.92
Cora	MOTION	60.14 ± 4.21	$\underline{63.24 \pm 3.91}$	$\mathbf{66.18 \pm 3.66}$	10.10
ogbn-arxiv	MOTION	44.38 ± 2.46	$\underline{47.86 \pm 2.32}$	$\mathbf{51.92 \pm 2.18}$	12.26

Sensitivity to training choices. The appendix sensitivity tables Tables 18, 19 and 32 to 35 test whether the gains depend on a narrow configuration. PRO-MAX remains close to its reported operating point for $K \in \{3, 5, 10\}$, while increasing the calibration set beyond 128 samples gives

only a small additional gain. Performance is more sensitive to excessive feature spread and to removing proximal regularization, whereas the tested auxiliary-loss weights give similar results. Using the stricter initialization based only on each client's first local data also preserves the ordering PRO-MAX $>$ PRO $>$ FedCBDR. Together, these results indicate a useful local operating range without implying robustness to every training configuration.

Computation and memory. Finally, Figure 4 shows the additional computational cost of PRO-MAX. Relative to PRO, it increases incremental runtime from 3.20 to 3.60 hours, peak GPU memory from 5.6 to 6.2 GB, and peak CPU RAM from 12.0 to 13.1 GB, while improving FAA by 2.84 points and reducing AF by 5.36 points. In these runs, both PRO variants remain below TARGET, MFCL, HR, and FedCBDR in runtime and peak GPU memory, but above FedAvg. The extra cost therefore accompanies memory maintenance and alignment, while the measurements themselves describe the simulated training pipeline rather than deployment-level efficiency.



(a) Runtime (hours). (b) GPU memory (GB). (c) CPU RAM (GB).

Figure 4: Incremental CIFAR-100 workstation cost from one sequential-client run per method. Details are in Tables 21 and 36.

5 Scope and Limitations

We study rehearsal-state maintenance in FCIL under controlled benchmark settings rather than real-world deployment. We use image, text, and node-classification streams to study class exposure and memory maintenance, while leaving network delays, different device capabilities, unstable connections, and long client absences for future work. Our protocol changes when and where classes appear rather than delaying model aggregation. We also test a stricter initialization that uses only each client's first local data, but we do not study random initialization without warmup, settings that completely exclude data from classes appearing later in the federation, naturally collected client streams, foundation models, or very large long-running federations.

We use compact class statistics and mean-only alignment, so PRO-MAX provides approximate memory maintenance rather than exact distribution transport. Our support and sensitivity experiments cover the tested settings but do not guarantee the same behavior under every data distribution or training configuration. Likewise, our resource measurements describe the simulated training pipeline rather than real deployment latency or energy use, and we evaluate one shared global model rather than personalized models or performance guarantees for every client. Class statistics and transport summaries may also reveal private information despite avoiding raw examples (Mahendran and Vedaldi, 2015; Fredrikson et al., 2015; Zhu et al., 2019). Appendix J discusses these scope, system, and privacy boundaries in more detail.

6 Conclusion

FCIL has one global model but no single class history. Historical evidence must therefore remain connected to adaptation and compatible with the evolving representation rather than merely remain stored. PRO coordinates READ, OPTIMIZE, and WRITE to expose available classes, rehearse before the adaptive module, and reconstruct statistics under the aggregated encoder. PRO-MAX adds MOVE only when direct reconstruction is unavailable, making its approximation and priority explicit. Controlled comparisons show that this coordination matters more than feature spread alone and that faithful replay cannot repair every weakness in the learned state. Across the evaluated streams, the lifecycle improves FAA and lowers AF under heterogeneous exposure.

AI USE STATEMENT

AI tools were used only to assist with language editing and improve the clarity of the manuscript. All scientific content, analyses, results, and conclusions were produced and verified by the authors.

ETHICS STATEMENT

This work does not involve new human-subject data collection or deployment decisions. We use existing research datasets and follow their established research-use protocols.

REPRODUCIBILITY STATEMENT

We provide the information needed to reproduce the method and reported evaluation. The ACE problem setup, exposure diagnostics, evaluation masks, and metrics are described in Appendices B, B.2, B.4 and B.7. The READ-OPTIMIZE-WRITE lifecycle, the MOVE extension, algorithmic summary, and optional adapted-space details are provided in Appendices C, C.4 and C.5. Dataset processing, task splits, encoders, schedules, initialization, optimization, baselines, replay objects, and alignment budgets are specified in Appendices D, D.2 and D.4. Complete reported tables, controls, scale-up results, sweeps, diagnostics, system measurements, and analytical derivations are given in Appendices E to I. Remaining run-artifact fields and evidence boundaries are documented in Appendices D.7 and J.

REFERENCES

Sara Babakniya, Zalan Fabian, Chaoyang He, Mahdi Soltanolkotabi, and Salman Avestimehr. A data-free approach to mitigate catastrophic forgetting in federated class incremental learning for vision tasks. In *Advances in Neural Information Processing Systems*, volume 36, pages 66408–66425, 2023. doi: 10.52202/075280-2899.

Keith Bonawitz, Vladimir Ivanov, Ben Kreuter, Antonio Marcedone, H. Brendan McMahan, Sarvar Patel, Daniel Ramage, Aaron Segal, and Karn Seth. Practical secure aggregation for privacy-preserving machine learning. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pages 1175–1191, 2017. doi: 10.1145/3133956.3133982.

Pietro Buzzega, Matteo Boschini, Angelo Porrello, Davide Abati, and Simone Calderara. Dark experience for general continual learning: a strong, simple baseline. *Advances in neural information processing systems*, 33:15920–15930, 2020.

Yatin Chaudhary, Pranav Rai, Matthias Schubert, Hinrich Schütze, and Pankaj Gupta. Federated continual learning for text classification via selective inter-client transfer. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4789–4799, 2022. doi: 10.18653/v1/2022.findings-emnlp.353.

Jiahua Dong, Lixu Wang, Zhen Fang, Gan Sun, Shichao Xu, Xiao Wang, and Qi Zhu. Federated class-incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10164–10173, 2022.

Jiahua Dong, Hongliu Li, Yang Cong, Gan Sun, Yulun Zhang, and Luc Van Gool. No one left behind: Real-world federated class-incremental learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(4):2054–2070, 2024. doi: 10.1109/TPAMI.2023.3334213.

Cynthia Dwork and Aaron Roth. *The Algorithmic Foundations of Differential Privacy*. Now Publishers, 2014. doi: 10.1561/0400000042.

Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, pages 1322–1333, 2015. doi: 10.1145/2810103.2813677.

Alex Gomez-Villa, Dipam Goswami, Kai Wang, Andrew D. Bagdanov, Bartlomiej Twardowski, and Joost van de Weijer. Exemplar-free continual representation learning via learnable drift compensation. In *Computer Vision – ECCV 2024*, volume 15065 of *Lecture Notes in Computer Science*, pages 473–490. Springer, 2025. doi: 10.1007/978-3-031-72667-5_27.

Zhipeng Guo, Yu Zhao, Yabin Zheng, Xiance Si, Zhiyuan Liu, and Maosong Sun. THUCTC: An efficient chinese text classifier. THU Natural Language Processing Lab repository, 2016. URL `https://github.com/thunlp/THUCTC`. The repository distributes the THUCNews Chinese news classification dataset.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

Weihua Hu, Matthias Fey, Marinka Zitnik, Yuxiao Dong, Hongyu Ren, Bowen Liu, Michele Catasta, and Jure Leskovec. Open graph benchmark: Datasets for machine learning on graphs. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL `https://proceedings.neurips.cc/paper/2020/hash/fb60d411a5c5b72b2e7d3527cfc84fd0-Abstract.html`.

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. SCAFFOLD: Stochastic controlled averaging for federated learning. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 5132–5143. PMLR, 2020.

Yoon Kim. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1746–1751, 2014. doi: 10.3115/v1/D14-1181. URL `https://aclanthology.org/D14-1181/`.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114 (13):3521–3526, 2017.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Stefan Larson, Anish Mahendran, Joseph J. Peper, Christopher Clarke, Andrew Lee, Parker Hill, Jonathan K. Kummerfeld, Kevin Leach, Michael A. Laurenzano, Lingjia Tang, and Jason Mars. An evaluation dataset for intent classification and out-of-scope prediction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 1311–1316, 2019. doi: 10.18653/v1/D19-1131.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. In *Proceedings of Machine Learning and Systems*, volume 2, pages 429–450, 2020.

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

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

Jinglin Liang, Jin Zhong, Hanlin Gu, Zhongqi Lu, Xingxing Tang, Gang Dai, Shuangping Huang, Lixin Fan, and Qiang Yang. Diffusion-driven data replay: A novel approach to combat forgetting in federated class continual learning. In *Computer Vision – ECCV 2024*, volume 15088 of *Lecture Notes in Computer Science*, pages 303–319. Springer, 2024. doi: 10.1007/978-3-031-73404-5_18.

Yuhang Ma, Zhongle Xie, Jue Wang, Ke Chen, and Lidan Shou. Continual federated learning based on knowledge distillation. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, pages 2182–2188, 2022. doi: 10.24963/ijcai.2022/303.

Aravindh Mahendran and Andrea Vedaldi. Understanding deep image representations by inverting them. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5188–5196, 2015.

H. Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 1273–1282. PMLR, 2017.

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

Milad Khademi Nori, Il-Min Kim, and Guanghui Wang. Federated class-incremental learning: A hybrid approach using latent exemplars and data-free techniques to address local and global forgetting. In *International Conference on Learning Representations*, 2025.

Grégoire Petit, Adrian Popescu, Hugo Schindler, David Picard, and Bertrand Delezoide. Fetril: Feature translation for exemplar-free class-incremental learning. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 3911–3920, 2023.

Zhuang Qi, Ying-Peng Tang, Lei Meng, Han Yu, Xiaoxiao Li, and Xiangxu Meng. Class-wise balancing data replay for federated class-incremental learning. In *Advances in Neural Information Processing Systems*, volume 38, pages 161317–161342, 2025. doi: 10.52202/085713-4863.

Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H. Lampert. iCaRL: Incremental classifier and representation learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2001–2010, 2017.

Xuankun Rong, Jianshu Zhang, Kun He, and Mang Ye. CAN: Leveraging clients as navigators for generative replay in federated continual learning. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 52025–52040. PMLR, 2025.

Prithviraj Sen, Galileo Mark Namata, Mustafa Bilgic, Lise Getoor, Brian Gallagher, and Tina Eliassi-Rad. Collective classification in network data. *AI Magazine*, 29(3):93–106, 2008. doi: 10.1609/AIMAG.V29I3.2157.

Donald Shenaj, Marco Toldo, Alberto Rigon, and Pietro Zanuttigh. Asynchronous federated continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 5055–5063, 2023.

Yue Tan, Guodong Long, Lu Liu, Tianyi Zhou, Qinghua Lu, Jing Jiang, and Chengqi Zhang. FedProto: Federated prototype learning across heterogeneous clients. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 8432–8440, 2022. doi: 10.1609/aaai.v36i8.20819.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=rJXMpikCZ.

Frank Wan, Fengyuan Ran, Ruikang Zhang, Wenke Huang, Xuankun Rong, Guibin Zhang, Yuxin Wu, Bo Du, and Mang Ye. MOTION: Multi-sculpt evolutionary coarsening for federated continual graph learning. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-2363. URL https://papers.nips.cc/paper_files/paper/2025/hash/659afd140447b7d50fadf8bde6a2240d-Abstract-Conference.html.

Fu-Yun Wang, Da-Wei Zhou, Han-Jia Ye, and De-Chuan Zhan. Foster: Feature boosting and compression for class-incremental learning. In *European conference on computer vision*, pages 398–414. Springer, 2022.

Jiayu Wu, Qixiang Zhang, and Guoxi Xu. Tiny imagenet challenge. Technical report, Stanford University, 2017. URL https://cs231n.stanford.edu/reports/2017/pdfs/930.pdf.

Abudukelimu Wuerkaixi, Sen Cui, Jingfeng Zhang, Kunda Yan, Bo Han, Gang Niu, Lei Fang, Changshui Zhang, and Masashi Sugiyama. Accurate forgetting for heterogeneous federated continual learning. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/dd1810bba6fd3d8b7a6f4aa7c263c591-Abstract-Conference.html.

Shipeng Yan, Jiangwei Xie, and Xuming He. Der: Dynamically expandable representation for class incremental learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3014–3023, 2021.

Jaehong Yoon, Wonyong Jeong, Giwoong Lee, Eunho Yang, and Sung Ju Hwang. Federated continual learning with weighted inter-client transfer. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 12073–12086. PMLR, 2021.

Lu Yu, Bartlomiej Twardowski, Xialei Liu, Luis Herranz, Kai Wang, Yongmei Cheng, Shangling Jui, and Joost van de Weijer. Semantic drift compensation for class-incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6982–6991, 2020. doi: 10.1109/CVPR42600.2020.00701. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Yu_Semantic_Drift_Compensation_for_Class-Incremental_Learning_CVPR_2020_paper.html.

Jie Zhang, Chen Chen, Weiming Zhuang, and Lingjuan Lyu. TARGET: Federated class-continual learning via exemplar-free distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4782–4793, 2023.

Fei Zhu, Xu-Yao Zhang, Chuang Wang, Fei Yin, and Cheng-Lin Liu. Prototype augmentation and self-supervision for incremental learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5871–5880, 2021.

Ligeng Zhu, Zhijian Liu, and Song Han. Deep leakage from gradients. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Yinlin Zhu, Miao Hu, and Di Wu. Federated continual graph learning. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 4203–4213, 2025. doi: 10.1145/3711896.3736956.

A Extended Related Work

The main paper positions Projected Rehearsal Orchestration (PRO) and its Memory Alignment Extension (PRO-MAX) around the continued usability of historical evidence in federated class-incremental learning (FCIL). This appendix expands the same three comparisons, distinguishing client-relative histories, historical evidence, and the representation reference used to construct or maintain that evidence.

A.1 Client-relative histories and heterogeneous federation

Federated continual learning already considers different client task histories. FedWeIT uses shared and task-specific knowledge for inter-client transfer, CFeD coordinates learning and historical review, and GLFC distinguishes local and global forgetting (Yoon et al., 2021; Ma et al., 2022; Dong et al., 2022). FedSeIT studies selective transfer in federated continual text classification, while FedSpace considers asynchronous continual progression (Chaudhary et al., 2022; Shenaj et al., 2023). Asynchronous class exposure (ACE) makes the role of a labeled class relative to each client explicit within this established setting. It is not a new delayed-aggregation protocol.

The evaluated sampler separates task support, local label coverage, and task-order perturbation. These settings can change available evidence and the objectives aggregated at a round, but they are not measurements of realized ACE. General federated learning (FL) methods such as FedProx and SCAFFOLD address optimization heterogeneity without providing a complete class-incremental retention mechanism (Li et al., 2020; Karimireddy et al., 2020). FedSSI and FedCBDR instead supply parameter-level retention or coordinated replay for continual federation (Li et al., 2025; Qi et al., 2025). Our comparisons ask whether the evaluated class-memory lifecycle is useful alongside these alternatives, not whether heterogeneous histories must always reduce accuracy.

A.2 Supplying and using historical evidence

Continual-learning methods preserve knowledge through parameter constraints, distillation, exemplars or replay, and architectural adaptation (Kirkpatrick et al., 2017; Li and Hoiem, 2017; Rebuffi et al., 2017; Buzzega et al., 2020; Yan et al., 2021; Wang et al., 2022). Representation-based approaches include exemplar-supported class means, prototype augmentation, and pseudo-feature construction (Rebuffi et al., 2017; Zhu et al., 2021; Petit et al., 2023). FedProto makes class representatives communication objects in heterogeneous federation (Tan et al., 2022). These precedents establish that compact feature evidence and prototype communication are not new by themselves.

TARGET and MFCL provide exemplar-free historical evidence through data-free replay. Other approaches use twin generators, diffusion, hybrid latent rehearsal, or client-aware replay (Zhang et al., 2023; Babakniya et al., 2023; Nguyen et al., 2024; Liang et al., 2024; Nori et al., 2025; Rong et al., 2025). Their mechanisms should not be treated as a single generator family with a common failure guarantee. FedCBDR and client-aware replay also address how historical evidence is allocated rather than merely how it is synthesized (Qi et al., 2025; Rong et al., 2025). The source-only intervention in this paper is a within-PRO comparison, not an experiment that changes only the generator of each baseline.

PRO uses constructed class moments and current features in an available-class-balanced surrogate. Its main distinction is the coordination of semantic support, the pre-adapter gradient path, and fresh construction under the aggregated encoder. Equal replay-object counts do not establish equal information, storage, communication, or computation, and the evaluated adaptations must be distinguished from unchanged native baseline implementations.

