FEDRALA: FEDERATED TEST-TIME OOD DETECTION VIA RELIABLE NEGATIVE-LABEL AGGREGATION AND LOCAL ADAPTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

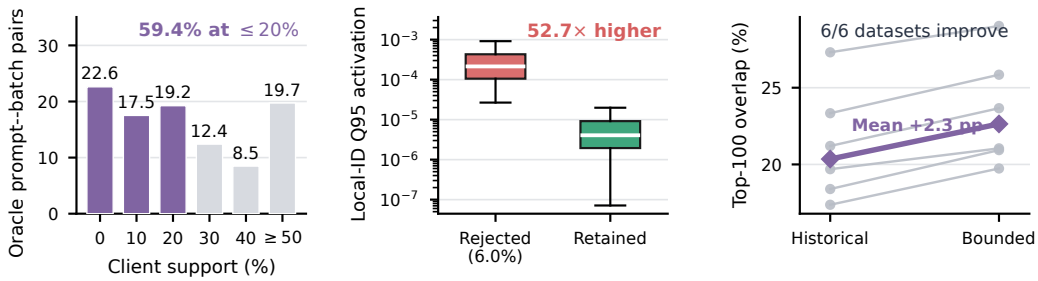
Vision-language models enable training-free out-of-distribution (OOD) detection by comparing image representations with in-distribution (ID) class prompts and negative textual labels. Recent work dynamically activates negative labels from unlabeled test streams to capture evolving OOD semantics, but mainly considers centralized test streams. In federated settings, OOD samples vary across clients, making the utility of negative labels dependent on local ID semantics and each client's current test distribution. Naive global averaging can suppress strong negative-label activations confined to a few clients and transfer labels that conflict with local ID semantics or respond weakly to local OOD samples. To address these challenges, we formulate federated test-time OOD detection through negative-label collaboration and propose FedRALA. FedRALA explicitly addresses three federation-specific failure modes: sparse cross-client OOD evidence, conflicts between shared negative labels and client-specific ID semantics, and temporal mismatch between historical shared evidence and the current local stream. FedRALA exchanges only prompt-level activation statistics, without sharing raw images or visual features. Experiments across multiple ID benchmarks and heterogeneous federated OOD streams validate its effectiveness. Notably, with TinyImageNet as the ID benchmark, FedRALA reduces the average FPR95 across six OOD datasets from 35.6% to 29.7% under severe client heterogeneity and disjoint OOD streams.

1 INTRODUCTION

Out-of-distribution (OOD) detection is essential for the reliable deployment of visual recognition systems in real-world environments, where models must distinguish known in-distribution (ID) samples from OOD samples (Hendrycks & Gimpel, 2017). Vision-language models (VLMs), such as CLIP (Radford et al., 2021), provide a promising solution by enabling OOD detection through image–text similarity without task-specific retraining. Early VLM-based OOD detectors compare image features with ID class prompts (Ming et al., 2022), whereas negative-label methods further compare them with textual concepts outside the ID label set (Jiang et al., 2024). Because a fixed negative-label bank cannot exhaustively cover the unforeseen semantics that may arise in an open-ended test stream, recent methods adapt visual scores, dictionaries, or negative labels from incoming unlabeled observations (Fan et al., 2024; Yang et al., 2025; Zhang et al., 2026; Ye et al., 2026; Zhu et al., 2026b).

However, these methods assume a centralized test stream processed by a single adaptive detector. A federated test environment is fundamentally different: clients retain their data locally, exhibit non-IID ID distributions and client-specific ID semantics, and may encounter different OOD samples at different times (McMahan et al., 2017; Kairouz et al., 2021). Although federated prompt learning has begun to address heterogeneous classification at test time (Zhu et al., 2026a), collaboration over dynamically activated negative labels remains underexplored. A straightforward solution is to average each negative label's activation statistics across clients. However, naive cross-client negative-label sharing exposes three key failure modes. First, global averaging can dilute useful OOD evidence supported by only a few clients. Second, a negative label useful to one client may conflict with another client's ID semantics. Third, historical evidence aggregated across clients may

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



(a) Sparse client support **(b)** Local-ID conflict **(c)** Stream alignment

Figure 1: **Diagnostics motivating FedRALA.** Results use TinyImageNet as ID, ten clients, and non-IID OOD streams with $\alpha = 0.1$. Panels **(a)**–**(c)** respectively show sparse cross-client support, local ID-semantic conflicts, and current-stream alignment under bounded adaptation. Ground-truth labels are used only for post-hoc analysis; all online methods operate on unlabeled streams.

be weakly aligned with a client's current local test stream. Such errors are especially consequential in online negative-label adaptation: the selected negative labels identify high-confidence OOD samples, which are stored in memory and subsequently used to initialize negative-label activations in the next round.

Figure 1 provides post-hoc diagnostic evidence for these three failure modes in an online federated stream. For diagnostic purposes only, after online inference and state updates are complete, we construct an oracle negative-label set by ranking candidate labels according to their ability to distinguish ground-truth OOD and ID samples in the current batch. This oracle is never used for negative-label selection, threshold estimation, memory updates, or OOD scoring. Among these oracle prompt–batch pairs, 59.4% appear in the active sets of at most 20% of clients, whereas only 19.7% are supported by at least half of the clients (Figure 1(a)). This sparse client support explains why global mean aggregation can dilute useful activations observed by only a few clients. Even highly activated OOD prompts may conflict with local ID semantics: the 6.0% rejected by the target-client ID check exhibit a $52.7\times$ higher median activation on local ID samples than the retained prompts (Figure 1(b)). Finally, even highly activated negative labels prioritized using local and cross-client histories may be weakly relevant to the client's current OOD samples. A bounded adjustment using the incoming unlabeled batch increases top-100 overlap with the oracle current-batch reference from 20.4% to 22.6%, with an average absolute rank change of only 4.24 positions (Figure 1(c)). Consistent with these observations, results averaged over all eight evaluation configurations show that naive global averaging degrades both AUROC and FPR95, whereas upper-tail aggregation and the successive local constraints recover the loss and ultimately surpass the local detector (Figure 3).

Motivated by these findings, we propose **FedRALA**, a training-free framework for reliable negative-label aggregation and local adaptation in federated test-time OOD detection. Its three components correspond directly to the three failure modes: prompt-wise top-K aggregation addresses sparse cross-client support, the local Q95 constraint filters labels that conflict with client-specific ID semantics, and bounded local reordering adapts the retained labels to the current test stream. FedRALA globally aggregates OOD evidence discovered across clients while preserving client-specific ID semantics and adapting the shared evidence to each client's current test stream. For federated aggregation, FedRALA replaces the global mean with a prompt-wise top-K average, reducing the dilution of strong activations supported by only a subset of clients. Each client penalizes labels with high upper-quantile activation on its local ID memory. A label therefore receives higher priority when it combines strong federated OOD support with low activation on the client's ID samples. For local adaptation, FedRALA calibrates the activation scores of the selected negative labels using their responses to the current unlabeled batch. The adjustment is bounded so that the current batch can refine each negative-label score without drastically changing the reliable scores accumulated from previous batches.

FedRALA communicates only prompt-level activation statistics. Raw images, visual features, detector thresholds, personalized ID prompts, and local memories remain on device, and no prompt retraining is required. We evaluate FedRALA across multiple ID benchmarks, heterogeneous OOD

streams, multiple CLIP backbones, and federations of up to 100 clients. In the particularly challenging setting with TinyImageNet as ID and disjoint OOD streams, FedRALA reduces the average FPR95 over six OOD datasets from 35.6% to 29.7%.

Our contributions are threefold:

- To our knowledge, we present the first systematic study of federated test-time OOD detection through negative-label collaboration. We identify sparse client support and target-ID conflict as two key obstacles to cross-client sharing, and further account for the temporal mismatch between aggregated negative-label statistics and evolving local test streams.
- We introduce FedRALA, which globally aggregates cross-client OOD evidence while preserving client-specific ID semantics and adapting the shared evidence to each local test stream. It communicates only prompt-level activation statistics, without additional training or sharing raw images or visual features.
- We evaluate FedRALA under Disjoint and Non-IID federated OOD protocols and demonstrate consistent gains over local adaptation and naive global aggregation across 10–100 clients.

2 Related Work

Vision-language OOD detection. Pretrained vision–language models (VLMs) enable training-free OOD detection through image–text alignment (Radford et al., 2021; Ming et al., 2022). Beyond matching an image to the ID label set, negative-label methods mine textual concepts outside that set and use their relative alignment as OOD evidence (Jiang et al., 2024). TANL further observes that only a subset of a corpus is activated by a particular test distribution and dynamically ranks negative labels from historical and current-batch responses (Zhang et al., 2026). FedRALA retains this training-free negative-label paradigm, but studies how activation evidence can be safely shared across heterogeneous clients.

Test-time OOD adaptation. Recent work calibrates OOD detectors from unlabeled observations at deployment. RTL fits a test-time linear relation between features and a base OOD score (Fan et al., 2024), whereas OODD maintains a priority-queue dictionary of representative test-time OOD features (Yang et al., 2025). For VLMs, TTL optimizes learnable OOD prompts from pseudo-labeled streams (Ye et al., 2026), and ANTS uses an external MLLM to construct adaptive negative textual spaces from cached test images (Zhu et al., 2026b). These methods build one adaptive state from a centralized stream. In contrast, FedRALA neither updates prompts nor generates new labels at test time: it communicates only prompt-wise activation statistics, preserves client-local data, and explicitly controls the reliability and target-ID compatibility of cross-client negative evidence.

Federated OOD detection and federated test-time adaptation. Federated OOD methods mainly learn detectors during training: FOSTER synthesizes virtual outliers, FOOGD learns score-based density models, and FOCoOp optimizes VLM prompts offline (Yu et al., 2023; Liao et al., 2024; 2025). Federated test-time adaptation instead targets classification under distribution shift; Latte exchanges class prototypes across clients, while controllable federated prompt learning adapts prompts at deployment (Bao et al., 2025; Zhu et al., 2026a). Neither line studies online cross-client negative-label adaptation with OOD detection as the objective. FedRALA fills this gap after offline prompt learning, exchanging only per-label activation statistics while keeping the VLM, ID prompts, samples, features, and memories local.

3 Methodology

3.1 Problem Setup and Overview

Consider J clients with frozen image and text encoders. Client k receives unlabeled batches $\{\mathcal{B}_{k,t}\}_{t\geq 1}$ from an unknown, time-varying mixture of its private ID distribution and local OOD distributions. Let $\{y_m^-\}_{m=1}^L$ be the shared negative-label candidate corpus. For visual feature $\mathbf{v}$,

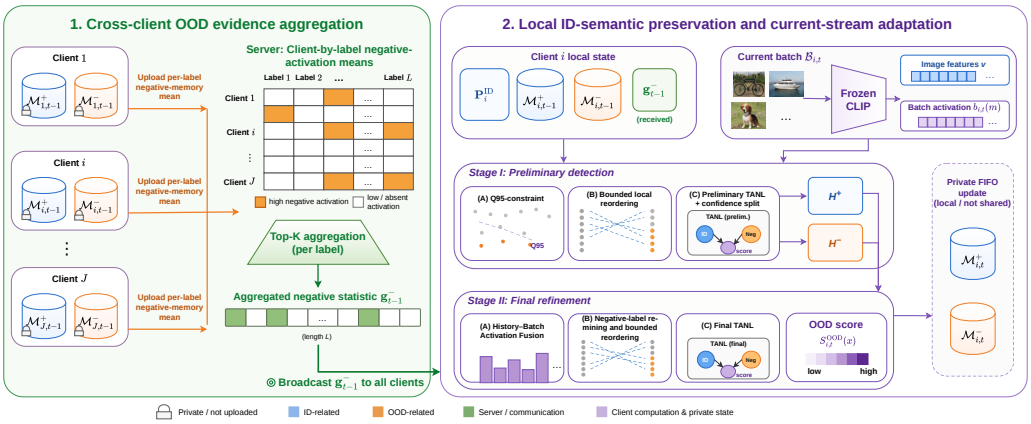


Figure 2: **Overview of FedRALA.** The server aggregates cross-client OOD evidence through label-wise top-K aggregation of client negative-memory means and broadcasts the resulting statistic. Each client preserves its ID semantics through a private Q95 constraint and adapts the shared evidence to its current stream in a bounded two-stage TANL loop: preliminary confidence splitting, followed by history–batch activation fusion, negative-label re-mining with bounded reordering, and final OOD scoring. The Stage-I confident samples are appended to the private FIFO memories only after final scoring. Only prompt-level activation statistics are communicated. In the diagram, P_i^{ID} denotes client i's fixed ID prompts, and H^+ and H^- denote its Stage-I confident ID-like and OOD-like subsets, respectively.

the activation of candidate m is its softmax probability over the union of client k's personalized ID prompts and the candidate corpus:

$$a_{k,m}(x) = \frac{\exp(\ell_k(\mathbf{v}, \mathbf{t}_m^-))}{\sum_{c=1}^{C_k} \exp(\ell_k(\mathbf{v}, \mathbf{t}_{k,c}^+)) + \sum_{r=1}^{L} \exp(\ell_k(\mathbf{v}, \mathbf{t}_r^-))}. \tag{1}$$

Each client stores activation vectors of confidently pseudo-labeled ID and OOD samples in private first-in-first-out (FIFO) memories $\mathcal{M}_{k,t}^+$ and $\mathcal{M}_{k,t}^-$, each holding at most N_{mem} entries. When full, a memory discards its oldest entry. The complete TANL score, adaptive threshold, fallback rules, and online memory update are given in Appendix A.

Figure 2 illustrates one synchronized test-time round. Personalized ID prompts are obtained before deployment and remain frozen. Each client uploads only the per-label mean of its previous negative memory, from which the server constructs and broadcasts the top-K aggregate $\mathbf{g}_{t-1}^-$. A receiving client combines this statistic with its private ID prompt, historical local memories, and current batch. Stage I forms confident ID-like and OOD-like subsets and computes their current-batch activation statistics. Stage II fuses these statistics with the historical statistics, re-mines and boundedly reorders the negative labels, and produces the final OOD score. The Stage-I confident samples are then appended to the private FIFO memories to form $\mathcal{M}_{k,t}^+$ and $\mathcal{M}_{k,t}^-$ for the next round. Raw samples, visual features, personalized prompts, thresholds, and FIFO entries remain local.

3.2 Reliable Negative-Label Aggregation

Client j summarizes candidate m's previous OOD-memory activation by $n_{j,t-1}^-(m) = |\mathcal{M}_{j,t-1}^-|^{-1} \sum_{\mathbf{z} \in \mathcal{M}_{j,t-1}^-} z_m$. For each label independently, the server averages the K largest client values:

$$\mathcal{S}_{m,t-1}^K = \mathrm{IndTopK}_{j \in \mathcal{J}}\, n_{j,t-1}^-(m), \qquad g_{t-1}^-(m) = \frac{1}{K} \sum_{j \in \mathcal{S}_{m,t-1}^K} n_{j,t-1}^-(m). \tag{2}$$

This recovers max aggregation at $K = 1$ and the global mean at $K = J$. Restricting K to an upper client tail reduces dilution of sparsely supported evidence, but choosing K too small can amplify an isolated noisy response.

