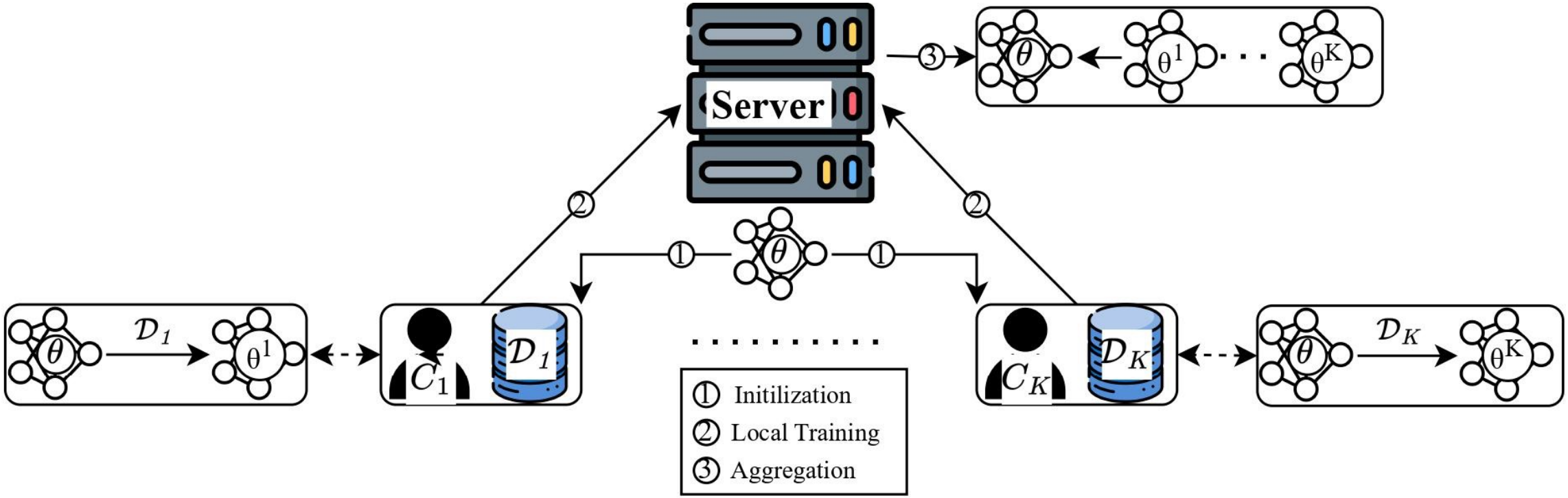
Ferrari: Federated Feature Unlearning via Optimizing Feature Sensitivity

Hanlin Gu[2*] Win Kent Ong[1*] Chee Seng Chan[1] Lixin Fan[2]

