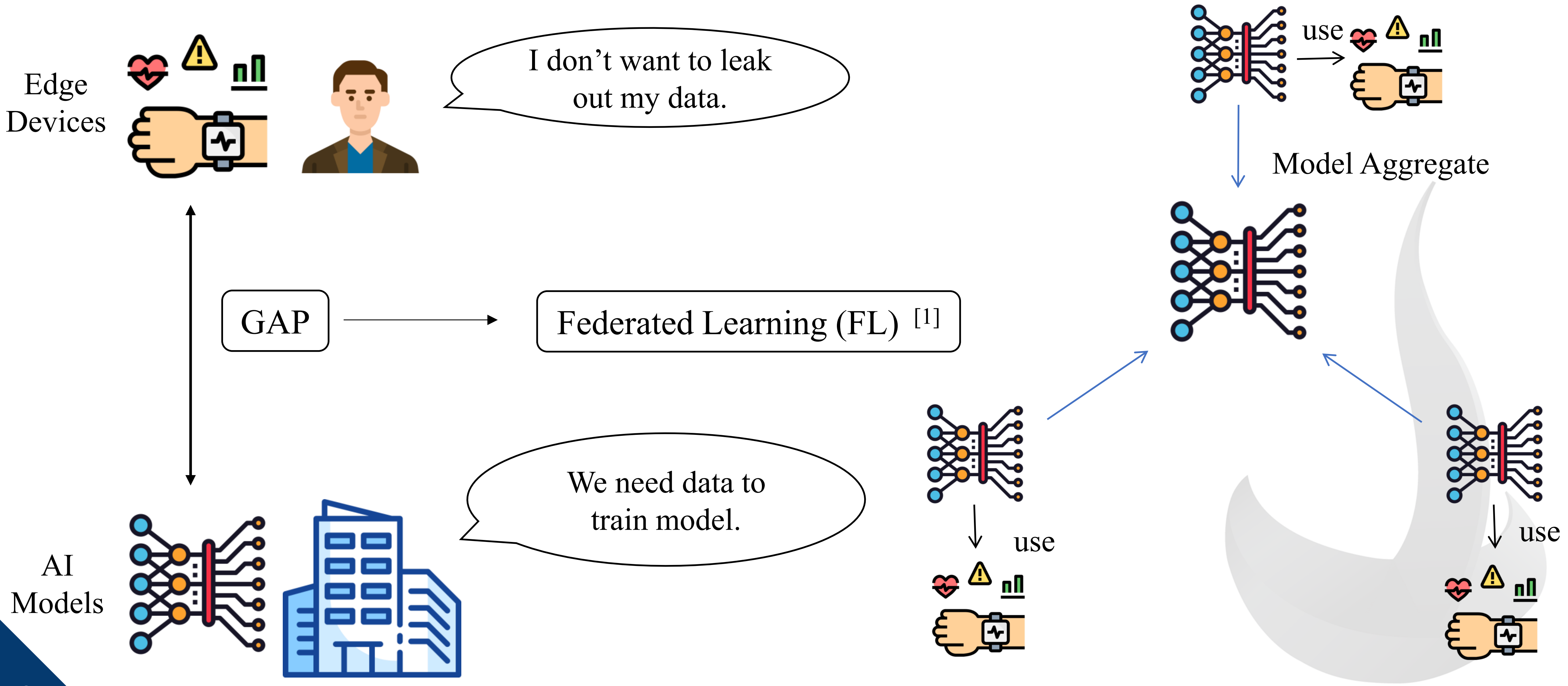
FedWCM: Unleashing the Potential of Momentum-based Federated Learning in Long-Tailed Scenarios

Tianle Li[1*], **Yongzhi Huang[2*]**,
Linshan Jiang[3], Qipeng Xie[2], Chang Liu[4], Wenfeng Du[1], Lu Wang[1]
and Kaishun Wu[2]

[1]Shenzhen University
[2]The Hong Kong University of Science and Technology (Guangzhou)
[3]National University of Singapore
[4]Nanyang Technological University
* Equal Contribution