Before using the broadcast vector, client k computes the empirical η-quantile $q^{+}_{k,t-1}(m)$ of candidate m's activations in its private ID memory. It forms

$$u^{\mathrm{rel}}_{k,t}(m) = g^{-}_{t-1}(m) - q^{+}_{k,t-1}(m), \qquad \mathcal{I}^{\mathrm{pre}}_{k,t} = \mathrm{IndTop}_{\beta}(\mathbf{u}^{\mathrm{rel}}_{k,t}), \tag{3}$$

so labels that overlap the receiving client's ID semantics are suppressed without transmitting its positive memory.

3.3 Bounded Local Adaptation

Historical reliability may not match the incoming batch, but replacing it with batch activation can let a transient batch erase useful history. For $m \in \mathcal{I}^{\mathrm{pre}}_{k,t}$, let $b_{k,t}(m)$ be the mean activation on $\mathcal{B}_{k,t}$, let $u^{\mathrm{rel}}_{k,t,(\beta+1)}$ be the selection boundary, and define

$$\begin{aligned} d_{k,t}(m) &= [u^{\mathrm{rel}}_{k,t}(m) - u^{\mathrm{rel}}_{k,t,(\beta+1)}]_{+}, \\ \widehat{b}_{k,t}(m) &= \frac{b_{k,t}(m)}{\max_{r \in \mathcal{I}^{\mathrm{pre}}_{k,t}} b_{k,t}(r)}, \\ r_{k,t}(m) &= d_{k,t}(m)(1 + \widehat{b}_{k,t}(m)). \end{aligned} \tag{4}$$

The denominator-zero case sets $\widehat{b} = 0$. FedRALA keeps the selected set fixed and sorts it by r, hence $d_{k,t}(m) \le r_{k,t}(m) \le 2d_{k,t}(m)$: the current batch can reorder similarly reliable labels but cannot replace the candidate set.

The resulting TANL scores select confident ID/OOD samples with mean activation vectors $\widehat{\mathbf{a}}^{+}_{k,t}$ and $\widehat{\mathbf{a}}^{-}_{k,t}$. To update full-corpus utilities, FedRALA preserves TANL's weighted combination of historical and current statistics:

$$\widetilde{\mathbf{p}}_{k,t} = \alpha \mathbf{q}^{+}_{k,t-1} + (1-\alpha)\widehat{\mathbf{a}}^{+}_{k,t}, \quad \widetilde{\mathbf{n}}_{k,t} = \alpha \mathbf{g}^{-}_{t-1} + (1-\alpha)\widehat{\mathbf{a}}^{-}_{k,t}, \quad \mathbf{u}^{\mathrm{fin}}_{k,t} = \widetilde{\mathbf{n}}_{k,t} - \widetilde{\mathbf{p}}_{k,t}. \tag{5}$$

Here, α is the history weight. Compared with local TANL, the historical positive and negative means are replaced by the target-local Q95 statistic and the global top-K negative statistic, respectively. The raw final top-β set is boundedly reordered once more. Algorithmic details, empty-memory fallbacks, and the cumulative TANL score appear in Appendix A.

3.4 Theoretical Guarantees

The analysis formalizes the three intended properties; complete assumptions, proofs, contamination extensions, and finite-batch bounds are deferred to Appendices B–C. Suppose label m has population client means $\mu_{j,m} = b_m + \Delta_m \mathbb{I}\{j \in \mathcal{S}_m\}$, with $|\mathcal{S}_m| = s_m$, and every uploaded coordinate has error at most ϵ. Then

$$\left|T_K(\mathbf{n}^{-}_m) - [b_m + \min\{1, s_m/K\}\Delta_m]\right| \le \epsilon, \qquad \left|T_J(\mathbf{n}^{-}_m) - [b_m + (s_m/J)\Delta_m]\right| \le \epsilon, \tag{6}$$

where T_K denotes the top-K average. Thus, when $K \le s_m$, the population top-K target retains the full sparse signal, whereas the global mean retains only s_m/J. Under nonempty conditionally i.i.d. FIFO memories, $\epsilon = \sqrt{\log(2JL/\delta)/(2N^{-}_{\min})}$ holds uniformly with probability $1-\delta$.

For Q95 selection, if ϵ_q uniformly bounds quantile-estimation error and the population utility gap $\Gamma_{k,\beta}$ at the top-β boundary exceeds $2(\epsilon + \epsilon_q)$, the empirical penalized set equals its population counterpart. Finally, $d_i > 2d_j \Rightarrow r_i > r_j$ for every current batch, giving a deterministic ordering certificate. Figure 1(a–c) empirically supports the three corresponding design principles: sparse cross-client support, target-local ID conflict, and bounded current-batch refinement. The results above are conditional sufficient guarantees; their complete assumptions and proofs are given in Appendices B–C.

4 Experiments

4.1 EXPERIMENTAL SETUP

Benchmarks and protocols. We use CIFAR-100 and TinyImageNet as ID benchmarks, each paired with six OOD datasets spanning objects, digits, scenes, textures, and near-OOD semantics. Ten-client ID partitions follow $\alpha_{\mathrm{ID}} \in \{0.5, 0.1\}$. OOD samples are assigned exactly once: *Disjoint* uses a balanced random partition, whereas *Non-IID* uses class-wise Dirichlet allocation (or 20 deterministic CLIP clusters when labels are unavailable) with $\alpha_{\mathrm{OOD}} = \alpha_{\mathrm{ID}}$. Each client's OOD subset is mixed with its private ID test stream. Complete dataset pairings and construction details are in Appendix A.

Methods and implementation. We compare standard frozen scores, fixed negative-label methods, test-time negative-label baselines, Local TANL, and federated aggregation variants. All methods use identical saved client-specific prompts, features, client partitions, and stream orders. The main model is CLIP ViT-B/16. TANL retains $\beta = 1{,}000$ active negative labels and uses batch size 128, memory size $N_{\mathrm{mem}} = 300$, history weight $\alpha = 0.95$, and $\gamma = 1$. FedRALA fixes $K = 2, \eta = 0.95$ in every main setting. We report client- and dataset-macro-averaged AUROC and FPR95. Full hyperparameters, baseline definitions, and seed controls are provided in Appendix A; sensitivity to the active-label count, batch size, memory size, history weight, and confidence gap is reported in Appendix G. Robustness to random partial client participation is evaluated in Appendix F. The cross-backbone study repeats the CIFAR-100 setting with $\alpha_{\mathrm{ID}} = 0.1$ using CLIP ViT-B/32, ViT-L/14, and RN101 while keeping all FedRALA hyperparameters fixed.

4.2 MAIN RESULTS

Table 1 compares the methods across eight configurations. Relative to Local, Global Mean lowers AUROC by 0.85 points and raises FPR95 by 1.16 points on average, showing that indiscriminate averaging can dilute negative-label activations supported by only a few clients. FedRALA improves both metrics in every configuration, with an average AUROC gain of 1.32 points and an FPR95 reduction of 3.96 points (9.3% relative). Its gains are larger for Disjoint OOD streams ($+1.84$ AUROC and -4.60 FPR95) than for Non-IID streams ($+0.81$ and -3.31, respectively), where each client already observes a broader range of OOD samples.

Fixed negative-label banks do not consistently outperform ID-only scores in this setting. Although all methods use the same personalized ID prompts, NegLabel and Global NegLabel often underperform MaxLogit, MCM, and Energy. Without local ID constraints or evidence from the current test stream, some candidate negative labels may overlap semantically with local ID classes, while others may be irrelevant to the OOD samples observed by a client. Test-time adaptation narrows this gap in some sparse-support settings, but remains sensitive to early pseudo-label errors and cross-client negative transfer.

4.3 ABLATION STUDIES

Reliable aggregation. Figure 3 summarizes the component-wise progression after averaging all eight evaluation configurations. Global Mean is worse than Local on both metrics, whereas Global Max and Global Top-2 substantially reduce FPR95. Their similar performance shows that the useful signal lies in the upper client tail rather than in global consensus. The broader sensitivity analysis in Figure 4 further shows that increasing K reintroduces dilution.

Local ID constraint. Adding Q95 to Top-2 improves AUROC from 87.44% to 87.77% and reduces FPR95 from 39.89% to 38.51%. This is the largest component-level FPR95 gain after upper-tail aggregation and verifies that federated OOD activation must be filtered against the receiving client's ID semantics.

Bounded local adaptation. Directly multiplying the reliability margin by current-batch activation degrades both metrics relative to Top-2+Q95, confirming that an unbounded response to a transient batch can overturn useful history. The bounded rule instead obtains the best average AUROC while preserving the Q95 variant's FPR95. Together with the small rank changes in Figure 1(c), this supports the intended role of local adaptation: refine the influence of historically reliable labels without changing their selected set or allowing current observations to dominate.

Table 1: **Main results.** Results are macro-averaged over six OOD datasets and ten clients. All methods use the same frozen client-specific ID prompts. Training-free test-time adaptation updates only online statistics, negative labels, or their rankings, without optimizing model parameters during OOD detection. Values are percentages; higher AUROC and lower FPR95 are better.

	$\alpha_{\text{ID}} = 0.5$				$\alpha_{\text{ID}} = 0.1$				Average	
	Disjoint		Non-IID		Disjoint		Non-IID			
Method	A↑	F↓	A↑	F↓	A↑	F↓	A↑	F↓	A↑	F↓
CIFAR-100 as ID										
Training-free and non-adaptive methods										
MaxLogit	84.20	53.03	84.36	52.01	91.89	32.79	92.10	31.02	88.14	42.21
MCM	80.06	56.38	80.08	55.66	85.63	44.56	85.53	44.08	82.83	50.17
Energy	84.01	53.32	84.18	52.09	91.97	32.70	92.21	30.87	88.09	42.25
NegLabel	82.24	55.47	82.37	54.75	91.69	34.30	91.85	32.68	87.04	44.30
Global NegLabel	82.00	56.02	82.15	55.18	91.76	33.97	91.96	32.41	86.97	44.40
Training-free test-time adaptation methods										
AdaNeg	82.63	55.34	82.48	54.96	89.24	37.97	89.58	36.60	85.98	46.22
ANTS	83.21	53.05	83.67	52.09	91.79	33.64	92.57	30.58	87.81	42.34
Local TANL	85.28	48.22	86.14	46.80	94.29	24.64	95.24	21.70	90.24	35.34
Global Mean	84.70	48.70	84.96	48.23	94.64	23.61	94.64	23.01	89.74	35.89
FedRALA	**86.89**	**45.05**	**87.46**	**43.31**	**95.27**	**20.34**	**95.61**	**18.55**	**91.31**	**31.81**
TinyImageNet as ID										
Training-free and non-adaptive methods										
MaxLogit	72.34	75.88	72.20	75.38	86.01	51.30	85.59	51.51	79.03	63.52
MCM	76.55	62.71	76.53	61.94	81.30	53.67	81.09	54.02	78.87	58.08
Energy	71.40	76.50	71.26	76.09	86.01	51.57	85.59	51.79	78.56	63.99
NegLabel	70.44	75.77	70.32	75.36	85.98	49.08	85.62	48.95	78.09	62.29
Global NegLabel	70.21	75.80	70.05	75.17	86.72	46.80	86.34	46.73	78.33	61.13
Training-free test-time adaptation methods										
AdaNeg	71.94	74.93	71.47	74.25	91.99	36.16	92.00	35.00	81.85	55.09
ANTS	**78.06**	63.78	**78.91**	62.44	89.77	40.51	90.84	37.46	84.39	51.05
Local TANL	73.74	67.07	75.91	64.41	90.43	35.59	92.17	31.15	83.06	49.56
Global Mean	73.64	66.98	73.26	66.90	90.73	34.72	89.79	36.70	81.86	51.33
FedRALA	76.55	**61.99**	77.24	**60.03**	**92.39**	**29.72**	**92.38**	**28.94**	**84.64**	**45.17**

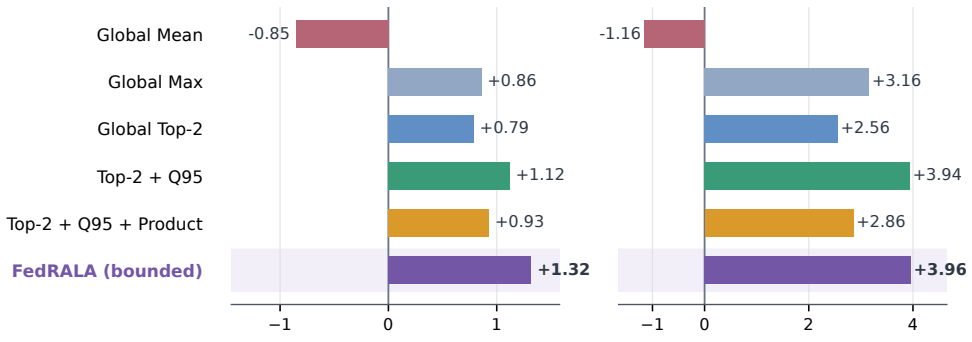


Figure 3: **Ablation of federated aggregation and local adaptation.** Results are averaged over all eight configurations in Table 1. We report percentage-point changes relative to Local TANL: **(a)** AUROC gain and **(b)** FPR95 reduction. Positive values indicate better performance, and bar-end numbers give the exact changes. FedRALA provides the strongest overall improvement, whereas naive global averaging degrades both metrics.

Table 2: **Scalability to 100 clients.** Results use TinyImageNet as ID and are macro-averaged over six OOD datasets and 100 clients. For Non-IID streams, $\alpha_{\text{OOD}} = \alpha_{\text{ID}}$. Offline prompt learning uses 100 rounds with 10 clients sampled per round; all 100 client states participate in test-time aggregation. FedRALA uses $K = 2$ throughout. Bold values indicate the best average performance within each ID-heterogeneity setting.

