IJCAI 2025
Guangzhou August 29-31

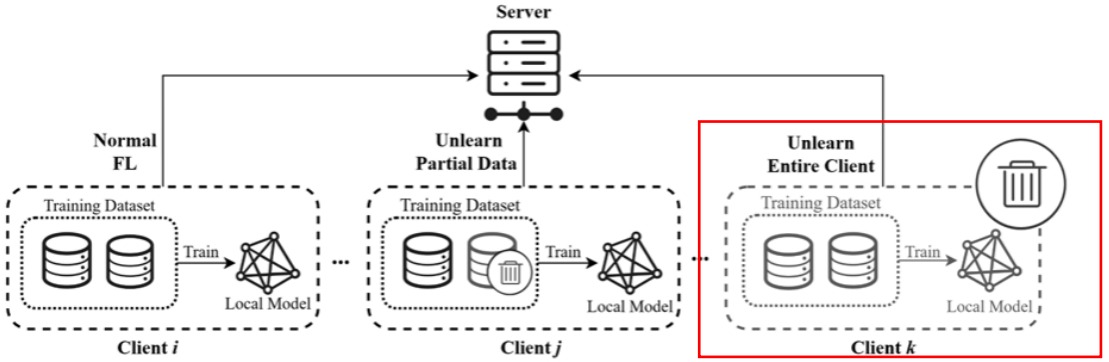
Zero-shot Federated Unlearning via Transforming from Data-Dependent to Personalized Model-Centric

Authors: Wenhan Wu, Huanghuang Liang, Yingling Yuan, Jiawei Jiang, Kanye Ye Wang, Chuang Hu, Dazhao Cheng

Affiliation: Wuhan University, Wuhan University of Technology, University of Macau

Background：Client-level Federated Unlearning

➢ Client-level Federated Unlearning (FU)



removing the influence of a specific client's data from a trained model without retraining from scratch

---

➢ FL process: $\mathcal{FL}$: $D \rightarrow \omega$

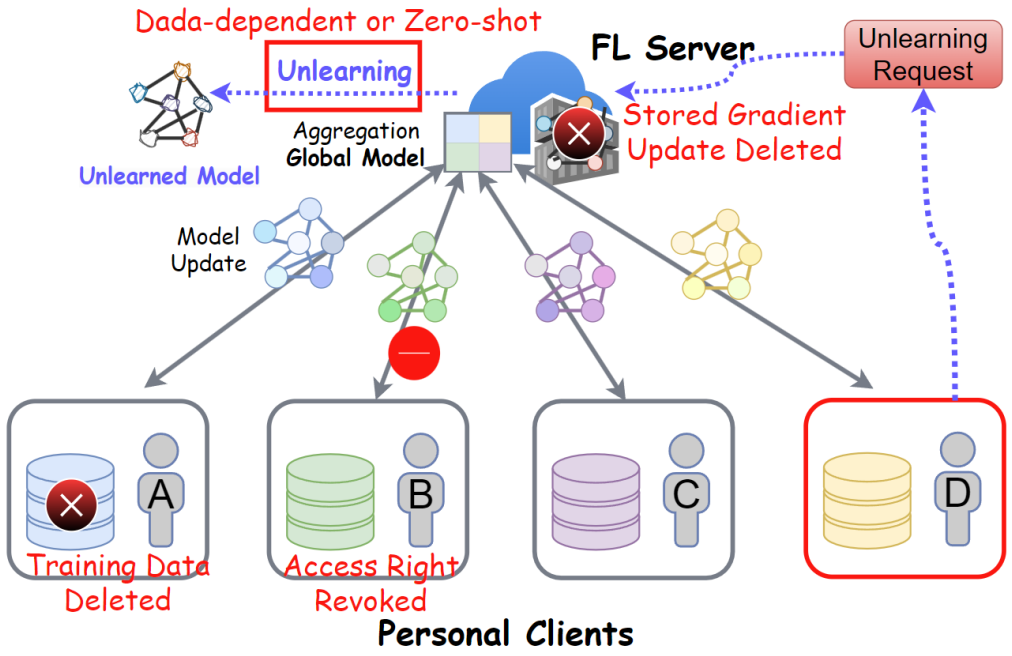
mapping the client data space D to the parameter space of the global model ω.