[1]Center of Image and Signal Processing, University of Malaya

[2]WeBank AI Lab, Shenzhen, China

Introduction – Federated Learning

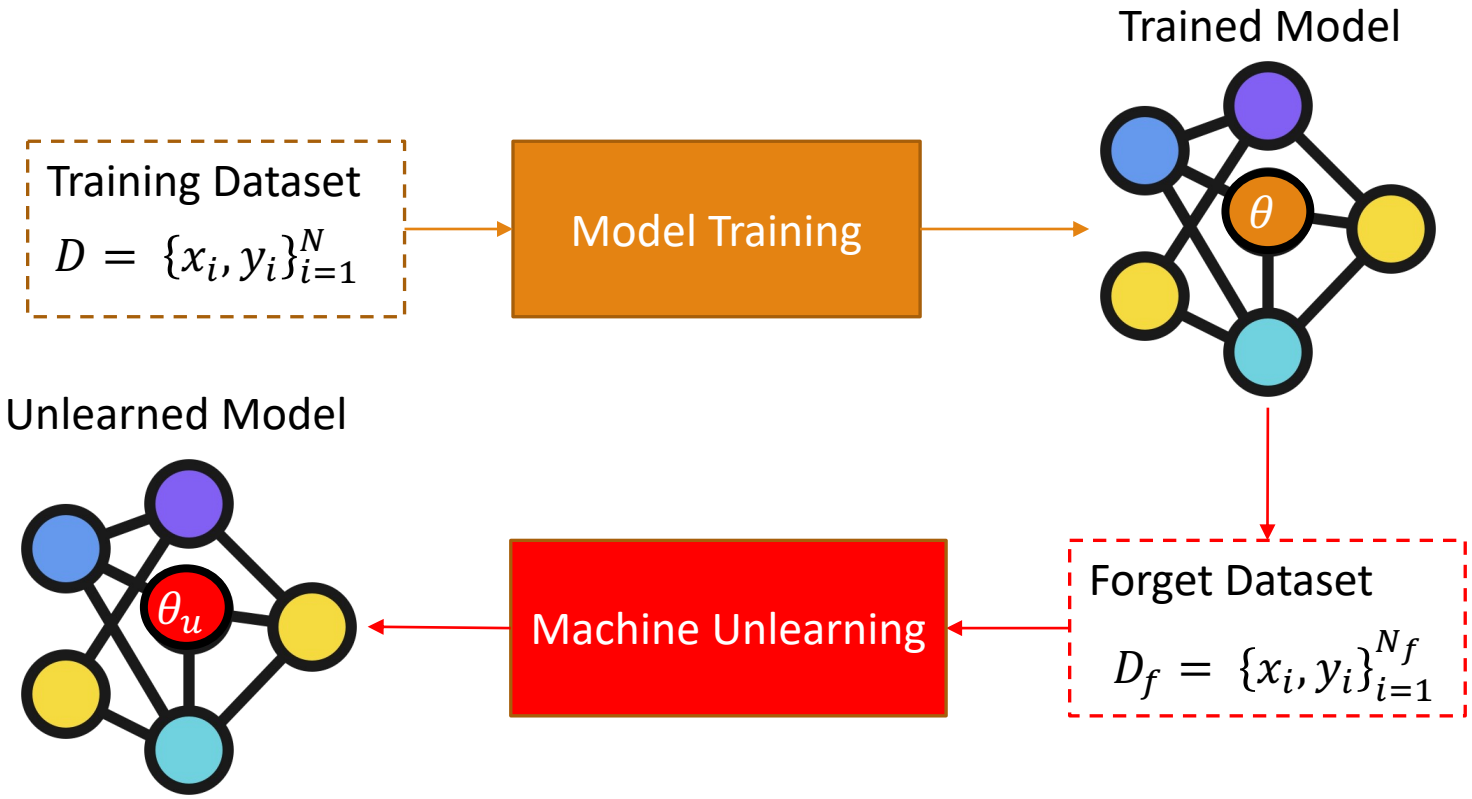


Machine Learning algorithm enables multiple parties to collaboratively train a model

- Without sharing private data, only sharing trained weights
- Better data privacy protection, reducing the risk of privacy leakage

Introduction – Machine Unlearning

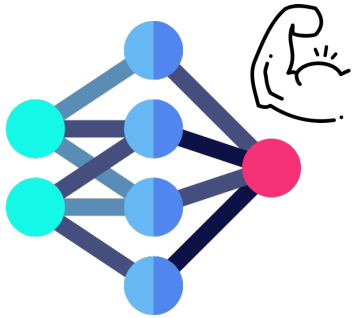
- Remove the influence of a subset of its training dataset from the trained neural network.



Introduction – Machine Unlearning

- PRIVACY REGULATION LAWS
  - California Consumer Privacy Act (CCPA)
  - General Data Protection Regulation (GDPR)
  - Consumer Privacy Protection Act (CPPA)
  - Secure the right to be forgotten

- REMOVE OUTDATED OR MISLABELLED TRAINING DATA
  - Improve model robustness



Motivation

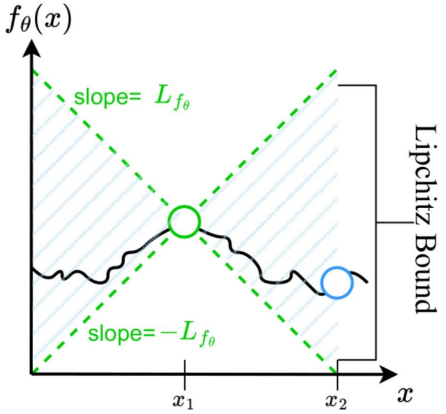
- Federated Unlearning
  - Current works focus on isolated data points
    - Client, sample or class level unlearning
- Feature Unlearning
  - Impractical for Federated Learning due to participation of all client (all datasets).
  - Difficulty in evaluating the effectiveness of feature unlearning.
    - Conventional method compared to the retrained model without the target feature **reduced model utility**.