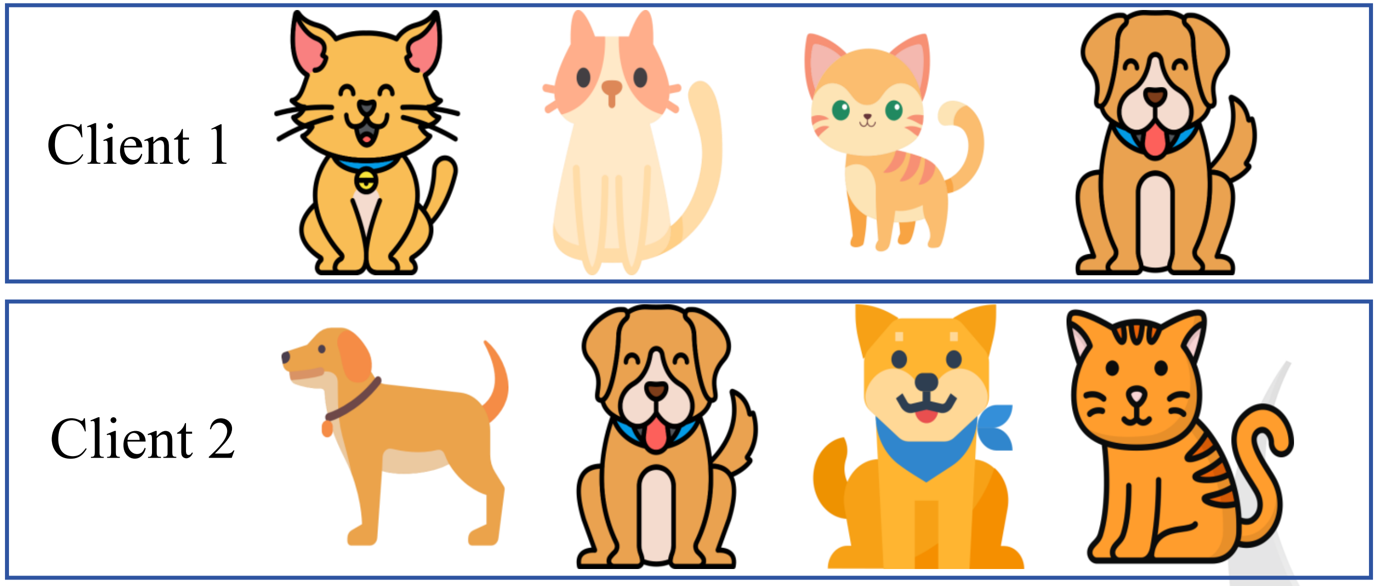
Motivation: Edge Intelligence Needs FL



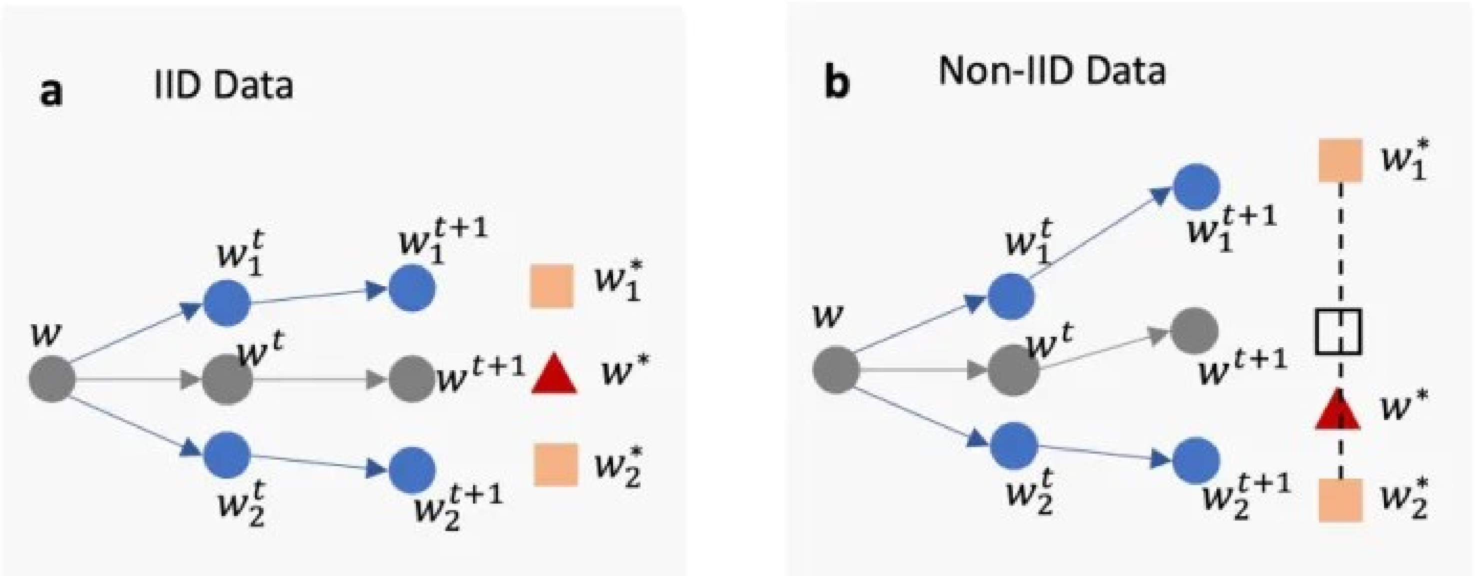
[1] Zhang T , Gao L , He C ,et al. Federated Learning for Internet of Things: Applications, Challenges, and Opportunities[J]. 2021.DOI:10.48550/arXiv.2111.07494.

Big headache in FL: Heterogeneity

Heterogeneity



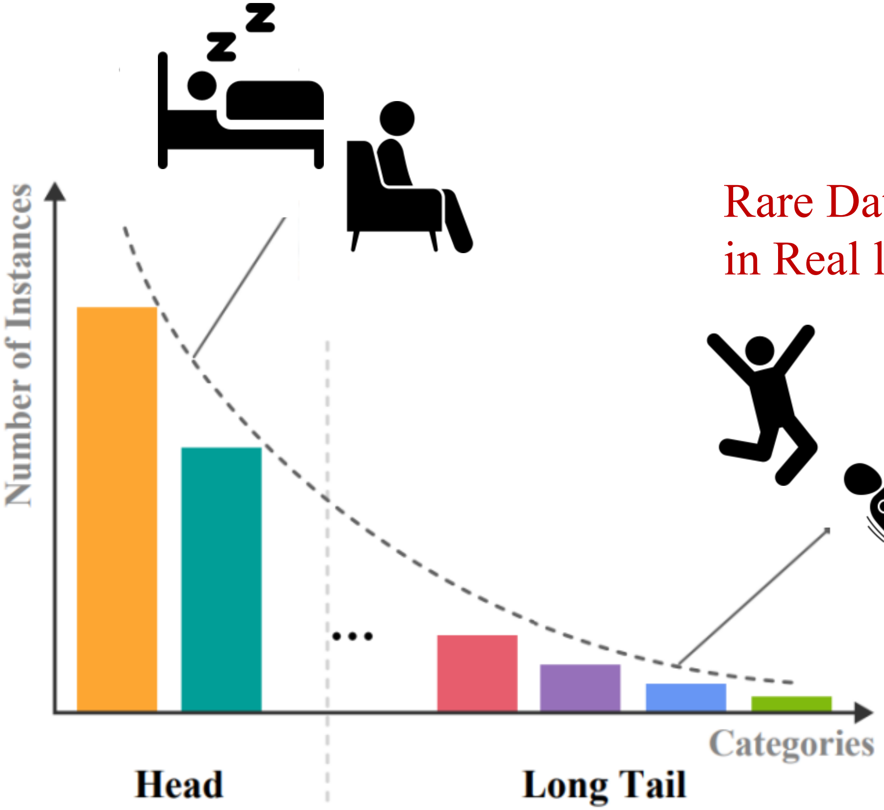
Client Drift[2]



[2] Karimireddy S P, Kale S, Mohri M, et al. Scaffold: Stochastic controlled averaging for federated learning[C]//International conference on machine learning. PMLR, 2020: 5132-5143.

The extreme case: Long-tail Distribution

Long-tail Distribution in reality[3]



Long-tailed heterogeneous data

[3] Yang L, Jiang H, Song Q, et al. A survey on long-tailed visual recognition[J]. International Journal of Computer Vision, 2022, 130(7): 1837-1872.

Existing Approaches for Heterogeneity & Long-Tail

Long-Tail

Federated Long-tailed (Local Corrector)

Calibrate classifiers, two-stage learning

Local-global Distributions misaligned

Federated Long-tailed (Global Corrector)

FedGrab[4], CLIP2FL[5], Creff[6]

Only correct the imbalance without speeding up

How to make FL both fast and fair under Heterogeneity & Long-Tail?

Heterogeneity

Centralized Long-tailed Approaches

Data layer: Enhancement, Balanced sampling

Training layer: Balance loss, Focal loss, Distribution-aware loss

Lost too much information in the client side

Why not introduce Momentum?