α_{ID}	Method	Disjoint		Non-IID		Average	
		AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓
0.5	Local TANL	67.65	72.03	68.30	69.88	67.98	70.96
	Global Mean	67.68	72.20	67.51	70.97	67.60	71.59
	FedRALA	68.01	70.51	68.32	68.64	**68.17**	**69.58**
0.2	Local TANL	82.52	50.62	83.14	48.08	82.83	49.35
	Global Mean	82.46	50.90	82.18	49.75	82.32	50.33
	FedRALA	83.08	48.40	83.16	46.62	**83.12**	**47.51**
0.1	Local TANL	91.07	32.50	91.38	29.42	91.23	30.96
	Global Mean	91.07	32.36	90.59	31.54	90.83	31.95
	FedRALA	91.53	30.21	91.31	28.92	**91.42**	**29.57**

4.4 Scalability to 100 Clients

Table 2 evaluates whether reliable negative-label aggregation remains effective when the federation grows from 10 to 100 clients at three levels of ID heterogeneity. Across the six configurations, FedRALA reduces FPR95 in every case and improves AUROC in five. Averaged over all six configurations, it raises AUROC by 0.23 points and reduces FPR95 by 1.54 points relative to Local TANL. Under the most heterogeneous Non-IID setting, FedRALA reduces FPR95 by 0.50 points while maintaining comparable AUROC (-0.07 points). In contrast, Global Mean degrades both metrics under Non-IID OOD at all three values of α_{ID}, confirming that indiscriminate averaging becomes unreliable in heterogeneous streams.

Communication and runtime overhead. FedRALA matches Global Mean's 279.44 kB upload and download per client per round, while adding 3.88 ms per client batch and 24.76 MiB over Local TANL; server aggregation takes 0.35 ms per round. The exact payload grows linearly with L, as quantified in Appendix E.

4.5 Sensitivity to Top-K Aggregation

Figure 4 in Appendix G.1 examines whether reliable aggregation merely favors the smallest possible client tail. In the 100-client setting, $K = 2$ performs best under both OOD protocols among the evaluated multi-client averages. Increasing K from 2 to 50 lowers the two-protocol average AUROC by 0.39 points and raises FPR95 by 1.61 points, showing that broad aggregation reintroduces weak-response dilution.

The 10-client results rule out a trivial "smaller K is always better" explanation. A per-label maximum ($K = 1$) is suboptimal under both protocols; $K = 5$ is best for Disjoint OOD, whereas $K = 2$ is best for Non-IID OOD. Hence, a small but non-degenerate client tail balances isolated noisy responses against excessive averaging. We retain $K = 2$ as a robust default across federation sizes and stream heterogeneity.

4.6 Generalization Across CLIP Backbones

FedRALA leads both metrics in all six backbone–protocol combinations in Table 3. Its overall 94.47% AUROC and 21.77% FPR95 improve Local TANL by 1.06 and 3.74 points. The two-protocol AUROC gains/FPR95 reductions are $1.41/4.70$, $1.24/4.78$, and $0.53/1.74$ points on ViT-B/32, ViT-L/14, and RN101, respectively. This consistency supports transfer across the evaluated Transformer and convolutional encoders, but does not remove dependence on the underlying VLM: all exchanged statistics are derived from its frozen image–text similarities, so absolute detection

Table 3: **Generalization across CLIP backbones.** Results use CIFAR-100 as ID with ten clients, $\alpha_{\mathrm{ID}} = 0.1$, and seed 1. For Non-IID streams, $\alpha_{\mathrm{OOD}} = 0.1$. Each entry is macro-averaged over six OOD datasets and all clients. FedRALA denotes the bounded method.

		Local TANL		Global Mean		FedRALA	
Backbone	OOD split	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓
ViT-B/32	Disjoint	93.00	26.32	93.11	25.81	**94.66**	**21.32**
	Non-IID	93.74	24.67	93.11	25.79	**94.90**	**20.28**
ViT-L/14	Disjoint	91.84	31.09	91.90	31.37	**93.48**	**25.16**
	Non-IID	92.86	27.68	91.69	31.21	**93.70**	**24.05**
RN101	Disjoint	94.03	23.10	94.16	22.53	**94.90**	**20.31**
	Non-IID	94.98	20.16	94.27	21.75	**95.17**	**19.47**

quality remains bounded by its representation and alignment quality. By contrast, Global Mean is worse than Local TANL on both metrics in all three Non-IID cases. Additional robustness results under alternative OOD assignments and extreme client heterogeneity are reported in Appendix D.

4.7 Mechanism Diagnostics

Ground-truth ID/OOD labels are used only after online inference to construct the diagnostic references in Figure 1; they are never used for negative-label selection, scoring, or memory updates. Panel (a) shows that 59.4% of useful negative-label–batch pairs are active at no more than 20% of clients, while only 19.7% are supported by at least half, explaining why broad averaging can suppress sparse activations. Panel (b) shows that Q95 excludes 6.0% of aggregated labels whose median local-ID activation is $52.7\times$ that of retained labels, confirming that cross-client activation alone does not ensure local compatibility. Finally, Panel (c) shows that bounded adaptation increases top-100 overlap with the current-batch oracle from 20.4% to 22.6%, while limiting the mean absolute rank change to 4.24 positions. Together with the component ablation in Figure 3, these diagnostics connect the observed gains to sparse-client aggregation, local-ID control, and bounded adaptation. The ablation further shows that reliable aggregation and local-ID control provide most of the improvement, while the bounded step acts as conservative current-stream calibration. Unlike direct multiplication by batch activation, which degrades both metrics relative to Top-2+Q95, FedRALA preserves the reliable candidate set and restricts adaptation to bounded within-set changes. Its role is therefore to introduce local responsiveness without allowing a transient batch to overwrite the cross-client historical statistic.

5 Conclusion and Limitations

FedRALA globally aggregates cross-client OOD evidence while preserving client-specific ID semantics and adapting the shared evidence to evolving local test streams. Across heterogeneous streams, 10–100 clients, and multiple CLIP backbones, it improves the AUROC–FPR95 trade-off over local and global adaptation. It also remains effective under random partial participation and dynamic client departure. Fully asynchronous aggregation with delayed or stale updates and formal privacy protection remain future work. FedRALA also inherits the semantic coverage and calibration limitations of its frozen VLM. It can aggregate and reweight available negative-label activations, but cannot recover OOD concepts that the backbone fails to align with their textual labels; under severe domain or language mismatch, unreliable activations may also be propagated across clients. The cross-backbone results show consistent relative gains across the evaluated CLIP encoders, rather than independence from VLM quality. Evaluation with broader and domain-specific VLMs, together with confidence-aware suppression of unreliable shared statistics, remains important future work.

AI USE STATEMENT

Generative AI tools were used only to assist with research brainstorming, theoretical and experimental-design drafting, code development and debugging, result analysis, literature summarization, figure preparation, language polishing, and manuscript editing. All AI-assisted content was independently reviewed and verified by the authors, who take full responsibility for the final paper.

ETHICS STATEMENT

This work uses public benchmark datasets and pretrained models and involves no human participants or personally identifiable information. Federated evaluation is simulated with sample-disjoint client partitions; no real user data are collected or transmitted. Although FedRALA keeps samples and features local, sharing activation statistics does not constitute a formal privacy guarantee. OOD detection errors can also create misplaced confidence in safety-critical deployments, so the detector should not be used as a sole safety mechanism. Moreover, biases and semantic-coverage gaps inherited from the pretrained VLM may produce uneven failure rates across domains or populations and can be amplified by cross-client aggregation. Deployment therefore requires backbone-specific validation, calibration monitoring, and an independent fallback mechanism.

REPRODUCIBILITY STATEMENT

Sections 3.2–3.3 specify the method, and Section 4.1 states the evaluation protocol and fixed hyperparameters. Appendix A gives the complete online algorithm and implementation details; Appendices B–C contain the assumptions and proofs; and Appendix H reports the detailed per-dataset results. An anonymized executable reference implementation of the core FedRALA online inference procedure, including a synthetic example and unit tests, is provided in the supplementary material. Public benchmark datasets and pretrained models should be obtained from their official sources. The complete experiment pipeline will be released upon acceptance.

REFERENCES

Wenxuan Bao, Ruxi Deng, Ruizhong Qiu, Tianxin Wei, Hanghang Tong, and Jingrui He. Latte: Collaborative test-time adaptation of vision-language models in federated learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 133–143, 2025.

Ke Fan, Tong Liu, Xingyu Qiu, Yikai Wang, Lian Huai, Zeyu Shangguan, Shuang Gou, Fengjian Liu, Yuqian Fu, Yanwei Fu, and Xingqun Jiang. Test-time linear out-of-distribution detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23752–23761, 2024.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations*, 2017.

Xue Jiang, Feng Liu, Zhen Fang, Hong Chen, Tongliang Liu, Feng Zheng, and Bo Han. Negative label guided OOD detection with pretrained vision-language models. In *International Conference on Learning Representations*, 2024.

Peter Kairouz, H. Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Keith Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *Foundations and Trends in Machine Learning*, 14(1–2): 1–210, 2021.

Xinting Liao, Weiming Liu, Pengyang Zhou, Fengyuan Yu, Jiahe Xu, Jun Wang, Wenjie Wang, Chaochao Chen, and Xiaolin Zheng. FOOGD: Federated collaboration for both out-of-distribution generalization and detection. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Xinting Liao, Weiming Liu, Jiaming Qian, Pengyang Zhou, Jiahe Xu, Wenjie Wang, Chaochao Chen, Xiaolin Zheng, and Tat-Seng Chua. FOCoOp: Enhancing out-of-distribution robustness in federated prompt learning for vision-language models. In *Proceedings of the 42nd International*

Conference on Machine Learning, volume 267 of *Proceedings of Machine Learning Research*, pp. 37528–37554, 2025.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *International Conference on Artificial Intelligence and Statistics*, pp. 1273–1282, 2017.

Yifei Ming, Ziyang Cai, Jiuxiang Gu, Yiyou Sun, Wei Li, and Yixuan Li. Delving into out-of-distribution detection with vision-language representations. In *Advances in Neural Information Processing Systems*, volume 35, pp. 35087–35102, 2022.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763, 2021.

Yifeng Yang, Lin Zhu, Zewen Sun, Hengyu Liu, Qinying Gu, and Nanyang Ye. OODD: Test-time out-of-distribution detection with dynamic dictionary. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 30630–30639, 2025.

Jinlun Ye, Jiang Liao, Runhe Lai, Xinhua Lu, Jiaxin Zhuang, Zhiyong Gan, and Ruixuan Wang. TTL: Test-time textual learning for OOD detection with pretrained vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27590–27599, 2026.

Shuyang Yu, Junyuan Hong, Haotao Wang, Zhangyang Wang, and Jiayu Zhou. Turning the curse of heterogeneity in federated learning into a blessing for out-of-distribution detection. In *International Conference on Learning Representations*, 2023.

Yabin Zhang, Maya Varma, Yunhe Gao, Jean-Benoit Delbrouck, Jiaming Liu, Chong Wang, and Curtis Langlotz. Activation matters: Test-time activated negative labels for OOD detection with vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17462–17473, 2026.

Rui Zhu, Liang Bai, Yanming Guo, Yirun Ruan, Tianyuan Yu, and Zhihe Lu. Controllable federated prompt learning at test time. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 39455–39465, 2026a.

Wenjie Zhu, Yabin Zhang, Xin Jin, Wenjun Zeng, and Lei Zhang. ANTS: Adaptive negative textual space shaping for OOD detection via test-time MLLM understanding and reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20–30, 2026b.

A Complete Method and Evaluation Protocol

A.1 TANL background

Let f_{img} and f_{txt} be the frozen VLM encoders and let $\ell_k(\mathbf{v}, \mathbf{t}) = s_k \mathbf{v}^\top \mathbf{t}$ denote the scaled image–text logit at client k. Client k has personalized ID text features $\{\mathbf{t}_{k,c}^+\}_{c=1}^{C_k}$ and an ordered active sequence of β negative labels $\mathcal{A}_{k,t}^- = (y_{k,t,1}^-, \dots, y_{k,t,\beta}^-)$. Following TANL, define

$$Z_k^+(x) = \sum_{c=1}^{C_k} \exp \ell_k(\mathbf{v}, \mathbf{t}_{k,c}^+), \qquad Z_{k,t,r}^-(x) = \sum_{j=1}^{\min(r\gamma,\beta)} \exp \ell_k(\mathbf{v}, \mathbf{t}_{k,t,j}^-), \tag{7}$$

where $R = \lceil \beta/\gamma \rceil$. Its cumulative ID and OOD scores are

$$S_{k,t}^{\text{ID}}(x) = \frac{1}{R} \sum_{r=1}^{R} \frac{Z_k^+(x)}{Z_k^+(x) + Z_{k,t,r}^-(x)}, \qquad S_{k,t}^{\text{OOD}}(x) = 1 - S_{k,t}^{\text{ID}}(x). \tag{8}$$

Earlier negative labels occur in more prefixes and therefore have greater influence.

Algorithm 1 FedRALA test-time round

Require: Clients $\mathcal{J}$, streams $\{\mathcal{B}_{k,t}\}$, candidate corpus, $K, \beta, \eta, \alpha, g, \gamma, N_{\mathrm{mem}}$
Initialize each client's frozen TANL state and FIFO memories
for $t = 1, \ldots, T$ **do**
  Snapshot $\{n^{-}_{j,t-1}(m)\}_{j,m}$ before processing round t
  Aggregate and broadcast $\mathbf{g}^{-}_{t-1}$ using Eq. 2
  for each client $k \in \mathcal{J}$ in parallel **do**
    Compute private Q95 utilities and $\mathcal{I}^{\mathrm{pre}}_{k,t}$ using Eq. 3
    Boundedly reorder the fixed preliminary set using Eq. 4
    Score $\mathcal{B}_{k,t}$, update $\tau_{k,t}$, and obtain confident ID/OOD subsets
    Compute their activation means $\widehat{\mathbf{a}}^{+}_{k,t}$ and $\widehat{\mathbf{a}}^{-}_{k,t}$
    Compute $\mathbf{u}^{\mathrm{fin}}_{k,t}$ using Eq. 5
    Re-mine all candidates and boundedly reorder the final top-β set
    Produce OOD scores and update private FIFO memories
  end for
end for
Ensure: Online OOD scores for all client streams

Client k stores the activation vectors in Eq. 1 for confidently pseudo-labeled ID and OOD samples in FIFO memories $\mathcal{M}^{+}_{k,t-1}$ and $\mathcal{M}^{-}_{k,t-1}$. Original local TANL initially orders candidates by $\bar{\mathbf{a}}^{-}_{k,t-1} - \bar{\mathbf{a}}^{+}_{k,t-1}$, where the two terms are the memory means. Its preliminary score updates an adaptive threshold $\tau_{k,t}$. With confidence gap g, samples satisfying

$$S^{\mathrm{ID}}_{k,t}(x) > \tau_{k,t} + g(1 - \tau_{k,t}) \quad \text{or} \quad S^{\mathrm{ID}}_{k,t}(x) < \tau_{k,t} - g\tau_{k,t}$$

form the confident ID and OOD subsets. If their activation means are $\widehat{\mathbf{a}}^{+}_{k,t}$ and $\widehat{\mathbf{a}}^{-}_{k,t}$, TANL applies momentum $\alpha\bar{\mathbf{a}}^{\pm}_{k,t-1} + (1-\alpha)\widehat{\mathbf{a}}^{\pm}_{k,t}$, re-mines the candidate corpus, scores the batch, and appends confident samples to the corresponding memories. If either subset is empty, the associated historical statistic is retained.