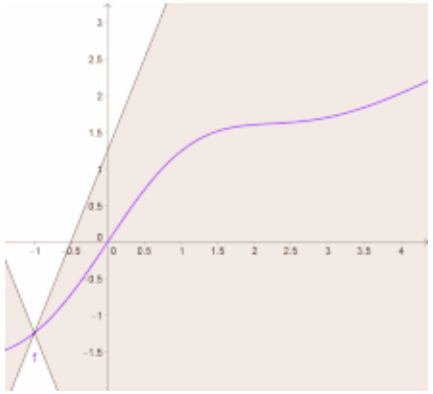
Contributions

- We define the Feature Sensitivity metric based on Lipschitz Continuity
- We proposed an effective **federated feature unlearning** framework
  - allowing clients to selectively unlearn specific features
  - without the participation of other clients
  - optimizing feature sensitivity locally
- We provide theoretical proof and extensive experimental results demonstrate the state-of-the-art **utility** and **effectiveness** of our proposed framework.

Lipschitz Continuity

Lipschitz continuity quantifies the sensitivity of a function, by quantifying how function values change with respect to variations in the independent variable





Exist a non-negative Lipschitz constant

$$||f_\theta(x_1) - f_\theta(x_2)||_Y = L_{f_\theta}||x_1 - x_2||_X, \forall (x_1, x_2) \in \mathcal{X}$$

Output Input

$$\sup_{x_1, x_2 \in \mathcal{X}, x_1 \neq x_2} \frac{||f_\theta(x_1) - f_\theta(x_2)||_Y}{||x_1 - x_2||_X} \leq L_{f_\theta}$$

Bounded Rate of Change - Average rate of change of the function bounded by Lipschitz bound.

$$-L_{f_\theta} \leq \frac{||f_\theta(x_1) - f_\theta(x_2)||_Y}{||x_1 - x_2||_X} \leq L_{f_\theta}$$

