RISK-GUIDED ASYNCHRONOUS RECOVERY FOR FEDERATED TEST-TIME ADAPTATION UNDER DISTRIBUTION SHIFTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Federated Learning (FL) enables decentralized model training without sharing raw data, yet deployed models often encounter distribution shifts. To maintain model performance under such evolving shifts, federated test-time adaptation (TTA) has emerged as a promising approach that adapts deployed models using unlabeled test data and periodically leverages cross-client collaboration to recover from local adaptation errors. However, as data distributions continue to shift after deployment, existing federated test-time adaptation methods suffer from: (1) errors accumulate during long-term TTA, (2) periodic collaboration incurs redundant communication, and (3) fixed layer recovery ignores degradation severity and may overwrite useful adapted parameters. To address these issues, we propose FedART, a risk-guided asynchronous recovery framework that efficiently mitigates error accumulation under continual distribution shifts. Compared to existing methods, FedART enables each client to initiate recovery asynchronously and adaptively determines recovery scope, thereby preserving beneficial local adaptations while improving efficiency. Specifically, FedART first introduces a local adaptation risk monitoring scheme to continuously estimate state risk of each client's local TTA, and designs a risk-triggered asynchronous collaboration mechanism to retrieve suitable client adaptation states to construct recovery candidates, avoiding error accumulation and reducing redundant communication. Then, we devise a risk-guided selective layer restoration mechanism that preserves useful adapted parameters instead of uniformly recovering layers, thereby improving model stability under continual distribution shifts. Extensive experiments across multiple benchmarks demonstrate that FedART achieves competitive or superior performance compared with strong baselines, while reducing the communication volume for TTA recovery by approximately 98% compared with periodic recovery.

1 INTRODUCTION

Federated Learning (FL) (McMahan et al., 2017) enables collaborative training across distributed clients without sharing raw data. Early FL research primarily focused on training a shared global model, while increasing attention to client heterogeneity led to federated optimization methods for non-IID data and diverse client conditions (Li et al., 2020; Karimireddy et al., 2020; Wang et al., 2020). In parallel, personalized FL allowed clients to benefit from collaborative learning while maintaining models tailored to their local distributions (Li et al., 2021; Shamsian et al., 2021). More recently, FL has expanded to dynamic and continual scenarios, where client data distributions may evolve over time (Jothimurugesan et al., 2023). Despite these advances, maintaining reliable model performance after deployment remains challenging, as client-side test distributions may deviate from those observed during training and continue to evolve. Test-time adaptation (TTA) (Wang et al., 2021; Liang et al., 2025b) addresses such deployment-time distribution shifts by adapting a pre-trained model directly to incoming unlabeled test data.

For example, FedTSA (Zhang et al., 2024) models temporal and inter-client correlations through temporal-spatial aggregation, while CoLA (Chen et al., 2024) accumulates and shares domain knowledge across devices for collaborative lifelong adaptation. These collaborative strategies improve the stability of local adaptation, enabling clients to adapt to evolving test distributions while leveraging

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.

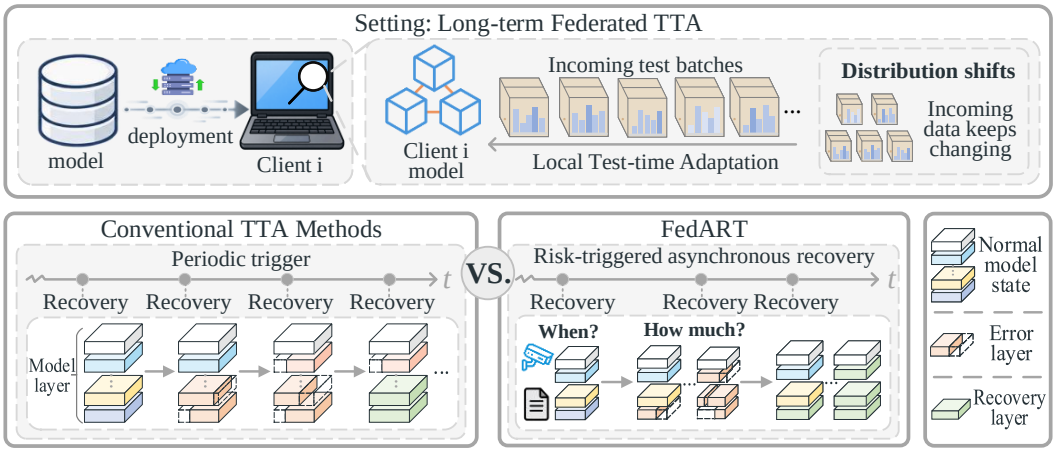


Figure 1: Comparison between conventional periodic recovery and FedART under long-term federated test-time adaptation. Conventional recovery uses periodic recovery and a fixed restoration scope, whereas FedART activates recovery only when needed and adapts restoration to degradation severity, reducing redundant communication while better preserving useful adapted parameters.

useful information from other clients. However, during long-term deployment under continual distribution shifts, federated TTA faces three key challenges: (1) Errors accumulate during long-term TTA, as online adaptation relies on model predictions from unlabeled test data, and erroneous updates can progressively degrade local models (Wang et al., 2022; Niu et al., 2022; 2023); (2) periodic cross-client collaboration triggers recovery at fixed intervals, leading to unnecessary communication for clients that do not require recovery; and (3) fixed recovery scope ignores degradation severity: excessive restoration may overwrite useful adapted parameters, while insufficient recovery may fail to correct severely degraded states.

To address these challenges, we propose FedART, a Risk-guided **A**synchronous **R**ecovery Framework for Federated **T**est-time Adaptation under continual distribution shifts. As illustrated in Fig. 1, our key idea is to use each client's adaptation risk to jointly determine ***when recovery should be triggered*** and ***how much of the model should be restored***. To realize this, FedART introduces a local adaptation risk monitoring scheme that continuously estimates adaptation risk to track changes in each client's local adaptation state during long-term TTA. Based on this risk signal, FedART performs risk-triggered asynchronous collaboration according to detected model degradation rather than predefined communication intervals, allowing recovery to be invoked only when needed. FedART further introduces a risk-guided selective layer restoration mechanism that adapts recovery scope to degradation severity instead of applying a fixed scope across different adaptation states. This severity-aware adjustment reduces overwriting of beneficial local adaptations while providing stronger correction. FedART adapts recovery to each client's evolving adaptation state instead of relying on predefined schedules and fixed recovery scopes, thereby mitigating long-term error accumulation while preserving beneficial local adaptations and reducing unnecessary recovery overhead.

Extensive experiments across multiple benchmarks demonstrate that FedART achieves competitive or superior performance against strong federated TTA baselines. Compared with periodic collaboration strategies, FedART reduces recovery communication volume by approximately 98%.

Our main contributions are summarized as follows:

- We identify a fundamental limitation of existing recovery strategies in federated TTA: predefined recovery schedules and fixed restoration scopes cannot adapt to dynamically evolving client states.
- We propose FedART, a risk-guided framework that uses client-specific adaptation risk to asynchronously trigger cross-client collaboration and adaptively determine recovery scope, effectively mitigating accumulated adaptation errors while preserving beneficial local adaptations in an efficient manner.
- Extensive experiments demonstrate that FedART reduces recovery communication while maintaining superior adaptation performance under continual distribution shifts.

2 Related Work

2.1 FL under Distribution Shift

Federated learning (McMahan et al., 2017) enables distributed clients to collaboratively train a shared model without sharing raw data. Classical FL methods (Li et al., 2020; Karimireddy et al., 2020; Wang et al., 2020) primarily address data heterogeneity during federated training. However, recent studies (Jothimurugesan et al., 2023) have shown that client data distributions may continue to change after deployment, with different clients experiencing distribution shifts at different times, rates, and degrees. Such continuously evolving distributions are largely neglected by classical FL methods. To address dynamic scenarios, recent work (Jothimurugesan et al., 2023) has begun to explicitly study federated learning under distributed concept drift, where client distributions evolve over time and may shift asynchronously across clients. Although such drift-aware approaches improve model robustness to temporal distribution changes, they typically rely on continued training and shared model updating, making it difficult to maintain deployed model performance under long-term distribution shifts.

2.2 Federated Test-Time Adaptation

Test-time adaptation (TTA) (Wang et al., 2021; Liang et al., 2025b) adapts a deployed model directly to distribution shifts using incoming unlabeled test data. Under continual test streams, repeated adaptation may lead to error accumulation, model drift, or forgetting, motivating methods that improve long-term adaptation stability (Wang et al., 2022; Niu et al., 2022; 2023; Yuan et al., 2023).

Recent studies extend TTA to federated settings, where clients adapt to their local test distributions while leveraging cross-client information to improve adaptation robustness (Zhang et al., 2026). FedTHE+ (Jiang & Lin, 2023) improves test-time personalization through adaptive head ensembling and tuning, while ATP (Bao et al., 2023) learns module-wise adaptation rates to accommodate heterogeneous client shifts. Other methods further exploit cross-client knowledge during deployment: FedTSA (Zhang et al., 2024) models temporal and inter-client correlations through temporal-spatial aggregation, CoLA (Chen et al., 2024) accumulates and shares domain knowledge across devices for collaborative lifelong adaptation, and FedCTTA (Rajib et al., 2025) performs similarity-aware knowledge aggregation for continual federated adaptation. TTA-FedDG (Liang et al., 2025a) introduces FedSPL to incorporate test-time adaptation into federated domain generalization.

These methods mainly focus on what adaptation knowledge should be personalized, aggregated, or shared across clients. Comparatively less attention has been paid to controlling the recovery process under long-term adaptation: *when* cross-client recovery should be invoked for an individual client and *how much* of its adapted state should be restored once degradation occurs. FedART addresses this complementary problem through risk-guided recovery control, using client-specific adaptation risk to trigger asynchronous collaboration and to determine a severity-adaptive restoration scope.

3 Method

3.1 Problem Formulation

We consider a federated test-time adaptation setting with K distributed clients initialized from a pre-trained source model f_{θ^0}. During deployment, each client asynchronously receives a sequence of unlabeled target batches. At local event t client i receives $\mathcal{B}_{i,t} = \{x_{i,t}^{(n)}\}_{n=1}^{N_{i,t}}$, where $N_{i,t}$ denotes the batch size. The target distribution $P_{i,t}(X, Y)$ may evolve across both clients and time, and client streams are not required to be synchronized. Let $\theta_{i,t-1}^{W}$ denote the working state of client i before processing $\mathcal{B}_{i,t}$. Without recovery, a standard local TTA operator $\mathcal{T}$ is defined as

$$\theta_{i,t}^{\text{cont}} = \mathcal{T}\left(\theta_{i,t-1}^{W}, \mathcal{B}_{i,t}\right). \tag{1}$$

Our goal is to maintain reliable online prediction under long-term, client-specific distribution shifts without target labels during deployment.