Heterogeneity Approaches

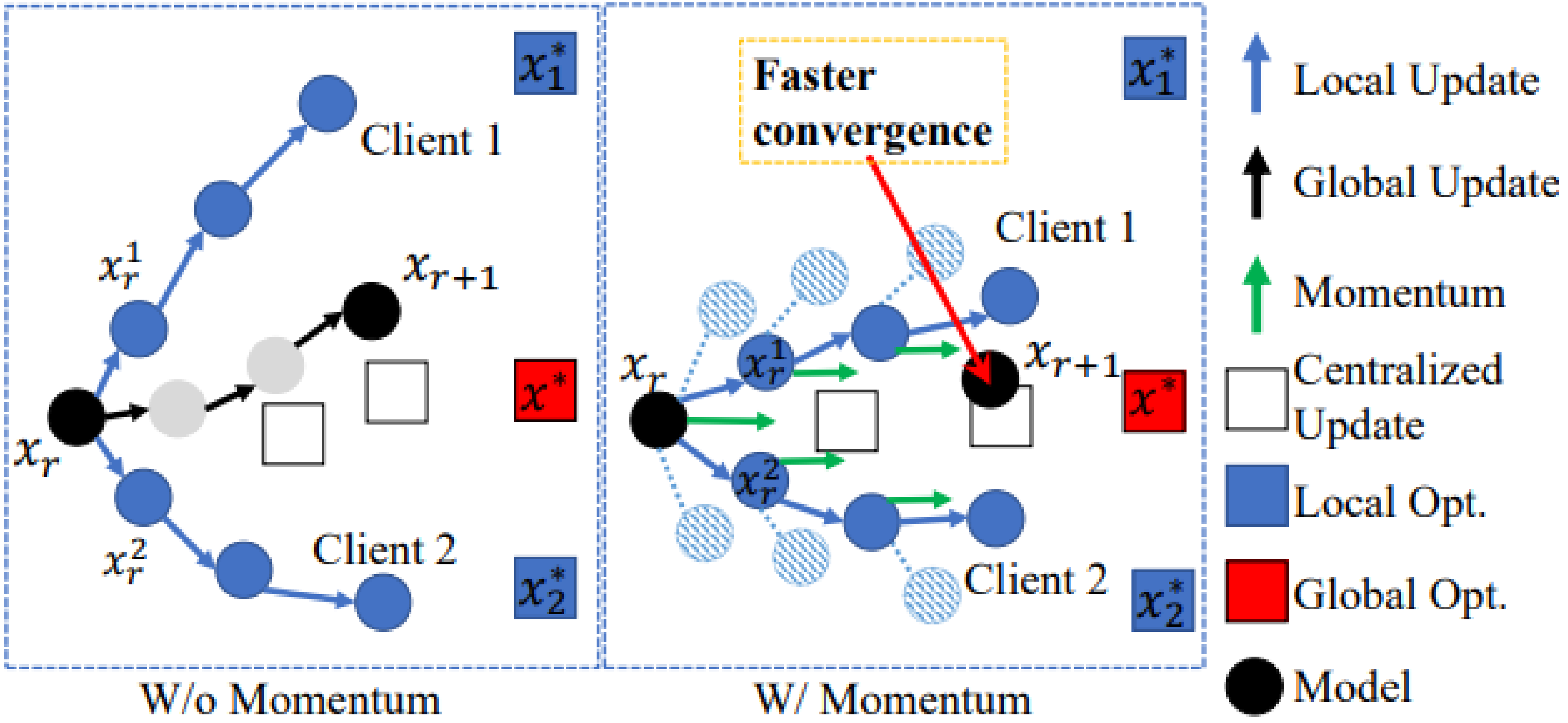
FedAvg, Scaffold, FedPro

Do not tackle the class imbalance

[4] Xiao Z, Chen Z, Liu S, et al. Fed-grab: Federated long-tailed learning with self-adjusting gradient balancer[J]. Advances in Neural Information Processing Systems, 2023, 36: 77745-77757.
[5] Shi J, Zheng S, Yin X, et al. Clip-guided federated learning on heterogeneity and long-tailed data[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2024, 38(13): 14955-14963.
[6] Shang X, Lu Y, Huang G, et al. Federated learning on heterogeneous and long-tailed data via classifier re-training with federated features[J]. arXiv preprint arXiv:2204.13399, 2022.

Momentum: A Powerful Accelerator

Momentum can stabilize and accelerate training[7][8]



[7] Xu J, Wang S, Wang L, et al. Fedcm: Federated learning with client-level momentum[J]. arXiv preprint arXiv:2106.10874, 2021.
[8] Cheng Z, Huang X, Wu P, et al. Momentum benefits non-iid federated learning simply and provably[J]. arXiv preprint arXiv:2306.16504, 2023.

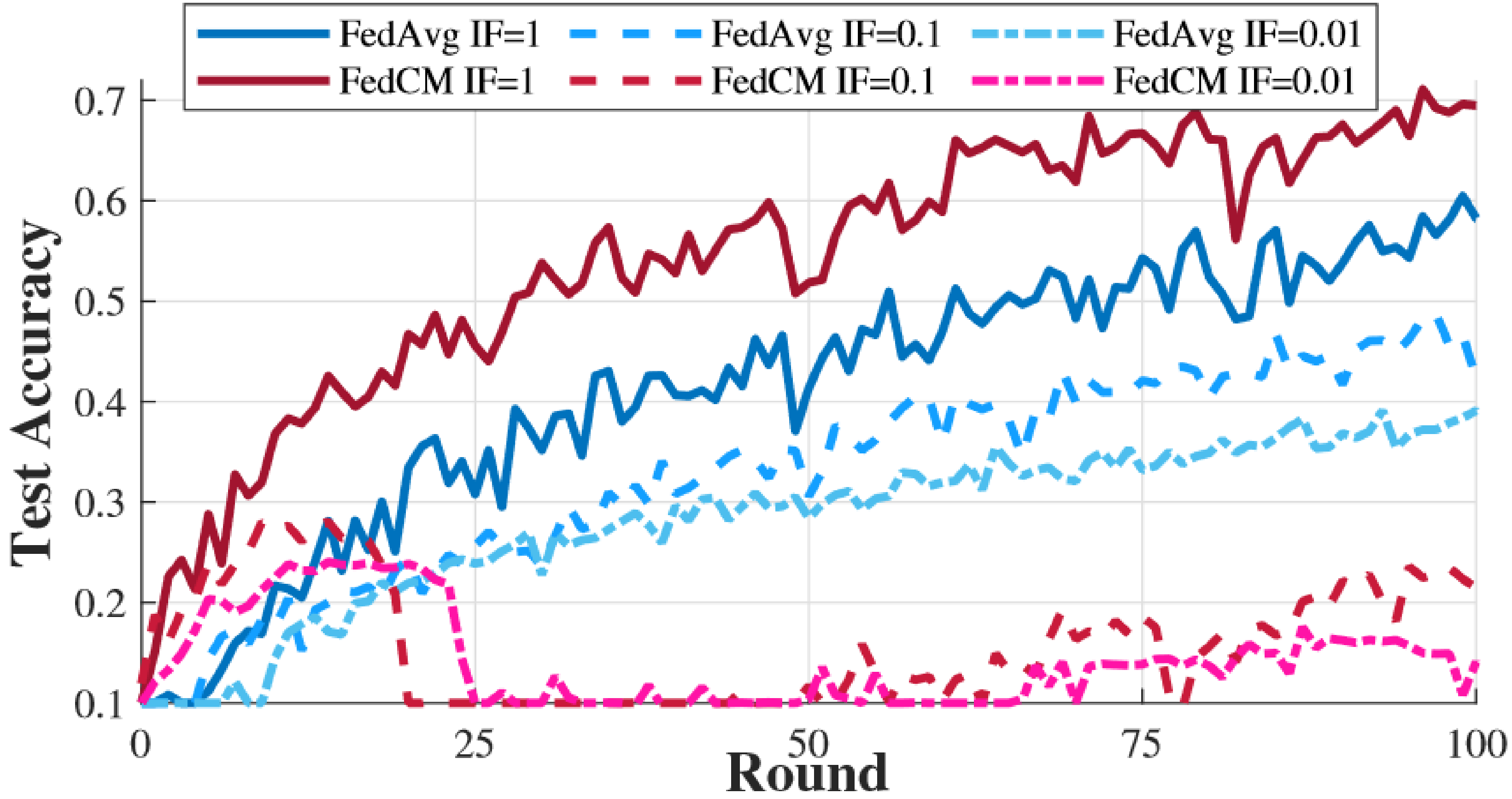
But …

It is a Double-Edged Sword.

$$IF = \frac{Least}{Most}$$

Distributed momentum, FedCM (in red series)

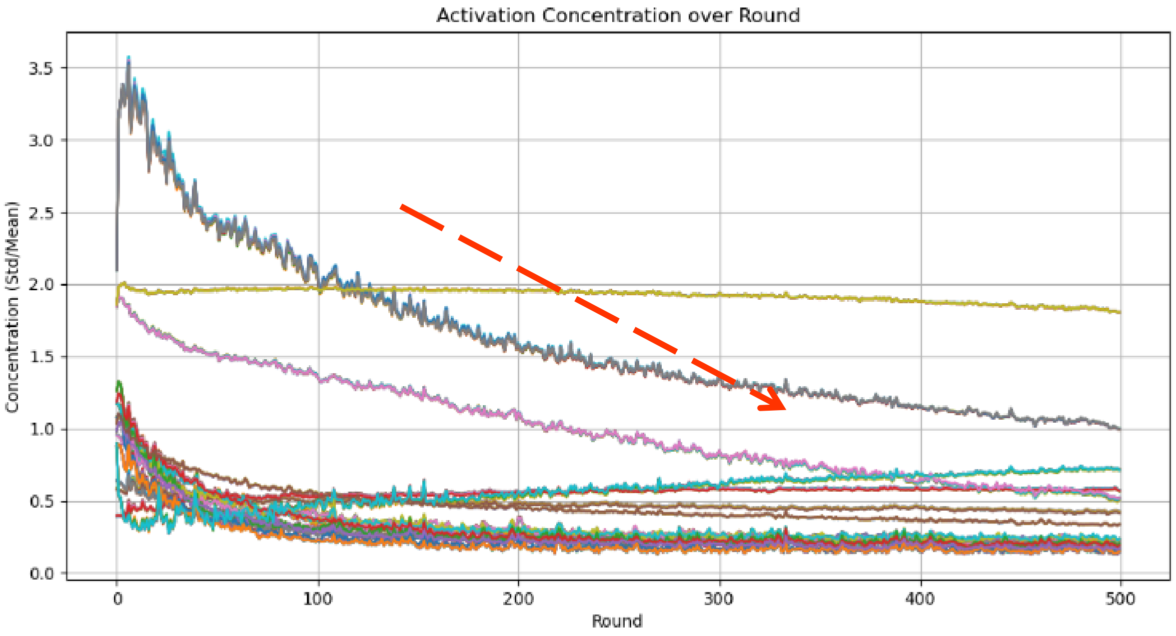
1. Fast convergence and high accuracy when the long tail is low (IF=1)
2. Failure to converge when the long tail is high (IF=0.1, 0.01)