A.2 Complete synchronized online procedure

At the beginning of round t, every test-time client snapshots its state before observing $\mathcal{B}_{k,t}$. The uploaded vectors are calculated from these snapshots, so no client can transmit evidence from the current batch before other clients in that round. The server applies Eq. 2 and broadcasts one aggregated L-dimensional vector. Q95 estimation, candidate selection, pseudo-labeling, thresholds, and FIFO updates remain private.

Empty memories. If an OOD memory is empty at initialization, the client retains its initialized TANL negative statistic for that round; if a confident current-batch subset is empty, the corresponding historical positive or federated negative statistic is retained. These fallback states are outside the clean nonempty-memory concentration statement in Appendix B.

Offline versus test-time participation. For the 10-client experiments, all clients participate in offline prompt training. In the 100-client experiment, offline prompt learning runs for 100 rounds and samples 10 clients without replacement per round; unselected clients retain their personalized prompt states. At test time, all 100 frozen client states participate in each synchronized aggregation round. Thus, clients-per-round controls offline training participation and does not subsample the test-time federation.

A.3 Datasets and federated streams

CIFAR-100 is paired with CIFAR-10, MNIST, SVHN, TinyImageNet, DTD, and Places365. TinyImageNet is paired with SSBHard, NINCO, iNaturalist, SUN, Places, and DTD. We follow the OpenOOD organization and use complete cached test features.

ID training samples are allocated by class-wise Dirichlet draws with $\alpha_{\mathrm{ID}} \in \{0.5, 0.1\}$ in the main 10-client experiments. Both OOD protocols are sample-disjoint. *Disjoint* randomly divides an OOD dataset approximately evenly among clients. *Non-IID* groups samples by dataset labels and uses class-wise Dirichlet allocation; when labels are unavailable, 20 deterministic clusters of cached CLIP features act as semantic groups. We use $\alpha_{\mathrm{OOD}} = \alpha_{\mathrm{ID}}$. Every OOD example belongs to exactly one client, and the union of client subsets covers the complete OOD dataset. Each subset is mixed with the client's original non-IID ID test samples and deterministically shuffled.

A.4 Baselines, hyperparameters, and metrics

Local TANL communicates nothing. *Global Mean*, *Global Max*, and *Global Top-q* replace its local negative-memory mean with the corresponding per-label cross-client aggregation. Component variants add Q95, semantic diversity, or current-batch adjustment. *FedRALA* always denotes Top-2, Q95 conflict control, and the bounded rule in Eq. 4; Batch Activation Product is only an unbounded ablation and is never used as the proposed method.

All methods use the same saved client-specific ID prompts and process identical client partitions and sample orders. The source prompt learner uses prompt length four, 25 communication rounds, and two local prompt epochs for the 10-client experiments. We use CLIP ViT-B/16, retain $\beta = 1{,}000$ active negative labels, and set the batch size to 128, $\alpha = 0.95$, $\gamma = 1$, confidence gap 0.2, and memory size $N_{\mathrm{mem}} = 300$. FedRALA fixes $K = 2$ and $\eta = 0.95$ across all main settings. No model or prompt is optimized during OOD detection.

The cross-backbone study uses CIFAR-100 with ten clients, $\alpha_{\mathrm{ID}} = 0.1$, and seed 1. It evaluates ViT-B/32, ViT-L/14, and RN101 with source prompt-learning batch sizes 128, 64, and 128, respectively; all negative-label and FedRALA hyperparameters remain unchanged.

Unless otherwise noted, the main and ablation results use training, ID-partition, stream, and OOD-partition seed 1. The assignment-robustness study changes only the OOD-partition seed: learned prompts, the ID client partition, and test-stream order remain fixed. AUROC and FPR95 are computed on each client's complete online stream and macro-averaged over clients and the six OOD datasets. Values in all tables are percentages.

B Full Theoretical Analysis

We analyze one test-time round. Sparse-support, FIFO, and quantile results take probability over memory samples, whereas bounded current-batch adaptation is conditional on the realized pre-batch state. The finite-memory statements are idealized and require nonempty FIFO memories whose entries are conditionally i.i.d. draws from the corresponding clean admission distributions. The analysis characterizes the negative-label set recovered by FedRALA and the stability of its subsequent local reordering; it does not require model or prompt optimization. For an integer $K \in \{1, \dots, J\}$ and $\mathbf{x} \in [0,1]^J$, define the upper-tail aggregation functional

$$T_K(\mathbf{x}) = \frac{1}{K} \sum_{j \in \mathrm{IndTopK}(\mathbf{x})} x_j.$$

Proposition 1 (Upper-tail aggregation interpretation)**.** *Let* $\mathcal{W}_K = \{\mathbf{w} \geq 0 : \sum_j w_j = 1,\ w_j \leq 1/K\}$. *Then*

$$T_K(\mathbf{x}) = \max_{\mathbf{w} \in \mathcal{W}_K} \mathbf{w}^\top \mathbf{x} = \min_{\tau \in \mathbb{R}} \left[\tau + \frac{1}{K} \sum_{j=1}^{J} (x_j - \tau)_+ \right]. \tag{9}$$

Hence Eq. 2 averages the largest K/J fraction of client statistics (an empirical upper-tail CVaR form), whereas $q^+_{k,m}$ is an upper-quantile, VaR-style ID-conflict statistic.

This view makes the aggregation objective explicit: a negative label need not be frequent across all clients, but it must receive strong evidence from a non-negligible subset, with no single client receiving weight above $1/K$. We next formalize the resulting protection against sparse-signal dilution.

Theorem 2 (Sparse-support preservation). *For a negative label m, let $1 \le K \le J$ and suppose its population negative-memory means obey*

$$\mu_{j,m} = b_m + \Delta_m \mathbb{I}\{j \in \mathcal{S}_m\}, \qquad |\mathcal{S}_m| = s_m, \quad 1 \le s_m \le J, \quad 0 \le b_m < b_m + \Delta_m \le 1, \tag{10}$$

and the uploaded estimates satisfy $\max_j |n_j^-(m) - \mu_{j,m}| \le \epsilon$. With $\kappa_m(K) = \min\{1, s_m/K\}$, FedRALA and Global Mean satisfy

$$\left|T_K(\mathbf{n}_m^-) - \left(b_m + \kappa_m(K)\Delta_m\right)\right| \le \epsilon, \tag{11}$$

$$\left|T_J(\mathbf{n}_m^-) - \left(b_m + (s_m/J)\Delta_m\right)\right| \le \epsilon. \tag{12}$$

Let $N_{\min}^- = \min_j |\mathcal{M}_j^-|$. If every negative FIFO is nonempty and its entries are conditionally i.i.d. draws from the client's clean negative-memory admission distribution, a union bound over J clients and L candidates gives $\epsilon = \sqrt{\log(2JL/\delta)/(2N_{\min}^-)}$ with probability at least $1-\delta$. Empty-memory fallback states are outside this idealized statement. The theorem exposes a scaling difference hidden by a global average.

Corollary 3 (Global-mean dilution). *Let $\theta_m = b_m + \Delta_m$ denote the full sparse-support strength. Whenever*

$$\Delta_m \left[\kappa_m(K) - \frac{s_m}{J}\right] > 2\epsilon, \tag{13}$$

T_K is closer to θ_m than Global Mean. In particular, when $K \le s_m$, the population Top-K target retains the complete signal Δ_m, whereas Global Mean retains only $(s_m/J)\Delta_m$. Thus, for fixed s_m and $K \le s_m$, increasing J dilutes the population global-mean signal but not the Top-K signal.

The same analysis explains why smaller K is not automatically better. Choosing $K \le s_m$ avoids population dilution, while Appendix C shows that sensitivity to c exceptional client coordinates scales as c/K; in contrast, $K > s_m$ attenuates the signal by s_m/K. A monotone variance claim would require stronger cross-client independence and support-stability assumptions and is not used here.

The local Q95 term turns support preservation into target-specific, ID-conflict-penalized selection. Let $q_{k,m}^+$ be the population η-quantile of label m's activation under client k's clean positive-memory admission distribution, and define

$$U_{k,m}^K = T_K(\boldsymbol{\mu}_m) - q_{k,m}^+.$$

Theorem 4 (Q95-penalized set recovery). *Suppose the bounds in Theorem 2 hold uniformly over the L candidates and $\max_m |\widehat{q}_{k,m}^+ - q_{k,m}^+| \le \epsilon_q$. Let $\Gamma_{k,\beta} = U_{k,(\beta)}^K - U_{k,(\beta+1)}^K$ be the population utility gap at the selection boundary. If*

$$\Gamma_{k,\beta} > 2(\epsilon + \epsilon_q), \tag{14}$$

then Eq. 3 recovers exactly the population top-β Q95-penalized negative-label set for client k.

The first bounded reordering changes only the order within this preliminary set. Afterward, Eq. 5 may re-mine a different final top-β set from the full corpus; the following proposition gives a margin condition under which that replacement cannot occur. The second bounded reordering again changes only the order within the resulting final set.

Proposition 5 (Bounded online stability). *Let d_i, d_j be the nonnegative reliability margins in Eq. 4. The adapted scores obey*

$$d_m \le r_m \le 2d_m, \qquad d_i > 2d_j \implies r_i > r_j$$

for every possible current batch. Moreover, writing $\mathbf{u}^{\text{batch}} = \widehat{\mathbf{a}}^- - \widehat{\mathbf{a}}^+$, the final utility satisfies

$$\mathbf{u}^{\text{fin}} = \alpha \mathbf{u}^{\text{rel}} + (1-\alpha)\mathbf{u}^{\text{batch}}, \qquad \|\mathbf{u}^{\text{fin}} - \mathbf{u}^{\text{rel}}\|_\infty \le 2(1-\alpha).$$

Consequently, if the reliable utility gap at the top-β boundary exceeds $4(1-\alpha)$, the momentum update cannot change the selected set in that round.

Together, these results separate the two roles of FedRALA: upper-tail aggregation and Q95 can recover a sparsely supported, ID-conflict-penalized candidate set under an explicit margin condition, while each bounded reordering controls its response to transient batches. Complete proofs, finite-batch ordering bounds, and contamination extensions are provided in Appendix C.

C Proofs and Additional Theoretical Analysis

This appendix proves the results in Section 3.4 and gives finite-memory, contamination, and score-stability extensions. The finite-memory probabilities are over clean FIFO samples after conditioning on the pre-batch admission policy and its induced admission distributions; the bounded current-batch results are conditional on the realized pre-batch state. These idealized finite-memory results require nonempty FIFOs with conditionally i.i.d. entries. Activations lie in $[0, 1]$, as they are softmax negative-label activation probabilities. Ties in Top-K, top-β, and quantile operations are resolved by a fixed deterministic rule.

C.1 Upper-tail representation

Proof of Proposition 1. The feasible set $\mathcal{W}_K$ is a polytope. A linear objective over this polytope is maximized at an extreme point that assigns weight $1/K$ to the K largest coordinates and zero to the others. Hence

$$\max_{\mathbf{w}\in\mathcal{W}_K} \mathbf{w}^\top \mathbf{x} = T_K(\mathbf{x}).$$

For every $\tau \in \mathbb{R}$ and $\mathbf{w} \in \mathcal{W}_K$,

$$\mathbf{w}^\top \mathbf{x} = \tau + \sum_j w_j(x_j - \tau) \le \tau + \frac{1}{K}\sum_j (x_j - \tau)_+.$$

Taking the maximum over $\mathbf{w}$ and then the minimum over τ gives one direction. If $K < J$, equality follows by choosing τ in the interval between the K-th and $(K+1)$-th largest order statistics and assigning weight $1/K$ to the top K coordinates. If $K = J$, choose any $\tau \le \min_j x_j$. This is the standard empirical upper-CVaR representation with tail mass K/J. □

An immediate consequence of the variational form is the Lipschitz property

$$|T_K(\mathbf{x}) - T_K(\mathbf{y})| \le \|\mathbf{x} - \mathbf{y}\|_\infty. \tag{15}$$

Indeed, every $\mathbf{w} \in \mathcal{W}_K$ is nonnegative and sums to one.

C.2 Sparse-support preservation and finite-memory concentration

Proof of Theorem 2. Under Eq. 10, the clean client vector contains s_m entries equal to $b_m + \Delta_m$ and $J - s_m$ entries equal to b_m. Its population Top-K average is therefore

$$T_K(\boldsymbol{\mu}_m) = b_m + \min\{1, s_m/K\}\Delta_m,$$

whereas $T_J(\boldsymbol{\mu}_m) = b_m + (s_m/J)\Delta_m$. Applying Eq. 15 to $\|\mathbf{n}_m^- - \boldsymbol{\mu}_m\|_\infty \le \epsilon$ proves Eq. 11. The case $K = J$ gives Eq. 12. □

Proof of Corollary 3. For $\theta_m = b_m + \Delta_m$, Theorem 2 gives

$$|T_K(\mathbf{n}_m^-) - \theta_m| \le [1 - \kappa_m(K)]\Delta_m + \epsilon,$$
$$|T_J(\mathbf{n}_m^-) - \theta_m| \ge [1 - s_m/J]\Delta_m - \epsilon.$$

The first upper bound is strictly smaller than the second lower bound under Eq. 13. When $K \le s_m$, $\kappa_m(K) = 1$, yielding the stated population scaling result. □

Lemma 6 (Uniform FIFO concentration)**.** *For client j, let the nonempty negative FIFO $\mathcal{M}_j^-$ contain N_j^- entries and set $N_{\min}^- = \min_j N_j^-$. Conditional on the clean negative-memory admission distributions, suppose each FIFO contains i.i.d. activation vectors in $[0, 1]^L$, with population coordinate means $\mu_{j,m}$. Then, with probability at least $1 - \delta$, simultaneously for all J clients and L labels,*

$$\max_{j,m} |n_j^-(m) - \mu_{j,m}| \le \epsilon_{\text{mem}}^-(\delta) := \sqrt{\frac{\log(2JL/\delta)}{2N_{\min}^-}}. \tag{16}$$

Proof. Hoeffding's inequality bounds the failure probability for coordinate (j, m) by $2\exp(-2N_j^-\epsilon^2)$, which is at most the same expression with $N_{\min}^-$. A union bound over JL coordinates proves the result. Neither cross-client nor cross-label independence is required. □

The two-level model isolates the effect of sparse semantic support. A more general vector whose supported entries lie within h_m of $b_m + \Delta_m$ and unsupported entries lie within h_m of b_m satisfies the same bounds after replacing ϵ with $\epsilon + h_m$.