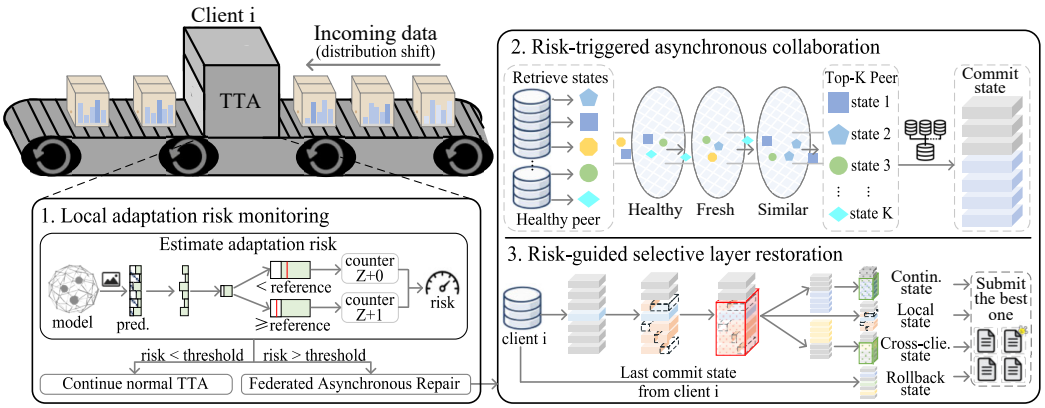


Figure 2: Overview of FedART. The framework consists of three components: local adaptation risk monitoring to track client degradation during test-time adaptation, risk-triggered asynchronous collaboration to retrieve suitable peer states only when recovery is needed, and risk-guided selective layer restoration to adapt the recovery scope to degradation severity, enabling efficient and minimally disruptive recovery under continual distribution shifts.

3.2 Framework Overview

FedART augments standard local TTA with three mechanisms, as illustrated in Fig. 2: local adaptation risk monitoring determines when recovery is needed, risk-triggered asynchronous collaboration retrieves suitable peer states on demand, and risk-guided selective layer restoration adjusts the restoration scope according to degradation severity. Together, these mechanisms replace periodic fixed-scope recovery with client-specific, severity-adaptive recovery.

3.3 Local Adaptation Risk Monitoring

During long-term TTA, local models are continuously updated using unlabeled test data, making it difficult to determine whether the current adaptation state remains reliable or has begun to degrade. To address this issue, we propose local adaptation risk monitoring, which continuously estimates client-specific adaptation risk from incoming test data and identifies unreliable adaptation states without requiring target labels.

Before deployment, client i uses a labeled source calibration set $\mathcal{D}_i^0$ to obtain an initial uncertainty threshold τ_i^0, a source-risk upper bound $U_{R,i}^0$, and a proxy false-positive upper bound $U_{F,i}^0$. During deployment, each incoming batch $\mathcal{B}_{i,t}$ is evaluated under the current working state $\theta_{i,t-1}^W$ before adaptation, using prediction uncertainty $u(x;\theta) = 1 - \max_c p_\theta(c \mid x)$ as the unlabeled risk proxy. The uncertainty-exceedance indicator and risk lower bound are defined as follows:

$$
\begin{aligned}
z_{i,t}^{(n)} &= \mathbb{I}\left[u\left(x_{i,t}^{(n)}; \theta_{i,t-1}^W\right) > \tau_{i,t}\right], \\
L_{R,i,t} &= \max\left(0, \widehat{q}_{i,t} - \sqrt{\frac{\log(1/\alpha_m)}{2M_{i,t}}} - U_{F,i}^0\right),
\end{aligned}
\tag{2}
$$

where $M_{i,t}$ denotes the number of samples in the monitoring sequence, $\widehat{q}_{i,t}$ is the empirical uncertainty-exceedance rate, and α_m denotes the confidence budget assigned to the current check.

To remain sensitive to emerging degradation, local adaptation risk monitoring maintains multiple monitoring sequences initialized at different times and uses the largest resulting risk lower bound as $L_{R,i,t}$. The confidence budgets are allocated across clients, monitor starts, and sequential checks, with the statistical details provided in Appendix A.3.1. Since continual adaptation may change the model's confidence scale, $\tau_{i,t}$ is periodically recalibrated using the current working model on the fixed source calibration set, without using target labels or resetting accumulated monitoring evidence.

Based on the estimated risk, local adaptation risk monitoring determines whether the current adaptation state requires recovery as follows:

$$A_{i,t} = \mathbb{I}\left[L_{R,i,t} > U^0_{R,i} + \epsilon_i\right], \tag{3}$$

where ϵ_i is the tolerated risk increase, and $A_{i,t} = 1$ triggers recovery.

Through continuous risk estimation and client-specific triggering, local adaptation risk monitoring enables timely recovery from emerging degradation while avoiding unnecessary recovery and communication caused by periodic recovery.

3.4 Risk-Triggered Asynchronous Collaboration

Cross-client collaboration can provide useful recovery information, but periodic collaboration incurs unnecessary communication and may retrieve stale or mismatched peer states. To address this issue, we propose risk-triggered asynchronous collaboration, which initiates communication according to client-specific recovery demand and retrieves peer states based on relevance, freshness, and reliability.

The mechanism maintains a lightweight server-side registry for each client j, including the timestamp of its latest committed state, prediction signature, risk upper bound, and shareability indicator. When $A_{i,t} = 1$, client i asynchronously requests peer states and is excluded from sharing. It computes its prediction signature and eligible peer set as follows:

$$\begin{aligned} \mathbf{s}_{i,t} &= \mathrm{Norm}\left(\frac{1}{N_{i,t}}\sum_{n=1}^{N_{i,t}} p_{\theta^W_{i,t-1}}\left(\cdot \mid x^{(n)}_{i,t}\right)\right), \\ \mathcal{P}_{i,t} &= \left\{j \neq i \;\middle|\; g_j = 1,\ 0 \le \xi_{i,t} - \xi^C_j \le S_{\max},\ \mathrm{sim}\left(\mathbf{s}_{i,t}, \mathbf{s}^C_j\right) \ge \gamma\right\}. \end{aligned} \tag{4}$$

where $\mathrm{sim}(\cdot,\cdot)$ denotes cosine similarity, $S_{\max}$ is the maximum staleness, γ is the similarity threshold, and $g_j = 1$ indicates that client j is shareable.

Risk-triggered asynchronous collaboration ranks eligible peer states according to prediction similarity, freshness, and reliability. Let $\Delta\xi_{ij} = \xi_{i,t} - \xi^C_j$ and $m_j = \max(\varepsilon_w, 1 - U^C_j)$. The corresponding peer weights are computed as follows:

$$\begin{aligned} \widetilde{w}_{ij} &= \mathrm{sim}\left(\mathbf{s}_{i,t}, \mathbf{s}^C_j\right)\exp\left(-\frac{\Delta\xi_{ij}}{T_s}\right) m_j, \\ w_{ij} &= \frac{\widetilde{w}_{ij}}{\sum_{k\in\mathcal{P}^{K_p}_{i,t}} \widetilde{w}_{ik}}, \qquad j \in \mathcal{P}^{K_p}_{i,t}. \end{aligned} \tag{5}$$

where $\varepsilon_w > 0$ prevents vanishing weights, T_s controls staleness decay, and $\mathcal{P}^{K_p}_{i,t}$ contains the top-K_p clients ranked by $\widetilde{w}_{ij}$.

By combining on-demand asynchronous communication with selective peer retrieval, risk-triggered asynchronous collaboration avoids unnecessary periodic communication while prioritizing recent, reliable, and distribution-relevant client states for recovery.

3.5 Risk-Guided Selective Layer Restoration

A fixed restoration scope may overwrite useful adapted parameters under mild degradation or provide insufficient correction under severe degradation. To address this issue, we propose risk-guided selective layer restoration, which adjusts the restoration scope according to degradation severity.

For client i, the continuation state is treated as the current locally adapted state, i.e., $\theta^{\mathrm{adapt}}_{i,t} = \theta^{\mathrm{cont}}_{i,t}$. The proposed mechanism estimates degradation severity from the prediction concentration of this

state and determines the corresponding restoration ratio as follows:

$$\begin{aligned}\widehat{\mathbf{p}}_{i,t} &= \mathrm{softmax}\left(\frac{1}{N_{i,t}}\sum_{n=1}^{N_{i,t}} f_{\theta_{i,t}^{\mathrm{adapt}}}\left(x_{i,t}^{(n)}\right)\right),\\ C_{i,t} &= \sum_{c}\widehat{p}_{i,t,c}\log\widehat{p}_{i,t,c},\\ r_{i,t} &= \min\left(r_{\max}, r_{\min}+\lambda_r\max\left(0, C_{i,t}-\overline{C}_{i,t-1}\right)\right),\end{aligned} \tag{6}$$

where $\overline{C}_{i,t-1}$ is the exponential moving average of prediction concentration over previously accepted adaptation states. Since $C_{i,t}$ approaches zero as predictions become more concentrated, a larger positive deviation from $\overline{C}_{i,t-1}$ yields a larger restoration ratio $r_{i,t}$.

Based on $r_{i,t}$, risk-guided selective layer restoration orders the adaptable layers from shallow to deep and selects the deepest $r_{i,t}$ proportion as the restoration set $\mathcal{L}_{i,t}^{\mathrm{rec}}$. The local restoration candidate is constructed as follows:

$$\theta_{i,t}^{\mathrm{loc},(l)} = \begin{cases}\theta^{0,(l)}, & l \in \mathcal{L}_{i,t}^{\mathrm{rec}},\\ \theta_{i,t}^{\mathrm{adapt},(l)}, & l \notin \mathcal{L}_{i,t}^{\mathrm{rec}},\end{cases} \tag{7}$$

where $\theta^{0,(l)}$ denotes the source-model parameters of layer l.

When eligible peer states are available, the mechanism first constructs the weighted peer state $\overline{\theta}_{i,t}^{\mathrm{peer},(l)} = \sum_{j\in\mathcal{P}_{i,t}^{K_p}} w_{ij}\theta_j^{C,(l)}$. The corresponding cross-client restoration candidate is then constructed as follows:

$$\theta_{i,t}^{\mathrm{col},(l)} = \begin{cases}\beta_{i,t}\theta^{0,(l)} + (1-\beta_{i,t})\overline{\theta}_{i,t}^{\mathrm{peer},(l)}, & l \in \mathcal{L}_{i,t}^{\mathrm{rec}},\\ \theta_{i,t}^{\mathrm{loc},(l)}, & l \notin \mathcal{L}_{i,t}^{\mathrm{rec}}.\end{cases} \tag{8}$$

where $\beta_{i,t} = 0.25 + 0.5r_{i,t}$ controls the contribution of the source model and peer states within the restored layers. Since $r_{i,t} \in [0,1]$, $\beta_{i,t} \in [0.25, 0.75]$, such that more severe degradation assigns a larger contribution to the source model.