A.3 Maintaining representation compatibility

Accurate Forgetting for Federated Continual Learning (AF-FCL) uses evolving learned feature evidence, while Hybrid Rehearsal (HR) maintains latent historical information and refreshes it through decoding and re-encoding (Wuerkaixi et al., 2024; Nori et al., 2025). Thus, feature replay and maintained latent memory already have direct precedents. PRO instead reconstructs fresh class-level moments from currently available inputs under the aggregated global encoder. This creates a common construction reference for participating writers without recovering every historical client population.

Semantic Drift Compensation (SDC) and Learnable Drift Compensation (LDC) study representation-drift compensation in centralized continual learning (Yu et al., 2020; Gomez-Villa et al., 2025). PRO-MAX estimates class-local movement from paired current calibration features and aggregates client summaries when direct WRITE is unavailable. The additional federated issue is that a client's local update and the aggregated global update need not produce the same feature displacement. Our analysis retains this discrepancy and limits MOVE to approximate mean maintenance. It does not claim the first use of drift compensation or exact global transport.

Pre-adapter feature rehearsal gives historical losses a path to the adapter and classifier, but still bypasses the base encoder. Other methods can preserve knowledge through parameter constraints or distillation without using this path. The four consistency criteria therefore organize the class-feature rehearsal design studied here rather than define universal requirements for all continual learners.

A.4 Modality and privacy scope

FedSeIT motivates a text-specific comparison, and POWER and MOTION provide federated continual graph-learning comparisons (Chaudhary et al., 2022; Zhu et al., 2025; Wan et al., 2025). The present graph experiments operate on node embeddings after a graph attention network (GAT) encoder. They do not preserve all graph topology through class moments, and they are not whole-graph classification experiments. The portability claim concerns reuse of the post-encoder memory interface after the evaluated encoders rather than universal transfer to every modality or architecture.

Class-level statistics are more compact than instance-level feature banks in their dependence on the number of stored objects, but compactness is not a privacy guarantee. Representations, model outputs, and gradients can disclose information about training data (Mahendran and Vedaldi, 2015; Fredrikson et al., 2015; Zhu et al., 2019). Neither the present comparisons nor the absence of raw historical exemplars measures the privacy benefit of class aggregation. Formal protections and leakage evaluation remain separate requirements.

B Complete ACE Protocol and Metrics

This section expands Section 3.1 using the same distinction between labeled exposure, local evidence, and constructed memory. Semantic tasks organize the stream, but their scheduled arrival does not by itself establish that every constituent class has been labeled by a participating client. The appendix reports final average accuracy (FAA), current-task accuracy (CTA), and average forgetting (AF), with formal definitions below.

B.1 Rounds, task positions, and labeled exposure

Let $\mathcal{C} = \{1, \dots, C\}$ and $\mathcal{T} = \{1, \dots, T\}$ denote clients and semantic tasks. The task class sets $\mathcal{Y}_t$ are disjoint. Client c follows permutation π_c, occupies local position $p_{c,r}$ at communication round r, and receives labeled data $\mathcal{D}_{c,r}$ with observed labels $\widetilde{\mathcal{Y}}_{c,r} \subseteq \mathcal{Y}_{\pi_c[p_{c,r}]}$. A scheduled position can have no labeled data when the client does not support that task. Empty observations add no labels to the local history.

For participants $\mathcal{S}_r$, the federation's labeled prediction support is

$$\mathcal{Y}_r^g = \bigcup_{s=0}^{r} \bigcup_{c \in \mathcal{S}_s} \widetilde{\mathcal{Y}}_{c,s}, \qquad \mathcal{H}_{c,r} = \bigcup_{s=0}^{r} \widetilde{\mathcal{Y}}_{c,s}, \qquad \mathcal{H}_{c,-1} = \varnothing. \tag{20}$$

The model predicts over $\mathcal{Y}_r^g$ without receiving a task identity. Local observation and contribution to the federation are different events when a client is not selected. The exposure record must therefore distinguish availability of a local label from the first participating-client contribution of that label. Memory keys satisfy $\mathcal{K}_r \subseteq \mathcal{Y}_r^g$, but equality is not assumed before a fresh WRITE has constructed the corresponding entries.

The exposure states in Section 3.1 are mutually exclusive. A class is current when it appears in current local labeled data, historical when it was labeled locally but is absent from current data, and global-only when it has been exposed elsewhere but has not been labeled locally. ACE holds

if two selected clients give different states to at least one $y \in \mathcal{Y}_r^g$. It does not require delayed communication.

B.2 Realized exposure diagnostics

Nominal order perturbation does not measure the disagreement realized among the clients selected in one round. For $|\mathcal{S}_r| \geq 2$ and $|\mathcal{Y}_r^g| > 0$, define

$$\mathrm{ACE}_r = \frac{\sum_{\substack{c,c' \in \mathcal{S}_r \\ c<c'}} \sum_{y \in \mathcal{Y}_r^g} \mathbf{1}[e_{c,r}(y) \neq e_{c',r}(y)]}{\binom{|\mathcal{S}_r|}{2} |\mathcal{Y}_r^g|}. \tag{21}$$

This is the probability of role disagreement for a uniformly chosen participant pair and exposed class. To distinguish disagreement from the breadth of previous labeled exposure, define

$$\mathrm{Supp}_r = \frac{\sum_{c \in \mathcal{S}_r} \sum_{y \in \mathcal{Y}_r^g} \mathbf{1}[y \in \mathcal{H}_{c,r}]}{|\mathcal{S}_r||\mathcal{Y}_r^g|}. \tag{22}$$

A stream average is taken over rounds on which the corresponding denominator is defined. These diagnostics are definitions, not additional measured response surfaces. The reported sweeps vary the sampler settings in Table 8. No measured ACE-accuracy regression is inferred from those percentages. Neither score alone determines difficulty because different clients can provide complementary evidence.

B.3 Support and order interventions

The task support set is $\mathcal{C}_t = \{c : |\mathcal{D}_{c,p_c(t)}| > 0\}$. Task support controls $|\mathcal{C}_t|/C$, while label coverage controls the fraction of $\mathcal{Y}_t$ assigned to a supporting client. The declared stream-wide feasibility condition is

$$\bigcup_{c \in \mathcal{C}_t} \widetilde{\mathcal{Y}}_{c,p_c(t)} = \mathcal{Y}_t, \qquad t \in \mathcal{T}. \tag{23}$$

It ensures eventual labeled support for the evaluated task classes, not support at every earlier round. The order intervention changes π_c while holding semantic class groups fixed. The supplied protocol does not specify every rounding rule, permutation perturbation procedure, or support-repair rule. These implementation fields remain necessary for exact replication. Lower task support also lowers total task evidence in the evaluated support intervention, so it is not a redistribution of a fixed total dataset.

The aligned reference has a common order, synchronous progression, full task support, and full local label coverage. Under these conditions the client-relative class roles coincide. Zero order perturbation alone is insufficient because clients may still have different labeled supports. The order sweep retains 40% task support and 50% label coverage, whereas the aligned reference uses 100% for both.

B.4 Semantic scheduling and evaluation-mask compatibility

For a common evaluation checkpoint r_p, let $\mathcal{T}_c^{(p)} = \{\pi_c[0], \ldots, \pi_c[p]\}$ and define the scheduled semantic set $\mathcal{Y}_{\mathrm{sem}}^{(p)} = \bigcup_c \bigcup_{t \in \mathcal{T}_c^{(p)}} \mathcal{Y}_t$. This set can be larger than $\mathcal{Y}_{r_p}^g$. The semantic-task convention in the earlier appendix is equivalent to the labeled-exposure convention in the main formulation only when every class of every scheduled task has actually been exposed through a participating client by that checkpoint. Stream-wide feasibility in Equation (23) does not establish this checkpoint-level condition.

The formulation and algorithms below use labeled exposure. A per-checkpoint audit confirms that the prediction masks and eligible test labels follow this convention for every reported stream position. The numerical records therefore retain their reported values without relabeling or recomputation. This check is especially relevant to CTA and AF, which depend on intermediate checkpoints.

Table 6: Evaluation-mask audit. Each reported stream position contributes one aggregate checkpoint. No prediction-mask or test-label violations were found.

Dataset	Evaluated checkpoints	Prediction-mask violations	Test-label violations
CIFAR-100	10	0	0
TinyImageNet	20	0	0
THUCNews-10	5	0	0
CLINC150	15	0	0
Cora	3	0	0
ogbn-arxiv	8	0	0

B.5 Current-only and cumulative objectives

To isolate the missing-evidence issue, fix the global encoder at round r and differentiate with respect to $v = (\psi, W)$. For the cumulative class-y feature distribution $P^{u,g}_{y,r}$ and cross-entropy (CE) loss $\ell_y(u; v) = \mathrm{CE}(W^\top a_\psi(u), y)$, define

$$F^{\mathrm{cum}}_r(v) = \frac{1}{|\mathcal{Y}^g_r|} \sum_{y \in \mathcal{Y}^g_r} \mathbb{E}_{P^{u,g}_{y,r}} \ell_y(u; v), \qquad F^{\mathrm{cur}}_{c,r}(v) = \sum_y \rho_{c,r}(y) \mathbb{E}_{P^{u,g}_{c,y,r}} \ell_y(u; v). \tag{24}$$

Here $\rho_{c,r}(y)$ is the local empirical class proportion and is zero when y is absent. With g_y and $g_{c,y}$ denoting the corresponding class gradients, $\bar{\rho}(y) = \sum_c \omega_c \rho_c(y)$, and normalized nonnegative client weights, the exact decomposition is

$$\begin{aligned} \sum_c \omega_c \nabla F^{\mathrm{cur}}_c - \nabla F^{\mathrm{cum}} = & \sum_y \left(\bar{\rho}(y) - \frac{1}{|\mathcal{Y}^g_r|} \right) g_y \\ & + \sum_c \omega_c \sum_y \rho_c(y)(g_{c,y} - g_y). \end{aligned} \tag{25}$$

The first term records class coverage and weighting, while the second records class-conditional mismatch. Missing support means the absence of a class-conditional risk term with that target label, not a claim that the corresponding classifier column receives no CE gradient from other classes. Reweighting cannot create a missing observation, but complementary client supports can reduce aggregate weighting error.

B.6 Rehearsal-consistency criteria

Exposure support. The available rehearsal set is $\mathcal{A}_{c,r} = \widetilde{\mathcal{Y}}_{c,r} \cup \mathcal{K}_r$. The surrogate places positive mass on each available class rather than assuming a memory exists for every globally exposed class. Newly exposed classes can be present in current data before they are available for shared-memory replay.

Gradient reachability. Historical replay must have a differentiable path to the trainable downstream modules that it is intended to constrain. For a stored feature $\widetilde{u}$ and $p = \mathrm{softmax}(W^\top a_\psi(\widetilde{u}))$, pre-adapter rehearsal gives

$$\nabla_\psi \mathcal{L}_{\mathrm{hist}} = J_\psi a_\psi(\widetilde{u})^\top W(p - e_y), \tag{26}$$

whereas detached post-adapter replay gives

$$\mathcal{L}^{\mathrm{post}}_{\mathrm{hist}} = \mathrm{CE}(W^\top \widetilde{z}, y), \qquad \nabla_\psi \mathcal{L}^{\mathrm{post}}_{\mathrm{hist}} = 0. \tag{27}$$

A path can exist even when its gradient vanishes at a particular parameter value. Neither path reaches the base encoder when historical features are supplied as fixed inputs.

Global-coordinate consistency. Fresh class statistics are computed from contributing inputs under the same aggregated encoder, as in Equations (12) and (13). This removes disagreement attributable solely to using different encoders on the same contributing populations. It does not make those populations representative of all historical clients.

Temporal consistency. For a fixed historical input population P^x_y, encoder updates change its pushforward distribution. A candidate feature map T_r satisfies

$$W_1\big(P^{u,g}_{y,r+1}, (T_r)_\# P^{u,g}_{y,r}\big) \le \mathbb{E}_{X \sim P^x_y} \| h_{\theta^g_{r+1}}(X) - T_r(h_{\theta^g_r}(X)) \|_2. \tag{28}$$

The coupling uses the same historical input before and after the update. The residual need not vanish, and a mean translation cannot preserve arbitrary changing covariance or multimodality. These four criteria guide this feature-rehearsal design rather than prescribing the components of every continual-learning method.

B.7 Evaluation metrics

The reported metrics evaluate one global single-head model, not personalized classifiers. Let r_p denote the checkpoint after the communication rounds allocated to position p. For a semantic task t, let $\mathcal{D}_t^{\text{test}}$ be its fixed test split and let $a_t^{(p)}$ be the global accuracy on that split using the declared prediction mask at r_p. A task identifier selects evaluation examples only and is never given to the predictor. Under the main labeled-exposure convention, evaluation of a complete semantic-task test split is appropriate only once all its labels belong to the exposed set. Earlier partially exposed splits require an explicit evaluator policy. The audit in Table 6 verifies this convention for the retained records.

FAA. For $\mathcal{D}_c^{(p),\text{test}} = \bigcup_{t \in \mathcal{T}_c^{(p)}} \mathcal{D}_t^{\text{test}}$, define its sample-weighted global accuracy $A_c^{(p)}$. The reported client-first convention is

$$\text{FAA} = \frac{1}{C} \sum_{c=1}^{C} A_c^{(T-1)}. \tag{29}$$

When every client follows a full permutation of all T tasks, their terminal test pools are identical, so this equals the accuracy of the final global model on the complete test union. This is not automatically an unweighted mean of task accuracies when task test counts differ. The distinction matters for unequal task sizes such as the Cora split.

CTA. CTA evaluates each scheduled semantic task at the checkpoint where it is current in the client's order,

$$\text{CTA} = \frac{1}{CT} \sum_{c=1}^{C} \sum_{p=0}^{T-1} a_{\pi_c[p]}^{(p)}. \tag{30}$$

It complements final utility by measuring current-task performance. Because the predictor is global, CTA also reflects knowledge transferred from other clients and is not an isolated estimate of one client's plasticity. Under sparse task support, a scheduled position need not contain local examples, so the implementation must state whether such positions are included as in this formula.

AF. Let $p_c(t)$ be the position of task t in π_c. The retained convention is

$$\text{AF} = \frac{1}{C(T-1)} \sum_{c=1}^{C} \sum_{t \neq \pi_c[T-1]} \left[\max_{p_c(t) \leq q \leq T-1} a_t^{(q)} - a_t^{(T-1)} \right]. \tag{31}$$

The maximum includes the final checkpoint, so each summand is nonnegative. The last scheduled task is excluded because this evaluation grid has no later position for it. Under ACE, the window is client-order-conditioned rather than anchored to a unique federation-wide first exposure. Low AF alone is not sufficient evidence of successful learning because a model that never learned a task well can also have a small decline. FAA, CTA, and AF must therefore be interpreted jointly. Accuracy and forgetting are reported as percentages, while differences are percentage points.

B.8 Statistical reporting

Unless explicitly marked mean-only, predictive tables report five-seed means and standard deviations. The targeted rehearsal controls now also report their seed-level standard deviations, while system tables use one run per method. Standard deviations are not confidence intervals, and differences of reported means do not provide paired uncertainty or significance. Per-seed tensors, evaluation checkpoints, class masks, and test counts are needed to recompute the metrics and their paired differences.

C Additional Method Details

This section expands the main READ-OPTIMIZE-WRITE lifecycle and the conditional MOVE extension. All algorithms use the current global model, constructed memory keys, and client-relative class evidence. They do not assume that every scheduled semantic class already has an available memory.

C.1 Shared initialization and controlled plasticity

All compared methods receive the same selected federated in-domain initialization before the supervised stream. The main comparisons use each client's complete local training-input pool with labels removed, including inputs scheduled for later local positions. The position-0-only control instead uses Equation (5). Neither protocol uses test examples or creates labeled class memories. A position-0-only pool is locally restricted but can still span different semantic tasks across clients.

The supplied warmup objective has the form

$$\mathcal{L}_{\mathrm{warm}} = \mathbb{E}_{x,\tilde{x}\sim\mathcal{A}(x)}\ell_{\mathrm{ssl}}\big(r_\omega(h_\theta(\tilde{x})), s(x,\tilde{x})\big). \tag{32}$$

The auxiliary head is discarded after initialization. Its precise modality-specific pretext target and augmentation configuration are required for independent replication. This generic expression alone does not fix them. Sharing initialization controls one comparison factor but does not establish that every remaining implementation difference is a single causal mechanism.