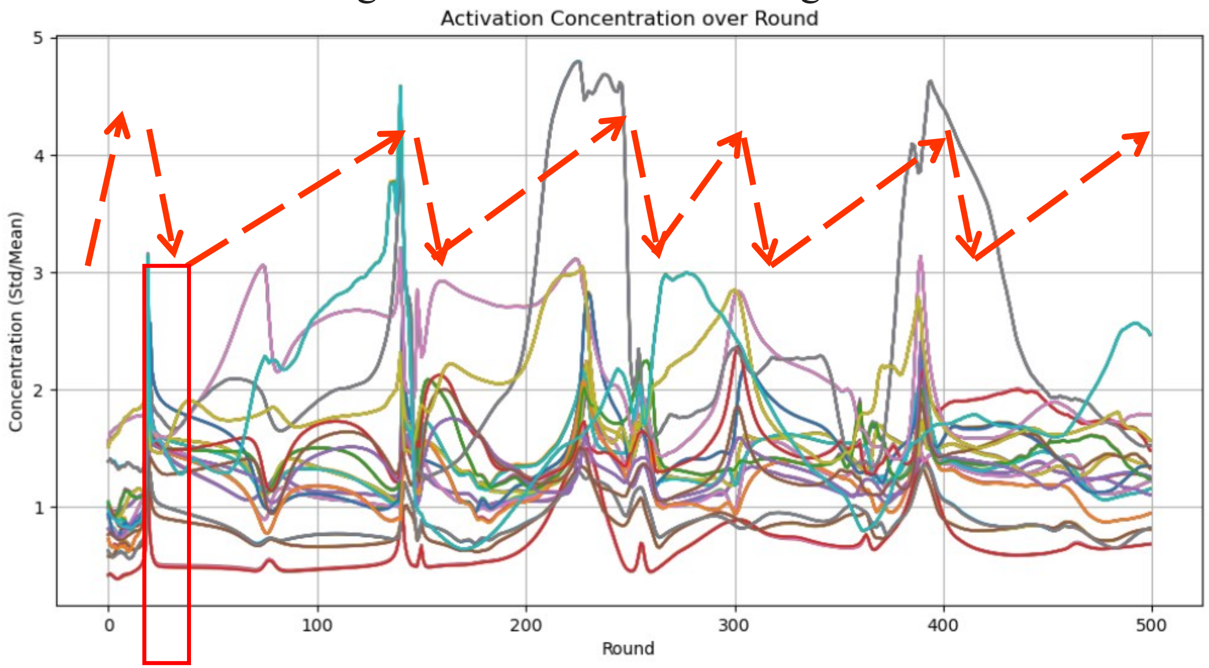


Why fail to converge?

Illustration: The Minority Collapse (Micro)

FedAvg (left) and FedCM (right) neural network layers' neuron concentration changes with rounds under long-tail distribution





Gradually changes, with an overall downward trend

Increases and then decreases sharply at a certain point

Up and down repeatedly

Major classes take advantage under Lon-tail. → Momentum exacerbates it. → Leading to the majority class gradient accumulation → The geometric structure of neurons is destroyed[9] → **Accuracy drops sharply**

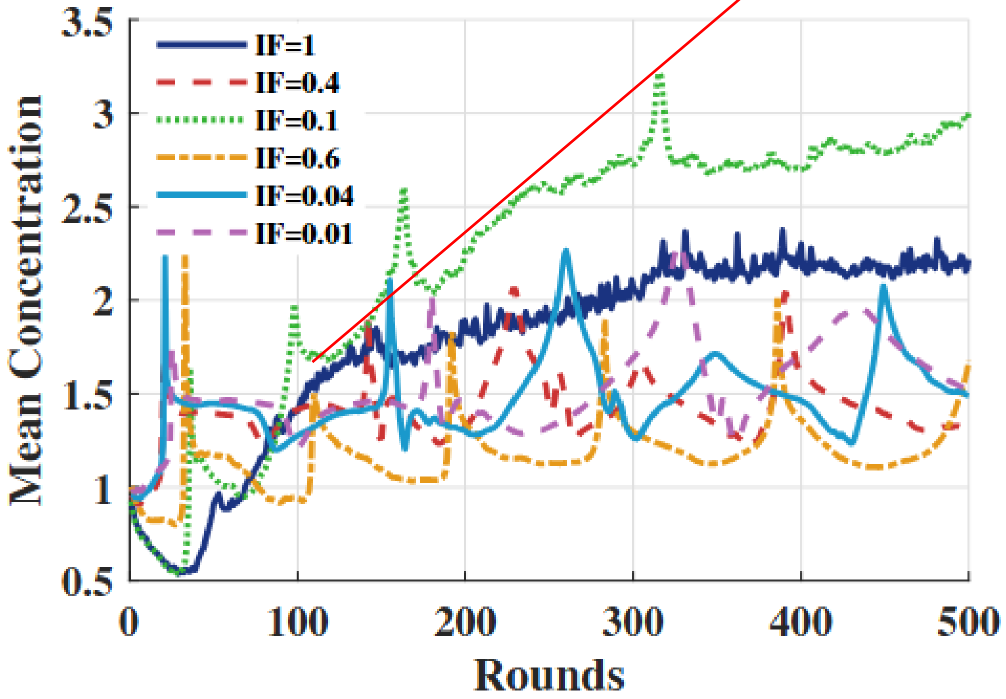
Similar to neural network poisoning

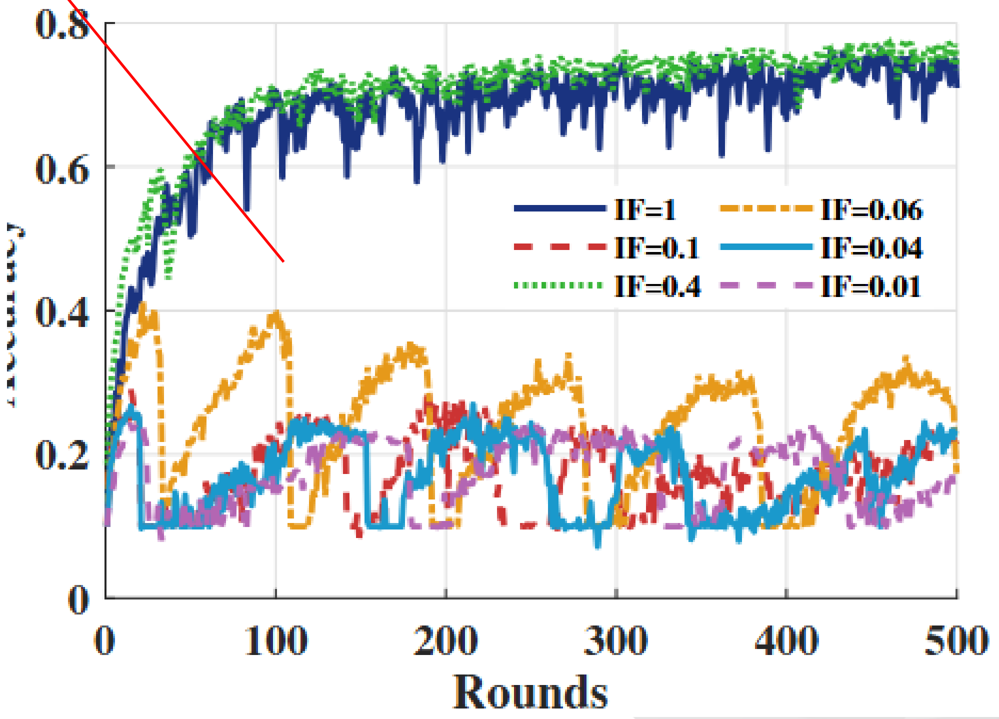
Similar to neural network disorder

[9] Yang Y, Chen S, Li X, et al. Inducing neural collapse in imbalanced learning: Do we really need a learnable classifier at the end of deep neural network?[J]. Advances in neural information processing systems, 2022, 35: 37991-38002.