➢ Unlearning process: $\mathcal{FU}$: $FL(D) \otimes D_r \otimes D_f \rightarrow \omega'$

- **Input:** the learned model $FL(D)$, retained data space D_r, and the forgotten data D_f on client C_f.
- **Output:** An unlearned model similar to the retrained model.
- **Objective:** $\Phi[\mathcal{FL}(D \backslash D_f)] = \Phi[\mathcal{FU}(\mathcal{FL}(D), D, D_f)]$, where $D \backslash D_f$ denotes the client data space excluding D_f, $\Phi[\cdot]$ indicates the probability distribution of the output.

Background：Why Zero-shot Federated Unlearning ?

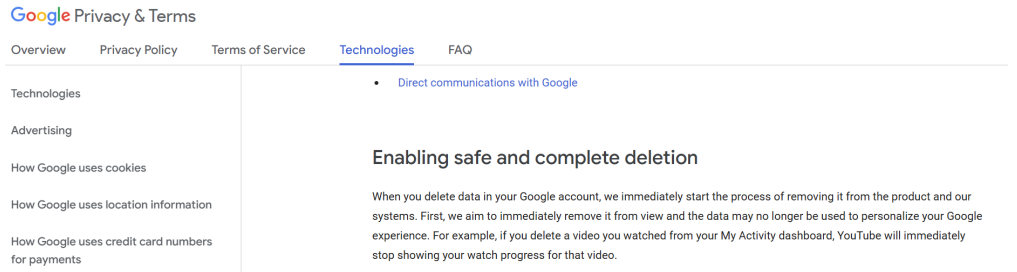
➢ Zero-shot: Unlearning with data or historical update unavailability



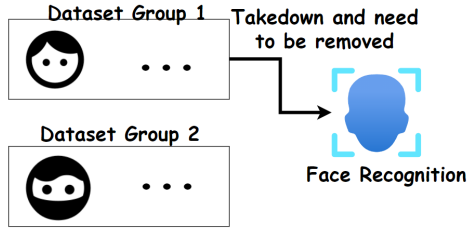
➢ Why zero-shot ?

Clients may move or delete training samples and historical updates, or lose access rights to the data after the model has been trained, due to privacy concerns, legal regulations, and organizational policies.

- ✓ Case 1：User deregistration from system like Google



- ✓ Case 2：Open-source dataset sanitization / sample withdrawal



Challenge: How to Distinguish Different Clients ?

- Existing Federated Client-level Unlearning Methods are Data Dependent.

Unlearning Attributes	FedEraser	FedRecovery	Knot	AdaClipping
All training samples?	✔	✘	✘	✘
Subset of all samples?	✘	✘	✔	✘
Historical data stored during training?	✔	✔	✘	✔
Data-free FU?	✘	✘	✘	✘

Data Dependent:

1. Raw data→Differences in local data distributions,

2. Historical Infomation→ Gradient signatures (e.g., direction, norm) or update trajectories of local models.

The information is naturally client-specific, enabling targeted removal of specific client

Zero-shot:

1. **Data-free:** How to distinguish and remove the target clients without data?
2. **Catastrophic Unlearning :** retained and forgotten clients may hold similar classes , with heterogeneous and overlapping data distributions, how to precise forget ?

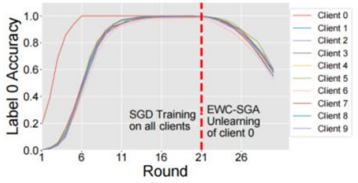
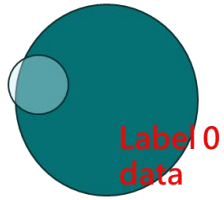


Fig. 1: Catastrophic Unlearning of the EWC-SGA Method on Non-IID MNIST Dataset for Label 0.

20 rounds of training and 10 rounds of unlearning



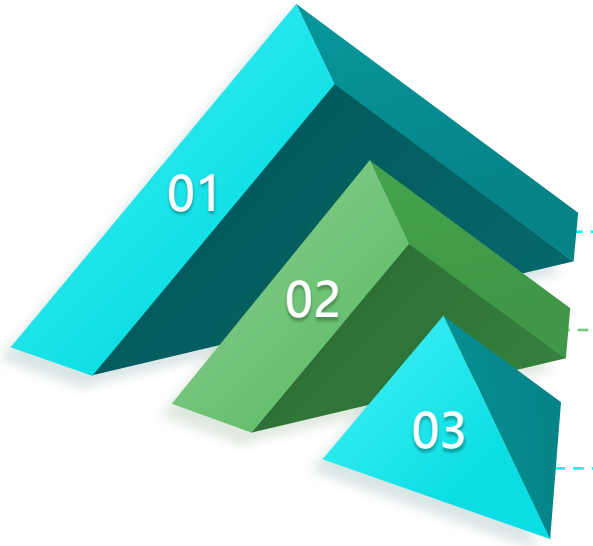
client 0 mainly holds a small portion of label 0 data from the dataset