During the incremental stream, clients use Equation (8). For two stochastic views of each current input, define $v_i^{(a)} = r_\omega(h_\theta(\tilde{x}_i^{(a)}))/(\|r_\omega(h_\theta(\tilde{x}_i^{(a)}))\|_2 + \epsilon)$. The recorded auxiliary loss is

$$\begin{aligned} \mathcal{L}_{\mathrm{aux}} &= \frac{1}{2|\mathcal{B}|}\sum_i \left[\ell_{\mathrm{ctr}}(v_i^{(1)}, v_i^{(2)}) + \ell_{\mathrm{ctr}}(v_i^{(2)}, v_i^{(1)})\right], \\ \ell_{\mathrm{ctr}}(v_i, v_i^+) &= -\log\frac{\exp(v_i^\top v_i^+/\tau_{\mathrm{ssl}})}{\sum_{j=1}^{|\mathcal{B}|}\exp(v_i^\top v_j^+/\tau_{\mathrm{ssl}})}. \end{aligned} \tag{33}$$

The supplied configuration uses $\tau_{\mathrm{ssl}} = 0.5$, with $\lambda_{\mathrm{aux}} = 0.05$ for images and 0.01 for text and graphs while the base encoder is trainable. The base is trainable during the first two rounds allocated to a new local task and is subsequently frozen for that task. It is also held fixed during every projected-rehearsal stage. The auxiliary term is disabled when the base is frozen. No additional proximal term is applied to the projected-rehearsal objective. Calibration for MOVE measures the base-encoder transition induced by current fitting, before adapter rehearsal.

C.2 Available evidence and fresh-statistic construction

Current features are recomputed under the locally fitted encoder. For every class in $\mathcal{A}_{c,r} = \widetilde{\mathcal{Y}}_{c,r} \cup \mathcal{K}_r$, the client uses the current-only, memory-only, or overlap mixture in Equation (6). The overlap coefficient follows Equation (7), and labels are sampled uniformly over $\mathcal{A}_{c,r}$. The legacy phrase "old-class replay" therefore refers to existing memory entries, including globally exposed classes that remain locally unseen, rather than to the task-index set $\mathcal{Y}^{(p-1)}$.

For reference, the projected sampling rule is

$$\tilde{u}_{y,j} = \mu_{y,r}^u + \gamma\sigma_{y,r}^u \odot \epsilon_j, \qquad \epsilon_j \sim \mathcal{N}(0, I), \qquad y \in \mathcal{K}_r. \tag{34}$$

Sampling is class-balanced at the label level even though the number of available current examples differs. The absence of a constructed memory is not repaired by sampling a label with no associated feature distribution.

After local fitting and rehearsal, the server uses

$$\Theta_{r+1}^g = \sum_{c\in\mathcal{S}_r}\omega_{c,r}\Theta_{c,r}, \qquad \omega_{c,r} = \frac{n_{c,r}}{\sum_{c'} n_{c',r}}. \tag{35}$$

At a write boundary, the contributing clients re-encode their available class examples under $h_{\theta^g_{r+1}}$. They upload the empirical mean, coordinate-wise second moment, and sample count. For $N_y = \sum_{c \in \mathcal{V}_{y,r}} n_{c,y,r}$, their pooled moments are

$$\mu^u_{y,r+1} = \frac{\sum_c n_{c,y,r} \widehat{\mu}^{u,g}_{c,y,r}}{N_y}, \qquad s^u_{y,r+1} = \frac{\sum_c n_{c,y,r} \widehat{s}^{u,g}_{c,y,r}}{N_y}, \qquad n_{y,r+1} = N_y. \tag{36}$$

The spread uses Equation (13), not an average of client standard deviations. A completed WRITE replaces the entry. Sharing coordinates justifies pooling the fresh contributions, whereas replacement is the specified update policy. Classes without a completed WRITE keep their previous entries in PRO. The record specifies task-position-boundary refresh for the evaluated schedule, while the general algorithm permits a configured set of write rounds. The exact handling of empty-data selected clients and partial boundary participation must be preserved in the run configuration.

C.3 Local and server MOVE

For each locally unsupported stored class, let $\mathcal{J}_{c,r} = \mathcal{K}_r \setminus \widetilde{\mathcal{Y}}_{c,r}$. The same temporary current inputs produce before- and after-fitting features. Neighbors are selected using the pre-update features, with the adaptive bandwidth and weights in Equation (15). The class-dependent bandwidth $\tau_{c,y}$ is not a fixed global τ_u.

The support score is $\kappa_{c,y} = \sum_{i \in \mathcal{N}_{c,y}} \alpha_{i,y}$, and the local displacement is Equation (16). It changes only a temporary client copy,

$$\tilde{\mu}^u_{c,y,r} = \mu^u_{y,r} + \Delta^u_{c,y}, \qquad y \in \mathcal{J}_{c,r}. \tag{37}$$

The stored spread and count remain unchanged. Clients upload the displacement and support score, not the paired calibration inputs. The symbol $q^u_{c,y}$ used for confidence in the earlier appendix is replaced by $\kappa_{c,y}$ to distinguish it from the rehearsal distribution $q^u_{c,y,r}$.

The server combines available summaries according to

$$\Delta^{u,g}_y = \frac{\sum_{c \in \mathcal{S}_r} \kappa_{c,y} \Delta^u_{c,y}}{\sum_{c \in \mathcal{S}_r} \kappa_{c,y} + \varepsilon}. \tag{38}$$

Absent summaries contribute zero weight. If the total support is zero, the aggregate displacement is zero. The source specifies support weighting rather than a calibrated-probability threshold. No additional confidence gate is introduced into the main algorithm by this appendix.

Let $\mathcal{W}_r$ be the classes with completed fresh WRITE at round r. The next server memory follows Equation (19): fresh WRITE takes precedence, and MOVE applies only to existing entries without that fresh replacement. This prevents an old-coordinate displacement from being added to a freshly reconstructed global-coordinate mean. The score can downweight weak kernel support, but adaptive bandwidth means it cannot certify absolute neighborhood proximity. The local-estimation and global-transition discrepancies are analyzed in Appendix I.

C.4 Algorithmic summary

The algorithms below follow the main specification. They condition training on available current data and a nonempty available rehearsal set. Exact empty-client sampling and cache-size fallback policies remain implementation fields to be released. Server aggregation is logical parallel federation, while the measured implementation executes clients sequentially on one graphics processing unit (GPU) and uses central processing unit (CPU) resources for support operations.

Algorithm 1 PRO with available-class rehearsal and scheduled global WRITE

Require: Client streams, participation and write schedules, initialized model Θ_0^g, plasticity schedule, replay settings

Apply the selected shared initialization and set $\mathcal{M}_0^u \leftarrow \varnothing$
for $r = 0, \ldots, R-1$ **do**
  Select $\mathcal{S}_r$ and broadcast $(\Theta_r^g, \mathcal{M}_r^u)$
  for all training participants $c \in \mathcal{S}_r$ **do**
    Initialize $\Theta_{c,r} \leftarrow \Theta_r^g$ and fit current data using Equation (8)
    Hold the encoder fixed and recompute current features
    READ $\mathcal{A}_{c,r} = \widetilde{\mathcal{Y}}_{c,r} \cup \mathcal{K}_r$ with Equation (6)
    OPTIMIZE (ψ, W) by class-balanced pre-adapter rehearsal
    Upload the local model and aggregation count
  Aggregate Θ_{r+1}^g using Equation (11) and copy $\mathcal{M}_r^u$ forward
  if a write boundary is reached **then**
    Broadcast the aggregated encoder to the available writing clients
    Re-encode available class examples without gradients and upload moments and counts
    WRITE completed entries using Equations (12) and (13)
  Keep every entry without completed WRITE unchanged

Algorithm 2 PRO-MAX with temporary local alignment and WRITE-priority server maintenance

Require: Inputs of Algorithm 1, current calibration subset and neighbor count K

Initialize as in PRO
for $r = 0, \ldots, R-1$ **do**
  Select clients and broadcast $(\Theta_r^g, \mathcal{M}_r^u)$
  for all training participants $c \in \mathcal{S}_r$ **do**
    Cache current calibration features under θ_r^g
    Perform current fitting and recompute features on the same calibration inputs
    for all $y \in \mathcal{K}_r \setminus \widetilde{\mathcal{Y}}_{c,r}$ with available calibration **do**
      Compute $\tau_{c,y}$, $\alpha_{i,y}$, $\kappa_{c,y}$, and $\Delta_{c,y}^u$
      Shift the temporary local mean only, retaining its spread and count
    READ available evidence and OPTIMIZE the adapter and classifier with the encoder fixed
    Upload the model, aggregation count, and transport summaries
  Aggregate Θ_{r+1}^g and available transport summaries
  At a scheduled boundary, obtain fresh global-coordinate WRITE entries $\mathcal{W}_r$
  Use fresh WRITE for $y \in \mathcal{W}_r$, otherwise MOVE existing means using Equation (19)
  Keep an entry unchanged when it has neither fresh WRITE nor a nonzero transport summary

C.5 Optional adapted-space summaries

The main equations specify the base-space bank $\mathcal{M}^u$. The supplied appendix also records an optional bank in $z = a_{\psi^g}(h_{\theta^g}(x))$, containing a class mean, radius, and count. Its radius is $\rho_y^z = \sqrt{\mathbb{E}[\|z - \mu_y^z\|_2^2 \mid y]/d_z}$. Paired adapted features can analogously define $\Delta_{c,y}^z$, and the archived optional mean update is $\mu_y^z \leftarrow \mathrm{norm}(\mu_y^z + \Delta_y^{z,g})$.

This optional computation is not a second historical-gradient path in the main method and must not be silently added to the base-space algorithm. The supplied record includes adapted-space alignment numbers but does not identify the adapted-bank enable flag for every accuracy and resource run. Those records are retained as auxiliary diagnostics. The per-run setting η_z and the timing of adapted-feature measurements must be checked before attributing their cost or accuracy to the base-only lifecycle.

D EXPERIMENTAL PROTOCOL AND REPRODUCIBILITY

This section supplies the implementation details behind the compact protocol in the main paper. Definitions of labeled exposure and evaluation metrics are centralized in Appendix B. The tables below give the reported configurations rather than introducing a second prediction-space convention.

D.1 DATASETS, TASK SPLITS, AND ENCODERS

CIFAR-100 (Krizhevsky, 2009) supports the controlled retention, support/order, lifecycle, and source-intervention comparisons. **TinyImageNet** (Wu et al., 2017) extends the visual stream to twenty ten-class tasks. **THUCNews-10** (Guo et al., 2016) and **CLINC150** (Larson et al., 2019) provide news and intent classification streams, while **Cora** (Sen et al., 2008) and **ogbn-arxiv** (Hu et al., 2020) provide node-classification streams. The main question is whether the class-memory interface remains useful after different encoders, not whether one input model works unchanged across modalities.

Images use ResNet-18 initialized without external ImageNet weights before shared in-domain warmup. Text uses the same TextCNN-style backbone on both datasets, and nodes use two-layer GAT encoders (He et al., 2016; Kim, 2014; Veličković et al., 2018). Projected graph memories contain node embeddings, not raw attributes or edges. The record does not specify every tokenizer, adapter dimension, dataset split identifier, or graph-visibility rule. These fields are listed below rather than inferred from a baseline paper.

D.2 CORE SCHEDULE AND EXPOSURE INTERVENTIONS

For each dataset and seed, compared methods share the class grouping, client partition, participation schedule, initialization protocol, and incremental training budget. The numerical schedules are in Table 7. In this appendix, rounds per task refer to communication rounds allocated to a local stream position, since semantic task identities can differ across clients at that position.

Table 7: Reported dataset-specific schedules. Participants denotes selected clients per communication round. Rounds are allocated per stream position. Graph batches use target nodes and their client-local computation graphs.

Dataset	Task split	Clients	Participants	Rounds/task	Local epochs	Batch size	Seeds
CIFAR-100	10×10 classes	20	10	20	5	64	5
TinyImageNet	20×10 classes	20	10	20	5	64	5
THUCNews-10	5×2 categories	20	10	10	3	64	5
CLINC150	15×10 intents	20	10	10	3	64	5
Cora	3 tasks $(2/2/3)$ classes	5	5	30	5	full local graph	5
ogbn-arxiv	8×5 classes	10	5	20	3	1024 target nodes	5

The aligned reference has full task support and label coverage with a common task order. The heterogeneous reference has 40% task support, 50% local label coverage, and 60% order perturbation. "Severe heterogeneity" in the retained tables denotes this reference setting, not a measured ACE value. Separate sweeps change support or order while holding the other listed settings fixed.

Table 8: Sampler settings for the aligned reference, heterogeneous reference, and separate sweeps. The percentages are not realized ACE scores.

Setting	Task support	Label coverage	Order disorder
Aligned	100%	100%	0%
Severe heterogeneity	40%	50%	60%
Support sweep	$\{80, 60, 40, 20\}\%$	50%	60%
Order sweep	40%	50%	$\{0, 30, 60, 90\}\%$

The support sweep changes total task evidence rather than only redistributing a fixed dataset. The zero-perturbation endpoint retains sparse support and label coverage and is not the aligned reference.

Feasibility repair must ensure the declared final class coverage without silently treating that condition as checkpoint-level labeled exposure.

D.3 Initialization and optimization

All methods use the same in-domain self-supervised initialization. The main input pool is the full local training-input pool without labels or test examples, including later-local-position inputs. The separate position-0-only comparison changes this access assumption, not the whole learning mechanism. Images use twenty warmup rounds, while text and graphs use ten, each with one local epoch. The warmup head is discarded before the incremental stream, as described in Appendix C.1. No no-warmup experiment is inferred from the position-0-only control.

Image optimization uses stochastic gradient descent (SGD) with momentum 0.9, weight decay 5×10^{-4}, and initial learning rate 0.05. Text uses Adam with learning rate 10^{-3} and weight decay 10^{-4}. Graphs use Adam with learning rate 5×10^{-3} and weight decay 5×10^{-4}. The supplied protocol states selection on held-out validation seeds followed by fixed settings for reported seeds, but does not provide the complete validation search ledger.

D.4 Replay objects and alignment budgets

Replay counts are per historical class, participating client, and local epoch. The smaller-count configuration provides twenty objects where applicable. Expanded replay gives TARGET, MFCL, HR, and FedCBDR one hundred objects while PRO and PRO-MAX retain twenty projected features. These are object-count configurations, not matched-byte, matched-information, or matched-compute regimes. Increasing the replay count alone does not prove that quantity is no longer limiting.

Table 9: Replay counts per historical class, participating client, and local epoch. Smaller and expanded denote object-count settings, not matched resources. Summary and parameter states are not sample counts.

Object	Smaller replay	Expanded replay
PRO/PRO-MAX projected pseudo features	20	20
TARGET/MFCL/HR replay samples	20	100
FedCBDR replay samples	20	100
FedProto/FedSpace prototype summaries	class-level	class-level
FedSSI parameter importance	parameter-level	parameter-level

PRO and PRO-MAX sample available memory entries and mix them with current examples according to Equation (6). The reported PRO-MAX settings are calibration size 128, $K = 5$, projected spread $\gamma = 0.10$, and proximal coefficient $\lambda_{\text{prox}} = 10^{-2}$. The supplied auxiliary weights and temperature are given with the auxiliary loss in Appendix C. The base encoder is trainable during the first two rounds of each new local task, then frozen. Projected rehearsal always updates only the adapter and classifier. The evaluated refresh schedule uses the final communication round of a task position. Fresh WRITE takes precedence over MOVE.

D.5 Baseline roles and adaptations

FedAvg, **FedProx**, and **FedProto** respectively provide current-data federation, optimization stabilization, and prototype-state controls (McMahan et al., 2017; Li et al., 2020; Tan et al., 2022). **TARGET**, **MFCL**, and **HR** provide distinct replay-oriented comparisons (Zhang et al., 2023; Babakniya et al., 2023; Nori et al., 2025). **FedSpace**, **FedSSI**, and **FedCBDR** examine asynchronous or heterogeneous continual federation, parameter-level retention, and class-wise replay coordination (Shenaj et al., 2023; Li et al., 2025; Qi et al., 2025). FedSSI state is parameter-level rather than a per-class replay summary.

Text comparisons additionally include **FedSeIT**, and graph comparisons include **POWER** and **MOTION** (Chaudhary et al., 2022; Zhu et al., 2025; Wan et al., 2025). Results apply to the evaluated adaptations under the common global-head protocol. Identical data and schedules control important

factors but do not make a comparison a single-component intervention. Baseline source revisions, adaptation patches, generator definitions, and complete tuning ranges are required to audit native-method fidelity.

D.6 Metrics, diagnostics, and statistical reporting

FAA, CTA, and AF use the definitions and mask qualifications in Appendix B.7. Five-seed means and standard deviations are reported where supplied. Targeted mean-only controls have no inferred uncertainty, and resource measurements use a single run per listed method. The reporting distinguishes differences of means from paired uncertainty estimates.

Lifecycle, transport, replay-source, initialization, and auxiliary-weight controls are collected in Appendix E. The calibration, neighbor, spread, and proximal sweeps in Appendix G.7 are one-factor-at-a-time comparisons, not a full factorial design or a test of exact theoretical scaling laws. Alignment is cosine similarity to fresh reference centers, not transport mean squared error (MSE). The maximum mean discrepancy (MMD) reporting convention requires the estimator audit in Appendix F.3.