Illustration: The Minority Collapse (Macro)

Minority Collapse[10] caused by majority class gradient accumulation.

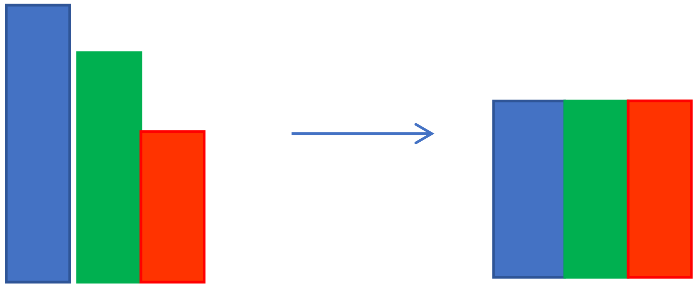


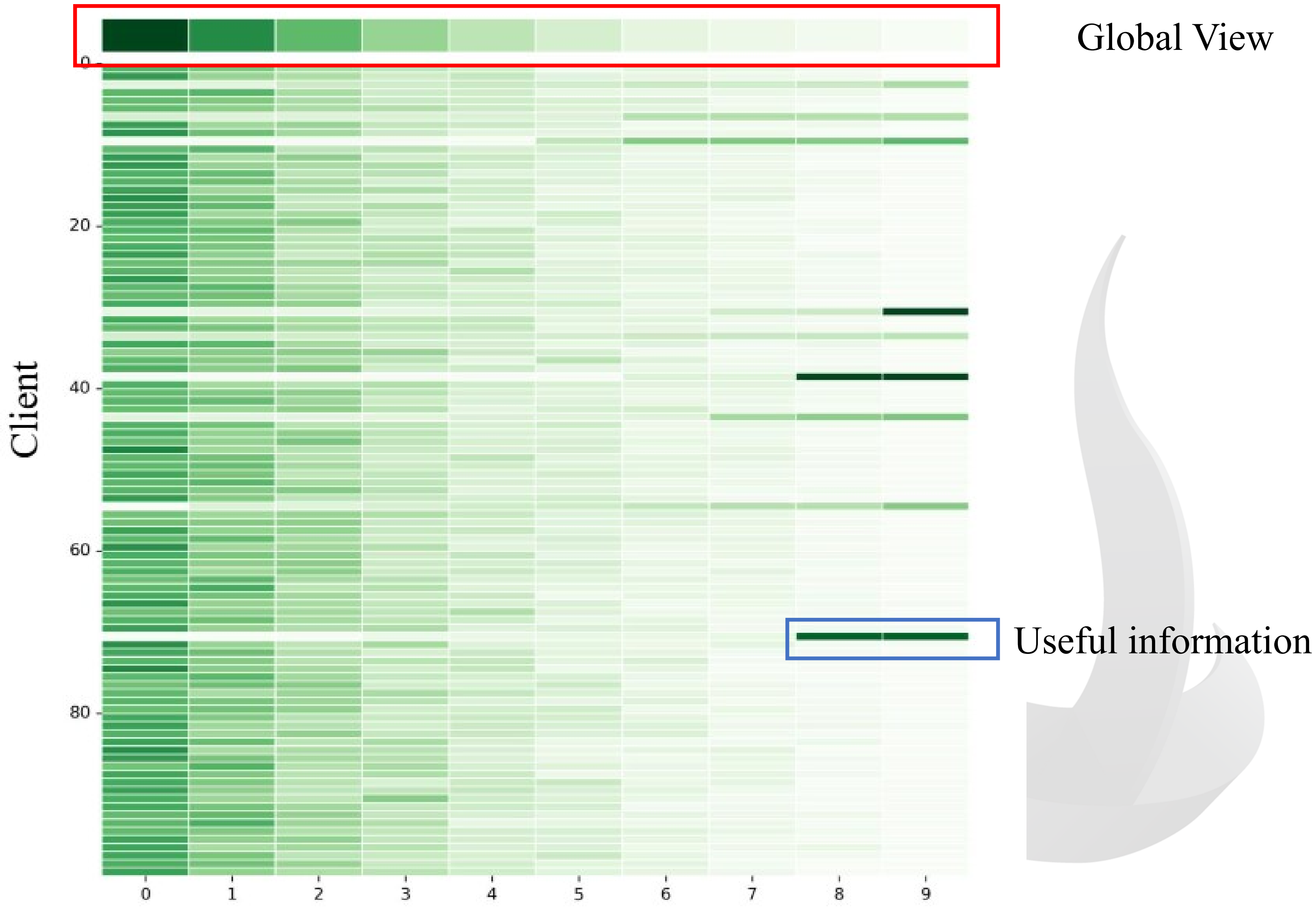


[10] Fang C, He H, Long Q, et al. Exploring deep neural networks via layer-peeled model: Minority collapse in imbalanced training[J]. Proceedings of the National Academy of Sciences, 2021, 118(43): e2103091118.

Why Not Just Fix Locally?

Local balance?