When eligible peers are available, FedART evaluates the continuation, local restoration, cross-client restoration, and rollback states using an unlabeled peer-consensus reference. A recovery candidate is committed only if it improves upon the continuation state by at least δ_{commit}; otherwise, the continuation state is retained. Details are provided in Appendix A.1.2.

4 Experiments

4.1 Experimental Setup

Setup. We evaluate FedART on CIFAR-10-C and CIFAR-100-C (Hendrycks & Dietterich, 2019), PACS (Li et al., 2017), and OfficeHome (Venkateswara et al., 2017) under long-term federated TTA. We compare with FedTHE+ (Jiang & Lin, 2023), ATP (Bao et al., 2023), FedTSA (Zhang et al., 2024), CoLA (Chen et al., 2024), FedCTTA (Rajib et al., 2025), and FedSPL (Liang et al., 2025a) under the same evaluation protocol. ResNet-18 is used as the backbone, with only BatchNorm affine parameters adapted and communicated during deployment. Target labels are unavailable during adaptation and used only for evaluation. Detailed experimental settings are provided in the Appendix.

4.2 Main Results

Overall Adaptation Performance. To evaluate whether FedART can maintain reliable performance over long-term federated TTA, we compare it with representative baselines across four benchmarks in Table 1. FedART achieves the highest mean online accuracy across four benchmarks, showing consistent performance across both continual distribution shifts and cross-domain shifts. The advantage is particularly pronounced on CIFAR-10-C and CIFAR-100-C, where FedART exceeds the strongest baseline by 0.84 and 1.40 percentage points, respectively. On PACS and OfficeHome, FedART also

Method	CIFAR-10-C		CIFAR-100-C		PACS		OfficeHome	
	Online	Worst	Online	Worst	Online	Worst	Online	Worst
FedTHE+	39.83±6.86	20.04	11.57±1.11	3.97	58.76±19.42	27.28	11.85±6.18	5.00
ATP	48.56±2.15	29.22	37.02±0.29	16.62	65.82±21.89	38.43	8.82±3.99	2.49
FedTSA	47.98±2.31	34.57	43.72±0.70	33.77	55.87±10.65	33.70	49.90±11.01	43.87
CoLA	33.29±8.79	15.86	10.79±1.35	3.99	54.95±25.97	33.12	10.61±6.55	4.59
FedCTTA	49.76±0.61	35.83	40.74±0.19	23.70	68.31±13.23	51.38	54.83±11.26	47.94
FedSPL	49.93±2.66	29.74	37.59±0.30	17.62	76.10±13.35	**59.60**	58.95±12.58	51.78
FedART(ours)	**50.77±0.63**	**38.75**	**45.12±0.29**	**33.80**	**76.34±12.93**	52.54	**58.96±11.22**	**53.13**

Table 1: Comparison of long-term adaptation performance with federated TTA baselines across four benchmarks. Online and Worst denote mean online and worst-client accuracy (%), respectively. Best results are shown in bold, and second-best results are underlined.

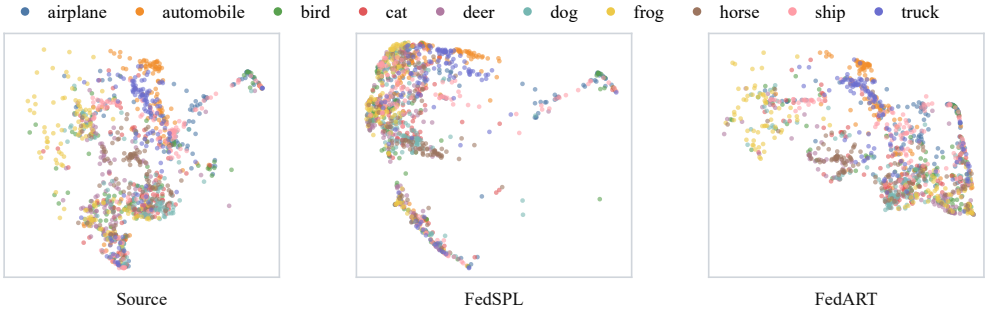


Figure 3: Feature representations after long-term adaptation on CIFAR-10-C. Colors denote ground-truth classes and are used only for offline visualization. The plots compare the representation structures of adapted models with that of the source model.

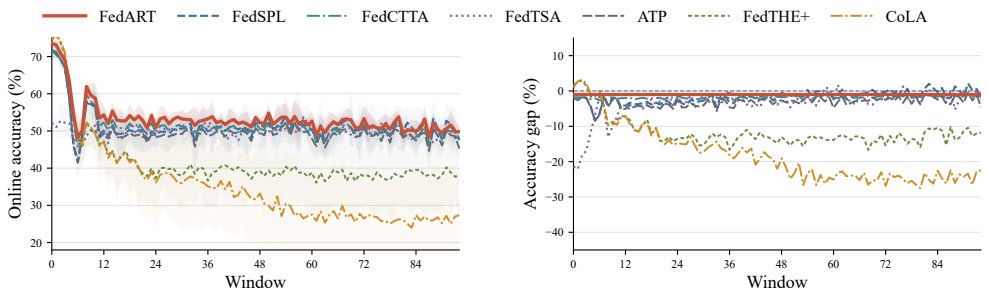
achieves the best mean online accuracy. Beyond average performance, FedART remains strong under client-specific shifts, achieving the highest worst-client accuracy on three of the four benchmarks and ranking second on PACS. These results show that FedART maintains reliable long-term adaptation across distribution shifts, with its clearest gains under continual distribution shifts.

Temporal Behavior under Continual Distribution Shifts. To examine whether the overall performance gains are sustained during long-term adaptation rather than driven by a few favorable windows, Fig. 4 tracks the adaptation trajectory on CIFAR-10-C. After the initial performance drop, FedART quickly stabilizes and maintains strong accuracy across most subsequent windows. This trend is further reflected in Fig. 4(b), where FedART consistently maintains a favorable gap against the competing methods as adaptation proceeds, while several baselines exhibit progressively larger performance deficits. These results show that FedART preserves its performance advantage through the long-term adaptation process, rather than relying on isolated high-performing windows.

Feature Representation after Long-Term Adaptation. To further examine how long-term adaptation affects the learned representation, Fig. 3 visualizes the feature space after the full adaptation stream on CIFAR-10-C. Compared with FedSPL, FedART exhibits a less fragmented feature structure and remains visually closer to the source representation. This observation is consistent with the temporal behavior in Fig. 4, where FedART maintains more stable performance during the adaptation process. Together, these results provide complementary qualitative evidence that FedART better preserves the representation structure during long-term adaptation.

4.3 Communication Efficiency

Accuracy-Communication Efficiency. To evaluate whether reliable long-term adaptation requires frequent cross-client communication, Table 2 compares the adaptation performance and communication overhead of FedART with representative federated TTA methods. Across all four benchmarks, FedART substantially reduces communication while maintaining the highest mean online accuracy. Compared with FedSPL, FedART reduces cross-client communication by 96.43%-99.56%, while the reduction relative to FedCTTA ranges from 91.06% to 98.85%. The contrast is particularly clear on PACS and OfficeHome, where FedART achieves comparable or slightly higher accuracy than FedSPL while reducing communication from about 40 MiB to only 0.220 and 0.182 MiB, respectively. Fig. 5



(a) Online accuracy under continual shifts (b) Window-wise accuracy gap to FedART

Figure 4: Temporal adaptation performance on CIFAR-10-C over sequential test windows. (a) Online accuracy of different federated TTA methods; (b) accuracy gap relative to FedART, where negative values indicate lower performance than FedART.

Method	CIFAR-10-C		CIFAR-100-C		PACS		OfficeHome	
	Acc.	Comm.	Acc.	Comm.	Acc.	Comm.	Acc.	Comm.
FedSPL	49.93±2.66	39.449	37.59±0.30	39.673	76.10±13.35	39.442	58.95±12.58	40.979
FedCTTA	49.76±0.61	15.770	40.74±0.19	15.844	68.31±13.23	15.767	54.83±11.26	15.815
FedART(ours)	**50.77±0.63**	**1.144**	**45.12±0.29**	**1.416**	**76.34±12.93**	**0.220**	**58.96±11.22**	**0.182**

Table 2: Accuracy-communication trade-off under long-term federated TTA across four benchmarks. Acc. denotes mean online accuracy (%), and Comm. denotes cross-client communication volume in MiB. Best results are shown in bold.

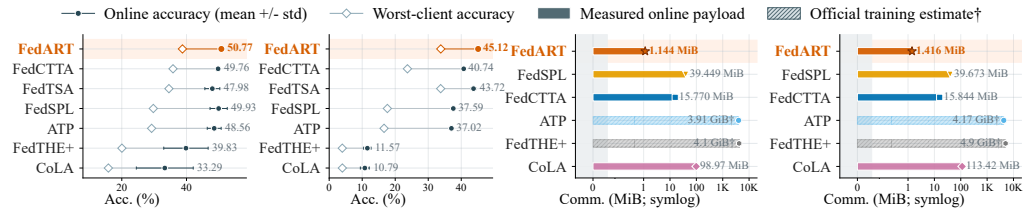
Variant	CIFAR-10-C				CIFAR-100-C			
	Easy	Medium	Hard	Avg.	Easy	Medium	Hard	Avg.
Full FedART	**64.93**	**56.12**	**50.30**	**57.12**	**56.46**	**45.20**	**36.19**	**45.95**
w/o Risk Monitoring	62.34	54.13	49.40	55.29	53.99	43.81	35.99	44.60
w/o Asynchronous Collaboration	62.88	55.23	48.60	55.57	55.57	44.47	35.58	45.21
w/o Selective Layer Restoration	63.07	54.75	48.76	55.53	55.79	44.95	36.08	45.61

Table 3: Ablation study of the three core recovery mechanisms in FedART under different severity levels on CIFAR-10-C and CIFAR-100-C. Accuracy is reported for Easy, Medium, and Hard severity settings, with the final column representing the equally weighted average across three severity levels.

further shows that the same efficiency advantage holds on CIFAR-10-C and CIFAR-100-C against a broader set of federated TTA baselines. These results show that FedART maintains strong long-term adaptation performance with substantially lower cross-client communication, yielding a favorable accuracy-communication balance across different distribution shifts.

4.4 Ablation Study

Contribution of the Core Mechanisms. To examine whether FedART requires the coordinated design of its three recovery mechanisms, Table 3 reports leave-one-component-out results under different severity levels. A consistent pattern emerges across both benchmarks: removing any component reduces accuracy under all Easy, Medium, and Hard settings, showing that each component contributes to the overall recovery performance. The most pronounced average degradation occurs when Local Adaptation Risk Monitoring is removed, with accuracy decreasing from 57.12% to 55.29% on CIFAR-10-C and from 45.95% to 44.60% on CIFAR-100-C, corresponding to drops of 1.83 and 1.35 percentage points, respectively. This result highlights the importance of reliably identifying when the current adaptation state requires recovery. Meanwhile, removing either risk-triggered asynchronous collaboration or risk-guided selective layer restoration consistently degrades performance across severity levels, showing that detecting degradation alone is insufficient: the subsequent recovery process also benefits from cross-client information and an adaptive restoration scope. Taken together, these ablations support the joint recovery design of FedART, where reliable triggering determines when recovery is needed, while asynchronous collaboration and selective restoration jointly determine how recovery is carried out.