D.7 Reproduction fields still requiring run artifacts

The source specifies the major schedules, optimizer families, replay counts, and principal PRO-MAX settings. Exact independent reproduction additionally requires sample and split identifiers, tokenizer and graph-visibility settings, client ownership and support-repair rules, task and participant schedules, evaluation masks, the seeds and validation ledger, adapter and auxiliary-head dimensions, augmentation settings, baseline commits and patches, and per-run optional-bank flags. Per-seed metric tensors and source-intervention checkpoint manifests are needed to verify derived metrics and paired effects. Message ledgers and memory-profiler definitions are needed for systems claims. These missing fields are not replaced with values borrowed from other methods or inferred from rounded aggregate tables.

E Complete Main-Result Tables and Controls

This section collects the numerical values to which the main text refers. The lifecycle figure shows FAA and AF, while the tables below also retain CTA where available. Main-only targeted controls remain mean-only. Reported seed aggregates do not establish paired significance, and the two source-substitution scopes are kept separate.

The restored tables use the previously supplied manuscript record. Values shared with the latest main and its appendix are retained without alteration. This restoration adds no training run, new uncertainty estimate, or recomputation under a different evaluation mask. The version and source mapping are documented in the accompanying revision audit.

E.1 COMPLETE CIFAR-100 COMPARISON

Table 10: CIFAR-100, mean ± standard deviation over five seeds. Aligned columns use expanded replay where applicable, and heterogeneous columns retain that budget. Both proposed variants keep 20 projected features per old class. Higher FAA and CTA are preferred, while lower AF is preferred. Bold marks the best value and underline the runner-up within each metric column.

Method	Aligned FAA	Aligned AF	Het. FAA	Het. CTA	Het. AF
FedAvg	10.84 ± 0.36	79.86 ± 0.81	7.86 ± 0.42	48.64 ± 1.58	82.74 ± 1.36
FedProx	11.27 ± 0.42	78.92 ± 0.77	8.92 ± 0.47	49.86 ± 1.51	80.96 ± 1.28
FedProto	20.65 ± 0.88	62.40 ± 1.42	18.74 ± 0.96	52.38 ± 1.43	66.82 ± 1.74
TARGET	43.18 ± 0.96	30.66 ± 1.28	27.86 ± 1.74	53.42 ± 1.86	57.84 ± 2.18
MFCL	44.02 ± 1.05	29.12 ± 1.22	29.42 ± 1.68	54.76 ± 1.79	54.63 ± 2.04
HR	45.34 ± 0.92	26.38 ± 1.13	32.18 ± 1.52	56.92 ± 1.64	49.76 ± 1.87
FedSpace	39.28 ± 1.08	35.12 ± 1.38	33.84 ± 1.39	55.18 ± 1.52	47.92 ± 1.76
FedSSI	40.58 ± 1.02	34.18 ± 1.32	34.62 ± 1.31	55.74 ± 1.44	46.38 ± 1.69
FedCBDR	$\underline{45.76 \pm 0.86}$	$\underline{25.84 \pm 1.05}$	38.96 ± 1.18	59.36 ± 1.27	39.72 ± 1.48
PRO	43.86 ± 0.82	27.36 ± 1.08	$\underline{43.28 \pm 0.94}$	$\underline{63.72 \pm 1.03}$	$\underline{31.54 \pm 1.18}$
PRO-MAX	$\mathbf{46.72 \pm 0.78}$	$\mathbf{23.64 \pm 0.97}$	$\mathbf{46.12 \pm 0.88}$	$\mathbf{65.06 \pm 0.96}$	$\mathbf{26.18 \pm 1.07}$

E.2 SEQUENTIAL REHEARSAL CONFIGURATIONS

Table 11: Sequential rehearsal controls on heterogeneous CIFAR-100. FedProto with balanced rehearsal (FedProto+BR) combines balanced sampling with pre-adapter mean rehearsal. PRO-Point adds global refresh, PRO adds spread, and PRO-MAX adds transport. Bold marks the best value and underline the runner-up within each metric column.

Configuration	FAA	CTA	AF
FedProto	18.74 ± 0.96	52.38 ± 1.43	66.82 ± 1.74
FedProto+BR	39.84 ± 1.18	61.92 ± 1.15	38.34 ± 1.42
PRO-Point	42.96 ± 1.01	63.28 ± 1.06	32.10 ± 1.25
PRO	$\underline{43.28 \pm 0.94}$	$\underline{63.72 \pm 1.03}$	$\underline{31.54 \pm 1.18}$
PRO-MAX	$\mathbf{46.12 \pm 0.88}$	$\mathbf{65.06 \pm 0.96}$	$\mathbf{26.18 \pm 1.07}$

E.3 TARGETED REHEARSAL CONTROLS

Table 12: Targeted rehearsal controls on heterogeneous CIFAR-100. The plot in Figure 3 reports their means, while this table provides the corresponding five-seed standard deviations. Higher FAA and lower AF are preferred. Rank styling is omitted because only two controls are compared.

Control	FAA	AF
Post-adapter replay	38.64 ± 1.10	41.38 ± 1.43
No global refresh	40.86 ± 1.04	36.82 ± 1.31

E.4 TRANSPORT CONTROLS

Table 13: Transport controls on heterogeneous CIFAR-100. Alignment is mean cosine similarity to fresh current-global class centers, not displacement MSE. Higher alignment, FAA, and CTA and lower AF are preferred. The zero-transport branch preserves the additional code path. Bold marks the best value and underline the runner-up within each metric column.

Variant	Alignment	FAA	CTA	AF
PRO	0.73 ± 0.03	43.28 ± 0.94	63.72 ± 1.03	31.54 ± 1.18
Zero transport	0.74 ± 0.03	43.32 ± 0.96	63.73 ± 1.04	31.50 ± 1.20
Shuffled class transport	0.68 ± 0.04	42.86 ± 1.08	63.19 ± 1.11	33.06 ± 1.36
Uniform neighbor weights	$\underline{0.81 \pm 0.03}$	$\underline{44.90 \pm 0.95}$	$\underline{64.43 \pm 1.01}$	$\underline{28.66 \pm 1.16}$
Full PRO-MAX	$\mathbf{0.86 \pm 0.02}$	$\mathbf{46.12 \pm 0.88}$	$\mathbf{65.06 \pm 0.96}$	$\mathbf{26.18 \pm 1.07}$

E.5 WITHIN-PRO REPLAY-SOURCE SUBSTITUTION

Table 14: Within-PRO replay-source intervention. Within each learning condition, the two branches share the starting model and memory. Only the historical source changes. MMD is a distribution diagnostic, not a measure of replay-only causality. Lower MMD and AF and higher FAA and CTA are preferred. Rank styling is omitted because each condition contains only two branches.

Condition	Replay source	MMD	FAA	CTA	AF
Normal	Projected	0.08 ± 0.01	43.28 ± 0.94	63.72 ± 1.03	31.54 ± 1.18
Normal	Fresh features	0.03 ± 0.01	43.40 ± 0.93	63.58 ± 1.01	31.24 ± 1.17
Weak	Projected	0.11 ± 0.02	42.22 ± 1.07	62.18 ± 1.15	33.02 ± 1.32
Weak	Fresh features	0.04 ± 0.01	42.46 ± 1.04	62.02 ± 1.13	32.50 ± 1.29

The projected and fresh branches share a starting model and memory within each support condition. Weak support also changes prior supervision. The table preserves the supplied MMD label. Its exact estimator transformation remains subject to the diagnostic definition audit in Appendix F.3.

E.6 DIRECT CONTRASTS UNDERLYING THE LIFECYCLE INTERPRETATION

The construction, placement, and movement comparisons answer different questions. Global refresh changes the coordinate reference of the supplied evidence, replay placement changes its gradient path, and MOVE maintains means without a completed fresh WRITE. The following contrasts summarize the reported means without treating them as independent factorial effects.

Table 15: Differences of reported means underlying the lifecycle interpretation. Positive FAA gain and AF reduction favor the second configuration. No paired uncertainty estimate is implied.

Comparison	FAA gain (pp)	AF reduction (pp)
FedProto+BR to **PRO-Point**	3.12	6.24
PRO-Point to **PRO**	0.32	0.56
PRO to **PRO-MAX**	2.84	5.36
Post-adapter control to **PRO**	4.64	9.84
No-refresh control to **PRO**	2.42	5.28
Zero transport to **PRO-MAX**	2.80	5.32

E.7 SIX-BENCHMARK NUMERICAL SUMMARY

Table 16: Heterogeneous six-benchmark summary. The comparator has the highest reported mean FAA among the evaluated non-PRO methods on that dataset. Bold marks the best FAA and underline the runner-up among the three listed methods. Differences are percentage-point differences of means, not paired significance estimates.

Dataset	Comparator	Its FAA	**PRO** FAA	**PRO-MAX** FAA	ΔFAA	AF reduction
CIFAR-100	FedCBDR	38.96 ± 1.18	$\underline{43.28 \pm 0.94}$	$\mathbf{46.12 \pm 0.88}$	7.16	13.54
TinyImageNet	FedCBDR	29.46 ± 1.18	$\underline{33.28 \pm 1.04}$	$\mathbf{37.54 \pm 0.98}$	8.08	13.76
THUCNews-10	FedSeIT	73.28 ± 1.18	$\underline{78.42 \pm 1.04}$	$\mathbf{81.62 \pm 0.96}$	8.34	9.44
CLINC150	FedSeIT	59.74 ± 1.14	$\underline{63.92 \pm 1.06}$	$\mathbf{67.84 \pm 0.98}$	8.10	10.92
Cora	MOTION	60.14 ± 4.21	$\underline{63.24 \pm 3.91}$	$\mathbf{66.18 \pm 3.66}$	6.04	10.10
ogbn-arxiv	MOTION	44.38 ± 2.46	$\underline{47.86 \pm 2.32}$	$\mathbf{51.92 \pm 2.18}$	7.54	12.26

E.8 COMPLETE THUCNEWS-10 AND CORA RESULTS

Table 17: Full THUCNews-10 and Cora results under aligned and heterogeneous streams. Predictive values are five-seed means and standard deviations. For FAA, CTA, and AF, bold marks the best value and underline the runner-up within each dataset and regime. The normalized communication columns retain the archived values and require the accounting qualification in Appendix H.

Method	Aligned				Heterogeneous			
	FAA ↑	CTA ↑	AF ↓	Comm. ↓	FAA ↑	CTA ↑	AF ↓	Comm. ↓
THUCNews-10								
FedAvg	68.54 ± 1.28	84.18 ± 1.04	31.82 ± 1.74	1.00×	61.08 ± 1.52	79.76 ± 1.21	41.62 ± 1.96	1.00×
FedProx	69.36 ± 1.21	84.06 ± 0.98	30.64 ± 1.68	1.00×	62.42 ± 1.47	80.13 ± 1.18	40.16 ± 1.87	1.00×
FedProto	72.78 ± 1.12	84.52 ± 0.96	25.92 ± 1.46	1.06×	66.24 ± 1.36	80.48 ± 1.10	34.82 ± 1.66	1.06×
FedSeIT	78.86 ± 0.97	85.42 ± 0.84	18.62 ± 1.24	1.25×	73.28 ± 1.18	82.14 ± 0.98	25.86 ± 1.39	1.25×
PRO	$\underline{82.94 \pm 0.82}$	$\underline{87.24 \pm 0.76}$	$\underline{13.82 \pm 1.02}$	1.10×	$\underline{78.42 \pm 1.04}$	$\underline{84.08 \pm 0.91}$	$\underline{20.84 \pm 1.26}$	1.10×
PRO-MAX	$\mathbf{85.16 \pm 0.74}$	$\mathbf{88.02 \pm 0.71}$	$\mathbf{10.96 \pm 0.88}$	1.14×	$\mathbf{81.62 \pm 0.96}$	$\mathbf{85.22 \pm 0.86}$	$\mathbf{16.42 \pm 1.10}$	1.14×
Cora								
FedAvg	46.18 ± 2.76	58.86 ± 2.94	53.84 ± 4.28	1.00×	39.84 ± 2.98	53.26 ± 3.16	61.82 ± 4.56	1.00×
FedProx	47.06 ± 2.68	59.12 ± 2.88	51.76 ± 4.12	1.00×	41.32 ± 2.87	54.08 ± 3.05	59.74 ± 4.38	1.00×
FedProto	52.46 ± 2.51	61.18 ± 2.72	44.28 ± 3.96	1.08×	46.82 ± 2.74	56.42 ± 2.93	51.86 ± 4.11	1.08×
POWER	65.84 ± 4.94	70.86 ± 3.38	28.64 ± 4.72	1.42×	58.76 ± 4.38	64.36 ± 3.72	38.82 ± 4.46	1.42×
MOTION	66.42 ± 4.72	71.28 ± 3.44	27.86 ± 4.51	1.36×	60.14 ± 4.21	65.08 ± 3.58	35.94 ± 4.29	1.36×
PRO	$\underline{68.86 \pm 4.06}$	$\underline{72.82 \pm 3.06}$	$\underline{23.84 \pm 4.06}$	1.12×	$\underline{63.24 \pm 3.91}$	$\underline{66.42 \pm 3.32}$	$\underline{30.86 \pm 3.94}$	1.12×
PRO-MAX	$\mathbf{70.42 \pm 3.82}$	$\mathbf{73.64 \pm 2.88}$	$\mathbf{20.96 \pm 3.74}$	1.16×	$\mathbf{66.18 \pm 3.66}$	$\mathbf{68.04 \pm 3.14}$	$\mathbf{25.84 \pm 3.62}$	1.16×

E.9 AUXILIARY-LOSS SENSITIVITY

Table 18: Additional PRO-MAX auxiliary-weight sensitivity on heterogeneous CIFAR-100, with $\lambda_{\text{prox}} = 10^{-2}$. Higher FAA and CTA and lower AF are preferred. Best values are bold, and runners-up are underlined.

λ_{aux}	FAA	CTA	AF
0	45.62 ± 0.98	64.41 ± 1.04	26.86 ± 1.16
0.01	$\underline{45.94 \pm 0.92}$	64.78 ± 1.00	$\underline{26.42 \pm 1.11}$
0.05	$\mathbf{46.12 \pm 0.88}$	$\underline{65.06 \pm 0.96}$	$\mathbf{26.18 \pm 1.07}$
0.10	45.83 ± 0.96	$\mathbf{65.19 \pm 1.01}$	26.71 ± 1.15

E.10 INITIALIZATION ACCESS CONTROL

Table 19: Shared transductive initialization versus client-local position-0-only initialization. A locally restricted pool is not necessarily globally class-unseen. Higher FAA and CTA and lower AF are preferred. Best values are bold, and runners-up are underlined across the listed configurations.

Method	Unlabeled pool	FAA	CTA	AF
FedCBDR	Full local pool	38.96 ± 1.18	59.36 ± 1.27	39.72 ± 1.48
FedCBDR	Position 0 only	38.10 ± 1.24	58.80 ± 1.31	40.45 ± 1.55
PRO	Full local pool	43.28 ± 0.94	63.72 ± 1.03	31.54 ± 1.18
PRO	Position 0 only	42.70 ± 1.02	63.05 ± 1.08	32.25 ± 1.24
PRO-MAX	Full local pool	$\mathbf{46.12 \pm 0.88}$	$\mathbf{65.06 \pm 0.96}$	$\mathbf{26.18 \pm 1.07}$
PRO-MAX	Position 0 only	$\underline{45.35 \pm 0.96}$	$\underline{64.35 \pm 1.02}$	$\underline{27.05 \pm 1.15}$

Position-0-only initialization restricts local input access but does not remove warmup and does not ensure globally future-class-excluded initialization when client orders differ.

E.11 SEPARATE RETAINED-INPUT REFERENCE

Table 20: Separately reported retained-input fresh-feature reference. This is a different intervention scope from Table 14. The two reference results are not pooled.

Historical source	FAA	CTA	AF
Projected memory	43.28 ± 0.94	63.72 ± 1.03	31.54 ± 1.18
Fresh retained-input features	45.10 ± 0.99	61.15 ± 1.12	24.90 ± 1.29

This retained-input reference is separate from the checkpoint-fork comparison in Table 14. Its FAA of 45.10 is not substituted for the 43.40 normal fresh-source result, because the supplied record assigns them different intervention scopes. The precise configuration differences require the corresponding run manifests.

E.12 INCREMENTAL CIFAR-100 RESOURCE SUMMARY

The resource tables use CPU, GPU, and random-access memory (RAM) as defined in this appendix.

Table 21: Incremental CIFAR-100 simulation cost, extracted from the same method rows as Table 37. One run is reported per method. Lower values are preferred. Best values are bold, and runners-up are underlined. Runtime excludes the separately reported shared initialization, and peak-memory conventions are discussed in Appendix H.

Method	Incremental time (h)	Peak GPU (GB)	Peak CPU RAM (GB)
FedAvg	**2.10**	**4.6**	**10.2**
TARGET	4.80	8.2	14.0
MFCL	5.40	8.6	15.2
HR	5.80	9.0	16.1
FedCBDR	4.20	7.4	14.3
PRO	<u>3.20</u>	<u>5.6</u>	<u>12.0</u>
PRO-MAX	3.60	6.2	13.1

F ADDITIONAL WEAK-TASK DIAGNOSTICS AND THEIR SCOPE

The main experiments test retention deterioration and the usefulness of the rehearsal lifecycle. The additional weak-task records examine associations between the learned state, replay mismatch, and later aggregate performance. They do not establish a generator-specific failure law or identify replay mismatch as the sole mediator of forgetting.