C.3 Q95 estimation and penalized-set recovery

For a fixed client k and label m, let $F_{k,m}^+$ be the CDF of the activation under its clean positive-memory admission distribution, and define the full quantile function

$$q_{k,m}^+(p) = \inf\{z \in [0,1] : F_{k,m}^+(z) \geq p\}, \qquad q_{k,m}^+ = q_{k,m}^+(\eta).$$

Suppose the nonempty positive FIFO $\mathcal{M}_k^+$ contains N_k^+ conditionally i.i.d. draws from this distribution, and let $\widehat{q}_{k,m}^+$ be the corresponding empirical quantile.

Lemma 7 (Uniform empirical-quantile control)**.** *Define*

$$\epsilon_{F,k}^+(\delta) = \sqrt{\frac{\log(2L/\delta)}{2N_k^+}}.$$

With probability at least $1-\delta$*, simultaneously over all* L *labels,*

$$q_{k,m}^+\Big(\eta - \epsilon_{F,k}^+(\delta)\Big) \leq \widehat{q}_{k,m}^+ \leq q_{k,m}^+\Big(\eta + \epsilon_{F,k}^+(\delta)\Big), \tag{17}$$

where quantile levels are clipped to $[0,1]$*. If each CDF has density at least* $f_{\min} > 0$ *between these two quantiles, then*

$$\max_m |\widehat{q}_{k,m}^+ - q_{k,m}^+| \leq \epsilon_{q,k}(\delta) := \epsilon_{F,k}^+(\delta)/f_{\min}. \tag{18}$$

Proof. The Dvoretzky–Kiefer–Wolfowitz inequality and a union bound over labels give $\max_m \|\widehat{F}_{k,m}^+ - F_{k,m}^+\|_\infty \leq \epsilon_{F,k}^+(\delta)$. Inverting the two CDF envelopes yields Eq. 17. The lower density assumption converts a quantile-level displacement into the value bound in Eq. 18. A statement simultaneous over every client instead replaces L by JL and N_k^+ by $N_{\min}^+ = \min_k N_k^+$. □

Proof of Theorem 4. Let

$$\widehat{U}_{k,m}^K = T_K(\mathbf{n}_m^-) - \widehat{q}_{k,m}^+.$$

Theorem 2 and the assumed quantile bound imply

$$\max_m |\widehat{U}_{k,m}^K - U_{k,m}^K| \leq E := \epsilon + \epsilon_q.$$

For any population-selected label i and non-selected label j, $U_{k,i}^K - U_{k,j}^K \geq \Gamma_{k,\beta}$. Therefore

$$\widehat{U}_{k,i}^K - \widehat{U}_{k,j}^K \geq \Gamma_{k,\beta} - 2E > 0.$$

No non-selected label can displace a selected one, proving exact recovery. □

C.4 Memory contamination and exceptional clients

The preceding results describe sampling and semantic heterogeneity. We next state an extension that makes clear that Top-K is support-preserving, but is not a Byzantine-robust aggregator.

Proposition 8 (Contamination-aware utility error)**.** *Suppose entries in the observed negative and positive FIFOs are conditionally i.i.d. from Huber mixtures* $\widetilde{P}^\pm = (1-\rho^\pm)P^\pm + \rho^\pm Q^\pm$ *of their clean admission distributions (client indices are suppressed). For every label* m*, at most* c *of the* J *uploaded client coordinates may additionally be replaced by arbitrary values in* $[0,1]$*. Assume each*

clean positive-memory CDF has density at least $f_{\min} > 0$ throughout the quantile interval whose levels span $\eta \pm [\epsilon^+_{F,k}(\delta/2) + \rho^+]$, clipped to $[0, 1]$. Then, with probability at least $1 - \delta$,

$$\max_m |\widehat{U}^K_{k,m} - U^K_{k,m}| \le E_K(\delta) := \epsilon^-_{\mathrm{mem}}(\delta/2) + \rho^- + \frac{c}{K} + \frac{\epsilon^+_{F,k}(\delta/2) + \rho^+}{f_{\min}}. \tag{19}$$

Consequently, a population top-β gap larger than $2E_K(\delta)$ still guarantees exact Q95-penalized set recovery.

Proof. Apply Hoeffding's inequality to samples from the observed negative-memory mixtures, uniformly over JL coordinates, and DKW to samples from the observed positive-memory mixture, uniformly over L coordinates. Assigning failure probability $\delta/2$ to each event gives the two sampling terms in Eq. 19. For any activation $A \in [0, 1]$, the Huber mixture satisfies

$$|\mathbb{E}_{\widetilde{P}^-} A - \mathbb{E}_{P^-} A| \le \rho^-, \qquad \|\widetilde{F}^+_{k,m} - F^+_{k,m}\|_\infty \le \rho^+.$$

Thus the negative mean differs from its clean target by at most the sampling error plus ρ^-. In addition to Eq. 15, the variational representation gives

$$|T_K(\mathbf{x}) - T_K(\mathbf{y})| \le \frac{1}{K}\|\mathbf{x} - \mathbf{y}\|_1.$$

Changing at most c uploaded client coordinates therefore contributes at most c/K. For the positive memory, the empirical CDF differs from the clean CDF by at most $\epsilon^+_{F,k}(\delta/2) + \rho^+$; the density assumption converts this level error into the final term in Eq. 19. Summing the errors and taking the union of the two concentration events proves the stated probability. The set-recovery statement follows from the same margin argument as Theorem 4. □

For the full sparse-support target $\theta_m - q^+_{k,m}$, Eq. 19 also contains the structural dilution term $[1 - \kappa_m(K)]\Delta_m$. Thus a larger K reduces the exceptional-client term c/K, but once $K > s_m$, it introduces a signal bias of $(1 - s_m/K)\Delta_m$. This model illustrates the trade-off examined in the Top-K sensitivity analysis.

C.5 Bounded local adaptation and finite-batch stability

Proof of Proposition 5. For every selected label, $0 \le \widehat{b}_m \le 1$, so $d_m \le r_m = d_m(1 + \widehat{b}_m) \le 2d_m$. If $d_i > 2d_j$, then $r_i \ge d_i > 2d_j \ge r_j$; hence no current batch can reverse the pair.

From Eq. 5,

$$\begin{aligned}\mathbf{u}^{\mathrm{fin}} &= \alpha(\mathbf{g}^- - \mathbf{q}^+) + (1 - \alpha)(\widehat{\mathbf{a}}^- - \widehat{\mathbf{a}}^+) \\ &= \alpha\mathbf{u}^{\mathrm{rel}} + (1 - \alpha)\mathbf{u}^{\mathrm{batch}}.\end{aligned}$$

Both utility vectors have coordinates in $[-1, 1]$, so their coordinate-wise difference is at most 2. This proves the ℓ_∞ drift bound. A perturbation of at most $e = 2(1 - \alpha)$ cannot change a top-β set whose boundary gap exceeds $2e$, by the same margin argument used above. □

Corollary 9 (Certified ranking changes)**.** *Let π_d and π_r be the orders induced by d_m and r_m on the fixed preliminary set, and write $i \prec_{\pi_d} j$ when i precedes j under the fixed tie-breaking rule. Define the set of unordered pairs*

$$\mathcal{C}(d) = \big\{\{i, j\} : i \prec_{\pi_d} j,\ d_i \le 2d_j\big\}.$$

Then the Kendall inversion count satisfies

$$D_{\mathrm{K}}(\pi_d, \pi_r) \le |\mathcal{C}(d)|. \tag{20}$$

Proof. Proposition 5 shows that a pair outside $\mathcal{C}(d)$ cannot invert. Every inversion is therefore contained in $\mathcal{C}(d)$. □

The fixed-set design also lowers the statistical complexity of current-batch adaptation from the full corpus size L to the selected size β.

Theorem 10 (Finite-batch ordering consistency)**.** *Condition on the history used to select $\mathcal{I} = \mathcal{I}_{k,t}^{\mathrm{pre}}$. Let the current batch contain B conditionally i.i.d. samples and define*

$$\mu_m = \mathbb{E}[a_{k,m}(X) \mid \mathcal{H}_{t-1}], \qquad \mu_* = \max_{m \in \mathcal{I}} \mu_m \geq \nu > 0.$$

Set

$$\epsilon_B(\delta) = \sqrt{\frac{\log(2\beta/\delta)}{2B}}.$$

With probability at least $1-\delta$, $\max_{m\in\mathcal{I}} |b_m - \mu_m| \leq \epsilon_B$. If $\epsilon_B < \nu$, let $h_m^* = \mu_m/\mu_*$, $r_m^* = d_m(1+h_m^*)$, and

$$\zeta_B = \frac{2\epsilon_B}{\nu - \epsilon_B}.$$

Then

$$\max_{m\in\mathcal{I}} |\widehat{b}_m - h_m^*| \leq \zeta_B, \qquad |r_m - r_m^*| \leq \zeta_B d_m. \tag{21}$$

Consequently, if $|r_i^* - r_j^*| > \zeta_B(d_i + d_j)$, the empirical and population adaptations order labels i and j identically.

Proof. Hoeffding's inequality and a union bound over the fixed β labels give the first event. On this event, the empirical maximum is at least $\nu - \epsilon_B$, and

$$\begin{aligned}\left|\frac{b_m}{\max_r b_r} - \frac{\mu_m}{\mu_*}\right| &\leq \frac{|b_m - \mu_m|}{\max_r b_r} + \frac{\mu_m |\max_r b_r - \mu_*|}{(\max_r b_r)\mu_*} \\ &\leq \frac{2\epsilon_B}{\nu - \epsilon_B}.\end{aligned}$$

Multiplication by d_m proves Eq. 21. For a pair (i,j), the sum of the two score errors is at most $\zeta_B(d_i + d_j)$, giving the ordering condition. □

Theorem 10 applies directly to the preliminary reordering because $\mathcal{I}_{k,t}^{\mathrm{pre}}$ is fixed before observing the current batch. The final set is re-mined using a utility that depends on that batch. Extending the same guarantee to its reordering therefore requires sample splitting; alternatively, uniform control over all L labels replaces $\log(2\beta/\delta)$ by $\log(2L/\delta)$.

C.6 Effect of reordering on the TANL score

The next result connects the internal ranking certificate to the final OOD score. It is stated for the default $\gamma = 1$.

Lemma 11 (OOD-score perturbation under bounded reordering)**.** *Let π_d and π_r be the pre- and post-adaptation orders of the same β labels. For an image x, let $w_m(x) = \exp(\ell_k(\mathbf{v}, \mathbf{t}_m^-))$, and let $Z^+(x) > 0$ be its positive-label partition term in Eq. 8. Then*

$$\left|S^{\mathrm{OOD}}(x;\pi_r) - S^{\mathrm{OOD}}(x;\pi_d)\right| \leq \min\left\{1, \frac{1}{\beta Z^+(x)} \sum_{(i,j)\in \mathrm{Inv}(\pi_d,\pi_r)} |w_i(x) - w_j(x)|\right\}. \tag{22}$$

By Corollary 9, every pair in the sum belongs to $\mathcal{C}(d)$.

Proof. Any permutation can be transformed into another through adjacent swaps, one for each inversion. Swapping adjacent labels i and j changes only the cumulative prefix ending at their first position. For $f(z) = Z^+/(Z^+ + z)$, $|f'(z)| = Z^+/(Z^+ + z)^2 \leq 1/Z^+$. The contribution of this adjacent swap to the average ID score is therefore at most $|w_i(x) - w_j(x)|/(\beta Z^+(x))$. Summing over the inversion sequence and using $S^{\mathrm{OOD}} = 1 - S^{\mathrm{ID}}$ proves the nontrivial bound; score values in $[0,1]$ give the outer minimum with one. □

These results deliberately provide guarantees for sparse-evidence recovery, Q95-penalized set selection, and online score stability rather than a direct AUROC/FPR95 guarantee. The latter would additionally require assumptions on the complete ID/OOD score distributions, adaptive threshold, and temporal pseudo-label feedback that are not needed by the FedRALA mechanism itself.

D Robustness to OOD Client Assignment

Table 4: **Robustness to OOD client assignment.** Results use TinyImageNet as ID with $\alpha_{\mathrm{ID}} = 0.1$ and report mean±standard deviation over three OOD-partition seeds. Only the assignment of the fixed OOD pool to clients is resampled; prompts, the ID partition, and test-stream order remain fixed at seed 1.

OOD split	Method	AUROC↑	FPR95↓
Disjoint	Local TANL	90.437 ± 0.035	35.497 ± 0.361
	Global Mean	90.714 ± 0.063	34.546 ± 0.388
	FedRALA	$\mathbf{92.385 \pm 0.051}$	$\mathbf{29.419 \pm 0.388}$
Non-IID	Local TANL	92.484 ± 0.272	29.857 ± 1.187
	Global Mean	90.348 ± 0.495	34.388 ± 2.008
	FedRALA	$\mathbf{92.779 \pm 0.361}$	$\mathbf{27.173 \pm 1.653}$

Table 4 isolates sensitivity to the OOD client assignment while holding the learned prompts, ID partition, and stream order fixed. Under Disjoint assignment, FedRALA improves Local TANL by 1.948 AUROC points and reduces FPR95 by 6.078 points. Under Non-IID assignment, the corresponding changes are $+0.295$ and -2.684 points, and both metrics improve for every evaluated partition seed. In contrast, Global Mean under Non-IID assignment lowers AUROC by 2.136 points and increases FPR95 by 4.531 points relative to Local TANL, further illustrating the risk of indiscriminate cross-client averaging.

D.1 Extreme Client Heterogeneity

Table 5: **Extreme client heterogeneity on CIFAR-100 with 10 clients.** Results are macro-averaged over six OOD datasets and reported in percent. Disjoint OOD uses $\alpha_{\mathrm{ID}} = 0.01$; Non-IID OOD uses $\alpha_{\mathrm{ID}} = \alpha_{\mathrm{OOD}} = 0.01$. Bold values indicate the best result.

	Disjoint		Non-IID	
Method	AUROC↑	FPR95↓	AUROC↑	FPR95↓
Local TANL	98.00	9.66	98.49	7.12
Global Mean	98.08	9.24	98.10	8.21
FedRALA	**98.43**	**7.82**	**98.50**	**6.80**

Table 5 extends the standard 10-client evaluation to an extreme concentration of 0.01. Relative to Local TANL, FedRALA raises AUROC by 0.43 points and reduces FPR95 by 1.85 points under Disjoint OOD. Under Non-IID OOD, it preserves essentially the same AUROC while reducing FPR95 by 0.32 points. It also outperforms Global Mean on both metrics under both protocols, showing that reliable cross-client aggregation remains effective under substantially stronger client heterogeneity.