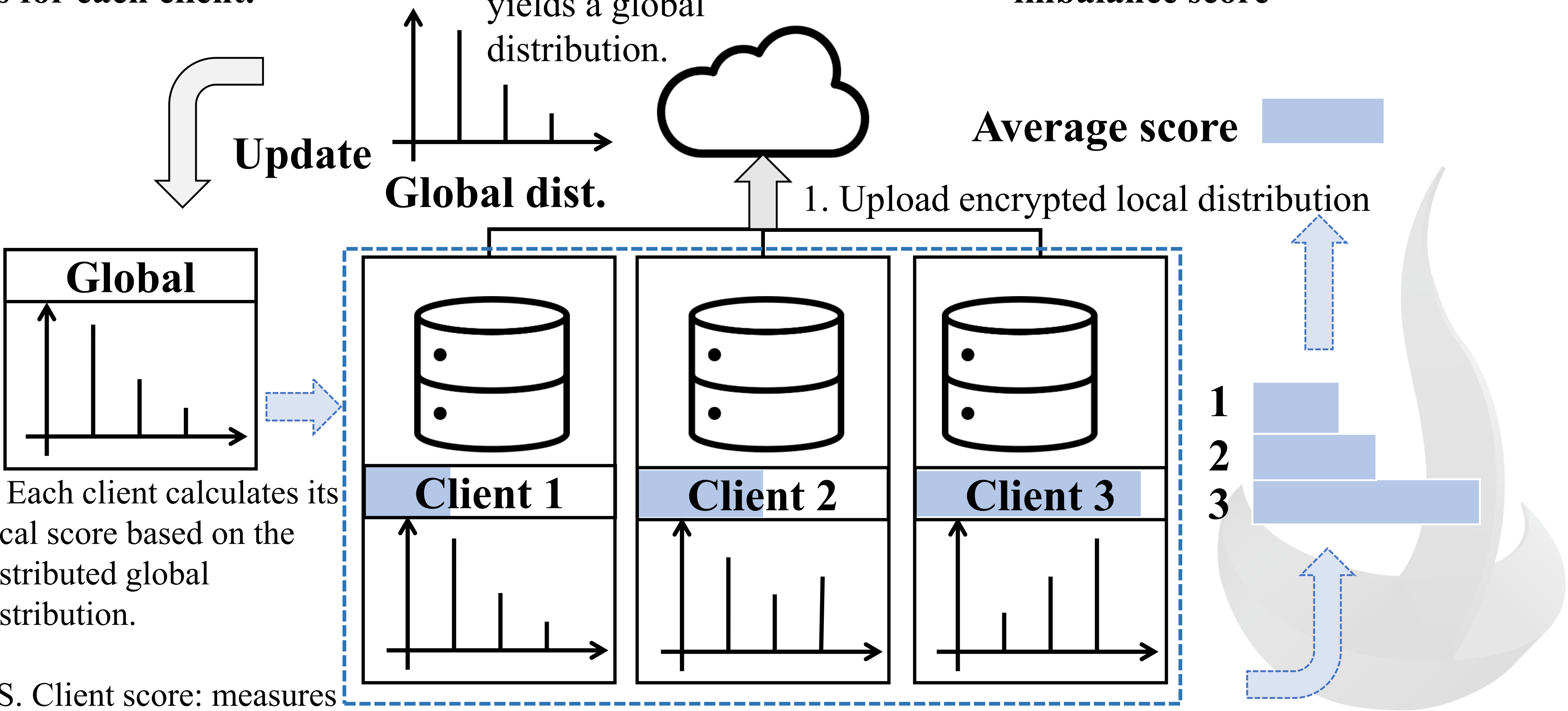




Obtaining the Global Distribution

Target 1. Data quality scores for each client.

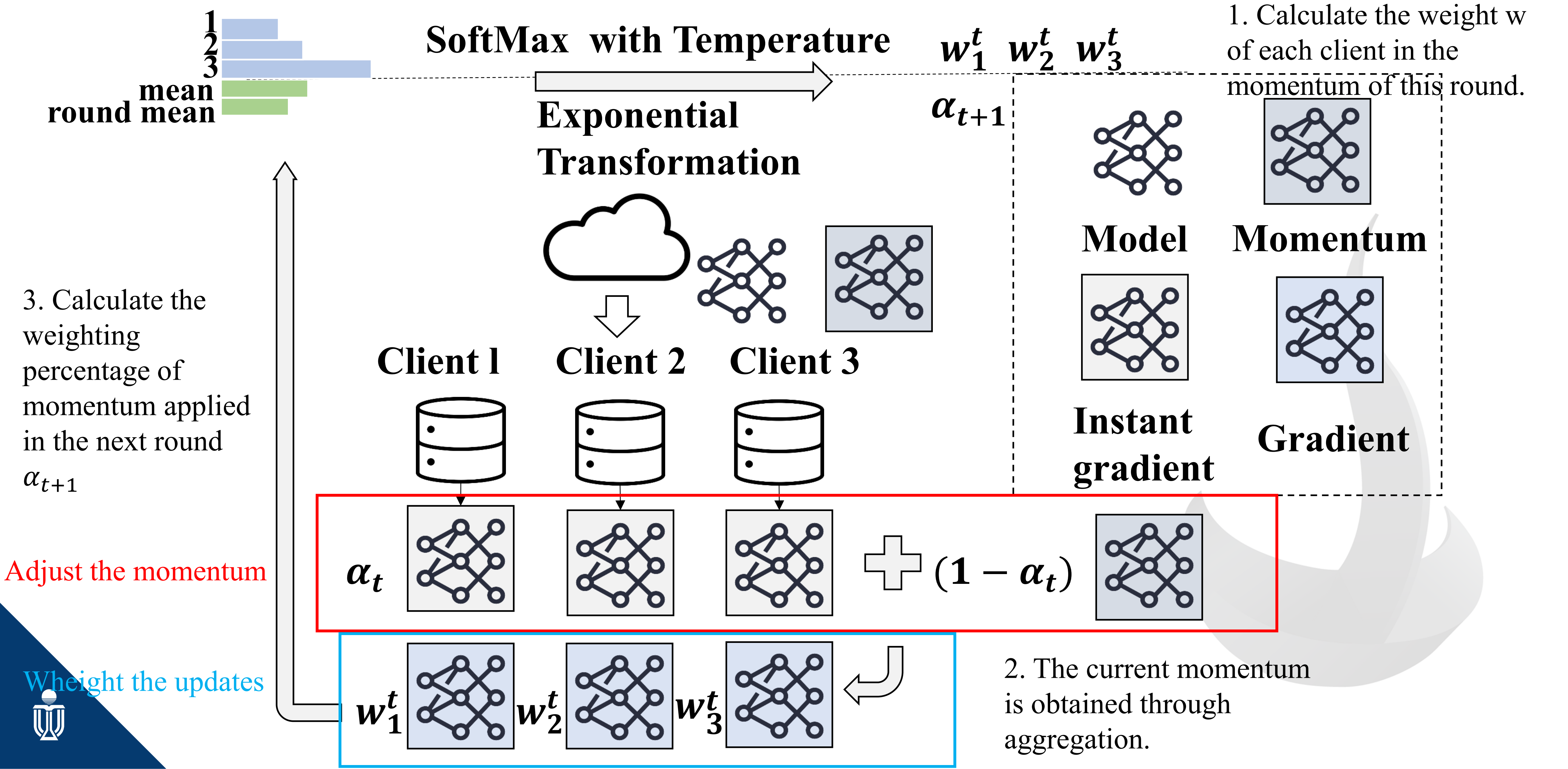
Target 2. An overall imbalance score



3. Each client calculates its local score based on the distributed global distribution.

P.S. Client score: measures the difference from a uniform distribution.

Our Solution: Weighted Client Momentum (FedWCM)



Method: How FedWCM Works

FedWCM workflow:

1. Estimate the global distribution.

2. Score each client — how much do they help the minority classes?

3. Aggregate their momentum with these scores.

4. Adjust the weight of momentum application for the next round

Algorithm 1 FedWCM Algorithm

Require: initial model x_0, global momentum Δ_0, $\alpha_0 = 0.1$, learning rates η_l, η_g, number of rounds R, local iterations B

Compute $\{s_k\}$ with D_g using Equation (3) 1 and 2

for $r = 0$ to $R-1$ **do**

  Sample subset $\mathcal{P}_r$ of clients

  for Each client $k \in \mathcal{P}_r$ **do**

    $x^r_{0,k} = x_r$

    for $b = 0$ to $B-1$ **do**

      Compute $g^r_{b,k} = \nabla f_k(x^r_{b,k}, D_{b,k})$

      $v^r_{b,k} = \alpha_r g^r_{b,k} + (1-\alpha_r)\Delta_r$ 4

      $x^r_{b+1,k} = x^r_{b,k} - \eta_l v^r_{b,k}$

    end for

    $\Delta^r_k = x^r_{B,k} - x_r$

  end for

  Compute w^r_k using Equation (4)

  Compute α_{r+1} using Equation (5)

  $\Delta_{r+1} = \frac{1}{\eta_l B} \sum_{k \in \mathcal{P}_r} w^r_k \Delta^r_k$ 3