F.1 Intervention and source-substitution distinction

The recorded weak-task diagnostic selects semantic task $t^\star = 4$. Its general description lists changes to supporting-client count, examples, or training budget, but the supplied aggregates do not identify every executed intervention strength and seed. In contrast, the main within-PRO source comparison forks a common checkpoint within each support condition and changes projected versus fresh historical features. The weak support condition reduces supporting clients while retaining examples per supporting client, so total supervision changes. These interventions must not be pooled or described as changing only a generator inside every baseline.

A weak task can influence later learning through its model state, representation, and replay source. Even a within-method correlation between mismatch and later performance leaves those paths entangled when the intervention changes prior learning. The causal scope of the main source substitution is limited to that within-PRO branch comparison.

F.2 Replay-real comparison space

For input replay, an evaluation encoder can map real and generated inputs to $r = \varphi(x)$ and $\tilde{r} = \varphi(\tilde{x})$. A projected feature u is not an input to φ merely because it has a feature representation. Cross-method comparison with PRO therefore requires either an explicitly common compatible feature space or a documented map $B_m : u_m \mapsto r$. The supplied prose asserts a shared evaluation space but does not specify this bridge for every method. The diagnostic table below retains the reported values without treating that missing bridge as established.

F.3 Mismatch definitions and event-time limits

For real features $\{r_i\}_{i=1}^N$ and replay features $\{\tilde{r}_j\}_{j=1}^M$, the supplied radial basis function (RBF) kernel expression is the unbiased squared-MMD estimator

$$\widehat{\mathrm{MMD}}_u^2 = \frac{\sum_{i \neq i'} k(r_i, r_{i'})}{N(N-1)} + \frac{\sum_{j \neq j'} k(\tilde{r}_j, \tilde{r}_{j'})}{M(M-1)} - \frac{2}{NM} \sum_{i,j} k(r_i, \tilde{r}_j). \tag{39}$$

The tables label their values "MMD". The source does not determine whether those recorded scalars are squared estimates, square roots after clipping, or another aggregation of the estimator. Kernel bandwidth, feature normalization, sample counts, estimator transformation, and checkpoint aggregation must be recovered from code before giving a more specific metric label. No numeric conversion is applied here. In particular, finite-sample unbiased squared-MMD estimates need not be nonnegative.

The additional reported diagnostics are empirical W_1 in the selected representation and cosine distance $1 - \langle \mu^{\mathrm{real}}, \mu^{\mathrm{rep}} \rangle / (\|\mu^{\mathrm{real}}\|_2 \|\mu^{\mathrm{rep}}\|_2)$. Transport solver settings, class aggregation, and zero-norm handling are not fully specified in the source. These diagnostics are not interchangeable with each other or with the feature-risk bounds.

The archived downstream statistic uses

$$\Delta_{\mathrm{down}} = \frac{1}{T - t^\star} \sum_{t > t^\star} \left(A_{m,t}^{\mathrm{clean}} - A_{m,t}^{\mathrm{weak}} \right). \tag{40}$$

With different client permutations, $t > t^\star$ means a larger semantic index, not necessarily later exposure. The value is therefore a semantic-index contrast unless the recorded schedule establishes the temporal interpretation. Likewise, "immediately after task $t^\star$" requires a specified checkpoint when that task appears at different positions across clients. The present aggregates cannot reconstruct that event alignment.

F.4 Retained weak-task aggregates

Table 22: Retained weak-task aggregates for the recorded semantic task. The MMD transformation, common feature-space bridge, and event-time interpretation require the audits described in the text. The semantic-index statistic is not a causal mediation estimate. Bold marks the best value and underline the runner-up within each metric column.

Method	$A_{t^\star}^{\mathrm{post}} \uparrow$	MMD $\downarrow$	$W_1 \downarrow$	Cos. Dist. $\downarrow$	$\Delta_{\mathrm{down}} \downarrow$	Final FAA $\uparrow$	Final AF $\downarrow$
TARGET	36.48 ± 1.84	0.19 ± 0.02	2.74 ± 0.21	0.32 ± 0.02	13.42 ± 1.06	27.86 ± 1.74	57.84 ± 2.18
MFCL	37.92 ± 1.76	0.17 ± 0.01	2.53 ± 0.19	0.30 ± 0.02	11.86 ± 0.98	29.42 ± 1.68	54.63 ± 2.04
HR	41.26 ± 1.62	0.15 ± 0.01	2.16 ± 0.17	0.25 ± 0.02	9.74 ± 0.92	32.18 ± 1.52	49.76 ± 1.87
FedCBDR	47.38 ± 1.48	0.12 ± 0.01	1.74 ± 0.14	0.21 ± 0.02	6.42 ± 0.74	38.96 ± 1.18	39.72 ± 1.48
PRO	$\underline{52.84 \pm 1.26}$	$\underline{0.08 \pm 0.01}$	$\underline{1.18 \pm 0.10}$	$\underline{0.15 \pm 0.01}$	$\underline{3.18 \pm 0.42}$	$\underline{43.28 \pm 0.94}$	$\underline{31.54 \pm 1.18}$
PRO-MAX	$\mathbf{55.62 \pm 1.18}$	$\mathbf{0.06 \pm 0.01}$	$\mathbf{0.96 \pm 0.08}$	$\mathbf{0.12 \pm 0.01}$	$\mathbf{1.94 \pm 0.31}$	$\mathbf{46.12 \pm 0.88}$	$\mathbf{26.18 \pm 1.07}$

Within these records, lower mismatch and smaller semantic-index degradation accompany higher final utility for the proposed methods. This association is compatible with useful memory maintenance, but also depends on prior representation and task learning. It does not isolate a generator mechanism, quantify a mediation effect, or prove a particular temporal propagation path.

F.5 Diagnostic figures and required provenance

The supplied source references a final-task heatmap, a real-versus-replay visualization, per-seed mismatch plots, and memory-evolution snapshots. Those visualizations need the task-accuracy arrays, feature samples, projection settings, seed identifiers, intervention metadata, and plot-generation code. They are retained as opt-in diagnostic figure slots rather than regenerated from the final means. Their default exclusion from this revision does not classify the original plots as fabricated. It prevents unverified point-level or temporal evidence from being silently treated as validated by this manuscript reconciliation.

G Additional Results and Diagnostics

The complete main-result tables are in Appendix E. Here we retain the larger benchmarks, full sampler sweeps, current-task summaries, alignment diagnostics, and sensitivity grids. These results use the supplied within-protocol comparisons rather than absolute comparisons to papers with different backbones or access assumptions. Normalized communication entries are archived accounting records, subject to Appendix H.

G.1 TinyImageNet scale-up

TinyImageNet extends the visual sequence to twenty tasks. The table uses the same listed method families and reports both aligned and heterogeneous configurations.

Table 23: TinyImageNet scale-up results under aligned and heterogeneous FCIL. We report FAA, CTA, AF, and normalized cumulative communication. For predictive metrics, bold marks the best value and underline the runner-up within each regime.

Method	Aligned				Heterogeneous			
	FAA $\uparrow$	CTA $\uparrow$	AF $\downarrow$	Comm. $\downarrow$	FAA $\uparrow$	CTA $\uparrow$	AF $\downarrow$	Comm. $\downarrow$
FedAvg	6.28 ± 0.52	42.72 ± 1.88	86.14 ± 1.42	$1.00\times$	3.72 ± 0.48	32.18 ± 2.12	90.12 ± 1.76	$1.00\times$
FedProx	7.36 ± 0.61	43.51 ± 1.79	84.27 ± 1.50	$1.00\times$	4.86 ± 0.54	34.62 ± 2.04	88.64 ± 1.68	$1.00\times$
FedProto	14.92 ± 0.88	45.03 ± 1.62	71.36 ± 1.94	$1.08\times$	11.48 ± 0.93	37.84 ± 1.96	76.72 ± 2.18	$1.08\times$
TARGET	26.84 ± 1.16	51.12 ± 1.42	50.68 ± 1.86	$2.80\times$	18.96 ± 1.62	40.76 ± 2.18	66.48 ± 2.36	$2.80\times$
MFCL	28.73 ± 1.24	52.08 ± 1.36	47.91 ± 1.74	$3.00\times$	20.42 ± 1.54	42.33 ± 2.06	63.84 ± 2.28	$3.00\times$
HR	31.25 ± 1.13	53.94 ± 1.31	44.36 ± 1.66	$3.20\times$	23.16 ± 1.46	44.25 ± 1.92	58.92 ± 2.16	$3.20\times$
FedSpace	30.72 ± 1.18	51.97 ± 1.38	46.28 ± 1.72	$2.25\times$	24.88 ± 1.35	43.76 ± 1.86	56.31 ± 2.05	$2.25\times$
FedSSI	31.94 ± 1.14	52.56 ± 1.34	44.91 ± 1.68	$2.15\times$	25.64 ± 1.28	44.18 ± 1.78	54.86 ± 1.98	$2.15\times$
FedCBDR	34.86 ± 1.02	55.84 ± 1.20	40.18 ± 1.42	$2.80\times$	29.46 ± 1.18	48.37 ± 1.62	49.72 ± 1.84	$2.80\times$
PRO	$\underline{37.42 \pm 0.96}$	$\underline{57.16 \pm 1.08}$	$\underline{35.72 \pm 1.26}$	$1.14\times$	$\underline{33.28 \pm 1.04}$	$\underline{51.96 \pm 1.42}$	$\underline{42.84 \pm 1.60}$	$1.14\times$
PRO-MAX	$\mathbf{40.36 \pm 0.91}$	$\mathbf{58.74 \pm 1.02}$	$\mathbf{30.84 \pm 1.18}$	$1.19\times$	$\mathbf{37.54 \pm 0.98}$	$\mathbf{54.18 \pm 1.34}$	$\mathbf{35.96 \pm 1.46}$	$1.19\times$

PRO-MAX reaches 37.54 FAA and 35.96 AF in the heterogeneous setting, compared with 33.28 and 42.84 for PRO. The difference supports the extension on this evaluated stream. Since the dataset and stream length both change relative to CIFAR-100, this comparison does not isolate scale as a causal source of difficulty or replay quality as a mediator.

G.2 CLINC150 SCALE-UP

The CLINC150 experiment tests the memory interface on a longer intent-classification sequence with text-compatible comparisons.

Table 24: CLINC150 scale-up results under aligned and heterogeneous FCIL. We report FAA, CTA, AF, and normalized cumulative communication. For predictive metrics, bold marks the best value and underline the runner-up within each regime.

Method	Aligned				Heterogeneous			
	FAA ↑	CTA ↑	AF ↓	Comm. ↓	FAA ↑	CTA ↑	AF ↓	Comm. ↓
FedAvg	47.36 ± 1.42	71.22 ± 1.36	54.85 ± 2.16	1.00×	39.76 ± 1.58	65.48 ± 1.52	63.16 ± 2.34	1.00×
FedProx	49.02 ± 1.36	71.84 ± 1.28	52.14 ± 2.04	1.00×	41.38 ± 1.49	66.12 ± 1.44	60.94 ± 2.20	1.00×
FedProto	55.48 ± 1.24	73.15 ± 1.22	43.68 ± 1.82	1.06×	48.72 ± 1.38	68.84 ± 1.36	50.68 ± 1.96	1.06×
FedSeIT	65.42 ± 1.02	78.16 ± 0.96	30.74 ± 1.40	1.25×	59.74 ± 1.14	74.08 ± 1.08	36.84 ± 1.54	1.25×
PRO	$\underline{68.92 \pm 0.96}$	$\underline{80.36 \pm 0.91}$	$\underline{25.42 \pm 1.26}$	1.10×	$\underline{63.92 \pm 1.06}$	$\underline{76.86 \pm 0.98}$	$\underline{31.24 \pm 1.38}$	1.10×
PRO-MAX	$\mathbf{72.36 \pm 0.88}$	$\mathbf{81.72 \pm 0.86}$	$\mathbf{20.18 \pm 1.12}$	1.14×	$\mathbf{67.84 \pm 0.98}$	$\mathbf{78.48 \pm 0.94}$	$\mathbf{25.92 \pm 1.26}$	1.14×

PRO-MAX improves the reported heterogeneous FAA and AF over PRO and the evaluated FedSeIT comparison. This supports the interface on the chosen TextCNN stream, not arbitrary text models or instruction-following tasks.

G.3 OGBN-ARXIV SCALE-UP

The ogbn-arxiv experiment tests node-classification streams with a larger label set and graph than the Cora comparison.

Table 25: ogbn-arxiv scale-up results under aligned and heterogeneous FCIL. We report FAA, CTA, AF, and normalized cumulative communication. For predictive metrics, bold marks the best value and underline the runner-up within each regime.

Method	Aligned				Heterogeneous			
	FAA ↑	CTA ↑	AF ↓	Comm. ↓	FAA ↑	CTA ↑	AF ↓	Comm. ↓
FedAvg	34.52 ± 2.28	51.84 ± 2.46	63.62 ± 3.62	1.00×	28.44 ± 2.46	46.76 ± 2.68	70.42 ± 3.88	1.00×
FedProx	35.96 ± 2.16	52.38 ± 2.38	61.74 ± 3.48	1.00×	30.16 ± 2.32	48.12 ± 2.54	68.18 ± 3.70	1.00×
FedProto	40.84 ± 2.04	55.42 ± 2.26	53.28 ± 3.28	1.08×	35.72 ± 2.18	50.84 ± 2.42	60.36 ± 3.46	1.08×
POWER	49.28 ± 2.36	61.74 ± 2.18	39.86 ± 3.12	1.42×	42.78 ± 2.54	56.48 ± 2.36	47.92 ± 3.34	1.42×
MOTION	50.76 ± 2.28	62.18 ± 2.10	37.92 ± 3.04	1.36×	44.38 ± 2.46	57.86 ± 2.24	45.12 ± 3.18	1.36×
PRO	$\underline{52.48 \pm 2.18}$	$\underline{64.32 \pm 2.02}$	$\underline{34.18 \pm 2.86}$	1.12×	$\underline{47.86 \pm 2.32}$	$\underline{59.24 \pm 2.18}$	$\underline{39.42 \pm 3.04}$	1.12×
PRO-MAX	$\mathbf{55.86 \pm 2.04}$	$\mathbf{66.24 \pm 1.94}$	$\mathbf{28.64 \pm 2.64}$	1.16×	$\mathbf{51.92 \pm 2.18}$	$\mathbf{61.78 \pm 2.06}$	$\mathbf{32.86 \pm 2.82}$	1.16×

PRO-MAX has the highest reported mean FAA and lowest AF among these evaluated methods. Node-level class statistics do not preserve all graph topology, and these aggregate comparisons do not identify graph-induced representation drift as the measured cause of the improvement.

G.4 COMPLETE SUPPORT AND ORDER SWEEPS

The main table selects endpoints for readability. These tables retain all four supplied levels and their standard deviations. Support changes the number of clients carrying each task while coverage and order remain fixed. Order changes permutations while support and coverage remain sparse.

Table 26: CIFAR-100 supervision-support sweep. Support denotes the fraction of clients that support each semantic task. Bold marks the best value and underline the runner-up within each metric column.