Feature Sensitivity

Feature Sensitivity: $s = \frac{\|f(x) - f(\bar{x})\|}{\|(x) - (\bar{x})\|}$

$$s = \frac{\|f(x) - f(x+\delta)\|}{\|(x) - (x+\delta)\|}$$

$$s = \frac{\|f(x) - f(x+\delta)\|}{\|\delta\|}$$

$x =$

$\bar{x} = x + \delta =$

Intuition Sensitivity-Guided Optimization

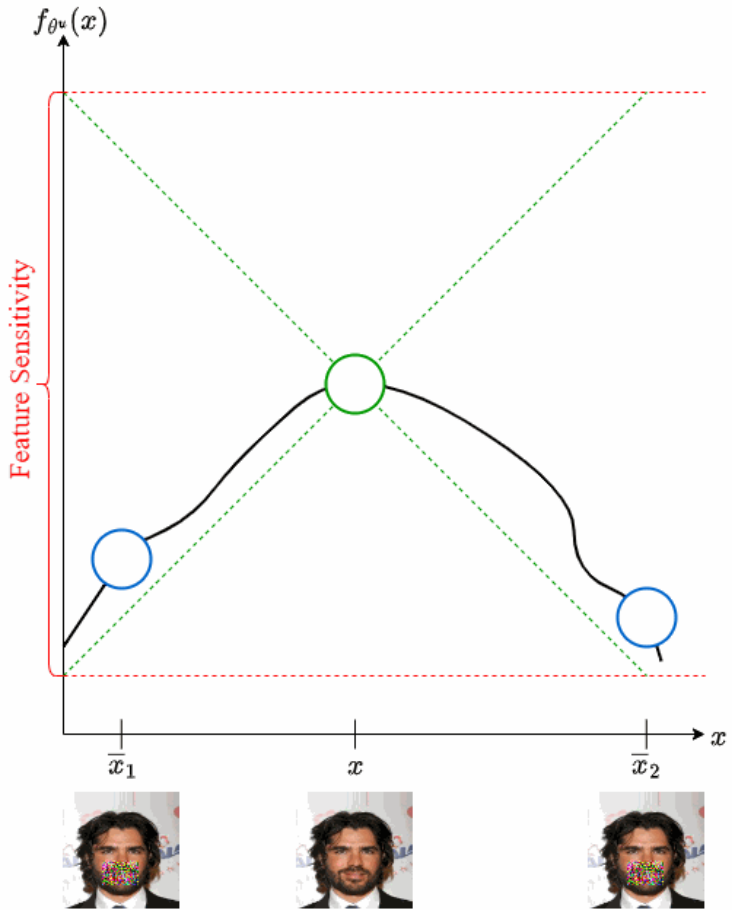
Core Idea: Optimize Feature Sensitivity via Guided Lipschitz Bound

$$\mathcal{L} = \frac{\|f(x) - f(x+\delta)\|}{\|\delta\|}, (x, y) \in D_u$$

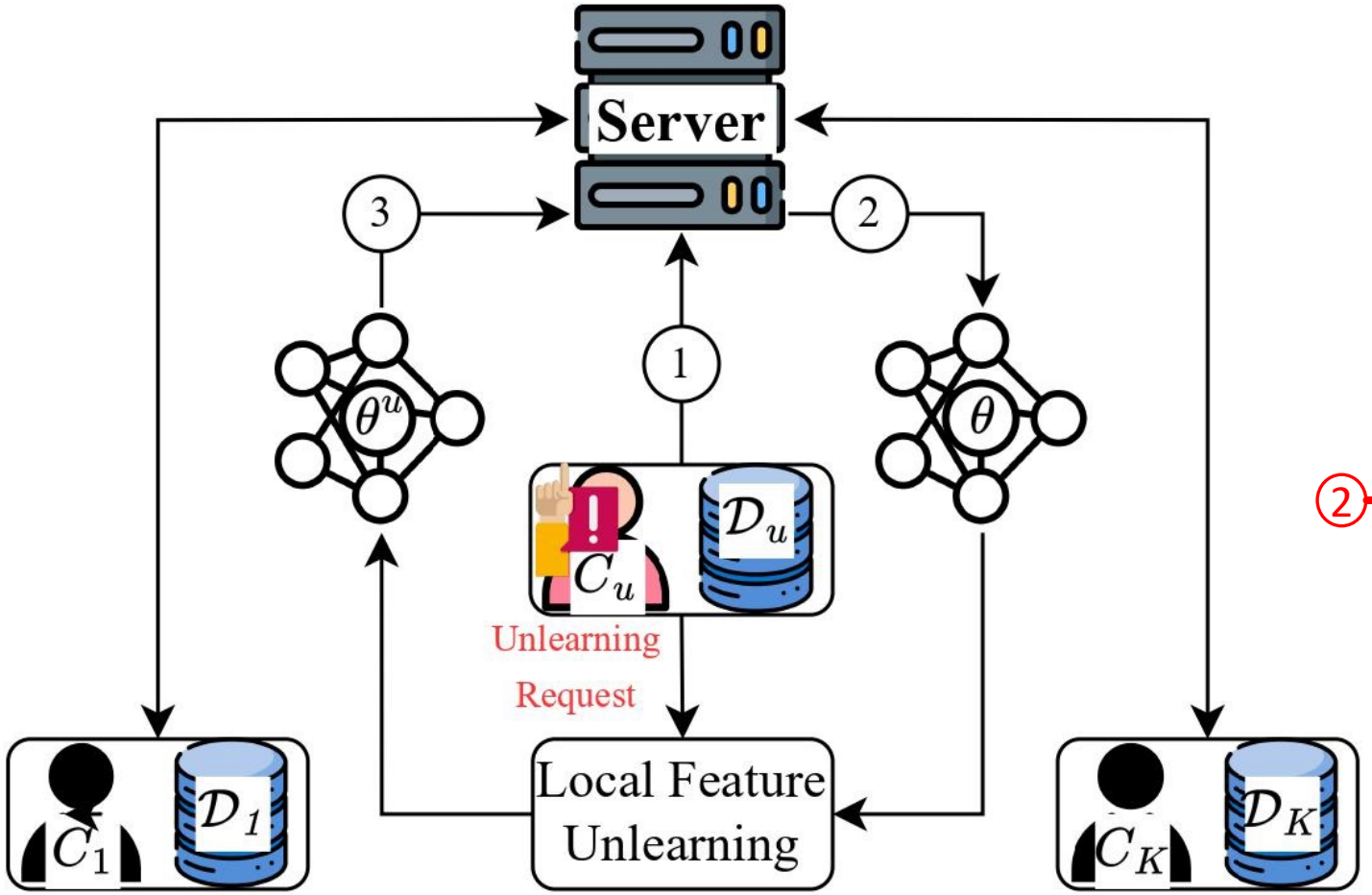
Feature Sensitivity as guided loss function to optimize the unlearn model θ^u via gradient descent