  $x_{r+1} = x_r - \eta_g \Delta_{r+1}$

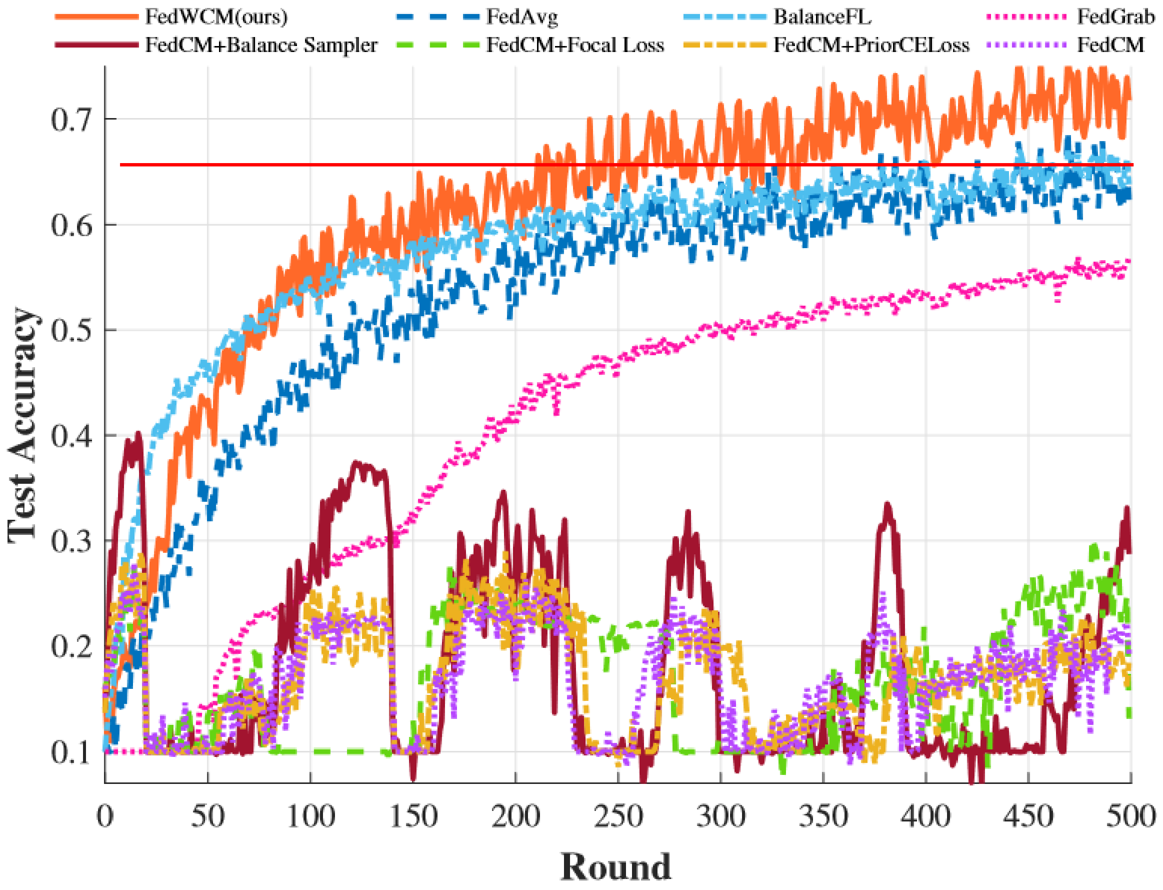
end for

Results: Higher Accuracy and Covergence

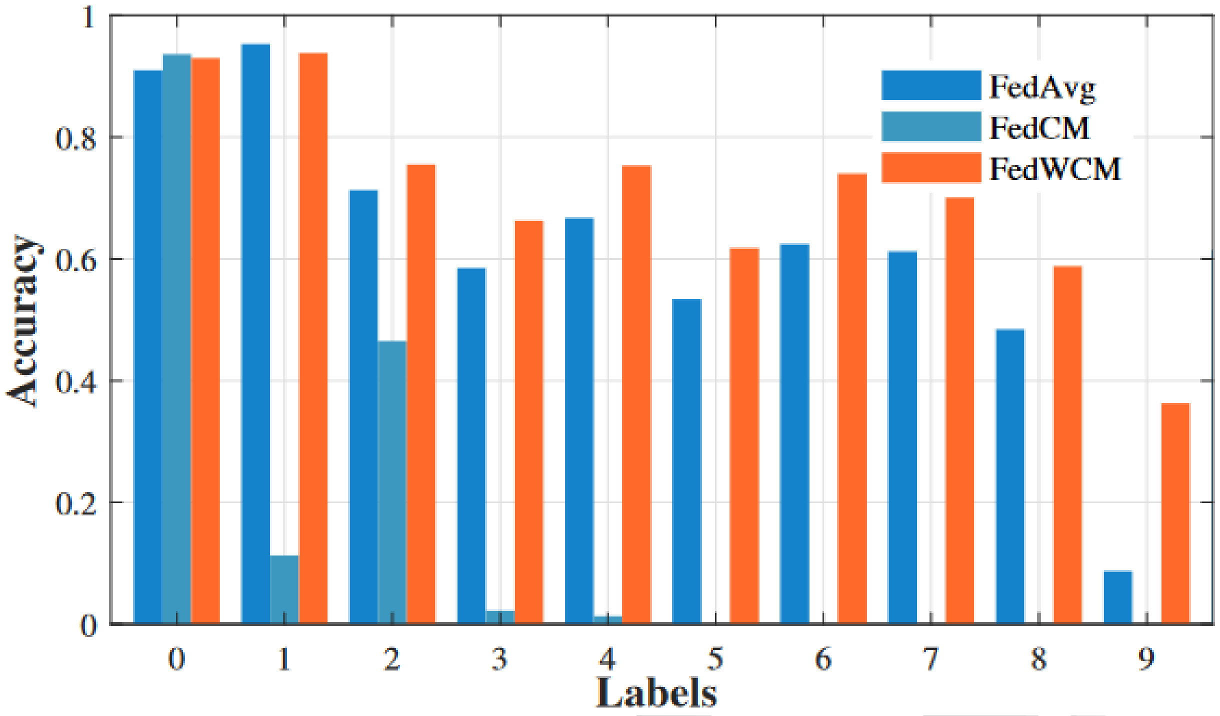
Federated long-tail Method | **Centralized Long Tail Improvement**