Potential Solution：Transforming to Model-centric

➢ Motivation:

transform from data-dependent to model centric ➡ **personalization** ➡ distinguish

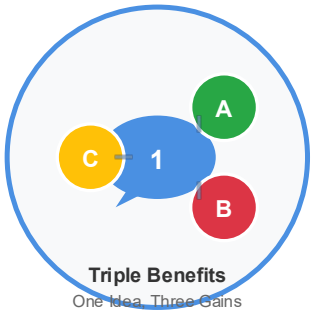
➢ Advantages:



Personalization embeds the client specific information into the model, enabling **zero-shot** ability without retraining.

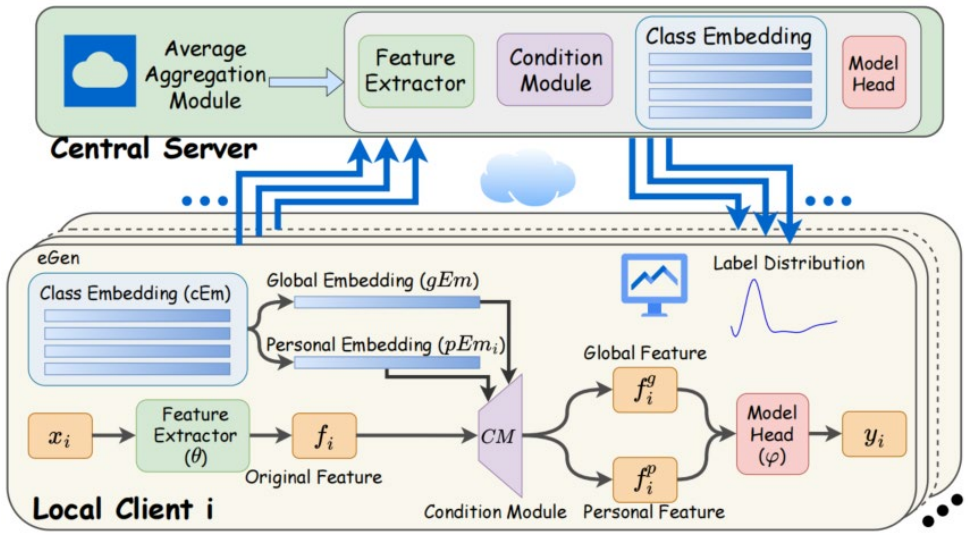
Maps samples with the same label to distinct decision spaces via personalization, effectively **avoiding catastrophic unlearning** of non-target knowledge.

Enhances learning performance under non-IID data.



We propose the first zero-shot FU framework.

Design：System Overview (Model Personalization)



Personalization with Condition Module

Objective: Training personalized $U-class$ classification models for each client by embedding client-specific features.

Embed the client-specific information into the personalized model.

1. **Feature Extractor:** extracts features: $f_i = \theta(x_i; \omega_\theta)$.

2. **Class Embedding Generation:** Global: $gEm = \frac{1}{U}\sum_{y=0}^{U-1} cEm_y$; Personal: $pEm_i = \sum_{u=0}^{U-1} cEm_u * \mathbb{E}_{D_i}\, \mathbb{I}(y_i, y)$.

3. **Feature Transformation with Conditional Module:**

Generates global (f_i^g) and personalized (f_i^p) features:

$$f_i^g = ReLU(CM_B(gEm;\ \omega^{CM}) + (CM_W(gEm;\ \omega^{CM}) + \mathbf{1}) \odot f_i),$$

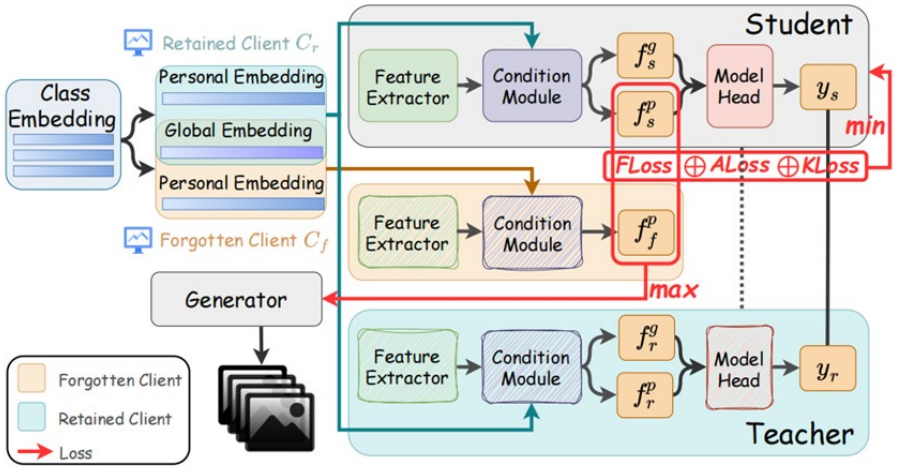
$$f_i^p = ReLU(CM_B(pEm_i;\ \omega^{CM}) + (CM_W(pEm_i;\ \omega^{CM}) + \mathbf{1}) \odot f_i).$$