E Communication and Runtime Overhead

Table 6 compares the detector-side overhead after frozen-CLIP feature encoding. FedRALA has the same communication cost as Global Mean, with 279.44 kB uploaded and downloaded per client and 5.5888 MB of total server traffic per round. Its server aggregation requires 0.3497 ms per round. Compared with Local TANL, FedRALA increases the client-side latency from 2.87 ms to 6.75 ms per 128-sample batch, an absolute increase of 3.88 ms. Its peak allocated memory increases by only 24.76 MiB (1.14%), from 2166.75 MiB to 2191.51 MiB. Relative to Global Mean, FedRALA preserves the communication cost while adding 4.23 ms of client computation and 0.297 ms of server computation per round.

The communication cost follows directly from the transmitted statistics. Each client uploads one negative-memory mean $\mathbf{n}^{-}_{k,t-1} \in \mathbb{R}^L$, and the server returns one aggregate $\mathbf{g}^{-}_{t-1} \in \mathbb{R}^L$. Let d_{stat}

denote the number of bytes used to encode one scalar statistic. Candidate identities and their ordering are fixed and shared in advance, so no label strings or indices are transmitted. Hence,

$$C_{\mathrm{up}} = d_{\mathrm{stat}} L, \qquad C_{\mathrm{down}} = d_{\mathrm{stat}} L, \qquad C_{\mathrm{total}} = J\big(C_{\mathrm{up}} + C_{\mathrm{down}}\big) = 2J d_{\mathrm{stat}} L, \tag{23}$$

where C_{up} and C_{down} are the per-client payloads, and C_{total} sums both directions over J clients. Thus, although FedRALA adds no communication over Global Mean, its per-client payload grows linearly with $d_{\mathrm{stat}} L$, and the total traffic grows as $\mathcal{O}(J d_{\mathrm{stat}} L)$ for large candidate corpora.

Table 6: **Communication and runtime overhead.** Measurements use TinyImageNet as ID with ten clients, $\alpha_{\mathrm{ID}} = 0.1$, a Non-IID SSBHard stream with $\alpha_{\mathrm{OOD}} = 0.1$, a batch size of 128, and one NVIDIA H20-3e GPU. Latencies are mean $\pm$ standard deviation over five repetitions, each containing 100 full client batches. Client latency measures post-encoding OOD scoring and local adaptation; for federated methods, it also includes construction of the uploaded activation statistic. Peak memory is the maximum CUDA memory allocated by the centralized ten-client simulator.

	Latency		Communication per round		
Method	Client (ms/batch)	Server (ms/round)	Upload / download (kB/client)	Server total (MB)	Peak allocated memory (MiB)
Local TANL	2.87 ± 0.60	0	0 / 0	0	2166.75
Global Mean	2.52 ± 0.02	0.0527 ± 0.0005	279.44 / 279.44	5.5888	2171.16
FedRALA	6.75 ± 0.01	0.3497 ± 0.0006	279.44 / 279.44	5.5888	2191.51

F Robustness to Partial Client Participation

The main evaluation uses round-based aggregation with all clients available. To test whether FedRALA depends on continuous client availability, we introduce random partial participation. Starting from $J = 10$ clients, each server round uniformly samples without replacement at most C_{round} clients from those whose test streams are unfinished. Selected clients process their next local batch and upload their negative-label statistics. Unselected clients pause without changing their local states or sample order, while clients leave the eligible set after completing their streams. If fewer than C_{round} clients remain, all available clients participate. Local TANL, Global Mean, and FedRALA share the same client-availability schedule and initial states. We use participation seed 1 and evaluate $C_{\mathrm{round}} \in \{8, 5\}$, corresponding to at most 80% and 50% of the federation being available per round. This protocol captures intermittent, asynchronous client availability within a round-based aggregator; it does not simulate delayed, stale, or out-of-order updates.

Table 7: **Robustness to random partial client participation.** Results use CIFAR-100 as ID and are macro-averaged over six OOD datasets and ten clients. C_{round} is the maximum number of unfinished clients selected per round. Local TANL is invariant to idle server rounds. Bold values are best in each row.

α_{ID}	OOD partition	C_{round}	Local TANL		Global Mean		FedRALA	
			AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓
0.1	Disjoint	8	94.29	24.64	94.53	23.84	**95.12**	**20.81**
	Disjoint	5	94.29	24.64	94.41	24.10	**94.98**	**21.05**
	Non-IID	8	95.24	21.70	94.66	23.01	**95.52**	**19.02**
	Non-IID	5	95.24	21.70	94.67	22.67	**95.48**	**19.21**
0.5	Disjoint	8	85.28	48.22	84.68	48.91	**86.55**	**45.50**
	Disjoint	5	85.28	48.22	84.76	48.89	**85.99**	**46.27**
	Non-IID	8	86.14	46.80	84.94	48.13	**87.01**	**44.16**
	Non-IID	5	86.14	46.80	85.25	48.02	**86.54**	**45.04**

Table 7 shows that FedRALA attains the highest AUROC and the lowest FPR95 in all eight partial-participation settings. Its advantage persists under the more restrictive $C_{\mathrm{round}} = 5$ setting: relative

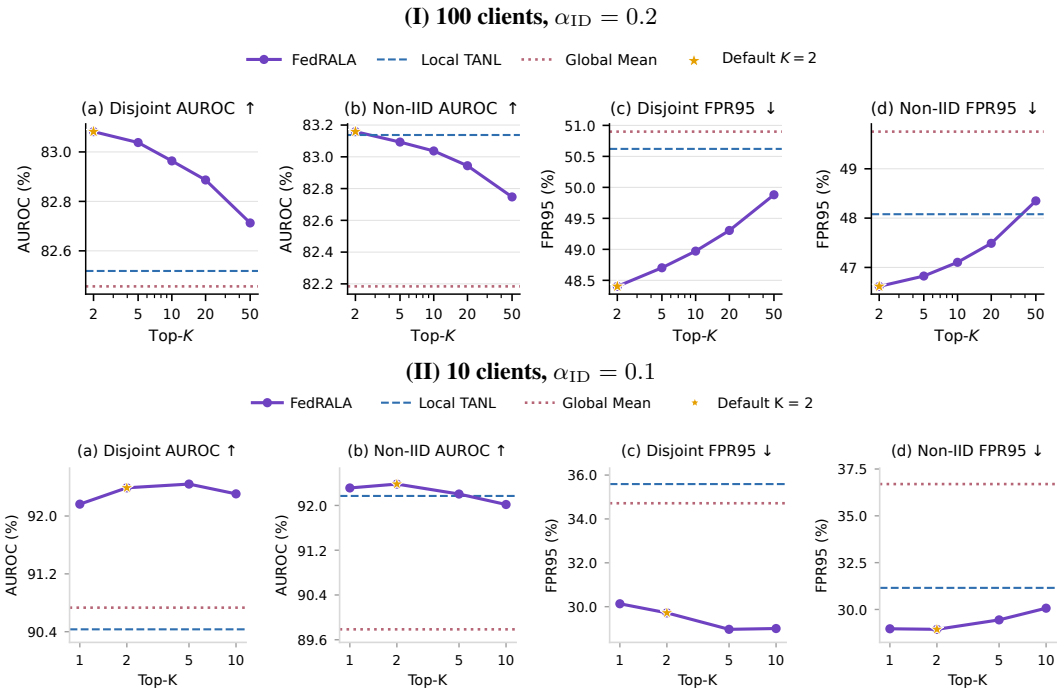


Figure 4: **Sensitivity to the number of aggregated clients K.** Both panels use TinyImageNet as ID and report Disjoint and Non-IID OOD streams. **(I)** The 100-client setting uses $\alpha_{\text{ID}} = 0.2$, $\alpha_{\text{OOD}} = 0.2$ for Non-IID streams, and $K \in \{2, 5, 10, 20, 50\}$. **(II)** The 10-client setting uses $\alpha_{\text{ID}} = 0.1$, $\alpha_{\text{OOD}} = 0.1$ for Non-IID streams, and $K \in \{1, 2, 5, 10\}$; $K = 1$ reduces the aggregation operator to a per-label maximum. Each point is macro-averaged over six OOD datasets and all clients. Horizontal lines denote Local TANL and Global Mean, and stars mark the default $K = 2$.

to Local TANL, FedRALA improves AUROC by 0.24–0.71 points and reduces FPR95 by 1.76–3.59 points across the four $C_{\text{round}} = 5$ settings. Global Mean, in contrast, is often worse than Local TANL, showing that partial availability does not remove the dilution and local-conflict problems of indiscriminate averaging. FedRALA therefore continues to benefit from cross-client negative-label collaboration without requiring every client to remain continuously online.

G Sensitivity Analyses

G.1 Number of Aggregated Clients

Figure 4 shows that reliable aggregation favors a small but non-degenerate upper client tail rather than obeying a monotonic "smaller K is always better" rule. In the 100-client setting, $K = 2$ is best under both OOD protocols among the evaluated multi-client averages. Increasing K from 2 to 50 reduces the two-protocol average AUROC by 0.39 points and increases FPR95 by 1.61 points, showing that broad aggregation reintroduces weak-response dilution.

The 10-client setting further separates this effect from a trivial preference for the smallest K. Under Disjoint streams, $K = 5$ achieves the best AUROC (92.44%) and FPR95 (28.97%), whereas under Non-IID streams, $K = 2$ is best on both metrics (92.38% AUROC and 28.94% FPR95). In particular, $K = 1$ is suboptimal in both protocols. Averaged across the two protocols, $K = 2$ gives the highest AUROC (92.39%), while $K = 5$ gives the lowest FPR95 (29.21%). These results indicate a broad optimum at small intermediate values: too few clients can overemphasize a single noisy response, whereas too many clients dilute sparse but reliable OOD evidence. We therefore retain $K = 2$ as a robust default across federation sizes and stream heterogeneity.

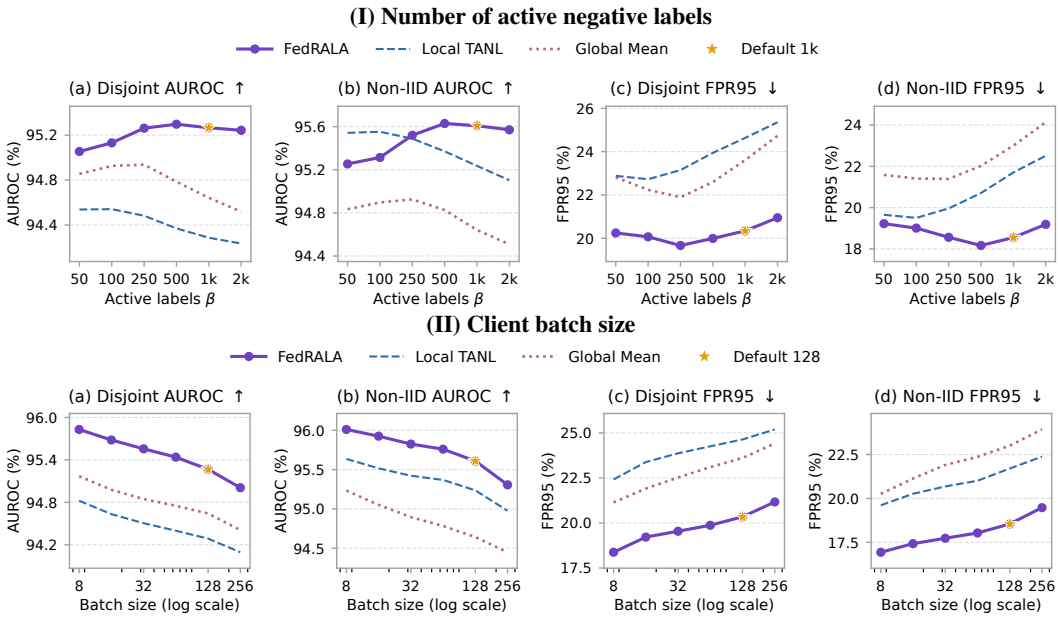


Figure 5: **Sensitivity to the active negative-label count and batch size.** Each row follows the four-panel layout and visual encoding of Figure 4. The batch-size axis is logarithmic; every marker is an evaluated value, while only selected values are labeled for legibility. FedRALA is the purple solid curve, Local TANL the blue dashed curve, and Global Mean the red dotted curve. Higher AUROC and lower FPR95 are better.

G.2 Online TANL Parameters

We evaluate whether FedRALA's advantage depends on the online TANL hyperparameters. The analysis uses CIFAR-100 as ID, ten clients, $\alpha_{\text{ID}} = 0.1$, CLIP ViT-B/16, and seed 1. Non-IID OOD streams use $\alpha_{\text{OOD}} = 0.1$. Every point is macro-averaged over six OOD datasets and all clients. We vary one parameter at a time while retaining the main defaults $\beta = 1{,}000$, batch size $B = 128$, memory size $N_{\text{mem}} = 300$, history weight $\alpha = 0.95$, and confidence gap $g = 0.2$ for the others. Local TANL, Global Mean, and FedRALA process identical client streams at every setting. Orange stars mark the defaults used in Table 1; the corresponding values exactly reproduce its CIFAR-100, $\alpha_{\text{ID}} = 0.1$ columns.

Figure 5(I) examines the active-label budget over $\beta \in \{50, 100, 250, 500, 1000, 2000\}$. Performance is non-monotonic: Local TANL peaks around $\beta = 100$, Global Mean around $\beta = 250$, while FedRALA obtains its best overall trade-off at $\beta = 250$–500. FedRALA achieves the lowest macro-average FPR95 at every tested value and leads both metrics throughout $\beta \geq 250$. Under Non-IID OOD with $\beta \in \{50, 100\}$, Local TANL has slightly higher AUROC while FedRALA retains lower FPR95, exposing an AUROC–FPR95 trade-off when the active-label budget is only 5–10% of the original TANL setting. This paired behavior is consistent with the methods' different selection objectives: Local TANL ranks labels using local mean activation margins, whereas FedRALA additionally penalizes target-ID upper-tail activation. We therefore do not claim uniform dominance in this extreme low-budget regime. Importantly, FedRALA with the original $\beta = 1{,}000$ still surpasses the strongest baseline selected over the entire grid: $(95.27, 20.34)$ versus $(94.94, 21.90)$ AUROC/FPR95 on Disjoint OOD, and $(95.61, 18.55)$ versus $(95.55, 19.50)$ on Non-IID OOD. This conservative comparison shows that the gain cannot be explained solely by active-set-size selection. We retain the original TANL default $\beta = 1{,}000$ in the main experiments without tuning on OOD test labels.

Figure 5(II) varies $B \in \{8, 16, 32, 64, 128, 256\}$; only $\{8, 32, 128, 256\}$ are labeled. All three methods generally improve with smaller batches, which refresh the online state more frequently and mix fewer temporally adjacent observations. FedRALA nevertheless leads both baselines on both metrics under both protocols at every tested batch size, so its advantage is not tied to a particular B.