(a) CIFAR-10-C: Acc. (b) CIFAR-100-C: Acc. (c) CIFAR-10-C: Comm. (d) CIFAR-100-C: Comm.

Figure 5: Performance and communication overhead on CIFAR-10-C and CIFAR-100-C. Panels (a-b) report online and worst-client accuracy, while panels (c-d) report communication volume.

Setting	Level	Online Acc. (%)		Worst Acc. (%)		Comm. (MiB)		Red. (%)
		FedSPL	FedART(ours)	FedSPL	FedART(ours)	FedSPL	FedART(ours)	
CIFAR-10-C								
Default	-	49.93	50.77	29.74	38.75	39.449	1.144	97.10
Dirichlet α	0.1	46.28	49.29	27.20	29.94	40.770	1.228	96.99
Dirichlet α	1.0	54.19	54.49	41.47	42.13	39.840	1.162	97.08
Availability	0.6	50.25	50.91	36.03	36.30	37.132	0.834	97.75
Availability	1.0	49.33	49.47	34.99	35.37	42.760	1.492	96.51
Clients	4	51.99	52.26	45.10	46.31	18.958	0.605	96.81
Clients	16	48.83	49.21	28.23	28.69	76.833	1.969	97.44
Max staleness	1	49.26	49.93	34.83	35.91	39.443	1.358	96.56
Max staleness	5	48.82	50.34	36.25	36.62	39.334	1.631	95.85
CIFAR-100-C								
Default	-	37.59	45.12	17.62	33.80	39.673	1.416	96.43
Dirichlet α	0.1	36.74	43.98	22.69	23.26	40.844	1.672	95.91
Dirichlet α	1.0	38.93	45.95	22.19	23.44	39.423	1.379	96.50
Availability	0.6	36.94	46.13	23.88	24.52	28.132	0.851	96.97
Availability	1.0	35.56	44.88	23.89	24.30	42.745	1.509	96.47
Clients	4	42.63	46.76	24.80	25.84	19.996	0.757	96.21
Clients	16	41.27	46.06	20.50	20.73	77.983	2.612	96.65
Max staleness	1	37.18	44.03	23.70	23.91	39.174	1.396	96.44
Max staleness	5	39.44	45.73	22.83	23.80	39.236	1.413	96.40

Table 4: Robustness analysis under varying data heterogeneity, client availability, federation size, and peer-state staleness. Online and Worst denote mean online and worst-client accuracy (%), respectively; Red. denotes communication reduction relative to FedSPL.

4.5 Robustness Analysis

Robustness under Different Federated Conditions. Table 4 evaluates FedART under varying data heterogeneity, client availability, federation size, and peer-state staleness. Across all settings, FedART outperforms FedSPL while reducing cross-client communication by 95.85%-97.75%. Under stronger data heterogeneity ($\alpha = 0.1$), FedART improves online accuracy by 3.01 and 7.24 percentage points on CIFAR-10-C and CIFAR-100-C, respectively. Its advantage remains stable across variations in client availability, federation size, and peer-state staleness. Overall, FedART maintains its accuracy-communication advantage across diverse federated conditions.

5 Conclusion

This work investigates long-term federated test-time adaptation under continual distribution shifts, where persistent local adaptation can accumulate errors while periodic, fixed-scope recovery is both communication-intensive and disruptive. We propose FedART, a risk-guided asynchronous recovery framework that uses client-specific adaptation risk to determine when recovery is needed and how much of the model should be restored. By coupling risk monitoring with on-demand cross-client collaboration and severity-adaptive layer restoration, FedART corrects degraded states while preserving useful local adaptations. Across four benchmarks, FedART maintains strong adaptation performance while substantially reducing recovery communication across varying federated conditions. These results suggest that reliable federated TTA does not require frequent uniform recovery, but rather recovery that is timely and proportional to the degradation.

AI USE STATEMENT

In this work, we did not use generative AI tools for any tasks with required disclosure. We have not used generative AI tools for generating synthetic datasets, developing theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, proposing or refining hypotheses, designing or providing feedback on research methodology or experiments, implementing methods, assisting with translation, cleaning and reformatting datasets, supporting qualitative and thematic data analysis, or interpreting results. No other required-disclosure tasks are applicable to this work.

Additionally, we used generative AI tools for a task with recommended disclosure: editing the research paper to improve readability, i.e., language polishing. The tools were used only to suggest improvements to grammar, wording, clarity, and style. They were not used to generate or alter the scientific content, research ideas, hypotheses, data, code, analyses, figures, references, or conclusions. We have reviewed all AI-assisted work. All AI-assisted edits were checked and revised by the authors to ensure accuracy, originality, and fidelity to the authors' intended meaning. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

Wenxuan Bao, Tianxin Wei, Haohan Wang, and Jingrui He. Adaptive test-time personalization for federated learning. In *Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS 2023), New Orleans, LA, USA, December 10–16, 2023*, volume 36 of *Advances in Neural Information Processing Systems*, 2023.

Guohao Chen, Shuaicheng Niu, Deyu Chen, Shuhai Zhang, Changsheng Li, Yuanqing Li, and Mingkui Tan. Cross-device collaborative test-time adaptation. In *Proceedings of the 38th Conference on Neural Information Processing Systems (NeurIPS 2024), Vancouver, BC, Canada, December 10–15, 2024*, volume 37 of *Advances in Neural Information Processing Systems*, 2024.

Dan Hendrycks and Thomas G. Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *Proceedings of the 7th International Conference on Learning Representations (ICLR 2019), New Orleans, LA, USA, May 6–9, 2019*. OpenReview.net, 2019.

Liangze Jiang and Tao Lin. Test-time robust personalization for federated learning. In *Proceedings of the 11th International Conference on Learning Representations (ICLR 2023), Kigali, Rwanda, May 1–5, 2023*. OpenReview.net, 2023.

Ellango Jothimurugesan, Kevin Hsieh, Jianyu Wang, Gauri Joshi, and Phillip B. Gibbons. Federated learning under distributed concept drift. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS 2023), Valencia, Spain, April 25–27, 2023*, volume 206 of *Proceedings of Machine Learning Research*, pp. 5834–5853. PMLR, 2023.

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank J. Reddi, Sebastian U. Stich, and Ananda Theertha Suresh. SCAFFOLD: Stochastic controlled averaging for federated learning. In *Proceedings of the 37th International Conference on Machine Learning (ICML 2020), Virtual Event, July 13–18, 2020*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5132–5143. PMLR, 2020.

Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M. Hospedales. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV 2017), Venice, Italy, October 22–29, 2017*, pp. 5543–5551. IEEE Computer Society, 2017.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. In *Proceedings of the 3rd Conference on Machine Learning and Systems (MLSys 2020), Austin, TX, USA, March 2–4, 2020*. MLSys, 2020.

Tian Li, Shengyuan Hu, Ahmad Beirami, and Virginia Smith. Ditto: Fair and robust federated learning through personalization. In *Proceedings of the 38th International Conference on Machine Learning (ICML 2021), Virtual Event, July 18–24, 2021*, volume 139 of *Proceedings of Machine Learning Research*, pp. 6357–6368. PMLR, 2021.

Haoyuan Liang, Xinyu Zhang, Shilei Cao, Guowen Li, and Juepeng Zheng. TTA-FedDG: Leveraging test-time adaptation to address federated domain generalization. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence (AAAI 2025), Philadelphia, PA, USA, February 25–March 4, 2025*, pp. 18658–18666. AAAI Press, 2025a.

Jian Liang, Ran He, and Tieniu Tan. A comprehensive survey on test-time adaptation under distribution shifts. *International Journal of Computer Vision*, 133(1):31–64, 2025b.

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 (AISTATS 2017), Fort Lauderdale, FL, USA, April 20–22, 2017*, volume 54 of *Proceedings of Machine Learning Research*, pp. 1273–1282. PMLR, 2017.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Shijian Zheng, Peilin Zhao, and Mingkui Tan. Efficient test-time model adaptation without forgetting. In *Proceedings of the 39th International Conference on Machine Learning (ICML 2022), Baltimore, MD, USA, July 17–23, 2022*, volume 162 of *Proceedings of Machine Learning Research*, pp. 16888–16905. PMLR, 2022.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Zhiquan Wen, Yaofo Chen, Peilin Zhao, and Mingkui Tan. Towards stable test-time adaptation in dynamic wild world. In *Proceedings of the 11th International Conference on Learning Representations (ICLR 2023), Kigali, Rwanda, May 1–5, 2023*. OpenReview.net, 2023.

Rakibul Hasan Rajib, Md. Akil Raihan Iftee, Mir Sazzat Hossain, A. K. M. Mahbubur Rahman, Sajib Mistry, M. Ashraful Amin, and Amin Ahsan Ali. FedCTTA: A collaborative approach to continual test-time adaptation in federated learning. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN 2025), Rome, Italy, June 30–July 5, 2025*, pp. 1–8. IEEE, 2025.

Aviv Shamsian, Aviv Navon, Ethan Fetaya, and Gal Chechik. Personalized federated learning using hypernetworks. In *Proceedings of the 38th International Conference on Machine Learning (ICML 2021), Virtual Event, July 18–24, 2021*, volume 139 of *Proceedings of Machine Learning Research*, pp. 9489–9502. PMLR, 2021.

Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017), Honolulu, HI, USA, July 21–26, 2017*, pp. 5385–5394. IEEE Computer Society, 2017.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno A. Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *Proceedings of the 9th International Conference on Learning Representations (ICLR 2021), Virtual Event, May 3–7, 2021*. OpenReview.net, 2021.

Jianyu Wang, Qinghua Liu, Hao Liang, Gauri Joshi, and H. Vincent Poor. Tackling the objective inconsistency problem in heterogeneous federated optimization. In *Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS 2020), Virtual Event, December 6–12, 2020*, volume 33 of *Advances in Neural Information Processing Systems*, 2020.

Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. Continual test-time domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2022), New Orleans, LA, USA, June 18–24, 2022*, pp. 7191–7201. IEEE, 2022.

Longhui Yuan, Binhui Xie, and Shuang Li. Robust test-time adaptation in dynamic scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2023), Vancouver, BC, Canada, June 17–24, 2023*, pp. 15922–15932. IEEE, 2023.

Chen Zhang, Ge Su, Huaxiao Zhou, Lu Hao, Feng Zhu, Jintai Chen, Yuxiang Cai, Jianwei Yin, and Hongxia Xu. Federated learning meets test-time adaptation: Methods, challenges, and future directions. *International Journal of Computer Vision*, 134(8):361, 2026.

Jiayuan Zhang, Xuefeng Liu, Yukang Zhang, Guogang Zhu, Jianwei Niu, and Shaojie Tang. Enabling collaborative test-time adaptation in dynamic environment via federated learning. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2024), Barcelona, Spain, August 25–29, 2024*, pp. 4191–4202. Association for Computing Machinery, 2024.