4. **Model Head Output:** $\hat{y}_i = \varphi([f_i^p;\ f_i^p];\ \omega^\varphi)$.

5. **Local Loss SGD:** $\mathcal{L}_i(\omega_i) = \mathcal{L}_i^{CE} + \mathcal{L}_i^{EM} + \lambda_1|\omega^{cGen}|_2^2 + \lambda_2|\omega^{CM}|_2^2$ · Cross-entropy loss: $\mathcal{L}_i^{CE}$ · class embedding guidance loss: $\mathcal{L}_i^{EM} = -\log\left(\frac{\exp(cos\_sim(f_i^g, cEm_{y_i}))}{\sum_{u=1}^{U}\exp(cos\_sim(f_i^g, cEm_u))}\right)$

6. **Aggregation:** FedAvg with $\omega_i^t = \{\theta, \varphi, eGen, CM\}$.

Design：Knowledge Distillation Based Unlearning



FU with Adversary Knowledge Distillation

Objective: Removing C_f's influence on the C_r's under a zero-shot setting by erasing personalized feature information.

Using forgotten information in personalized model to achieve unlearning

1. Data Generation: $x = G(z;\ \omega_G)$. x feed into teacher R (x,ω_r) on C_r, student S, and forgetting model $F(x,\omega_f)$ on C_f .

2. Parameter Freezing: freezing: $\omega_S^{CM} = \omega_r^{CM}, \omega_S^{cGEn} = \omega_r^{cGen}$.

3. Zero-shot Knowledge Distillation :

- **Forgetting Personalized Features:** Minimize personalized feature's cosine similarity between S and F :

$$FLoss = 1 - \cos\_sim(f_s^p, f_f^p),$$

- **Maintaining Knowledge on C_r:** Align outputs of R and S:

KL divergence: $KLoss = \tau^2 \sum_{i=0}^{U-1} \sigma\left(\frac{\widehat{y_r}}{\tau}\right)_i log\left(\frac{\widehat{y_r}}{\tau}\right)_i / log\left(\frac{\widehat{y_r}}{\tau}\right)_i$.

Attention Patterns: $ALoss = \frac{1}{|\mathcal{N}_l|} \Sigma \left\| f\left(A_l^{(r)}\right) / \left\| f\left(A_l^{(r)}\right) \right\|_2 - f\left(A_l^{(s)}\right) / \left\| f\left(A_l^{(s)}\right) \right\|_2 \right\|_2$

4. Student and Generator Co-Optimization :

- S minimizes: $\min_{\omega_S^\theta, \omega_S^\varphi} F_{stu} = FLoss + \beta\ KLoss + \gamma\ ALoss.$
- G adversarily strengthens the forgetting signal: $\max_{\omega^G} FLoss$.

Evaluations： Partial Evaluation Results

➢ Metrics——Model Accuracy:

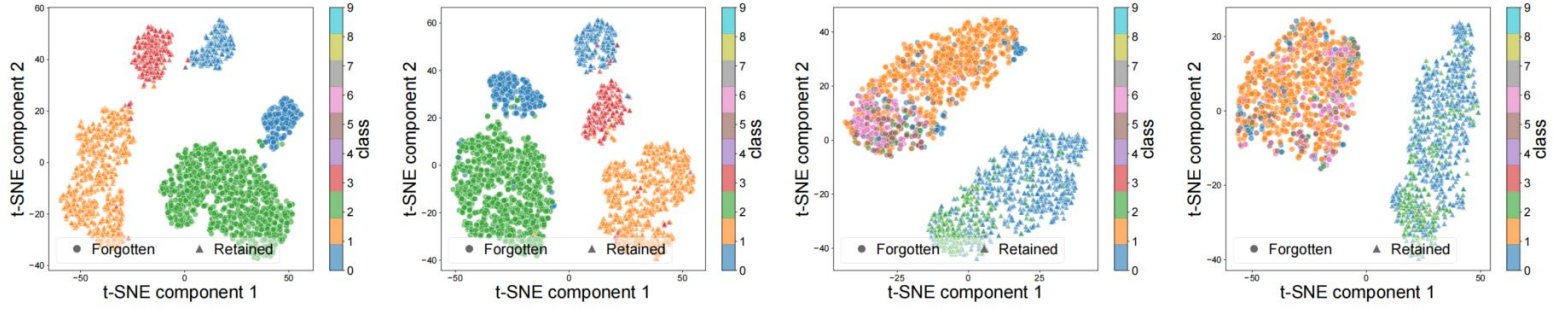
The accuracy of D_f and D_r for C_f and C_r should be closer to that of the retrained model, as the expected behavior of the unlearned model should resemble the retrained model.