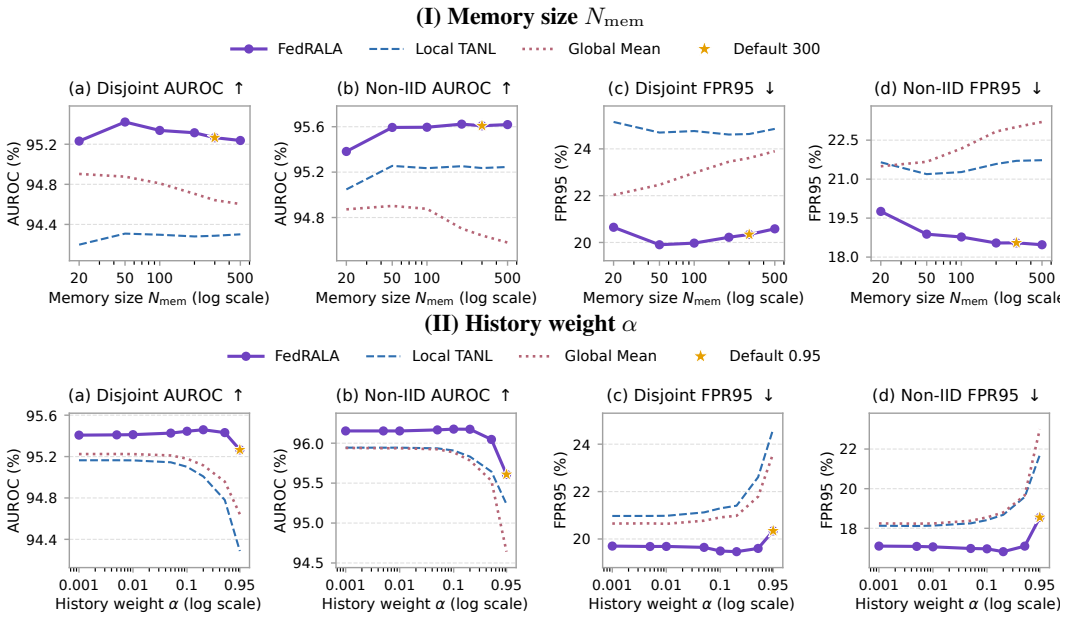


Figure 6: **Sensitivity to memory size and history weight.** The protocol and visual encoding follow Figure 5. Both horizontal axes are logarithmic; every marker is an evaluated value, while only selected values are labeled. The history weight α controls the contribution of previous-round statistics relative to current-batch activations. FedRALA uses the target-local Q95 and global top-K negative statistics as its historical terms. It is neither optimizer momentum nor a Dirichlet heterogeneity parameter.

Although $B = 8$ gives the best absolute performance, it requires 16 times as many online update and aggregation rounds as $B = 128$ for a fixed-length stream. We therefore retain $B = 128$ as the main performance–communication setting.

Figure 6(I) varies $N_{\text{mem}} \in \{20, 50, 100, 200, 300, 500\}$; only $\{20, 50, 100, 500\}$ are labeled, while the star identifies the default $N_{\text{mem}} = 300$. The small memory $N_{\text{mem}} = 20$ consistently underperforms, indicating that insufficient storage discards useful activation history. Performance improves sharply at $N_{\text{mem}} = 50$ and then enters a broad plateau: over $N_{\text{mem}} \in \{50, 100, 200, 300, 500\}$, FedRALA's two-protocol average varies by less than 0.08 AUROC points and 0.16 FPR95 points, while remaining better than both baselines throughout. The main default $N_{\text{mem}} = 300$ lies within this stable region. Figure 6(II) plots $\alpha \in \{0.001, 0.005, 0.01, 0.05, 0.1, 0.2, 0.5, 0.7, 0.9, 0.95, 1.0\}$. For readability, only $\{0.001, 0.01, 0.1, 0.95\}$ are labeled on the horizontal axis; every marker denotes an evaluated value. Across the full sweep, all three methods improve as α decreases from 1.0 to 0.1–0.2, suggesting that excessive historical weighting can retain stale negative-label statistics. Below 0.1, the changes become small, indicating diminishing sensitivity rather than continued benefit from an almost memoryless update. Crucially, FedRALA leads both baselines on both metrics at all eleven values and under both OOD protocols. We retain the original TANL value $\alpha = 0.95$ in the main experiments rather than selecting a value from OOD test performance.

Figure 7 varies $g \in \{0.001, 0.005, 0.01, 0.05, 0.1, 0.2, 0.3, 0.4\}$. For readability, only $\{0.001, 0.01, 0.1, 0.4\}$ are labeled on the horizontal axis; every marker denotes an evaluated value, and the star marks the default $g = 0.2$. FedRALA leads both baselines on AUROC and FPR95 under this 400-fold range, its AUROC changes by at most 0.12 points and its FPR95 by at most 0.39 points within either protocol. Performance broadly plateaus for $g \leq 0.1$, indicating that the gain does not rely on a narrowly selected near-zero gap. Smaller gaps admit more observations near the adaptive threshold, whereas larger gaps impose stricter memory admission. We retain the original TANL default $g = 0.2$ in the main comparison without tuning on OOD test labels. Together, the five analyses show that the reported gains are not tied to a narrowly tuned active-set size, batch size, memory capacity, history weight, or confidence gap.

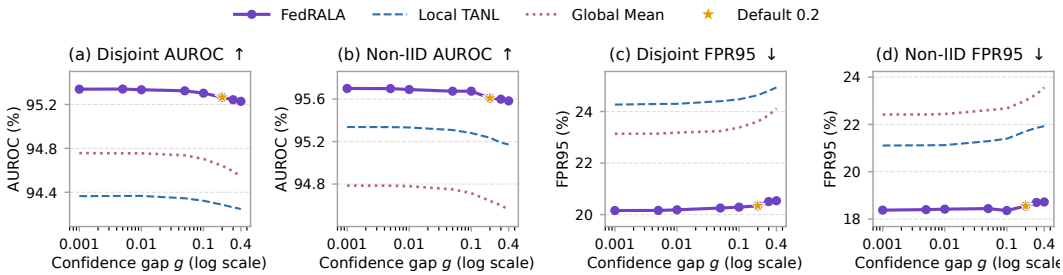


Figure 7: **Sensitivity to the TANL confidence gap.** The four panels and visual encoding follow Figure 5. The horizontal axis is logarithmic. The gap g controls the confidence interval around TANL's adaptive threshold: larger values impose stricter admission to the confident ID and OOD FIFO memories. It is not a fixed probability threshold.

Table 8: **Detailed per-OOD results with CIFAR-100 as ID, $\alpha_{\text{ID}} = 0.5$, and Disjoint OOD streams.** Results are macro-averaged over ten clients. A and F denote AUROC and FPR95, respectively; the final pair is averaged over the six OOD datasets. All entries are percentages.

	CIFAR-10		MNIST		SVHN		Tiny-IN		Texture		Places365		Average	
Method	A↑	F↓	A↑	F↓	A↑	F↓	A↑	F↓	A↑	F↓	A↑	F↓	A↑	F↓
Training-free and non-adaptive methods														
MaxLogit	90.07	40.36	82.97	45.11	92.42	21.90	80.48	66.39	74.84	80.64	84.45	63.81	84.20	53.03
MCM	83.38	50.84	77.85	53.40	90.83	29.75	78.28	61.67	76.62	69.06	73.44	73.58	80.06	56.38
Energy	90.10	40.65	83.15	44.52	91.60	22.03	80.06	67.51	74.04	81.11	85.10	64.06	84.01	53.32
NegLabel	89.72	43.24	88.43	32.15	90.93	24.95	75.63	73.08	67.87	87.62	80.82	71.77	82.24	55.47
Global NegLabel	89.75	43.90	88.18	32.79	90.62	25.50	75.50	73.37	67.49	88.19	80.46	72.37	82.00	56.02
Training-free test-time adaptation methods														
AdaNeg	89.29	44.94	89.31	30.68	91.49	24.08	75.95	73.60	67.53	88.58	82.19	70.19	82.63	55.34
ANTS	91.41	39.19	91.21	28.48	92.18	22.53	70.54	77.94	71.58	83.81	82.33	66.35	83.21	53.05
Local TANL	88.57	46.07	99.01	4.63	97.32	9.99	70.45	76.59	76.11	81.49	80.20	70.53	85.28	48.22
Global Mean	89.22	45.11	99.02	4.62	97.17	10.45	72.37	75.72	74.33	80.94	76.11	75.35	84.70	48.70
FedRALA	88.98	46.33	99.05	4.49	97.37	9.83	70.76	75.48	81.48	74.28	83.73	59.87	**86.89**	**45.05**

H Detailed Per-Dataset Results

Protocol. The first eight tables retain the methods and metrics of Table 1; each entry is macro-averaged over ten clients, and only the final Average columns are bolded. The first four tables use CIFAR-100 as ID and the following four use TinyImageNet as ID. The final six tables give the detailed 100-client results from Table 2; their captions state the corresponding protocol.

H.1 CIFAR-100 as ID

H.2 TinyImageNet as ID

Table 9: **Detailed per-OOD results with CIFAR-100 as ID, $\alpha_{\text{ID}} = 0.5$, and Non-IID OOD streams.** Results are macro-averaged over ten clients. A and F denote AUROC and FPR95, respectively; the final pair is averaged over the six OOD datasets. All entries are percentages.

	CIFAR-10		MNIST		SVHN		Tiny-IN		Texture		Places365		Average	
Method	A↑	F↓	A↑	F↓	A↑	F↓	A↑	F↓	A↑	F↓	A↑	F↓	A↑	F↓
Training-free and non-adaptive methods														
MaxLogit	90.32	39.16	83.16	43.48	92.19	22.94	80.64	65.62	75.06	77.58	84.77	63.26	84.36	52.01
MCM	83.86	49.92	77.84	51.72	90.35	30.50	78.21	62.03	76.58	66.80	73.65	72.98	80.08	55.66
Energy	90.32	39.62	83.34	42.96	91.46	22.79	80.23	66.24	74.33	77.90	85.43	63.04	84.18	52.09
NegLabel	89.98	42.04	88.50	31.46	90.87	25.49	75.74	72.77	68.06	85.62	81.07	71.12	82.37	54.75
Global NegLabel	90.06	42.13	88.23	32.38	90.57	25.91	75.58	72.99	67.71	85.76	80.74	71.91	82.15	55.18
Training-free test-time adaptation methods														
AdaNeg	89.40	43.26	89.33	30.20	91.46	24.43	76.30	73.22	66.63	87.26	81.78	71.35	82.48	54.96
ANTS	92.07	36.25	90.89	28.45	92.08	23.41	71.89	77.21	71.89	82.16	83.17	65.05	83.67	52.09
Local TANL	89.16	44.60	98.98	4.76	97.58	9.75	70.80	75.04	78.73	78.93	81.58	67.73	86.14	46.80
Global Mean	89.60	43.95	98.99	4.66	97.15	10.46	72.18	75.26	75.44	80.06	76.39	75.00	84.96	48.23
FedRALA	89.61	43.70	99.01	4.58	97.23	10.36	71.35	75.26	83.36	68.53	84.21	57.41	**87.46**	**43.31**

Table 10: **Detailed per-OOD results with CIFAR-100 as ID, $\alpha_{\text{ID}} = 0.1$, and Disjoint OOD streams.** Results are macro-averaged over ten clients. A and F denote AUROC and FPR95, respectively; the final pair is averaged over the six OOD datasets. All entries are percentages.

	CIFAR-10		MNIST		SVHN		Tiny-IN		Texture		Places365		Average	
Method	A↑	F↓	A↑	F↓	A↑	F↓	A↑	F↓	A↑	F↓	A↑	F↓	A↑	F↓
Training-free and non-adaptive methods														
MaxLogit	91.66	34.80	89.33	30.86	97.09	10.73	88.86	46.17	90.68	42.25	93.70	31.94	91.89	32.79
MCM	86.04	45.11	84.41	39.53	93.77	22.63	84.24	50.39	84.48	50.51	80.87	59.17	85.63	44.56
Energy	91.68	35.32	89.45	30.46	96.98	10.05	88.74	46.83	90.79	42.22	94.20	31.35	91.97	32.70
NegLabel	91.51	38.30	93.12	21.10	97.06	11.22	86.27	52.84	88.99	47.18	93.20	35.15	91.69	34.30
Global NegLabel	91.61	37.82	92.75	21.82	97.14	10.93	86.07	53.59	89.14	46.66	93.84	33.02	91.76	33.97
Training-free test-time adaptation methods														
AdaNeg	77.70	64.09	96.15	15.89	98.98	4.73	76.58	66.28	91.95	40.59	94.10	36.26	89.24	37.97
ANTS	92.29	36.11	94.70	17.34	97.38	10.58	82.37	61.66	90.53	42.32	93.49	33.82	91.79	33.64
Local TANL	91.33	39.70	99.33	3.33	99.45	2.19	86.10	50.01	93.69	31.18	95.82	21.39	94.29	24.64
Global Mean	91.66	38.72	99.33	3.34	99.44	2.33	87.33	48.75	94.05	28.56	96.05	19.97	94.64	23.61
FedRALA	92.00	36.94	99.41	2.93	99.59	1.55	87.78	45.07	96.16	18.19	96.64	17.37	**95.27**	**20.34**

Table 11: **Detailed per-OOD results with CIFAR-100 as ID, $\alpha_{\text{ID}} = 0.1$, and Non-IID OOD streams.** Results are macro-averaged over ten clients. A and F denote AUROC and FPR95, respectively; the final pair is averaged over the six OOD datasets. All entries are percentages.

	CIFAR-10		MNIST		SVHN		Tiny-IN		Texture		Places365		Average	
Method	A↑	F↓	A↑	F↓	A↑	F↓	A↑	F↓	A↑	F↓	A↑	F↓	A↑	F↓
Training-free and non-adaptive methods														
MaxLogit	91.68	35.63	89.65	29.47	97.10	10.38	89.96	39.92	90.42	40.25	93.80	30.49	92.10	31.02
MCM	85.65	44.68	84.42	39.24	93.45	22.28	84.91	47.21	83.99	51.94	80.73	59.13	85.53	44.08
Energy	91.75	35.86	89.84	28.78	96.99	9.86	89.86	40.66	90.51	40.30	94.30	29.76	92.21	30.87
NegLabel	91.39	38.51	93.42	20.25	97.15	10.11	87.40	47.57	88.68	44.77	93.06	34.90	91.85	32.68
Global NegLabel	91.48	38.38	93.06	21.09	97.27	9.79	87.20	48.53	88.96	44.16	93.77	32.50	91.96	32.41
Training-free test-time adaptation methods														
AdaNeg	78.90	63.00	90.12	20.88	99.25	3.76	84.56	55.32	89.38	47.44	95.25	29.22	89.58	36.60
ANTS	92.45	35.04	94.99	16.35	97.48	9.41	85.83	52.94	90.58	39.07	94.09	30.64	92.57	30.58
Local TANL	92.56	35.70	99.34	3.34	99.45	2.21	88.40	42.55	95.38	26.68	96.29	19.75	95.24	21.70
Global Mean	91.40	38.33	99.37	3.06	99.40	2.55	87.93	44.66	94.22	28.79	95.53	20.63	94.64	23.01
FedRALA	92.36	34.13	99.42	2.85	99.57	1.59	88.85	40.27	96.75	16.07	96.71	16.38	**95.61**	**18.55**

Table 12: **Detailed per-OOD results with TinyImageNet as ID, $\alpha_{\text{ID}} = 0.5$, and Disjoint OOD streams.** Results are macro-averaged over ten clients. A and F denote AUROC and FPR95, respectively; the final pair is averaged over the six OOD datasets. All entries are percentages.