A Appendix

A.1 Algorithmic Details

A.1.1 FedART Online Adaptation Procedure

Algorithm 1 summarizes the complete online procedure of FedART. For each incoming target batch, FedART first evaluates the current adaptation risk. Recovery is triggered only when sufficient degradation evidence is detected, after which eligible peer states are retrieved asynchronously and the recovery scope is determined according to degradation severity.

Algorithm 1: FedART Online Adaptation with Risk-Guided Asynchronous Recovery

Require: Source model θ^0, clients $\{1, \dots, K\}$, local TTA operator $\mathcal{T}$, source calibration sets $\{\mathcal{D}_i^0\}_{i=1}^K$, risk tolerance ϵ_i, maximum staleness $S_{\max}$, similarity threshold γ, top-K_p peers, commit margin δ_{commit}.
Ensure: Updated working states $\{\theta_i^W\}$ and committed states $\{\theta_i^C\}$.

for each client i **do**
  Initialize $\theta_i^W \leftarrow \theta^0$ and $\theta_i^C \leftarrow \theta^0$.
  Calibrate τ_i^0, $U_{R,i}^0$, and $U_{F,i}^0$ using $\mathcal{D}_i^0$.
  Initialize risk monitor $\mathcal{M}_i$, concentration baseline $\bar{C}_i$, and registry entry.
end for
for each asynchronous event (i, t) with $\mathcal{B}_{i,t}$ **do**
  Update uncertainty threshold $\tau_{i,t}$ if required.
  Compute uncertainty exceedances and update $\mathcal{M}_i$.
  Obtain $L_{R,i,t}$ and $A_{i,t} \leftarrow \mathbb{I}[L_{R,i,t} > U_{R,i}^0 + \epsilon_i]$.
  Compute continuation state $\theta_{i,t}^{\text{cont}} \leftarrow \mathcal{T}(\theta_{i,t-1}^W, \mathcal{B}_{i,t})$ and set $\theta_{i,t}^{\text{adapt}} \leftarrow \theta_{i,t}^{\text{cont}}$.
  Compute prediction signature $s_{i,t}$.
  if $A_{i,t} = 0$ **then**
    $\theta_i^W, \theta_i^C \leftarrow \theta_{i,t}^{\text{cont}}$ and update its lightweight registry metadata.
  else
    Revoke the temporary sharing eligibility of client i.
    Retrieve eligible peers $\mathcal{P}_{i,t}$ using Eq. (4).
    Rank peers and compute weights $\{w_{ij}\}$ using Eq. (5).
    Compute concentration $C_{i,t}$ and restoration ratio $r_{i,t}$ using Eq. (6).
    Construct local recovery candidate $\theta_{i,t}^{\text{loc}}$ using Eq. (7).
    Set rollback candidate $\theta_{i,t}^{\text{rb}} \leftarrow \theta_{i,t-1}^C$.
    if $\mathcal{P}_{i,t} \neq \emptyset$ **then**
      Construct cross-client candidate $\theta_{i,t}^{\text{col}}$ using Eq. (8).
      Compute peer consensus and score all candidates.
      Select the best candidate satisfying the commit-margin criterion.
    else
      $\theta_{i,t}^* \leftarrow \theta_{i,t}^{\text{cont}}$.
    end if
    $\theta_i^W, \theta_i^C \leftarrow \theta_{i,t}^*$ and update the registry.
  end if
end for

A.1.2 Unlabeled Candidate Evaluation

Since target labels are unavailable during deployment, a restoration candidate cannot be selected according to its true prediction accuracy. FedART therefore uses the predictions of eligible peer states as an unlabeled reference and compares each recovery candidate against the continuation state before commitment. Algorithm 2 summarizes this procedure.

Algorithm 2: Unlabeled Candidate Evaluation and Conservative Commit

Require: Current target batch $\mathcal{B}_{i,t}$, continuation state $\theta_{i,t}^{\text{cont}}$, local restoration state $\theta_{i,t}^{\text{loc}}$, cross-client restoration state $\theta_{i,t}^{\text{col}}$, previous committed state $\theta_{i,t-1}^{C}$, eligible peer set $\mathcal{P}_{i,t}^{K_p}$, peer weights $\{w_{ij}\}$, and commit margin δ_{commit}.
Ensure: Selected state $\theta_{i,t}^{*}$.

if $\mathcal{P}_{i,t}^{K_p} = \emptyset$ **then**
  Keep the continuation state: $\theta_{i,t}^{*} \leftarrow \theta_{i,t}^{\text{cont}}$.
  Update $\theta_{i,t}^{W}, \theta_{i,t}^{C} \leftarrow \theta_{i,t}^{*}$.
  return $\theta_{i,t}^{*}$.
end if
Build the unlabeled peer reference from eligible peer predictions.
Score all candidate states by agreement with the peer reference.
Select the highest-scoring recovery candidate.
if its score exceeds the continuation state by at least δ_{commit} **then**
  Commit the selected recovery candidate as $\theta_{i,t}^{*}$.
else
  Keep the continuation state: $\theta_{i,t}^{*} \leftarrow \theta_{i,t}^{\text{cont}}$.
end if
Update $\theta_{i,t}^{W}, \theta_{i,t}^{C} \leftarrow \theta_{i,t}^{*}$.
return $\theta_{i,t}^{*}$.

A.2 Experimental Details

A.2.1 Datasets and Protocol

We evaluate FedART on CIFAR-10-C and CIFAR-100-C (Hendrycks & Dietterich, 2019), PACS (Li et al., 2017), and OfficeHome (Venkateswara et al., 2017) under a unified long-term federated TTA protocol. CIFAR-10-C and CIFAR-100-C provide continual corruption streams, while PACS and OfficeHome introduce cross-domain distribution shifts, allowing us to evaluate FedART under two different forms of deployment-time distribution change. Each target stream contains 96 sequential windows with $K = 8$ clients. Client availability is set to 0.85, and client-level non-IID data are generated using a Dirichlet distribution with $\alpha = 0.3$. We use five stream seeds and three source-model seeds, cyclically assigned across runs.

A.2.2 Baselines and Implementation

We compare FedART with FedTHE+ (Jiang & Lin, 2023), ATP (Bao et al., 2023), FedTSA (Zhang et al., 2024), CoLA (Chen et al., 2024), FedCTTA (Rajib et al., 2025), and FedSPL (Liang et al., 2025a) under the same source models, client trajectories, and evaluation protocol. ResNet-18 is used as the backbone, and only BatchNorm affine parameters are adapted and communicated during deployment. Target labels remain unavailable throughout online adaptation and are used only for offline evaluation.

For local adaptation risk monitoring, the initial uncertainty threshold is selected on the labeled source calibration set by searching candidate thresholds and maximizing the F1 score of uncertainty exceedance against prediction errors, with recall and the smaller threshold used for tie breaking. The source-risk and proxy false-positive upper bounds are computed from their empirical calibration rates with Hoeffding-type penalties.

A.2.3 Hyperparameter Settings

The default hyperparameters used in our experiments are summarized in Table 5.

Parameter	Value	Description
α_{sys}	0.2	System-level confidence budget
ϵ_i	0.05	Tolerated risk increase
$r_{\min}$	0.5	Minimum restoration ratio
λ_r	20.0	Restoration-ratio scaling factor
$r_{\max}$	1.0	Maximum restoration ratio
$S_{\max}$	3	Maximum peer-state staleness
γ	0.2	Minimum prediction similarity
K_p	3	Maximum selected peer states
δ_{commit}	0.0	Candidate commit margin

Table 5: Default hyperparameter settings used in the experiments.

A.2.4 Evaluation Metrics

We report mean online accuracy with sample standard deviation to measure overall long-term adaptation performance, and worst-client accuracy to assess performance under unfavorable client-specific shifts. Communication volume is measured in MiB to quantify cross-client recovery overhead.

A.2.5 Computing Environment

All experiments were conducted on a server with the following hardware and software configurations:

- **GPU:** NVIDIA A800 80GB PCIe, with 80GB GPU memory and compute capability 8.0.
- **CPU:** Intel Xeon Platinum 8358 @ 2.60GHz, with 2 sockets, 64 physical cores, and 128 logical threads.
- **System Memory:** Approximately 1.0 TiB.
- **Operating System:** Ubuntu 22.04.3 LTS (Linux kernel 5.15.0-60-generic, x86_64).
- **Software Environment:** Python 3.10.20, PyTorch 2.7.1+cu118, torchvision 0.22.1+cu118, CUDA 11.8, and cuDNN 9.1.0.
- **Main Libraries:** NumPy 2.2.6, SciPy 1.15.3, scikit-learn 1.7.2, and Pillow 12.2.0.

A.3 Theoretical Analysis

A.3.1 Statistical Analysis of Sequential Risk Monitoring

We provide a statistical analysis of the sequential risk monitoring mechanism introduced in Section 3.3. Because FedART allows both the working model and the uncertainty threshold to evolve during deployment, the observations used by the risk monitor are generally adaptive rather than independent. We therefore formulate the monitoring process conditionally on the deployment history.

The analysis establishes three properties. First, the empirical uncertainty exceedance rate yields a valid lower confidence bound on the latent prediction risk under an explicit proxy-validity condition. Second, the confidence budget can be allocated across multiple monitor starts and sequential checks without requiring independence among overlapping monitoring sequences. Third, the resulting recovery alarm admits a system-wide false-trigger interpretation relative to the calibrated source-risk level.

Conditional risk formulation. Consider client i and the s-th monitoring sequence started during deployment. At its m-th sequential check, let $M_{i,s,m}$ denote the number of target samples observed by this sequence. For its n-th observed sample, define the uncertainty-exceedance indicator as

$$z_{i,s}^{(n)} = \mathbb{I}\left[u\left(x_{i,s}^{(n)}; \theta_{i,s,n-1}^{W}\right) > \tau_{i,s,n}\right],$$

where $\theta_{i,s,n-1}^{W}$ and $\tau_{i,s,n}$ denote the working model and uncertainty threshold available immediately before the current observation.

The empirical exceedance rate is $\widehat{q}_{i,s,m} = \frac{1}{M_{i,s,m}} \sum_{n=1}^{M_{i,s,m}} z_{i,s}^{(n)}$.

For theoretical analysis, define the latent 0-1 prediction error

$$\ell_{i,s}^{(n)} = \mathbb{I}\left[\arg\max_{c} p_{\theta^W_{i,s,n-1}}\left(c \mid x_{i,s}^{(n)}\right) \neq y_{i,s}^{(n)}\right].$$

The target label $y_{i,s}^{(n)}$ is unavailable to FedART during deployment and is introduced only to characterize the latent prediction risk.