Method	Support 80%		Support 60%		Support 40%		Support 20%	
	FAA ↑	AF ↓	FAA ↑	AF ↓	FAA ↑	AF ↓	FAA ↑	AF ↓
FedAvg	9.62 ± 0.44	80.42 ± 1.26	8.71 ± 0.46	81.38 ± 1.31	7.86 ± 0.42	82.74 ± 1.36	5.92 ± 0.50	86.18 ± 1.54
FedProx	10.47 ± 0.48	78.86 ± 1.18	9.62 ± 0.50	79.72 ± 1.24	8.92 ± 0.47	80.96 ± 1.28	6.84 ± 0.54	84.12 ± 1.46
FedProto	21.36 ± 0.92	63.42 ± 1.58	19.86 ± 0.94	65.18 ± 1.66	18.74 ± 0.96	66.82 ± 1.74	14.92 ± 1.12	72.64 ± 2.08
TARGET	40.12 ± 1.22	39.26 ± 1.44	35.74 ± 1.42	47.62 ± 1.68	27.86 ± 1.74	57.84 ± 2.18	24.72 ± 1.84	63.18 ± 2.44
MFCL	41.06 ± 1.18	37.92 ± 1.38	37.12 ± 1.35	44.76 ± 1.60	29.42 ± 1.68	54.63 ± 2.04	25.98 ± 1.78	60.46 ± 2.32
HR	42.84 ± 1.10	34.78 ± 1.32	39.36 ± 1.24	40.18 ± 1.52	32.18 ± 1.52	49.76 ± 1.87	28.84 ± 1.66	55.84 ± 2.18
FedSpace	40.12 ± 1.15	34.96 ± 1.38	37.42 ± 1.24	40.18 ± 1.52	33.84 ± 1.39	47.92 ± 1.76	29.16 ± 1.58	56.84 ± 2.06
FedSSI	41.28 ± 1.10	33.84 ± 1.34	38.46 ± 1.18	39.12 ± 1.48	34.62 ± 1.31	46.38 ± 1.69	30.48 ± 1.49	54.72 ± 1.98
FedCBDR	43.84 ± 1.03	31.48 ± 1.25	41.26 ± 1.12	35.72 ± 1.36	38.96 ± 1.18	39.72 ± 1.48	34.42 ± 1.35	48.64 ± 1.72
PRO	$\underline{44.26 \pm 0.98}$	$\underline{29.86 \pm 1.18}$	$\underline{43.92 \pm 0.96}$	$\underline{30.74 \pm 1.15}$	$\underline{43.28 \pm 0.94}$	$\underline{31.54 \pm 1.18}$	$\underline{39.86 \pm 1.16}$	$\underline{38.76 \pm 1.52}$
PRO-MAX	$\mathbf{47.12 \pm 0.90}$	$\mathbf{23.84 \pm 1.04}$	$\mathbf{46.74 \pm 0.88}$	$\mathbf{24.72 \pm 1.02}$	$\mathbf{46.12 \pm 0.88}$	$\mathbf{26.18 \pm 1.07}$	$\mathbf{42.68 \pm 1.08}$	$\mathbf{32.42 \pm 1.34}$

All listed methods have lower mean FAA as support decreases across the reported grid. PRO and PRO-MAX lose less FAA than the selected replay comparator at the displayed endpoints. Because total task evidence also decreases, the pattern does not isolate redistribution heterogeneity or a generator-specific mechanism.

Table 27: CIFAR-100 order-perturbation sweep. Perturbation denotes the fraction of task positions changed per client. All columns in this sweep keep task support at 40% and per-client label coverage at 50%. Only task-order perturbation is varied. Thus, the 0% perturbation column is not the aligned setting because supervision remains sparse. Bold marks the best value and underline the runner-up within each metric column.

Method	Perturbation 0%		Perturbation 30%		Perturbation 60%		Perturbation 90%	
	FAA ↑	AF ↓	FAA ↑	AF ↓	FAA ↑	AF ↓	FAA ↑	AF ↓
FedAvg	10.12 ± 0.40	80.52 ± 1.10	9.76 ± 0.41	80.94 ± 1.02	7.86 ± 0.42	82.74 ± 1.36	6.58 ± 0.49	85.62 ± 1.58
FedProx	10.74 ± 0.43	79.32 ± 1.05	10.18 ± 0.46	79.84 ± 0.96	8.92 ± 0.47	80.96 ± 1.28	7.14 ± 0.52	84.18 ± 1.46
FedProto	20.12 ± 0.90	63.48 ± 1.50	19.86 ± 0.92	64.28 ± 1.58	18.74 ± 0.96	66.82 ± 1.74	15.48 ± 1.08	71.84 ± 2.04
TARGET	39.42 ± 1.18	39.86 ± 1.46	36.84 ± 1.36	43.72 ± 1.64	27.86 ± 1.74	57.84 ± 2.18	23.46 ± 1.92	65.28 ± 2.46
MFCL	40.18 ± 1.15	38.42 ± 1.40	38.92 ± 1.28	40.84 ± 1.56	29.42 ± 1.68	54.63 ± 2.04	24.94 ± 1.82	62.12 ± 2.34
HR	41.72 ± 1.06	34.58 ± 1.32	40.68 ± 1.12	36.28 ± 1.42	32.18 ± 1.52	49.76 ± 1.87	27.62 ± 1.66	57.48 ± 2.16
FedSpace	37.84 ± 1.12	38.96 ± 1.44	36.42 ± 1.22	41.24 ± 1.56	33.84 ± 1.39	47.92 ± 1.76	30.68 ± 1.54	54.74 ± 2.04
FedSSI	39.12 ± 1.08	37.86 ± 1.40	37.68 ± 1.16	40.46 ± 1.50	34.62 ± 1.31	46.38 ± 1.69	31.24 ± 1.48	53.36 ± 1.96
FedCBDR	43.20 ± 1.00	30.96 ± 1.20	42.68 ± 1.02	33.86 ± 1.34	38.96 ± 1.18	39.72 ± 1.48	35.18 ± 1.40	46.92 ± 1.80
PRO	$\underline{43.76 \pm 0.88}$	$\underline{28.64 \pm 1.10}$	$\underline{43.64 \pm 0.90}$	$\underline{29.74 \pm 1.12}$	$\underline{43.28 \pm 0.94}$	$\underline{31.54 \pm 1.18}$	$\underline{40.84 \pm 1.10}$	$\underline{36.58 \pm 1.44}$
PRO-MAX	$\mathbf{46.62 \pm 0.80}$	$\mathbf{24.22 \pm 0.98}$	$\mathbf{46.48 \pm 0.84}$	$\mathbf{24.92 \pm 1.02}$	$\mathbf{46.12 \pm 0.88}$	$\mathbf{26.18 \pm 1.07}$	$\mathbf{43.96 \pm 1.02}$	$\mathbf{31.24 \pm 1.30}$

The zero-perturbation column still uses 40% task support and 50% local label coverage. Reported FAA decreases and AF increases across these order settings, but perturbation percentages are not measured ACE. The result is sensitivity to this sampler intervention rather than a universal monotonicity law for temporal disagreement.

G.5 CTA SUMMARIES

These tables report stream-averaged CTA, not per-round accuracy trajectories. They complement FAA and AF by checking that retention is not accompanied by lower current-task performance in the evaluated comparisons. CTA uses the global model and client-relative schedule in Appendix B.7.

Table 28: CTA for image benchmarks under severe heterogeneity. Bold marks the best value and underline the runner-up within each dataset.

Method	CIFAR-100	TinyImageNet
FedAvg	48.64 ± 1.58	32.18 ± 2.12
FedProx	49.86 ± 1.51	34.62 ± 2.04
FedProto	52.38 ± 1.43	37.84 ± 1.96
TARGET	53.42 ± 1.86	40.76 ± 2.18
MFCL	54.76 ± 1.79	42.33 ± 2.06
HR	56.92 ± 1.64	44.25 ± 1.92
FedSpace	55.18 ± 1.52	43.76 ± 1.86
FedSSI	55.74 ± 1.44	44.18 ± 1.78
FedCBDR	59.36 ± 1.27	48.37 ± 1.62
PRO	$\underline{63.72 \pm 1.03}$	$\underline{51.96 \pm 1.42}$
PRO-MAX	$\mathbf{65.06 \pm 0.96}$	$\mathbf{54.18 \pm 1.34}$

Table 29: CTA for text benchmarks under severe heterogeneity. Bold marks the best value and underline the runner-up within each dataset.

Method	THUCNews-10	CLINC150
FedAvg	79.76 ± 1.21	65.48 ± 1.52
FedProx	80.13 ± 1.18	66.12 ± 1.44
FedProto	80.48 ± 1.10	68.84 ± 1.36
FedSeIT	82.14 ± 0.98	74.08 ± 1.08
PRO	$\underline{84.08 \pm 0.91}$	$\underline{76.86 \pm 0.98}$
PRO-MAX	$\mathbf{85.22 \pm 0.86}$	$\mathbf{78.48 \pm 0.94}$

Table 30: CTA for graph benchmarks under severe heterogeneity. Bold marks the best value and underline the runner-up within each dataset.

Method	Cora	ogbn-arxiv
FedAvg	53.26 ± 3.16	46.76 ± 2.68
FedProx	54.08 ± 3.05	48.12 ± 2.54
FedProto	56.42 ± 2.93	50.84 ± 2.42
POWER	64.36 ± 3.72	56.48 ± 2.36
MOTION	65.08 ± 3.58	57.86 ± 2.24
PRO	$\underline{66.42 \pm 3.32}$	$\underline{59.24 \pm 2.18}$
PRO-MAX	$\mathbf{68.04 \pm 3.14}$	$\mathbf{61.78 \pm 2.06}$

Across the listed modality tables, PRO-MAX has higher reported mean CTA than PRO and the included baselines. This supports joint current-task performance and retention, but does not supply event-level plasticity curves or client-specific causal attribution.

G.6 Memory-center alignment

The recorded alignment statistic compares stored or moved historical means with fresh current-global reference means computed for evaluation. These oracle centers are diagnostic references, not training inputs supplied by WRITE. For the evaluated historical class set,

$$\mathrm{Align}^u = \frac{1}{|\mathcal{Y}^{\mathrm{old}}|} \sum_{y \in \mathcal{Y}^{\mathrm{old}}} \frac{\langle \tilde{\mu}_y^u, \mu_{y,\mathrm{oracle}}^u \rangle}{\|\tilde{\mu}_y^u\|_2 \|\mu_{y,\mathrm{oracle}}^u\|_2}. \tag{41}$$

The source must additionally specify reference data, checkpoint aggregation, class selection, and zero-norm handling for exact replication. Optional adapted-space values are retained without assuming the optional bank was enabled in every main accuracy or resource run.

Table 31: Memory alignment diagnostic for PRO and PRO-MAX. Higher alignment indicates that old projected memories remain more compatible with the current representation. Rank styling is omitted because each dataset contains only two methods.

Dataset	Method	Align^u ↑	Align^z ↑	AF ↓
CIFAR-100	**PRO**	0.73 ± 0.03	0.76 ± 0.03	31.54 ± 1.18
CIFAR-100	**PRO-MAX**	0.86 ± 0.02	0.88 ± 0.02	26.18 ± 1.07
THUCNews-10	**PRO**	0.82 ± 0.02	0.84 ± 0.02	20.84 ± 1.26
THUCNews-10	**PRO-MAX**	0.91 ± 0.01	0.92 ± 0.01	16.42 ± 1.10
Cora	**PRO**	0.68 ± 0.04	0.70 ± 0.04	30.86 ± 3.94
Cora	**PRO-MAX**	0.79 ± 0.03	0.81 ± 0.03	25.84 ± 3.62

The reported PRO-MAX means have higher alignment and lower AF than PRO on these datasets. Cosine center agreement neither measures the displacement error in the transport bound nor establishes that alignment alone determines retention. The main transport controls add a more targeted comparison of zero, shuffled, and uniform-weight movement.

G.7 Hyperparameter sensitivity

The following one-factor-at-a-time grids use the reported heterogeneous CIFAR-100 PRO-MAX configuration. They vary projected spread, neighborhood size, proximal regularization, and calibration count. The auxiliary-weight grid and initialization restriction are in Tables 18 and 19.

Table 32: Sensitivity to projected-memory noise scale γ on CIFAR-100 severe heterogeneity. Higher FAA and lower AF are preferred. Best values are bold, and runners-up are underlined.

γ	FAA ↑	AF ↓
0.00	43.46 ± 1.02	31.28 ± 1.36
0.05	45.18 ± 0.94	27.84 ± 1.18
0.10	$\mathbf{46.12 \pm 0.88}$	$\mathbf{26.18 \pm 1.07}$
0.20	$\underline{45.76 \pm 0.91}$	$\underline{26.92 \pm 1.12}$
0.40	42.68 ± 1.16	33.46 ± 1.54

The moderate spread settings have higher reported FAA than the endpoints of this grid. The comparison shows sensitivity to sampling spread. It does not directly measure pseudo-feature corruption or independent component effects.

Table 33: Sensitivity to the number of neighbors K in PRO-MAX on CIFAR-100 severe heterogeneity. Higher FAA and lower AF are preferred. Best values are bold, and runners-up are underlined.

K	FAA ↑	AF ↓
1	43.92 ± 1.10	30.86 ± 1.42
3	45.38 ± 0.96	27.42 ± 1.20
5	$\mathbf{46.12 \pm 0.88}$	$\mathbf{26.18 \pm 1.07}$
10	$\underline{45.94 \pm 0.92}$	$\underline{26.74 \pm 1.14}$
20	44.26 ± 1.04	29.62 ± 1.33

The reported means are higher for intermediate neighbor counts than for 1 or 20. Locality and averaging noise provide possible interpretations under the analysis assumptions, but the grid does not estimate either error term separately.

Table 34: Sensitivity to the proximal coefficient λ_{prox} on CIFAR-100 severe heterogeneity. Higher FAA and lower AF are preferred. Best values are bold, and runners-up are underlined.

λ_{prox}	FAA ↑	AF ↓
0	39.72 ± 1.15	39.16 ± 1.41
10^{-4}	43.84 ± 1.04	31.62 ± 1.28
10^{-3}	$\underline{45.36 \pm 0.96}$	$\underline{28.74 \pm 1.18}$
10^{-2}	$\mathbf{46.12 \pm 0.88}$	$\mathbf{26.18 \pm 1.07}$
10^{-1}	44.18 ± 1.02	30.46 ± 1.31

Removing the proximal penalty lowers the reported FAA, while an excessively large tested coefficient also lowers it relative to the selected setting. These observations characterize this grid, not a guaranteed optimum or convergence rate.

Table 35: Sensitivity to calibration subset size $|\mathcal{B}_{c,p}|$ in PRO-MAX on CIFAR-100 severe heterogeneity. Higher FAA and lower AF are preferred. Best values are bold, and runners-up are underlined.

$\lvert\mathcal{B}_{c,p}\rvert$	FAA ↑	AF ↓	Extra client memory
16	43.84 ± 1.12	31.04 ± 1.48	low
32	44.96 ± 1.02	28.72 ± 1.30	low
64	45.68 ± 0.96	27.36 ± 1.18	moderate
128	$\underline{46.12 \pm 0.88}$	$\underline{26.18 \pm 1.07}$	moderate
256	$\mathbf{46.24 \pm 0.87}$	$\mathbf{26.02 \pm 1.06}$	higher

Increasing calibration from 128 to 256 gives a small additional FAA gain in the reported means. The low, moderate, and higher cache labels are qualitative descriptions rather than measured memory sizes. The result does not establish an exact inverse-support law.

G.8 Optional memory-evolution visualization

The archived snapshot layout is retained only when diagnostic provenance has been verified. A shared principal component analysis (PCA) projection does not itself establish shared encoder coordinates, and a semantic task number is not a universal time boundary under ACE. Snapshot checkpoints, feature sources, projection fitting, and before/after memory identities are needed before using the figure to support a temporal claim.

H System Measurements and Analytical Accounting

The main resource comparison uses incremental wall-clock time and peak memory from one sequential-client workstation run per method. It is not a physical-device latency, energy, or hardware-robustness experiment. This appendix distinguishes those run measurements from analytical payload counts and unresolved accounting fields.

H.1 Platform and execution

Client updates run sequentially on one GPU, while loading, aggregation, memory accounting, and logging use CPU resources. The supplied platform description is retained below. Software family names are not exact patch-version manifests.

Table 36: Compute platform used in the experiments. CPU information is obtained from `/proc/cpuinfo`, and GPU information is obtained from `lspci` and `nvidia-smi`.

Component	Specification
CPU	2× Intel Xeon Gold 6242 @ 2.80 GHz
CPU cores / threads	32 physical cores / 64 logical threads
CPU cache	22 MB L3 cache per socket
CPU instruction support	AVX2, AVX-512, AVX512-VNNI
GPU	NVIDIA RTX A5000
GPU memory	24 GB GDDR6
System memory	256 GB DDR4 RAM
Storage	2 TB local solid-state drive
Software	Python 3.10, PyTorch 2.x, CUDA 12.x, cuDNN
Execution mode	Single-GPU sequential FL simulation

H.2 Incremental runtime and warmup entries

Tables 37 and 38 report incremental times, peak memories, and warmup entries for every listed method. Each dataset uses the same recorded warmup duration because the shared initialization protocol is common across its methods.

The reported subtotal is the incremental time plus the corresponding warmup time. These wall-clock records support a full-pipeline time comparison within each dataset. GPU memory remains a peak measurement whose allocator convention is separate from the timing ledger.

Table 37: Recorded incremental times, warmup entries, and peak memory for one run per method. Lower incremental time, subtotal, and peak memory are preferred. Best values are bold, and runners-up are underlined within each dataset. Tied best values are all bold, while the next distinct value is underlined. Shared warmup entries are unranked. The reported subtotal is the full-pipeline time obtained by adding the recorded warmup duration to the incremental time.