	SSBHard		NINCO		iNat.		SUN		Places		DTD		Average	
Method	A↑	F↓	A↑	F↓	A↑	F↓	A↑	F↓	A↑	F↓	A↑	F↓	A↑	F↓
Training-free and non-adaptive methods														
MaxLogit	68.64	91.07	62.11	92.10	80.79	54.52	80.34	65.36	74.76	74.56	67.38	77.67	72.34	75.88
MCM	69.46	79.62	75.06	65.04	80.60	47.10	77.11	65.74	76.04	65.18	81.02	53.56	76.55	62.71
Energy	68.51	91.15	60.58	92.73	80.31	56.05	80.25	65.47	74.16	74.96	64.59	78.64	71.40	76.50
NegLabel	70.04	87.61	56.77	93.36	84.76	50.38	81.41	61.63	68.67	79.35	60.98	82.27	70.44	75.77
Global NegLabel	69.68	87.77	56.53	93.09	84.47	50.98	81.32	61.50	68.69	79.15	60.56	82.33	70.21	75.80
Training-free test-time adaptation methods														
AdaNeg	65.28	95.08	57.55	95.69	89.30	42.36	83.60	58.72	71.52	77.75	64.37	79.97	71.94	74.93
ANTS	75.19	77.92	63.37	87.73	93.08	29.83	88.88	45.75	76.87	68.97	70.97	72.47	**78.06**	63.78
Local TANL	71.12	79.93	55.49	90.11	92.89	24.25	81.94	55.98	68.34	80.93	72.64	71.22	73.74	67.07
Global Mean	71.59	79.86	55.86	89.26	94.19	23.10	82.68	53.94	65.88	82.76	71.65	72.94	73.64	66.98
FedRALA	72.46	77.68	59.29	86.15	94.17	19.94	84.66	45.20	72.39	77.94	76.33	65.03	76.55	**61.99**

Table 13: **Detailed per-OOD results with TinyImageNet as ID, $\alpha_{\text{ID}} = 0.5$, and Non-IID OOD streams.** Results are macro-averaged over ten clients. A and F denote AUROC and FPR95, respectively; the final pair is averaged over the six OOD datasets. All entries are percentages.

	SSBHard		NINCO		iNat.		SUN		Places		DTD		Average	
Method	A↑	F↓	A↑	F↓	A↑	F↓	A↑	F↓	A↑	F↓	A↑	F↓	A↑	F↓
Training-free and non-adaptive methods														
MaxLogit	68.27	89.72	61.97	91.37	80.68	55.40	79.98	64.80	75.26	72.46	67.05	78.50	72.20	75.38
MCM	69.47	78.66	75.08	64.87	80.52	47.79	76.97	63.90	76.44	62.53	80.69	53.86	76.53	61.94
Energy	68.11	89.77	60.44	91.84	80.17	57.30	79.88	65.05	74.65	73.14	64.30	79.44	71.26	76.09
NegLabel	69.57	86.47	56.58	92.99	84.72	50.71	81.29	61.74	68.89	78.18	60.89	82.08	70.32	75.36
Global NegLabel	69.21	86.45	56.25	92.70	84.36	50.36	81.12	61.64	68.95	77.67	60.42	82.23	70.05	75.17
Training-free test-time adaptation methods														
AdaNeg	65.58	90.49	55.79	93.89	89.34	42.44	82.83	61.47	71.56	76.30	63.68	80.91	71.47	74.25
ANTS	77.51	73.80	65.07	85.55	92.97	29.13	88.96	45.86	77.43	67.68	71.49	72.65	**78.91**	62.44
Local TANL	72.85	76.67	61.13	87.53	93.54	22.33	83.86	53.80	69.66	76.85	74.44	69.29	75.91	64.41
Global Mean	70.10	79.74	56.08	88.21	94.12	22.80	81.98	56.70	65.71	81.58	71.58	72.37	73.26	66.90
FedRALA	71.90	76.98	61.69	83.50	94.24	19.37	84.84	44.21	73.53	74.11	77.24	62.03	77.24	**60.03**

Table 14: **Detailed per-OOD results with TinyImageNet as ID, $\alpha_{\text{ID}} = 0.1$, and Disjoint OOD streams.** Results are macro-averaged over ten clients. A and F denote AUROC and FPR95, respectively; the final pair is averaged over the six OOD datasets. All entries are percentages.

	SSBHard		NINCO		iNat.		SUN		Places		DTD		Average	
Method	A↑	F↓	A↑	F↓	A↑	F↓	A↑	F↓	A↑	F↓	A↑	F↓	A↑	F↓
Training-free and non-adaptive methods														
MaxLogit	82.19	68.49	79.68	67.53	91.51	31.91	90.66	39.72	88.56	45.47	83.46	54.68	86.01	51.30
MCM	75.58	69.50	79.69	56.61	83.85	42.28	81.84	54.12	81.25	54.48	85.59	45.01	81.30	53.67
Energy	82.56	68.55	79.30	68.26	91.85	32.35	91.16	39.24	88.93	45.42	82.23	55.62	86.01	51.57
NegLabel	84.06	63.22	76.79	70.37	95.64	21.02	92.72	32.13	86.65	49.89	80.03	57.86	85.98	49.08
Global NegLabel	84.29	62.01	77.16	69.52	96.85	16.70	93.48	29.04	87.86	47.40	80.69	56.15	86.72	46.80
Training-free test-time adaptation methods														
AdaNeg	89.66	57.03	79.26	67.21	99.58	1.04	96.76	21.96	95.09	29.61	91.59	40.10	91.99	36.16
ANTS	87.33	53.35	81.18	64.92	97.91	10.26	95.58	21.95	89.85	45.89	86.79	46.69	89.77	40.51
Local TANL	89.74	45.41	81.40	59.20	98.17	6.88	94.65	20.46	88.96	42.16	89.67	39.43	90.43	35.59
Global Mean	89.69	44.87	82.16	56.86	98.41	6.28	94.95	19.33	88.83	42.86	90.34	38.09	90.73	34.72
FedRALA	90.68	41.36	85.37	49.43	98.51	4.89	96.03	15.29	91.30	35.76	92.45	31.59	**92.39**	**29.72**

Table 15: **Detailed per-OOD results with TinyImageNet as ID, $\alpha_{\text{ID}} = 0.1$, and Non-IID OOD streams.** Results are macro-averaged over ten clients. A and F denote AUROC and FPR95, respectively; the final pair is averaged over the six OOD datasets. All entries are percentages.

	SSBHard		NINCO		iNat.		SUN		Places		DTD		Average	
Method	A↑	F↓	A↑	F↓	A↑	F↓	A↑	F↓	A↑	F↓	A↑	F↓	A↑	F↓
Training-free and non-adaptive methods														
MaxLogit	81.17	68.18	78.78	67.26	91.58	30.82	90.97	37.76	88.35	45.87	82.69	59.21	85.59	51.51
MCM	76.03	67.11	78.47	60.31	83.68	42.42	82.33	51.40	81.08	53.38	84.97	49.48	81.09	54.02
Energy	81.43	68.43	78.46	67.96	91.97	31.01	91.46	37.26	88.72	45.66	81.51	60.41	85.59	51.79
NegLabel	83.04	62.82	76.04	68.37	95.70	19.81	93.08	32.24	86.48	48.32	79.36	62.15	85.62	48.95
Global NegLabel	83.26	61.98	76.39	66.82	96.92	15.47	93.75	29.58	87.73	45.82	79.99	60.71	86.34	46.73
Training-free test-time adaptation methods														
AdaNeg	87.83	59.49	83.33	59.53	99.67	0.68	97.12	19.04	95.01	29.00	89.06	42.28	92.00	35.00
ANTS	90.65	44.69	82.89	58.54	97.75	11.23	96.36	19.30	90.98	39.74	86.38	51.28	90.84	37.46
Local TANL	93.60	32.61	83.69	57.65	98.51	6.11	96.23	15.33	90.79	34.66	90.19	40.56	92.17	31.15
Global Mean	88.78	45.98	79.58	61.30	98.47	5.98	95.04	18.35	87.96	42.74	88.89	45.82	89.79	36.70
FedRALA	89.69	43.46	85.51	46.45	98.54	4.64	96.33	13.09	91.86	32.87	92.37	33.12	**92.38**	**28.94**

Table 16: **Detailed 100-client results with TinyImageNet as ID, $\alpha_{\text{ID}} = 0.5$, and Disjoint OOD streams.** Results are macro-averaged over 100 clients. A and F denote AUROC and FPR95, respectively; the final pair is averaged over the six OOD datasets. All entries are percentages.

	SSBHard		NINCO		iNat.		SUN		Places		DTD		Average	
Method	A↑	F↓	A↑	F↓	A↑	F↓	A↑	F↓	A↑	F↓	A↑	F↓	A↑	F↓
Local TANL	75.40	73.91	51.91	89.83	72.05	57.70	73.89	62.84	65.74	72.76	66.93	75.12	67.65	72.03
Global Mean	75.93	74.20	51.91	89.83	72.33	58.25	73.60	63.08	65.40	72.73	66.93	75.12	67.68	72.20
FedRALA	76.33	69.66	51.89	89.77	72.53	56.17	74.04	60.95	66.42	71.54	66.87	74.98	**68.01**	**70.51**

Table 17: **Detailed 100-client results with TinyImageNet as ID, $\alpha_{\text{ID}} = 0.5$, and Non-IID OOD streams.** Results are macro-averaged over 100 clients. A and F denote AUROC and FPR95, respectively; the final pair is averaged over the six OOD datasets. All entries are percentages.

	SSBHard		NINCO		iNat.		SUN		Places		DTD		Average	
Method	A↑	F↓	A↑	F↓	A↑	F↓	A↑	F↓	A↑	F↓	A↑	F↓	A↑	F↓
Local TANL	76.11	70.56	52.30	86.61	72.45	57.36	75.30	61.48	66.56	71.94	67.05	71.35	68.30	69.88
Global Mean	73.83	73.39	52.18	86.78	72.46	58.86	74.25	62.73	65.38	72.59	66.98	71.46	67.51	70.97
FedRALA	76.28	68.01	52.42	86.47	72.25	56.30	75.05	59.49	66.79	70.49	67.11	71.07	**68.32**	**68.64**

Table 18: **Detailed 100-client results with TinyImageNet as ID, $\alpha_{\text{ID}} = 0.2$, and Disjoint OOD streams.** Results are macro-averaged over 100 clients. A and F denote AUROC and FPR95, respectively; the final pair is averaged over the six OOD datasets. All entries are percentages.

	SSBHard		NINCO		iNat.		SUN		Places		DTD		Average	
Method	A↑	F↓	A↑	F↓	A↑	F↓	A↑	F↓	A↑	F↓	A↑	F↓	A↑	F↓
Local TANL	86.62	53.38	70.38	74.32	86.78	35.43	87.25	37.50	82.42	49.10	81.67	53.99	82.52	50.62
Global Mean	86.58	53.59	70.35	74.31	86.99	36.07	87.06	38.00	82.10	49.41	81.65	54.01	82.46	50.90
FedRALA	88.10	47.99	70.49	74.07	87.35	32.86	87.71	34.87	83.20	46.82	81.65	53.80	**83.08**	**48.40**

Table 19: **Detailed 100-client results with TinyImageNet as ID, $\alpha_{\text{ID}} = 0.2$, and Non-IID OOD streams.** Results are macro-averaged over 100 clients. A and F denote AUROC and FPR95, respectively; the final pair is averaged over the six OOD datasets. All entries are percentages.

	SSBHard		NINCO		iNat.		SUN		Places		DTD		Average	
Method	A↑	F↓	A↑	F↓	A↑	F↓	A↑	F↓	A↑	F↓	A↑	F↓	A↑	F↓
Local TANL	87.44	49.92	71.08	67.30	86.44	35.61	88.15	37.29	83.07	48.14	82.64	50.21	83.14	48.08
Global Mean	85.16	54.68	70.53	67.82	86.12	37.20	87.20	38.62	81.84	49.85	82.26	50.36	82.18	49.75
FedRALA	87.18	48.81	71.34	66.64	86.63	33.41	87.92	35.16	83.21	46.33	82.66	49.35	**83.16**	**46.62**

Table 20: **Detailed 100-client results with TinyImageNet as ID, $\alpha_{\text{ID}} = 0.1$, and Disjoint OOD streams.** Results are macro-averaged over 100 clients. A and F denote AUROC and FPR95, respectively; the final pair is averaged over the six OOD datasets. All entries are percentages.

	SSBHard		NINCO		iNat.		SUN		Places		DTD		Average	
Method	A↑	F↓	A↑	F↓	A↑	F↓	A↑	F↓	A↑	F↓	A↑	F↓	A↑	F↓
Local TANL	92.92	34.34	83.18	56.87	93.90	19.01	94.09	20.90	91.52	29.38	90.83	34.48	91.07	32.50
Global Mean	92.82	34.18	83.24	56.77	94.04	19.29	94.04	20.87	91.42	28.62	90.85	34.42	91.07	32.36
FedRALA	93.97	29.17	83.57	55.91	94.26	17.40	94.33	18.32	92.06	26.55	90.98	33.90	**91.53**	**30.21**

Table 21: **Detailed 100-client results with TinyImageNet as ID, $\alpha_{\text{ID}} = 0.1$, and Non-IID OOD streams.** Results are macro-averaged over 100 clients. A and F denote AUROC and FPR95, respectively; the final pair is averaged over the six OOD datasets. All entries are percentages.

	SSBHard		NINCO		iNat.		SUN		Places		DTD		Average	
Method	A↑	F↓	A↑	F↓	A↑	F↓	A↑	F↓	A↑	F↓	A↑	F↓	A↑	F↓
Local TANL	92.77	31.02	83.90	47.19	94.22	19.46	94.47	20.49	91.67	28.52	91.26	29.85	**91.38**	29.42
Global Mean	90.96	37.19	83.25	48.27	93.85	20.64	93.83	21.99	90.68	30.74	91.00	30.38	90.59	31.54
FedRALA	92.19	32.62	84.21	45.78	94.28	18.27	94.31	19.56	91.56	28.03	91.29	29.26	91.31	**28.92**