Let $\mathcal{F}_{i,s,n-1}$ denote the deployment history available immediately before observing $x_{i,s}^{(n)}$. Both $\theta^W_{i,s,n-1}$ and $\tau_{i,s,n}$ are measurable with respect to $\mathcal{F}_{i,s,n-1}$. Therefore, they may depend on the complete previous adaptation history while remaining fixed before the current observation.

Define

$$\begin{aligned}
\mu_{i,s}^{(n)} &= \mathbb{E}\left[z_{i,s}^{(n)} \mid \mathcal{F}_{i,s,n-1}\right], \\
r_{i,s}^{(n)} &= \mathbb{E}\left[\ell_{i,s}^{(n)} \mid \mathcal{F}_{i,s,n-1}\right], \\
f_{i,s}^{(n)} &= \mathbb{E}\left[z_{i,s}^{(n)}\left(1-\ell_{i,s}^{(n)}\right) \mid \mathcal{F}_{i,s,n-1}\right].
\end{aligned}$$

Their averages up to the m-th sequential check are

$$\begin{aligned}
\overline{\mu}_{i,s,m} &= \frac{1}{M_{i,s,m}} \sum_{n=1}^{M_{i,s,m}} \mu_{i,s}^{(n)}, \\
R_{i,s,m} &= \frac{1}{M_{i,s,m}} \sum_{n=1}^{M_{i,s,m}} r_{i,s}^{(n)}, \\
F_{i,s,m} &= \frac{1}{M_{i,s,m}} \sum_{n=1}^{M_{i,s,m}} f_{i,s}^{(n)}.
\end{aligned}$$

For every observation, $z_{i,s}^{(n)} \leq \ell_{i,s}^{(n)} + z_{i,s}^{(n)}(1-\ell_{i,s}^{(n)})$. Taking conditional expectations and averaging over the monitoring sequence gives $R_{i,s,m} \geq \overline{\mu}_{i,s,m} - F_{i,s,m}$.

Assumption 1 (deployment validity of the uncertainty proxy). Let $U_{F,i}^0$ denote the conservative false-positive bound obtained from the labeled source calibration set before deployment. We assume that, for every active monitoring sequence and sequential check, $F_{i,s,m} \leq U_{F,i}^0$.

Assumption 1 does not require the target distribution to remain identical to the source distribution. The target prediction risk, uncertainty distribution, and model state may all evolve over time. The assumption only requires the false-positive component of the uncertainty-exceedance proxy–high uncertainty on correctly predicted samples–to remain bounded by the conservative calibration level used by FedART.

Under Assumption 1, the risk decomposition above gives $R_{i,s,m} \geq \overline{\mu}_{i,s,m} - U_{F,i}^0$.

Proposition 1 (conditional lower bound on adaptation risk). For any fixed client i, monitor start s, and sequential check m, let $\alpha_{i,s,m} \in (0,1)$ denote its confidence level. Define

$$L_{R,i,s,m} = \max\left(0,\, \widehat{q}_{i,s,m} - \sqrt{\frac{\log(1/\alpha_{i,s,m})}{2M_{i,s,m}}} - U_{F,i}^0\right).$$

Under Assumption 1,

$$\Pr\left[R_{i,s,m} < L_{R,i,s,m}\right] \leq \alpha_{i,s,m}.$$

Proof. $z_{i,s}^{(n)} \in \{0,1\}$ and $\mu_{i,s}^{(n)} = \mathbb{E}[z_{i,s}^{(n)} \mid \mathcal{F}_{i,s,n-1}]$, the sequence $z_{i,s}^{(n)} - \mu_{i,s}^{(n)}$ is a bounded martingale-difference sequence. The conditional Hoeffding inequality therefore gives

$$\Pr\left[\overline{\mu}_{i,s,m} < \widehat{q}_{i,s,m} - \sqrt{\frac{\log(1/\alpha_{i,s,m})}{2M_{i,s,m}}}\right] \leq \alpha_{i,s,m}.$$

Combining this inequality with the proxy-risk bound above yields, with probability at least $1-\alpha_{i,s,m}$,

$$R_{i,s,m} \geq \widehat{q}_{i,s,m} - \sqrt{\frac{\log(1/\alpha_{i,s,m})}{2M_{i,s,m}}} - U^0_{F,i}.$$

Finally, $R_{i,s,m} \geq 0$, so clipping the lower bound at zero preserves its validity.

Proposition 1 permits the working model and uncertainty threshold to depend arbitrarily on the previous deployment history. In particular, independence between sequential target observations is not required; the key requirement is that the model and threshold used for the current observation are determined before that observation is evaluated.

System-wide confidence allocation for multi-start monitoring. FedART may maintain multiple monitoring sequences initialized at different times. To account for repeated checks, let $c = \frac{6}{\pi^2}$, so that $\sum_{k=1}^{\infty} \frac{c}{k^2} = 1$.

Let α_{sys} denote a system-level confidence budget. We allocate half of this budget to online sequential monitoring and half to calibration of the source-risk reference.

For client i, the online-monitoring budget is $\alpha_i^{\text{mon}} = \frac{\alpha_{\text{sys}}}{2K}$.

The s-th monitor start receives $\alpha_{i,s} = \alpha_i^{\text{mon}} c/s^2$, and the m-th sequential check within this monitor receives $\alpha_{i,s,m} = \alpha_{i,s} c/m^2$.

Equivalently,

$$\boxed{\alpha_{i,s,m} = \frac{\alpha_{\text{sys}}}{2K}\left(\frac{6}{\pi^2}\right)^2 \frac{1}{s^2 m^2}}.$$

Proposition 2 (system-wide validity of multi-start monitoring). Under Assumption 1,

$$\Pr\left[\exists\, i, s, m : R_{i,s,m} < L_{R,i,s,m}\right] \leq \frac{\alpha_{\text{sys}}}{2}.$$

Proof. Applying Proposition 1 and the union bound gives

$$\begin{aligned}
&\Pr\left[\exists\, i, s, m : R_{i,s,m} < L_{R,i,s,m}\right] \\
&\leq \sum_{i=1}^{K} \sum_{s=1}^{\infty} \sum_{m=1}^{\infty} \alpha_{i,s,m} \\
&= K \frac{\alpha_{\text{sys}}}{2K} \left(\frac{6}{\pi^2}\right)^2 \left(\sum_{s=1}^{\infty} \frac{1}{s^2}\right) \left(\sum_{m=1}^{\infty} \frac{1}{m^2}\right) \\
&= \frac{\alpha_{\text{sys}}}{2}.
\end{aligned}$$

No independence is required among overlapping monitor starts or sequential checks because the argument relies only on the marginal validity of Proposition 1 and the union bound.

Implication for recovery triggering. Let $\mathcal{S}_{i,t}$ denote the set of active monitor starts for client i at event t, and let $m_{i,s,t}$ denote the current sequential check index of monitor start s. FedART reports $L^{\max}_{R,i,t} = \max_{s \in \mathcal{S}_{i,t}} L_{R,i,s,m_{i,s,t}}$ and triggers recovery when $A_{i,t} = \mathbb{I}[L^{\max}_{R,i,t} > U^0_{R,i} + \epsilon_i]$.

Let R_i^0 denote the corresponding latent source prediction risk used as the pre-deployment reference. Suppose its calibrated upper confidence bound $U^0_{R,i}$ satisfies $\Pr[R_i^0 > U^0_{R,i}] \leq \alpha_{\text{sys}}/(2K)$.

Proposition 3 (system-wide validity of recovery alarms). Under Assumption 1, with probability at least $1-\alpha_{\text{sys}}$, every recovery alarm raised by FedART is supported by at least one active monitoring sequence whose latent prediction risk exceeds the calibrated source-risk level by more than ϵ_i. In particular,

$$A_{i,t} = 1 \quad \Longrightarrow \quad \exists\, s \in \mathcal{S}_{i,t} : R_{i,s,m_{i,s,t}} > R_i^0 + \epsilon_i$$

simultaneously across all clients and monitoring times, except on an event of probability at most α_{sys}.

Proof. Proposition 2 shows that all online risk lower bounds are simultaneously valid except on an event of probability at most $\alpha_{\text{sys}}/2$.

Applying a union bound to the source-calibration bound above over the K clients gives $\Pr[\exists\, i : R_i^0 > U_{R,i}^0] \le \alpha_{\text{sys}}/2$.

Therefore, except on an event of total probability at most α_{sys}, both $R_{i,s,m} \ge L_{R,i,s,m}$ for every active monitoring sequence and $R_i^0 \le U_{R,i}^0$ for every client.

If $A_{i,t} = 1$, then by the recovery-alarm rule above there exists an active monitor start s such that $L_{R,i,s,m_{i,s,t}} > U_{R,i}^0 + \epsilon_i$. Consequently,

$$R_{i,s,m_{i,s,t}} \ge L_{R,i,s,m_{i,s,t}} > U_{R,i}^0 + \epsilon_i \ge R_i^0 + \epsilon_i,$$

which proves the result.

Interpretation and limitation. The above analysis provides a statistical interpretation of FedART's risk-triggered recovery mechanism. The lower confidence bound remains valid under adaptive test-time updates, while the hierarchical confidence allocation accounts for repeated testing across overlapping monitoring sequences. As a result, a recovery alarm is not merely an increase in empirical uncertainty: under Assumption 1, it constitutes statistical evidence that the latent prediction risk of at least one active monitoring window exceeds the calibrated source-risk reference by more than the tolerated margin ϵ_i.

The guarantee is deliberately one-sided. It controls the validity of triggered alarms but does not guarantee that every degradation event will be detected immediately. Its validity also depends on Assumption 1, which makes explicit the required relationship between the uncertainty proxy and latent prediction errors under deployment shift.

A.3.2 Properties of Risk-Guided Selective Restoration

We next analyze the structural properties of the risk-guided selective restoration mechanism introduced in Section 3.5. Once the risk monitor activates recovery, FedART uses the deviation of prediction concentration from its historical baseline to characterize the severity of the detected degradation and determine the restoration scope. We show that this mechanism yields a bounded and monotonic restoration scope, introduces no larger parameter perturbation than full restoration, and progressively strengthens source anchoring as the restoration scope increases.

Severity-dependent restoration scope. Recall that FedART computes

$$r_{i,t} = \min\left(r_{\max}, r_{\min} + \lambda_r \max\left(0, C_{i,t} - \overline{C}_{i,t-1}\right)\right).$$

Define the concentration deviation as $\Delta C_{i,t} = C_{i,t} - \overline{C}_{i,t-1}$.

Since $C_{i,t} = \sum_c \widehat{p}_{i,t,c} \log \widehat{p}_{i,t,c}$ is the negative entropy of the batch-level predictive distribution, $C_{i,t}$ approaches zero as predictions become increasingly concentrated on fewer classes. Therefore, a larger positive $\Delta C_{i,t}$ indicates stronger prediction concentration relative to the historical baseline maintained over previously accepted adaptation states.

Proposition 4 (bounded and monotonic restoration scope). Assume $0 \le r_{\min} \le r_{\max} \le 1$ and $\lambda_r \ge 0$. Then $r_{\min} \le r_{i,t} \le r_{\max}$, and $r_{i,t}$ is monotonically non-decreasing with respect to $\Delta C_{i,t}$.