Dataset	IF	FedAvg		BalanceFL		FedCM		FedCM + Focal Loss		FedCM + Balance Loss		FedCM + Balance Sampler		FedWCM	
		0.6	0.1	0.6	0.1	0.6	0.1	0.6	0.1	0.6	0.1	0.6	0.1	0.6	0.1
Fashion-MNIST	1	**0.8800**	0.8074	0.8795	**0.8443**	0.8419	0.7604	0.8246	0.6931	0.7451	0.6907	0.8546	0.7821	0.8625	0.8181
	0.5	**0.8688**	0.8079	0.8638	0.8462	0.8601	**0.8544**	0.8363	0.7058	0.7622	0.6737	0.8476	0.7906	0.8659	0.8366
	0.1	0.8450	0.8313	**0.8497**	**0.8475**	0.8211	0.8268	0.8161	0.8065	0.7873	0.8002	0.8245	0.8252	0.8469	0.8328
	0.05	0.8318	0.8408	**0.8520**	**0.8545**	0.3914	0.4975	0.1945	0.1967	0.4335	0.4064	0.5474	0.5273	0.8499	0.8426
	0.01	0.7871	0.7894	**0.8192**	**0.8126**	0.7378	0.7524	0.7188	0.7265	0.8039	0.8011	0.8027	0.8030	0.7882	0.7947
SVHN	1	0.9361	0.8986	**0.9370**	0.9146	0.9246	0.8836	0.9242	0.8749	0.8928	0.8312	0.9310	0.8911	0.9355	**0.9276**
	0.5	0.9251	0.9271	0.9261	0.9243	0.7137	0.6981	0.6068	0.5961	0.5594	0.5870	0.7085	0.6761	**0.9324**	**0.9284**
	0.1	0.8681	0.8741	0.8979	0.8976	0.1594	0.0670	0.1959	0.1959	0.0762	0.1322	0.1959	0.1976	**0.9057**	**0.9024**
	0.05	0.8594	0.8647	0.8675	0.8709	0.3100	0.0723	0.0670	0.1107	0.0969	0.0909	0.1802	0.0932	**0.8759**	**0.8836**
	0.01	0.7884	0.7803	0.7901	0.7954	0.0670	0.0670	0.0670	0.0751	0.0760	0.0759	0.1736	0.2427	**0.7998**	**0.8408**
CIFAR-10	1	0.7906	0.6881	0.7629	0.6813	0.8126	0.7092	0.8040	0.6937	0.7931	0.7169	0.8065	0.7198	**0.8242**	**0.7337**
	0.5	0.7535	0.7183	0.7539	0.7429	0.6793	0.6686	0.6565	0.6319	0.6877	0.6924	0.6968	0.6590	**0.7926**	**0.7968**
	0.1	0.6232	0.6775	0.6380	0.6541	0.2175	0.2393	0.1311	0.3095	0.1864	0.3016	0.2871	0.3994	**0.6905**	**0.7207**
	0.05	0.5715	0.5642	0.5652	0.5535	0.2274	0.2358	0.2005	0.1413	0.2680	0.2525	0.1427	0.1315	**0.6006**	**0.6132**
	0.01	0.4567	0.4600	0.4731	0.4616	0.1865	0.2312	0.1687	0.2023	0.2087	0.2405	0.1249	0.1584	**0.4983**	**0.5012**
CIFAR-100	1	0.4297	0.3731	0.3691	0.3232	0.4129	0.2400	0.3990	0.2357	0.3630	0.2089	0.3599	0.2339	**0.4545**	**0.3858**
	0.5	0.3545	0.3882	0.3203	0.3639	0.2996	0.4200	0.3058	0.3853	0.2694	0.3722	0.2835	0.3790	**0.4195**	**0.4202**
	0.1	0.2839	0.2744	0.2440	0.2407	0.2948	0.3135	0.3014	0.3166	0.2952	0.3156	0.2952	0.2955	**0.3150**	**0.3235**
	0.05	0.2155	0.2300	0.2070	0.2157	0.1130	0.2695	0.0100	0.2806	0.1000	0.2786	0.0930	0.2721	**0.2573**	**0.2832**
	0.01	0.1663	0.1885	0.1565	0.1609	0.0116	0.1035	0.0109	0.1027	0.0100	0.1286	0.0100	0.0723	**0.1985**	**0.2005**
ImageNet	1	0.2760	0.2290	0.2292	0.1947	0.2479	0.1408	0.2438	0.1222	0.2082	0.1024	0.2134	0.1155	**0.3094**	**0.2462**
	0.5	0.2154	0.2140	0.1628	0.2124	0.1045	0.0392	0.0923	0.0695	0.0928	0.0544	0.1154	0.1067	**0.2598**	**0.2198**
	0.1	0.1631	0.1535	0.1124	0.1161	0.1796	0.1738	0.1864	0.1763	0.1796	0.1788	0.1528	0.1521	**0.1923**	**0.1874**
	0.05	0.1458	0.1355	0.0915	0.0998	0.0052	0.1597	0.1355	0.1448	0.1471	0.1576	0.1130	0.1542	**0.1626**	**0.1660**
	0.01	0.0882	0.1123	0.0627	0.0612	0.0050	0.1137	0.0063	0.1354	0.0050	0.1209	0.0052	0.1217	**0.0974**	**0.1383**

Efficiency: Faster Covergence and Fairness

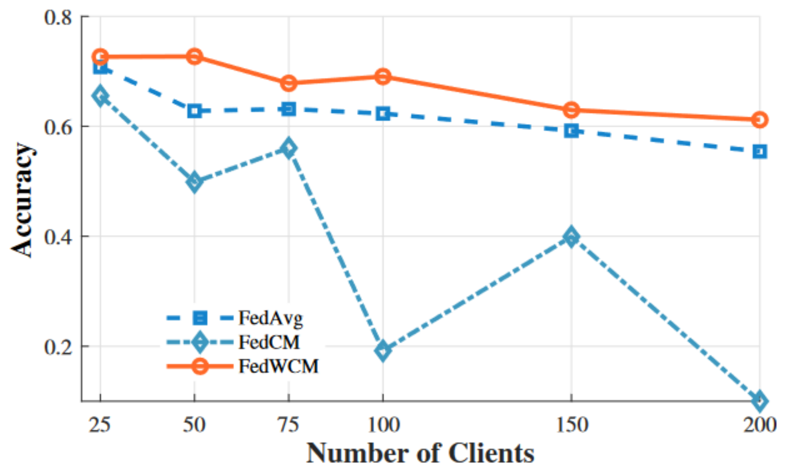


Overall performance comparison

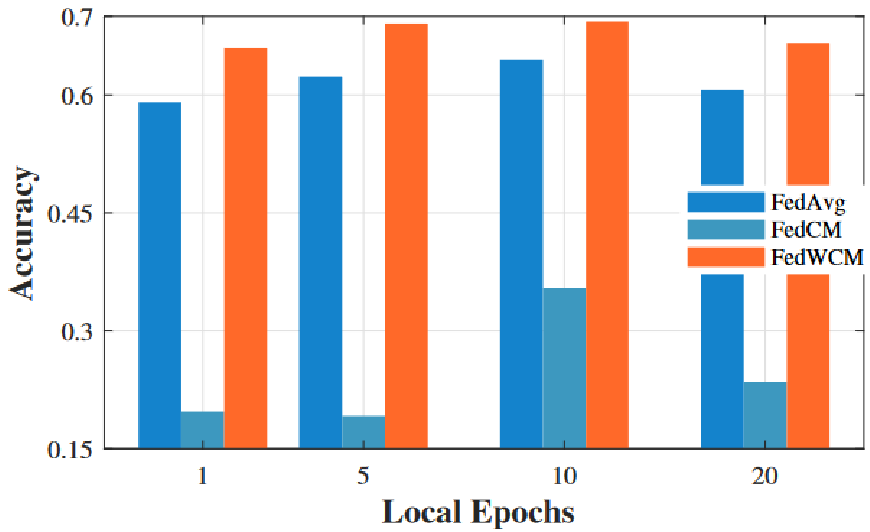


Per-label accuracy

Robustness: Stable under Diverse Conditions



Sampling Rate	FedAvg	FedCM	FedWCM
5%	0.6865	0.3130	**0.7127**
10%	0.6232	0.1918	**0.6905**
20%	0.6450	0.3006	**0.7164**
40%	0.6418	0.2268	**0.6933**
80%	0.6441	0.1000	**0.6980**



$\beta = 0.1$	IF	1	0.4	0.1	0.06	0.04	0.01
FedAvg		0.6859	0.7059	0.6228	0.6295	0.5358	0.4838
FedCM		0.7179	0.7394	0.2346	0.2077	0.2206	0.2283
FedWCM		**0.7337**	**0.7735**	**0.6629**	**0.6538**	**0.5972**	**0.5078**
$\beta = 0.6$	**IF**	**1**	**0.4**	**0.1**	**0.06**	**0.04**	**0.01**
FedAvg		0.7912	0.7294	0.6232	0.5801	0.5543	0.4637
FedCM		0.8104	0.7363	0.1918	0.2616	0.1894	0.2399
FedWCM		**0.8426**	**0.7969**	**0.6905**	**0.6216**	**0.6042**	**0.5164**

Privacy Concerns

Homomorphic encryption[11]

1. Key Generation: The server picks one client to generate a key pair and share the public key with all clients.

2. Encrypted Upload: Each client encrypts its local class counts and sends them to the server.

3. Aggregation & Decryption: The server aggregates the encrypted counts and sends the result back to the key-owning client, who decrypts it and uploads the global statistics to the server.

One-time overhead from encryption

User	Plaintext (Byte)	Ciphertext (Byte)
10	136	88556
20	216	88554
50	456	88631
100	856	88548
150	1256	88576

Table 2: Plaintext and ciphertext values for different users.

[11] Zhang C, Li S, Xia J, et al. {BatchCrypt}: Efficient homomorphic encryption for {Cross-Silo} federated learning[C]//2020 USENIX annual technical conference (USENIX ATC 20). 2020: 493-506.

Conclusion and Outlook

1. Problem: Heterogeneity + long-tail → hard for federated learning

2. Accelator: Momentum is fast, but can fail under extreme imbalance

3. Our method: FedWCM:
- ➢ keeps speed ✅
- ➢ prevents collapse ❌
- ➢ protects minority classes 👥

4. Result: Momentum: simple & efficient → huge potential

5. Future: explore new domains 🌟

Thank you!

yhuang849@connect.hkust-gz.edu.cn