DataSet	ζ	C_r	C_f	Origin		Retrained		ZeroFU		FedMM		FedGKT		FedBadT	
				D_r	D_f	D_r	D_f	D_r	D_f	D_r	D_f	D_r	D_f	D_r	D_f
MNIST	0.01	0	1	96.40	99.22	96.30	0.24	92.63	0.00	73.11	0.00	71.14	0.00	88.75	0.00
		8	9	98.54	99.91	98.81	0.01	97.05	0.01	89.11	0.00	89.11	0.00	89.11	1.13
	0.1	4	5	99.10	99.26	99.03	57.18	96.03	38.95	20.98	0.00	60.70	0.12	82.66	27.53
		3	6	96.02	97.29	94.41	83.79	91.04	83.54	44.89	65.19	54.60	0.00	85.00	73.23
SVHN	0.01	0	1	96.02	95.13	96.02	0.00	96.02	0.00	96.02	0.00	96.02	0.00	96.02	4.96
		8	9	95.99	99.99	95.84	0.00	95.99	0.00	95.99	0.00	93.44	0.00	95.99	8.95
	0.1	4	5	98.13	70.87	98.13	60.26	98.13	46.60	98.13	36.98	98.13	39.98	98.13	38.10
		3	6	88.23	59.93	86.08	54.60	88.26	45.41	88.16	0.00	88.16	0.00	88.16	8.40
FMNIST	0.01	0	1	99.45	99.06	99.45	6.27	99.45	8.69	100.00	12.70	71.95	0.00	100.00	96.28
		8	9	99.98	99.44	99.50	12.96	99.98	12.13	99.99	0.00	99.98	0.00	99.98	84.92
	0.1	4	5	83.12	85.91	88.13	13.35	79.78	16.11	63.12	16.11	59.41	16.11	59.40	83.48
		3	6	76.74	98.11	88.41	13.88	84.81	17.21	49.61	0.00	21.04	0.00	73.84	95.41
CIFAR10	0.01	0	1	98.78	100.00	98.78	0.00	98.78	0.00	98.78	0.00	98.78	0.00	98.78	100.00
		8	9	80.36	99.96	77.88	0.00	78.76	0.00	21.28	0.00	75.68	0.00	80.13	99.96
	0.1	4	5	86.08	80.27	79.77	28.22	78.10	29.01	71.01	7.07	48.70	2.30	71.11	12.11
		3	6	87.19	91.70	80.15	3.76	76.67	2.01	47.48	0.00	41.20	9.58	65.30	92.21

Table 1: Forgetting Accuracy Results Comparison under Different Datasets with the red numbers the optimal results.

- The unlearned models closely resemble retrained models in feature space distribution
- It avoids catastrophic unlearning of the same label in other clients

➢ Visualization——personalized feature t-SNE Visualization



(a) MNIST Retrained (b) MNIST Unlearned (c) CIFAR10 Retrained (d) CIFAR10 Unlearned

Figure 5: t-SNE visualizations of personalized features for unlearned and retained clients.

ZeroFU has the capability to map the same label data from different clients to distinct feature spaces.

Take the Way

- We propose a novel zero-shot federated unlearning framework, designed to handle scenarios where neither training data nor historical updates are accessible.
- We propose utilizing personalized client information embedded in the model during training to effectively achieve unlearning by extracting and obfuscating client specific features, thereby tranforming FU from data-dependent to personalized model-centric.
- We deploy and evaluate ZeroFU on real-world datasets and evaluated its performance on both forgotten effectiveness and knowledge retention.

Thank You !

Email: wenhanwu@whu.edu.cn

IJCAI 2025
Guangzhou August 29-31