Proof. If $\Delta C_{i,t} \le 0$, then $\max(0, \Delta C_{i,t}) = 0$ and hence $r_{i,t} = r_{\min}$.

If $\Delta C_{i,t} > 0$, before the upper bound is reached, $r_{i,t} = r_{\min} + \lambda_r \Delta C_{i,t}$ with $\partial r_{i,t}/\partial \Delta C_{i,t} = \lambda_r \ge 0$. Once the value reaches $r_{\max}$, the outer minimum keeps the restoration ratio fixed at $r_{\max}$. Therefore, $r_{i,t}$ is bounded in $[r_{\min}, r_{\max}]$ and monotonically non-decreasing with respect to $\Delta C_{i,t}$.

Proposition 4 formalizes the role of the concentration deviation after a recovery event has been activated. Small or absent positive deviations lead to the minimum restoration scope, whereas stronger concentration progressively increases the fraction of restored layers until the predefined maximum is reached.

Parameter perturbation induced by selective restoration. Let $\mathcal{L}$ denote the complete set of adaptable layers, and let $\mathcal{L}_{i,t}^{\mathrm{rec}} \subseteq \mathcal{L}$ denote the deepest $r_{i,t}$ proportion selected by FedART. The local restoration candidate is

$$\theta_{i,t}^{\mathrm{loc},(l)} = \begin{cases} \theta^{0,(l)}, & l \in \mathcal{L}_{i,t}^{\mathrm{rec}}, \\ \theta_{i,t}^{\mathrm{adapt},(l)}, & l \notin \mathcal{L}_{i,t}^{\mathrm{rec}}. \end{cases}$$

Its squared deviation from the locally adapted state is

$$\left\| \theta_{i,t}^{\mathrm{loc}} - \theta_{i,t}^{\mathrm{adapt}} \right\|_2^2 = \sum_{l \in \mathcal{L}_{i,t}^{\mathrm{rec}}} \left\| \theta^{0,(l)} - \theta_{i,t}^{\mathrm{adapt},(l)} \right\|_2^2.$$

For comparison, let $\theta_{i,t}^{\mathrm{full}}$ denote full restoration of all adaptable layers, i.e., $\theta_{i,t}^{\mathrm{full},(l)} = \theta^{0,(l)}$ for $l \in \mathcal{L}$. Then

$$\left\| \theta_{i,t}^{\mathrm{full}} - \theta_{i,t}^{\mathrm{adapt}} \right\|_2^2 = \sum_{l \in \mathcal{L}} \left\| \theta^{0,(l)} - \theta_{i,t}^{\mathrm{adapt},(l)} \right\|_2^2.$$

Proposition 5 (perturbation bound of selective restoration). For any $\mathcal{L}_{i,t}^{\mathrm{rec}} \subseteq \mathcal{L}$,

$$\left\| \theta_{i,t}^{\mathrm{loc}} - \theta_{i,t}^{\mathrm{adapt}} \right\|_2 \leq \left\| \theta_{i,t}^{\mathrm{full}} - \theta_{i,t}^{\mathrm{adapt}} \right\|_2.$$

Moreover, for two nested restoration sets $\mathcal{L}_1 \subseteq \mathcal{L}_2$,

$$\left\| \theta^{\mathrm{rest}}(\mathcal{L}_1) - \theta_{i,t}^{\mathrm{adapt}} \right\|_2 \leq \left\| \theta^{\mathrm{rest}}(\mathcal{L}_2) - \theta_{i,t}^{\mathrm{adapt}} \right\|_2.$$

Proof. Each restored layer contributes the non-negative quantity $\|\theta^{0,(l)} - \theta_{i,t}^{\mathrm{adapt},(l)}\|_2^2$. Because $\mathcal{L}_{i,t}^{\mathrm{rec}} \subseteq \mathcal{L}$, the sum over the selected subset cannot exceed the corresponding sum over all adaptable layers. The same argument applies to any nested restoration sets $\mathcal{L}_1 \subseteq \mathcal{L}_2$.

Proposition 5 gives a direct parameter-space interpretation of selective restoration. Compared with full restoration, FedART modifies a subset of the adaptable parameters and therefore introduces no larger deviation from the current adapted state. As the restoration scope increases, the perturbation can increase only through the additional non-negative contributions of newly restored layers.

This result characterizes the restoration behavior in parameter space rather than target performance. The final decision to replace the continuation state is made separately by the unlabeled candidate-evaluation mechanism analyzed in Appendix A.3.3.

Severity-dependent source anchoring. When eligible peers are available, FedART constructs the weighted peer state $\overline{\theta}_{i,t}^{\mathrm{peer},(l)} = \sum_{j \in \mathcal{P}_{i,t}^{K_p}} w_{ij} \theta_j^{C,(l)}$ and defines the collaborative restoration candidate on the selected layers as

$$\theta_{i,t}^{\mathrm{col},(l)} = \beta_{i,t} \theta^{0,(l)} + (1 - \beta_{i,t}) \overline{\theta}_{i,t}^{\mathrm{peer},(l)}, \qquad l \in \mathcal{L}_{i,t}^{\mathrm{rec}},$$

where $\beta_{i,t} = 0.25 + 0.5 r_{i,t}$.

Proposition 6 (monotonic source anchoring). Let

$$\beta_{i,t} = 0.25 + 0.5 r_{i,t}.$$

Then $\beta_{i,t}$ is monotonically increasing with respect to $r_{i,t}$ and satisfies

$$0.25 \leq \beta_{i,t} \leq 0.75.$$

By Proposition 4, $r_{i,t}$ is monotonically non-decreasing with respect to the positive concentration deviation. Therefore, the source-model contribution within the restored layers increases monotonically with degradation severity.

Proof. Since

$$\frac{\partial \beta_{i,t}}{\partial r_{i,t}} = 0.5 > 0,$$

$\beta_{i,t}$ is monotonically increasing in $r_{i,t}$. Together with $0 \le r_{i,t} \le 1$, this gives $0.25 \le \beta_{i,t} \le 0.75$. Combining this result with Proposition 4 establishes the monotonic dependence of source anchoring on the positive concentration-deviation signal.

A.3.3 Unlabeled Consensus and Commit Guarantee

We next analyze the unlabeled candidate-evaluation and commit mechanism introduced in Section 3.5. When eligible peers are available, FedART constructs a weighted predictive reference from their committed states and evaluates each candidate according to its agreement with this reference. The continuation state serves as the baseline: a recovery candidate is committed only if it improves the consensus criterion by at least the predefined margin δ_{commit}.

The following analysis characterizes the consensus objective optimized by FedART during candidate selection.

Peer consensus distribution. Let $\mathcal{P}_{i,t}^{K_p}$ denote the selected eligible peers for client i at event t. Their normalized weights satisfy $w_{ij} \ge 0$ and $\sum_{j \in \mathcal{P}_{i,t}^{K_p}} w_{ij} = 1$.

For an unlabeled sample x, FedART defines $q_{i,t}(x) = \sum_{j \in \mathcal{P}_{i,t}^{K_p}} w_{ij} p_{\theta_j^C}(\cdot \mid x)$.

Proposition 7 (validity of the peer consensus reference). For every input x, $q_{i,t}(x)$ is a valid categorical probability distribution.

Proof. Each $p_{\theta_j^C}(\cdot \mid x)$ has non-negative entries summing to one. Since the peer weights are non-negative and normalized, $q_{i,t,c}(x) = \sum_{j \in \mathcal{P}_{i,t}^{K_p}} w_{ij} p_{\theta_j^C}(c \mid x) \ge 0$ for every class c, and $\sum_c q_{i,t,c}(x) = \sum_{j \in \mathcal{P}_{i,t}^{K_p}} w_{ij} \sum_c p_{\theta_j^C}(c \mid x) = 1$.

Consensus divergence and candidate score. For a candidate state ϑ, define its average divergence from the peer consensus on the current unlabeled target batch as

$$D_{i,t}(\vartheta) = \frac{1}{|\mathcal{B}_{i,t}|} \sum_{x \in \mathcal{B}_{i,t}} D_{\text{KL}}\left(q_{i,t}(x) \| p_\vartheta(\cdot \mid x)\right).$$

The score used by FedART is $S_{i,t}(\vartheta) = -D_{i,t}(\vartheta)$.

Proposition 8 (equivalence of score maximization and consensus matching). For every candidate state ϑ, $S_{i,t}(\vartheta) \le 0$ and $\arg\max_\vartheta S_{i,t}(\vartheta) = \arg\min_\vartheta D_{i,t}(\vartheta)$.

Furthermore, $S_{i,t}(\vartheta) = 0$ if and only if $p_\vartheta(\cdot \mid x) = q_{i,t}(x)$ for every $x \in \mathcal{B}_{i,t}$.

Proof. The KL divergence is non-negative, so $D_{i,t}(\vartheta) \ge 0$. Because $S_{i,t}(\vartheta) = -D_{i,t}(\vartheta)$, the score is upper-bounded by zero and maximizing the score is equivalent to minimizing the corresponding consensus divergence. Equality holds exactly when the KL divergence is zero for every sample in the evaluation batch.

Thus, the candidate score does not estimate target accuracy directly. It ranks available states according to predictive agreement with the weighted behavior of the selected eligible peers.

Commit rule relative to the continuation state. Let $\vartheta_{i,t}^*$ denote the highest-scoring non-continuation candidate among $\theta_{i,t}^{\text{loc}}$, $\theta_{i,t}^{\text{col}}$, and $\theta_{i,t-1}^C$.

FedART replaces the continuation state only if

$$S_{i,t}\left(\vartheta_{i,t}^*\right) \ge S_{i,t}\left(\theta_{i,t}^{\text{cont}}\right) + \delta_{\text{commit}},$$

where $\delta_{\text{commit}} \ge 0$. Otherwise, the continuation state is retained.

Proposition 9 (margin-certified consensus improvement). If a non-continuation candidate $\vartheta^*_{i,t}$ is committed, then

$$D_{i,t}\left(\vartheta^*_{i,t}\right) \leq D_{i,t}\left(\theta^{\mathrm{cont}}_{i,t}\right) - \delta_{\mathrm{commit}}.$$

Proof. From the commit rule above, $S_{i,t}(\vartheta^*_{i,t}) \geq S_{i,t}(\theta^{\mathrm{cont}}_{i,t}) + \delta_{\mathrm{commit}}$. Using $S_{i,t}(\vartheta) = -D_{i,t}(\vartheta)$ gives $-D_{i,t}(\vartheta^*_{i,t}) \geq -D_{i,t}(\theta^{\mathrm{cont}}_{i,t}) + \delta_{\mathrm{commit}}$, which yields the stated divergence-improvement bound.

Proposition 9 gives the operational interpretation of δ_{commit}. A recovery candidate is not accepted merely because it has the highest score among the recovery options; it must outperform the normal continuation state by a specified margin under the same unlabeled consensus criterion.