Dataset	Method	Incremental time	Warmup time	Reported subtotal	Peak GPU mem.	Peak CPU mem.
CIFAR-100	FedAvg	**2.10** h	0.60 h	**2.70** h	**4.6** GB	**10.2** GB
CIFAR-100	FedProx	2.20 h	0.60 h	2.80 h	4.7 GB	10.4 GB
CIFAR-100	FedProto	2.40 h	0.60 h	3.00 h	4.9 GB	11.1 GB
CIFAR-100	TARGET	4.80 h	0.60 h	5.40 h	8.2 GB	14.0 GB
CIFAR-100	MFCL	5.40 h	0.60 h	6.00 h	8.6 GB	15.2 GB
CIFAR-100	HR	5.80 h	0.60 h	6.40 h	9.0 GB	16.1 GB
CIFAR-100	FedSpace	3.00 h	0.60 h	3.60 h	5.4 GB	12.1 GB
CIFAR-100	FedSSI	2.80 h	0.60 h	3.40 h	5.2 GB	11.8 GB
CIFAR-100	FedCBDR	4.20 h	0.60 h	4.80 h	7.4 GB	14.3 GB
CIFAR-100	**PRO**	3.20 h	0.60 h	3.80 h	5.6 GB	12.0 GB
CIFAR-100	**PRO-MAX**	3.60 h	0.60 h	4.20 h	6.2 GB	13.1 GB
THUCNews-10	FedAvg	**0.32** h	0.12 h	**0.44** h	**1.8** GB	**5.0** GB
THUCNews-10	FedProx	0.34 h	0.12 h	0.46 h	**1.8** GB	5.1 GB
THUCNews-10	FedProto	0.38 h	0.12 h	0.50 h	1.9 GB	5.3 GB
THUCNews-10	FedSeIT	0.74 h	0.12 h	0.86 h	2.6 GB	7.0 GB
THUCNews-10	**PRO**	0.52 h	0.12 h	0.64 h	2.2 GB	6.0 GB
THUCNews-10	**PRO-MAX**	0.60 h	0.12 h	0.72 h	2.4 GB	6.4 GB
Cora	FedAvg	**0.18** h	0.08 h	**0.26** h	**1.5** GB	**4.0** GB
Cora	FedProx	0.19 h	0.08 h	0.27 h	**1.5** GB	**4.0** GB
Cora	FedProto	0.23 h	0.08 h	0.31 h	1.6 GB	4.3 GB
Cora	POWER	0.54 h	0.08 h	0.62 h	2.8 GB	7.1 GB
Cora	MOTION	0.58 h	0.08 h	0.66 h	3.0 GB	7.4 GB
Cora	**PRO**	0.42 h	0.08 h	0.50 h	2.1 GB	5.2 GB
Cora	**PRO-MAX**	0.48 h	0.08 h	0.56 h	2.4 GB	5.6 GB

Table 38: Recorded incremental times, warmup entries, and peak memory for one run per method. Lower incremental time, subtotal, and peak memory are preferred. Best values are bold, and runners-up are underlined within each dataset. Tied best values are all bold, while the next distinct value is underlined. Shared warmup entries are unranked. The reported subtotal is the full-pipeline time obtained by adding the recorded warmup duration to the incremental time.

Dataset	Method	Incremental time	Warmup time	Reported subtotal	Peak GPU mem.	Peak CPU mem.
TinyImageNet	FedAvg	**7.40** h	1.40 h	**8.80** h	**6.8** GB	**18.0** GB
TinyImageNet	FedProx	7.80 h	1.40 h	9.20 h	6.9 GB	18.2 GB
TinyImageNet	FedProto	8.40 h	1.40 h	9.80 h	7.2 GB	19.0 GB
TinyImageNet	TARGET	18.60 h	1.40 h	20.00 h	11.8 GB	24.0 GB
TinyImageNet	MFCL	20.40 h	1.40 h	21.80 h	12.4 GB	26.0 GB
TinyImageNet	HR	21.80 h	1.40 h	23.20 h	13.0 GB	28.0 GB
TinyImageNet	FedSpace	10.60 h	1.40 h	12.00 h	8.0 GB	21.0 GB
TinyImageNet	FedSSI	9.80 h	1.40 h	11.20 h	7.8 GB	20.0 GB
TinyImageNet	FedCBDR	15.20 h	1.40 h	16.60 h	10.2 GB	23.0 GB
TinyImageNet	**PRO**	11.00 h	1.40 h	12.40 h	8.1 GB	20.0 GB
TinyImageNet	**PRO-MAX**	12.40 h	1.40 h	13.80 h	8.8 GB	21.0 GB
CLINC150	FedAvg	**1.40** h	0.30 h	**1.70** h	**2.2** GB	**7.0** GB
CLINC150	FedProx	1.50 h	0.30 h	1.80 h	**2.2** GB	7.2 GB
CLINC150	FedProto	1.70 h	0.30 h	2.00 h	2.4 GB	8.0 GB
CLINC150	FedSeIT	2.80 h	0.30 h	3.10 h	3.4 GB	10.0 GB
CLINC150	**PRO**	2.10 h	0.30 h	2.40 h	2.8 GB	8.2 GB
CLINC150	**PRO-MAX**	2.40 h	0.30 h	2.70 h	3.1 GB	9.0 GB
ogbn-arxiv	FedAvg	**2.80** h	0.60 h	**3.40** h	**4.6** GB	**18.0** GB
ogbn-arxiv	FedProx	3.00 h	0.60 h	3.60 h	4.7 GB	18.4 GB
ogbn-arxiv	FedProto	3.40 h	0.60 h	4.00 h	4.9 GB	19.0 GB
ogbn-arxiv	POWER	5.60 h	0.60 h	6.20 h	7.4 GB	24.0 GB
ogbn-arxiv	MOTION	6.20 h	0.60 h	6.80 h	7.8 GB	25.0 GB
ogbn-arxiv	**PRO**	4.80 h	0.60 h	5.40 h	6.2 GB	21.0 GB
ogbn-arxiv	**PRO-MAX**	5.40 h	0.60 h	6.00 h	6.8 GB	22.0 GB

H.3 Base-space communication model

For b bytes per scalar and P transmitted model scalars, a full model upload or download contributes bP. Let $\mathcal{V}_{c,r}$ be the current classes for which client c actually uploads fresh statistics and $\mathcal{J}_{c,r}$ the classes for which it uploads transport summaries. The primary-bank payloads are

$$D_{c,r}^{\mathrm{mem}} = b|\mathcal{K}_r|(2d_u + 1), \tag{42}$$

$$U_{c,r}^{\mathrm{stat}} = b|\mathcal{V}_{c,r}|(2d_u + 1), \tag{43}$$

$$U_{c,r}^{\mathrm{move}} = b|\mathcal{J}_{c,r}|(d_u + 1). \tag{44}$$

These sets refer to constructed and actually transmitted classes rather than all semantic labels scheduled before a task position. Statistics are uploaded only at triggered write boundaries. Any extra aggregated-encoder broadcast at such a boundary must be counted separately from the ordinary next-round model broadcast.

A complete serialized accounting sums model, memory, statistic, transport, boundary, and metadata traffic over selected clients and communication rounds. Scalar expressions alone exclude identifiers, headers, retransmissions, and representation-specific serialization. They are not measurements of network delay or actual compressed message size. Optional adapted-space state adds its declared mean/radius/count and transport payload only for runs with the corresponding enable flag.

H.4 Persistent and temporary memory

The primary bank uses $b|\mathcal{K}_r|(2d_u + 1)$ payload bytes, excluding the resident model and metadata. PRO-MAX retains the same persistent order but needs temporary calibration and transport storage. One feature snapshot takes $bB_{\mathrm{cal}}d_u$ bytes. Holding both before and after snapshots simultaneously takes twice that amount before neighbor indices, distances, model state, and work buffers. Streaming can change the peak, so a one-snapshot expression is not a full peak-memory formula. The optional adapted bank adds $b|\mathcal{K}_r|(d_z + 2)$ bytes under the stated uniform-scalar convention.

Table 39: Analytical paired-feature payload for FP32 features with $d_u = 512$. A paired snapshot stores the calibration features before and after an encoder transition. These values exclude indices, distances, gradients, optimizer state, model replicas, and allocator overhead.

Calibration count	One snapshot (KiB)	Paired snapshots (KiB)
16	32	64
32	64	128
64	128	256
128	256	512
256	512	1024

H.5 ARCHIVED COMMUNICATION RECORDS

The following directional totals and normalized ratios are retained from the supplied source. Their per-message ledgers, exact scalar types, optional-bank flags, and boundary broadcast counts are not supplied, so they are not treated as independently audited byte measurements. A normalized ratio is specific to a dataset and its FedAvg denominator. It does not directly measure comparability of raw communication cost across different datasets. Small discrepancies between rounded directional totals and displayed ratios can arise from rounding and must be checked using unrounded logs.

Table 40: Bidirectional communication for image benchmarks under the severe-heterogeneity setting. Lower communication is preferred. Best values are bold, and runners-up are underlined within each dataset and direction. Tied best values are all bold, while the next distinct value is underlined. $S \to C$ denotes server-to-client communication, and $C \to S$ denotes client-to-server communication. Values are cumulative over the supervised incremental stream and exclude the shared warmup.

Method	CIFAR-100			TinyImageNet		
	$S \to C$	$C \to S$	Total / FedAvg	$S \to C$	$C \to S$	Total / FedAvg
FedAvg	**89.4** GB	**89.4** GB	**1.00×**	**178.8** GB	**178.8** GB	**1.00×**
FedProx	**89.4** GB	**89.4** GB	**1.00×**	**178.8** GB	**178.8** GB	**1.00×**
FedProto	94.8 GB	98.3 GB	1.08×	189.5 GB	196.7 GB	1.08×
TARGET	272.6 GB	228.0 GB	2.80×	545.0 GB	455.0 GB	2.80×
MFCL	286.1 GB	250.3 GB	3.00×	572.0 GB	500.8 GB	3.00×
HR	299.5 GB	272.6 GB	3.20×	599.0 GB	545.0 GB	3.20×
FedSpace	111.8 GB	111.7 GB	1.25×	402.3 GB	402.3 GB	2.25×
FedSSI	101.8 GB	103.6 GB	1.15×	384.4 GB	384.4 GB	2.15×
FedCBDR	263.7 GB	237.0 GB	2.80×	527.4 GB	474.0 GB	2.80×
PRO	110.6 GB	89.7 GB	1.12×	226.8 GB	180.9 GB	1.14×
PRO-MAX	110.6 GB	96.8 GB	1.16×	226.8 GB	198.7 GB	1.19×

Table 41: Bidirectional communication for text and graph benchmarks under the severe-heterogeneity setting. Values are cumulative over the full stream, and lower communication is preferred. Best values are bold, and runners-up are underlined within each dataset and direction. Tied best values are all bold, while the next distinct value is underlined.

Method	THUCNews-10			CLINC150			Cora			ogbn-arxiv		
	S→C	C→S	Total / FedAvg	S→C	C→S	Total / FedAvg	S→C	C→S	Total / FedAvg	S→C	C→S	Total / FedAvg
FedAvg	**4.20**	**4.20**	**1.00×**	**12.60**	**12.60**	**1.00×**	**0.18**	**0.18**	**1.00×**	**1.80**	**1.80**	**1.00×**
FedProx	**4.20**	**4.20**	**1.00×**	**12.60**	**12.60**	**1.00×**	**0.18**	**0.18**	**1.00×**	**1.80**	**1.80**	**1.00×**
FedProto	4.41	4.49	1.06×	13.20	13.50	1.06×	0.19	0.20	1.08×	1.93	1.96	1.08×
FedSeIT	5.88	4.62	1.25×	17.60	13.90	1.25×	-	-	-	-	-	-
POWER	-	-	-	-	-	-	0.27	0.24	1.42×	2.70	2.41	1.42×
MOTION	-	-	-	-	-	-	0.26	0.23	1.36×	2.56	2.34	1.36×
PRO	4.58	4.66	1.10×	13.70	14.00	1.10×	0.20	0.20	1.12×	1.98	2.05	1.12×
PRO-MAX	4.62	4.96	1.14×	13.90	14.80	1.14×	0.20	0.22	1.16×	2.00	2.18	1.16×

H.6 WARMUP ACCOUNTING

Warmup is a shared initialization stage, while the main time comparison excludes it. Full-pipeline reporting must state whether an initialization checkpoint is reused and count the appropriate warmup model messages and computation. Sharing a warmup schedule does not justify assigning missing timing entries a value of zero or copying a measured time into unmeasured baseline rows.

H.7 REPLAY-COUNT CONFIGURATIONS

The smaller and expanded settings in Table 9 change the number of replay objects. They do not certify matched communication, bytes, information, or compute, and the source does not provide a budget-matching optimization that would establish such a frontier. The source's earlier resource-matched label is therefore not used for the retained object-count setting.

H.8 AGGREGATE COMPUTE LEDGER

The aggregate ledger below sums each reported experiment group once. The eight reported groups total 1976 GPU-hours, and the additional 320 preliminary-run hours give 2296 GPU-hours overall.

Table 42: Approximate total compute used in the reported experiments. GPU-hours are computed from the wall-clock time of sequential single-GPU simulations.

Experiment group	Number of seeds	Approx. GPU-hours
Main CIFAR-100 aligned and heterogeneous runs	5	395
Generator-failure stress-test runs	5	136
THUCNews-10 and Cora modality-portability runs	5	55
CIFAR-100 ablation runs	5	75
TinyImageNet scale-up runs	5	790
CLINC150 scale-up runs	5	63
ogbn-arxiv scale-up runs	5	192
Sensitivity and diagnostic runs	5	270
Preliminary tuning and failed runs	-	320
Reported experiments only	-	1976
Total including preliminary runs	-	2296

The listed GPU-hours refer to elapsed single-GPU simulation time rather than integrated GPU utilization, energy, or CPU-hours. The main cost statement therefore remains a per-method runtime comparison rather than a matched-resource or energy claim.

I ANALYSIS OF AVAILABLE REHEARSAL AND MEMORY MAINTENANCE

The analysis follows the main lifecycle rather than treating replay mismatch as a generator-specific explanation. It first identifies the available rehearsal objective, then separates construction and maintenance errors, and finally bounds the stabilized MOVE estimator relative to the actual global transition. These statements concern fixed-stage surrogates under explicit assumptions. They do not prove convergence of the complete nonconvex training trajectory or a monotone relation between ACE and accuracy.

I.1 FIXED-STAGE ASSUMPTIONS

Fix a rehearsal stage and differentiate only with respect to $v = (\psi, W)$. Current and remembered feature distributions are held fixed during that derivative, as in the main rehearsal implementation. Let $J_r = |\mathcal{Y}_r^g|$ and assume $\mathcal{A}_{c,r}$ is nonempty.

Assumption 1 (Rehearsal regularity)**.** *The feature distributions have finite first moments. In the analyzed parameter region, $\ell_y(u; v)$ and $\nabla_v \ell_y(u; v)$ are respectively L_u- and L_g-Lipschitz in u, and differentiation and expectation can be interchanged. The reference class risks and gradient norms are*

bounded by B and G. Bounded loss, when used below, is an assumption on this region and support rather than a global property of neural CE loss.

I.2 Available support, class weighting, and feature mismatch

Define $\rho_c^{\mathrm{rep}}(y) = \mathbf{1}[y \in \mathcal{A}_{c,r}]/|\mathcal{A}_{c,r}|$, assigning zero weight to unavailable classes. Sampling a class uniformly over the available set and then sampling its real or remembered feature has the following property.

Lemma 1 (Unbiased gradient of the available surrogate)**.** *Under Assumption 1,*

$$\mathbb{E}_{y\sim\mathrm{Unif}(\mathcal{A}_{c,r}),\, u\sim q^u_{c,y,r}}[\nabla_v \ell_y(u;v)] = \nabla_v \widetilde{F}_{c,r}(v). \tag{45}$$

Proof. Expand the expectation over the finite available class set and interchange differentiation and the class-conditional expectation. The result is exactly Equation (9). No unavailable class distribution is introduced. □

Unbiasedness for this surrogate does not imply unbiasedness for the cumulative population objective. To describe the discrepancy, define

$$D_c = \sum_{y\in\mathcal{A}_{c,r}} \rho_c^{\mathrm{rep}}(y) W_1(P^{u,g}_{y,r}, q^u_{c,y,r}). \tag{46}$$

Proposition 1 (Coverage, weighting, and feature mismatch)**.** *Under the same assumptions,*

$$|F_r^{\mathrm{cum}} - \widetilde{F}_{c,r}| \le B\|\rho_c^{\mathrm{rep}} - \mathbf{1}/J_r\|_1 + L_u D_c, \tag{47}$$

$$\|\nabla_v F_r^{\mathrm{cum}} - \nabla_v \widetilde{F}_{c,r}\|_2 \le G\|\rho_c^{\mathrm{rep}} - \mathbf{1}/J_r\|_1 + L_g D_c. \tag{48}$$

Proof. Add and subtract the reference class risks weighted by ρ_c^{rep}. Bounded reference risks control the change in weights. A class-conditional coupling and Lipschitzness bound each change in expectation by $L_u W_1$. Apply the same argument to vector-valued gradients and use the triangle inequality for the second inequality. □

If $\mathcal{A}_{c,r} = \mathcal{Y}_r^g$, uniform class sampling removes the weighting term but not the feature term. Across clients, use $\bar{\rho} = \sum_c \omega_c \rho_c^{\mathrm{rep}}$ and $\sum_c \omega_c D_c$ for the analogous aggregate bound. Complementary supports can reduce aggregate weighting error. Consequently, ACE alone determines neither the gradient error nor the eventual accuracy. The omission of an additional rehearsal regularizer is a design choice, not a consequence that follows from unbiased sampling.

I.3 Replay mismatch as a source-independent bound

Proposition 2 (Class-conditional replay mismatch)**.** *For fixed matching class weights $\lambda_y \ge 0$ and Lipschitz losses on a common feature space,*

$$\left|\sum_y \lambda_y \left(\mathbb{E}_{P_y}\ell_y - \mathbb{E}_{Q_y}\ell_y\right)\right| \le L_u \sum_y \lambda_y W_1(P_y, Q_y). \tag{49}$$

If $0 \le \ell_y \le B$ on the relevant support, replace $L_u W_1(P_y, Q_y)$ by $B\,\mathrm{TV}(P_y, Q_y)$, with $\mathrm{TV}(P,Q) = \sup_A |P(A) - Q(A)|$.

Proof. Bound the difference for each class using a coupling and Lipschitzness, or the bounded-loss total-variation inequality, and sum with the nonnegative weights. □

The bound applies to generated inputs after an appropriate common representation, stored features, and reconstructed features. It is an upper bound, so larger mismatch does not by itself prove a larger realized risk gap or identify the cause of forgetting. Comparisons involving different class priors also require the weighting term rather than only the feature-distribution term. MMD values are not substituted for W_1 in these inequalities.

I.4 Separating memory construction from later maintenance

Fix a historical input population and a direct-write round s. Let P_s denote its global feature distribution, P_s^{write} the population represented by current writers, $Q_s^\star$ that writer population's diagonal summary, $\widehat{Q}_s^g$ its empirical summary in global coordinates, and Q_s the stored summary. Define

$$\begin{aligned} \epsilon^{\mathrm{pop}} &= W_1(P_s, P_s^{\mathrm{write}}), & \epsilon^{\mathrm{family}} &= W_1(P_s^{\mathrm{write}}, Q_s^\star), \\ \epsilon^{\mathrm{est}} &= W_1(Q_s^\star, \widehat{Q}_s^g), & \epsilon^{\mathrm{coord}} &= W_1(\widehat{Q}_s^g, Q_s). \end{aligned} \tag{50}$$

For an L_T-Lipschitz candidate map T from write-time features to a later round r, set $\epsilon^{\mathrm{dyn}} = W_1(P_r, T_\# P_s)$ and $\epsilon^{\mathrm{maint}} = W_1(T_\# Q_s, Q_r^{\mathrm{used}})$.

Proposition 3 (Construction and maintenance error)**.**

$$\begin{aligned} W_1(P_r, Q_r^{\mathrm{used}}) \le{}& \epsilon^{\mathrm{dyn}} + \epsilon^{\mathrm{maint}} \\ &+ L_T(\epsilon^{\mathrm{pop}} + \epsilon^{\mathrm{family}} + \epsilon^{\mathrm{est}} + \epsilon^{\mathrm{coord}}). \end{aligned} \tag{51}$$

Proof. Insert $T_\# P_s$ and $T_\# Q_s$ between the two endpoint distributions. Apply the triangle inequality and $W_1(T_\# P, T_\# Q) \le L_T W_1(P, Q)$. Insert the three intermediate construction distributions to bound $W_1(P_s, Q_s)$. □

WRITE addresses the encoder-choice coordinate term for contributing data at construction. It does not remove incomplete-population coverage, finite-support estimation, or the limitations of a diagonal summary. MOVE approximates the later maintenance term. This decomposition is not a factorial attribution of accuracy improvements, and no operation is guaranteed to reduce every term.

Proposition 4 (Diagonal-moment estimation)**.** *For population and empirical diagonal summaries with the same γ,*

$$W_1(Q_y^\star, \widehat{Q}_y) \le \|\widehat{\mu}_y - \mu_y\|_2 + \gamma \|\widehat{\sigma}_y - \sigma_y\|_2. \tag{52}$$

Proof. Couple both Gaussians using the same standard-normal vector. Bound the expected displacement by the norm of the mean difference plus the root mean squared norm of the scaled noise difference. □

For independent observations at a fixed encoder, suitable sub-Gaussian and bounded-fourth-moment assumptions, and coordinate standard deviations bounded away from zero, the estimation expression has rate $O_p(\sqrt{d_u/n_y})$. A common positive variance floor can instead yield floor-dependent constants. The actual moment bound remains applicable without claiming this rate for arbitrary correlated client data or treating stored construction counts as certified effective sample sizes.

I.5 Global-coordinate construction

Assumption 2 (Local parameter-to-feature regularity)**.** *For the compared inputs and encoders, $\|h_\theta(x) - h_{\theta'}(x)\|_2 \le L_h \|\theta - \theta'\|_2$ on the analyzed parameter region.*

For fixed writer populations and identical normalized writer weights ν_c, local-coordinate pooling differs from global-coordinate pooling by at most

$$\left\| \sum_c \nu_c \left[\mathbb{E}_{P^x_{c,y}} h_{\theta_c}(X) - \mathbb{E}_{P^x_{c,y}} h_{\theta^g}(X) \right] \right\|_2 \le L_h \sum_c \nu_c \|\theta_c - \theta^g\|_2. \tag{53}$$

The inequality follows by moving the norm inside each expectation and applying Lipschitzness. Re-encoding under the same global encoder removes this particular discrepancy. It neither reconstructs missing historical populations nor proves that actual coordinate error grows monotonically with model dispersion.

I.6 CONTROLLED LOCAL PLASTICITY

Proposition 5 (Proximal control of a local update)**.** *Let F_c be the non-proximal part of the fitting objective and let $\bar{\Theta} = \Pi_r(\Theta)$. Suppose $\|\nabla_{\bar{\Theta}} F_c\|_2 \le G$ and the returned local iterate is ξ-stationary for $F_c + \lambda_{\text{prox}}\|\bar{\Theta} - \bar{\Theta}^g\|_2^2$, with $\lambda_{\text{prox}} > 0$. Then*

$$\|\bar{\Theta}_c - \bar{\Theta}^g\|_2 \le \frac{G + \xi}{2\lambda_{\text{prox}}}. \tag{54}$$

If the base parameters are included, Assumption 2 gives the corresponding feature bound $L_h(G + \xi)/(2\lambda_{\text{prox}})$.

Proof. Rearrange the approximate stationarity condition, apply the triangle inequality, and divide by $2\lambda_{\text{prox}}$. □

Freezing a client's encoder eliminates its own within-round base update. It does not imply that the aggregated encoder is unchanged when other participating clients update their base parameters. Consequently, unsupported historical memories can still require maintenance under different client plasticity schedules. This bound assumes the stated stationarity condition and is not a convergence guarantee for a fixed number of local SGD steps.

I.7 STABILIZED LOCAL TRANSPORT

The estimator here is the stabilized estimator in Equation (16), with the adaptive weights from Equation (15). For $\kappa_{c,y} > 0$, set

$$a_{i,y} = \frac{\alpha_{i,y}}{\kappa_{c,y}}, \qquad s_{c,y} = \frac{\kappa_{c,y}}{\kappa_{c,y} + \varepsilon}, \qquad n_{c,y}^{\text{eff}} = \frac{\kappa_{c,y}^2}{\sum_{i \in \mathcal{N}_{c,y}} \alpha_{i,y}^2} = \frac{1}{\sum_i a_{i,y}^2}. \tag{55}$$

Let $\mathscr{G}$ fix the selected pre-update neighborhoods, weights, displacement fields, and global reference used in the comparison.

Assumption 3 (Conditional transport regularity)**.** *On the analyzed neighborhood, $d_i = \delta_c(u_i^-) + \xi_i$, with $\mathbb{E}[\xi_i \mid \mathscr{G}] = 0$ and*

$$\mathbb{E}\left[\left\|\sum_i w_i \xi_i\right\|_2^2 \middle| \mathscr{G}\right] \le \sigma_{\delta,c}^2 \sum_i w_i^2 \tag{56}$$

for weights fixed given $\mathscr{G}$. The displacement field is β_c-Lipschitz between the selected calibration features and the memory center. Where the old center lies outside current support, this assumes an appropriate local extension; it is not identified from current observations alone.

Because neighbors and weights depend on the calibration sample, centering conditional only on an individual feature is not enough to justify the weighted-noise calculation. The stronger conditional assumption above is stated explicitly and is not an automatic property of reusing training examples for calibration.

Theorem 1 (Local displacement error with stabilization)**.** *For $r_{c,y} = \max_{i \in \mathcal{N}_{c,y}} \|u_i^- - \mu_{y,r}^u\|_2$, the assumptions give*

$$\begin{aligned}\mathbb{E}[\|\Delta_{c,y}^u - \delta_c(\mu_{y,r}^u)\|_2^2 \mid \mathscr{G}] \le\ & \left[s_{c,y}\beta_c r_{c,y} + (1 - s_{c,y})\|\delta_c(\mu_{y,r}^u)\|_2\right]^2 \\ & + \frac{s_{c,y}^2 \sigma_{\delta,c}^2}{n_{c,y}^{\text{eff}}}.\end{aligned} \tag{57}$$

Proof. Subtract $\delta_c(\mu_{y,r}^u)$ from the stabilized weighted estimate. The result is $s_{c,y}\sum_i a_{i,y}[\delta_c(u_i^-) - \delta_c(\mu_{y,r}^u)] - (1 - s_{c,y})\delta_c(\mu_{y,r}^u) + s_{c,y}\sum_i a_{i,y}\xi_i$. Conditional centering removes the bias-noise cross term. Bound the deterministic part by Lipschitzness and the triangle inequality, and the noise term by Equation (56). □

The bound separates locality bias, shrinkage from the stabilizer, and averaging noise. For $\varepsilon = 0$, its variance term is $\sigma^2_{\delta,c}/n^{\text{eff}}_{c,y}$. Since $0 \le \alpha_{i,y} \le 1$, $\sum_i \alpha^2_{i,y} \le \kappa_{c,y}$ and hence $1/n^{\text{eff}}_{c,y} \le 1/\kappa_{c,y}$. Thus a $1/\kappa$ expression is a looser bound under the same assumptions, not an identity between support score and effective sample size.

Neither expression makes the support score a calibrated confidence. Away from the bandwidth floor, multiplying every neighborhood distance and its median bandwidth by the same positive constant leaves all Gaussian weights, and therefore $\kappa_{c,y}$, unchanged while increasing the absolute radius. Kernel support does not certify proximity or transferability.

I.8 Client estimates versus the actual global transition

Exact historical transport would require the population mean displacement under the actual global transition,

$$\Delta^{g,\star}_{y,r} = \mathbb{E}_{x \sim P^x_y}[h_{\theta^g_{r+1}}(x) - h_{\theta^g_r}(x)]. \tag{58}$$

Define $b_{c,y} = \kappa_{c,y}/(\sum_j \kappa_{j,y} + \varepsilon)$, let $B_{c,y}$ be the square root of the local bound, and set

$$\epsilon^{\text{agg}}_{y,r} = \left\| \sum_c b_{c,y} \delta_c(\mu^u_{y,r}) - \Delta^{g,\star}_{y,r} \right\|_2 . \tag{59}$$

Conditional Minkowski's inequality yields

$$\left(\mathbb{E}[\|\Delta^{u,g}_y - \Delta^{g,\star}_{y,r}\|_2^2 \mid \mathscr{G}]\right)^{1/2} \le \sum_c b_{c,y} B_{c,y} + \epsilon^{\text{agg}}_{y,r}. \tag{60}$$

To verify this, insert $\sum_c b_{c,y}\delta_c(\mu^u_{y,r})$ between the estimate and global target and apply the triangle inequality in conditional L_2. Client-error independence is unnecessary for this bound. The residual retains client-to-global update differences, model versus transport aggregation weights, server normalization shrinkage, and pointwise versus class-average movement. It is not assumed zero or treated as empirically measured by cosine center alignment.

I.9 Persistent state and communication order

The primary bank therefore scales as

$$\text{MemoryBank}_{\text{PRO}} = O(|\mathcal{K}_r|(2d_u + 1)) . \tag{61}$$

Proposition 6 (Class-memory payload and server operations)**.** *At round r, the primary bank stores $|\mathcal{K}_r|(2d_u + 1)$ scalars. Its full downlink per selected client has the same scalar count. For locally estimated classes $\mathcal{J}_{c,r}$, MOVE adds $|\mathcal{J}_{c,r}|(d_u + 1)$ uploaded scalars. Parameter averaging and dense class-statistic aggregation require $O(|\mathcal{S}_r||\Theta| + |\mathcal{S}_r||\mathcal{K}_r|d_u)$ arithmetic operations, apart from metadata handling and any optional adapted bank.*

Proof. Each primary entry contains a mean, diagonal spread, and construction count. Each transport summary contains a vector and weight. Sum the storage and arithmetic requirements over stored classes and participating clients. None of these server operations differentiates through a model. □

These counts exclude models resident in memory, optimizer state, calibration buffers, serialization, and extra boundary encoder broadcasts. Optional adapted-space state adds its declared payload but is not silently included in the main base-space method. Appendix H distinguishes these analytical counts from the retained workstation and communication records.

J Scope, Assumptions, and Further Validation

The purpose of this study is to evaluate how shared rehearsal evidence remains useful as client class histories and representations evolve. The controlled protocols support comparisons of rehearsal construction, placement, and maintenance, while the cross-modal experiments examine reuse of the same class-memory interface. The following boundaries distinguish what these results establish from the settings, guarantees, and extensions that require further evidence.

Controlled federation and operational deployment. The experiments use benchmark-derived client partitions and prescribed exposure schedules. This allows task support, label coverage, and task order to vary under a common protocol, but does not establish that the sampler reproduces a particular deployed population. ACE concerns disagreement in client class histories and is distinct from communication asynchrony. Delayed updates, persistent client unavailability, heterogeneous devices, and unreliable networks are separate deployment questions that may require additional scheduling, stale-update handling, and memory-maintenance mechanisms.

Initialization and data access. The main comparisons share an in-domain label-free initialization protocol that includes later-local training inputs without their labels. The position-0-only control narrows this access and tests whether the reported ordering persists under a restricted local pool. Because clients can begin with different semantic tasks, that control is not equivalent to globally excluding future-class inputs or starting without shared initialization. The results should therefore be interpreted under the reported access assumptions rather than as initialization-independent performance.

Model, task, and scale coverage. The evaluated encoders are ResNet-18, TextCNN, and two-layer GAT, and the prediction tasks are image, text, and graph node classification. These experiments support reuse of the post-encoder class-memory interface across the tested representations. They do not establish performance for foundation models, generative objectives, whole-graph prediction, or naturally occurring client streams. The scale-up benchmarks increase class-space size and stream length, but very large client populations and indefinitely expanding label spaces remain outside the evaluated range. Such extensions may require changes to memory transmission or representation rather than only rerunning the current configuration.

Scope of the memory model and analysis. The evaluated memory summarizes each class with a mean and coordinate-wise spread, while MOVE updates only the mean when fresh reconstruction is unavailable. This is a compact maintenance design rather than a model of arbitrary distributional evolution. The analysis distinguishes approximation and estimation errors under stated regularity assumptions and retains the discrepancy between local displacement estimates and the global encoder transition. It does not imply exact transport or convergence of the entire nonconvex federated trajectory. The support and sensitivity studies characterize behavior within the reported grids rather than uniform robustness to all evidence distributions.

Mechanistic and statistical coverage. Sequential and targeted controls examine alternatives concerning rehearsal placement, global-coordinate refresh, and class-associated movement. They do not form a full factorial experiment or identify independent additive contributions. The reported sweeps also do not exhaust interactions between exposure conditions, optimization choices, and calibration budgets. Five-seed means and standard deviations characterize variation in the predictive experiments, while mean-only controls retain their stated uncertainty limitations. These comparisons support the evaluated design choices without establishing an exhaustive causal account of continual-learning performance.

Resource measurements and deployment guarantees. Runtime and peak-memory measurements describe sequential-client simulations on the reported workstation. They quantify the overhead of the evaluated training pipeline but are not measurements of deployed latency, energy consumption, network contention, or hardware-independent efficiency. Persistent class-memory size and communication accounting must also be distinguished from transient training buffers and execution time. Broader systems claims require measurements under the corresponding hardware, networking, and participation conditions.

Privacy, security, and distribution of benefits. Class statistics and displacement summaries remain privacy-sensitive despite the absence of raw-example transmission (Mahendran and Vedaldi, 2015; Fredrikson et al., 2015; Zhu et al., 2019). The present study does not establish formal privacy guarantees, adversarial robustness, or the utility-privacy trade-off of an integrated protection mechanism. Secure aggregation and differential privacy require their own protocol design and accounting rather than following automatically from class-level memory (Bonawitz et al., 2017; Dwork and Roth, 2014). The predictive evaluation targets one global model and does not establish personalized utility or guarantees for the worst-served clients.