When $\delta_{\mathrm{commit}} > 0$, the committed recovery state must yield a strict reduction in consensus divergence. When $\delta_{\mathrm{commit}} = 0$, a recovery candidate can replace the continuation state only if its consensus divergence is no larger.

Proposition 10 (baseline preservation under the consensus objective). Let $\theta^*_{i,t}$ denote the state finally deployed by FedART. Whenever eligible peers are available,

$$S_{i,t}\left(\theta^*_{i,t}\right) \geq S_{i,t}\left(\theta^{\mathrm{cont}}_{i,t}\right),$$

or equivalently,

$$D_{i,t}\left(\theta^*_{i,t}\right) \leq D_{i,t}\left(\theta^{\mathrm{cont}}_{i,t}\right).$$

Proof. There are two possible cases.

If no non-continuation candidate satisfies the commit criterion, FedART retains $\theta^*_{i,t} = \theta^{\mathrm{cont}}_{i,t}$, and equality holds.

If a non-continuation candidate is committed, Proposition 9 gives

$$\begin{aligned} D_{i,t}\left(\theta^*_{i,t}\right) &\leq D_{i,t}\left(\theta^{\mathrm{cont}}_{i,t}\right) - \delta_{\mathrm{commit}} \\ &\leq D_{i,t}\left(\theta^{\mathrm{cont}}_{i,t}\right), \end{aligned}$$

because $\delta_{\mathrm{commit}} \geq 0$. Thus, the divergence-preservation inequality holds in both cases.

When no eligible peer is available, FedART directly retains $\theta^{\mathrm{cont}}_{i,t}$. Hence, recovery cannot replace the continuation state without an available consensus reference.

Interpretation and scope of the guarantee. Propositions 7-10 establish three properties of the unlabeled candidate selection mechanism.

First, the selected peers define a valid predictive reference distribution. Second, the candidate score ranks states according to their KL divergence from this reference. Third, the continuation state acts as an explicit baseline, so a recovery state is deployed only when it improves the consensus objective by the required margin.

Accordingly, the commit mechanism provides a *baseline-preserving guarantee with respect to the unlabeled peer-consensus objective*, which may differ from unknown target accuracy because peer consensus may itself be imperfect.

The role of the commit rule is therefore narrower but operationally important: even after the risk monitor has triggered recovery, FedART does not automatically overwrite the normal TTA trajectory. A recovery candidate must also pass an independent unlabeled evidence test before it is committed.

A.3.4 Communication and Computational Complexity

We finally analyze the communication and computational complexity of FedART. The key property of the proposed framework is that recovery-specific operations are activated only at detected degradation events rather than at every predefined communication interval.

In the current implementation, only BatchNorm affine parameters are adapted and exchanged during deployment. Let P_{BN} denote the number of such parameters.

Communication-cost decomposition. Let T denote the total number of online client events in a deployment stream, and let A denote the number of events at which the local risk monitor triggers recovery. We define the empirical recovery-event rate as $\rho = A/T$.

At the a-th triggered recovery event, let $d_a = |\mathcal{P}_{i,t}^{K_p}| \leq K_p$ denote the number of selected eligible peers.

It is useful to distinguish two communication components.

First, the system maintains a lightweight registry describing the latest committed client states through their timestamps, prediction signatures, risk information, and sharing status. Let $\mathcal{C}_{\mathrm{reg}}$ denote the communication required to maintain this registry over the complete stream. This term depends on the implementation and publication schedule and is separate from recovery-triggered peer retrieval.

Second, when a recovery event is triggered, FedART retrieves the selected peer states required by the recovery procedure. Ignoring constant-size control metadata, the recovery-specific model-state communication is

$$\mathcal{C}_{\mathrm{rec}}^{\mathrm{FedART}} = \mathcal{O}\left(\sum_{a=1}^{A}(d_a + c_{\mathrm{comm}})P_{\mathrm{BN}}\right),$$

where c_{comm} is a constant accounting for requester-side and returned recovery-state payloads.

Since $d_a \leq K_p$,

$$\mathcal{C}_{\mathrm{rec}}^{\mathrm{FedART}} = \mathcal{O}\left(A(K_p+1)P_{\mathrm{BN}}\right) = \mathcal{O}\left(AK_pP_{\mathrm{BN}}\right)$$

for fixed $K_p \geq 1$.

Accordingly, the complete model-state communication can be written as

$$\mathcal{C}_{\mathrm{total}}^{\mathrm{FedART}} = \mathcal{C}_{\mathrm{reg}} + \mathcal{O}\left(AK_pP_{\mathrm{BN}}\right).$$

This decomposition makes explicit that the event-sparse property concerns the *recovery-specific* communication introduced by cross-client recovery.

Proposition 11 (event-sparse recovery communication). Suppose a periodic recovery scheme invokes the same cross-client recovery procedure once every H online events and uses at most K_p peers per invocation. Its number of recovery invocations is approximately $N_{\mathrm{periodic}} \approx T/H$, whereas FedART invokes recovery $N_{\mathrm{FedART}} = A = \rho T$ times.

Therefore, the recovery-invocation ratio is

$$\frac{N_{\mathrm{FedART}}}{N_{\mathrm{periodic}}} \approx \frac{A}{T/H} = \frac{AH}{T} = \rho H.$$

If the two schemes exchange model states of the same order per recovery invocation, their recovery-specific communication costs satisfy

$$\mathcal{C}_{\mathrm{rec}}^{\mathrm{FedART}} = \mathcal{O}\left(AK_pP_{\mathrm{BN}}\right),$$

and

$$\mathcal{C}_{\mathrm{rec}}^{\mathrm{periodic}} = \mathcal{O}\left(\frac{T}{H}K_pP_{\mathrm{BN}}\right).$$

Thus, when $\rho H < 1$, FedART invokes recovery less frequently than the corresponding periodic schedule, with the recovery-count ratio governed by ρH.

Proof. The periodic scheme invokes recovery approximately once per H events and therefore performs approximately T/H recovery operations. FedART instead performs one recovery operation at each of its A detected degradation events. Their invocation ratio is consequently $A/(T/H) = \rho H$.

Both methods communicate at most K_p peer states per recovery invocation, so multiplying the per-recovery payload $\mathcal{O}(K_pP_{\mathrm{BN}})$ by their respective invocation counts yields the stated communication costs.

Proposition 11 shows that FedART's communication benefit is determined by the empirical sparsity of recovery events rather than by a different asymptotic order per recovery. When ρH is small, the same type of recovery operation is invoked substantially fewer times than under a periodic schedule.

A strictly asymptotic reduction follows only under the stronger condition $A = o(T/H)$, which is not required for the practical communication savings studied here. The empirical results instead quantify the finite-stream reduction produced by the observed recovery-event sparsity.

Communication with increasing federation size. An additional property of top-K_p peer retrieval is that the recovery-specific payload of an individual event does not directly scale with the total number of clients K. Since $d_a \leq K_p$, the communicated peer-state payload of one triggered recovery remains $\mathcal{O}(K_p P_{\mathrm{BN}})$ when K_p is fixed.

Therefore, increasing the federation size enlarges the candidate peer pool but does not require a requesting client to retrieve states from every client. This statement concerns recovery-state transfer; server-side peer search or registry-maintenance costs may depend separately on the implementation.

Computational complexity. We next separate the computation performed at every online event from the additional computation activated only during recovery.

Let C_{TTA} denote the cost of the standard local TTA update for one event, C_{mon} the cost of uncertainty evaluation and risk-monitor maintenance, C_{rec} the cost of severity estimation and selective restoration during recovery, C_{peer} the cost of evaluating one selected peer on the current unlabeled target batch, and C_{cand} the cost of evaluating the recovery candidates.

For a normal event without triggered recovery, the computation is $C_{\mathrm{normal}} = \mathcal{O}(C_{\mathrm{TTA}} + C_{\mathrm{mon}})$.

At a triggered recovery event using d_a peers, the additional recovery-specific computation is $C_{\mathrm{alarm}} = \mathcal{O}(C_{\mathrm{rec}} + d_a C_{\mathrm{peer}} + C_{\mathrm{cand}})$.

FedART evaluates only a constant number of candidate states: continuation, local restoration, cross-client restoration, and rollback. Hence, $C_{\mathrm{cand}} = \mathcal{O}(C_{\mathrm{forward}})$ up to a constant factor.

Across the complete deployment stream,

$$\mathcal{C}^{\mathrm{comp}}_{\mathrm{FedART}} = \mathcal{O}\left(T\left(C_{\mathrm{TTA}} + C_{\mathrm{mon}}\right) + \sum_{a=1}^{A}\left(C_{\mathrm{rec}} + d_a C_{\mathrm{peer}} + C_{\mathrm{cand}}\right)\right).$$

Using $d_a \leq K_p$ gives

$$\mathcal{C}^{\mathrm{comp}}_{\mathrm{FedART}} = \mathcal{O}\left(T\left(C_{\mathrm{TTA}} + C_{\mathrm{mon}}\right) + A\left(C_{\mathrm{rec}} + K_p C_{\mathrm{peer}} + C_{\mathrm{cand}}\right)\right).$$

Proposition 12 (event-sparse recovery computation). Define the per-event recovery-specific cost as $C_{\mathrm{recovery}} = C_{\mathrm{rec}} + K_p C_{\mathrm{peer}} + C_{\mathrm{cand}}$.

The recovery-specific computational overhead of FedART is $\mathcal{O}(A C_{\mathrm{recovery}})$, whereas a periodic scheme invoking the same recovery procedure every H events incurs $\mathcal{O}((T/H) C_{\mathrm{recovery}})$.

The ratio of recovery invocations is again $A/(T/H) = \rho H$.

Proof. Recovery-specific operations are executed only at the A triggered events in FedART. Under periodic recovery, the corresponding operations are executed approximately T/H times. Multiplying the per-recovery computational cost by the respective invocation counts gives the two expressions above.

Proposition 12 concerns only the additional recovery computation. FedART still incurs the standard local TTA cost and the risk-monitoring cost at every online event. Hence, the event-sparse design reduces the frequency of the expensive recovery branch rather than eliminating the stream-wide cost of test-time adaptation itself.

Summary. The complexity analysis separates the always-active adaptation path from the event-triggered recovery path. Standard local TTA and risk monitoring scale with the number of online

events T, whereas recovery-specific peer communication, peer evaluation, selective restoration, and candidate evaluation scale with the number of triggered events A.

Consequently, with respect to recovery invocations,

$$\boxed{N_{\text{periodic}} \approx \frac{T}{H}, \qquad N_{\text{FedART}} = A = \rho T, \qquad \frac{N_{\text{FedART}}}{N_{\text{periodic}}} \approx \rho H.}$$

Thus, FedART converts recovery from a fixed-schedule operation into an event-triggered operation whose practical overhead is governed by the observed degradation-event rate. When degradation events are sparse relative to the periodic recovery schedule, cross-client recovery communication and recovery-specific computation are correspondingly reduced.