$$\theta^u \leftarrow \theta^u - \eta \cdot \nabla_{\theta^u}(\mathcal{L})$$

$$\nabla_{\theta^u}(\mathcal{L}) = \frac{\partial \mathcal{L}}{\partial \theta_u}$$



Federated Feature Unlearning Framework



Algorithm 1 Federated Feature Unlearning

Input: Unlearned client u, Local dataset $\mathcal{D}_u$ with data size n_u, Unlearn feature $\{\mathcal{F}_i\}_{i=1}^N$, Global model parameters θ, Gaussian noise σ, Learning rate η, Sample number N

Output: Unlearned model parameters θ^u

1: ▷ *The unlearned client k performs:*
2: **for** $(x, \mathcal{F}_i)$ in $(\mathcal{D}_u, \{\mathcal{F}_i\}_{i=1}^N)$ **do**
3: $\theta^u = \theta$
4: **for** $i = 1$ **to** N **do**
5: Sample $\delta_{\mathcal{F}_i}$ according to Eq. (4)
6: Compute $L_i = \frac{\|f_{\theta^u}(x) - f_{\theta^u}(x+\delta_{\mathcal{F},i})\|_2}{\|\delta_{\mathcal{F},i}\|_2}$
7: **end for**
8: $L = \frac{1}{N}\sum_{i=1}^N L_i$
9: $\theta^u \leftarrow \theta^u - \eta \cdot \nabla_{\theta^u}(L)$
10: **end for**
11: Upload θ^u to the server ③
12: ▷ *The server performs:*
13: Replace the global model θ with the θ^u
14: **return** θ^u

② (lines 2–10)

Theoretical Analysis

$$\ell_1 = \min_{\|\delta_{\mathcal{F}}\| \geq C} \mathbb{E}_{(x,y)\in\mathcal{D}} \min_{\theta} \ell\big(f_\theta(x+\delta_{\mathcal{F}}), y\big)$$

$$\ell_2 = \max_{\|\delta_{\mathcal{F}}\| \leq C} \mathbb{E}_{(x,y)\in\mathcal{D}} \min_{\theta} \ell\big(f_\theta(x+\delta_{\mathcal{F}}), y\big)$$

Assumption 1. *Assume* $\ell_2 \leq \ell_1$

larger perturbations would naturally lead to greater utility loss

Assumption 2. *Suppose the federated model achieves zero training loss.*

Theorem 1. *If Assumption 1 and Assumption 2 hold, the utility loss of unlearned model obtained by Algorithm 1 is less than the utility loss with unlearning successfully, i.e.*

$$\ell_u \leq \ell_1, \tag{3.10}$$

where $\ell_u = \mathbb{E}_{(x,y)\in\mathcal{D}} \ell(f_{\theta^u}(x), y)$

$$\begin{aligned}
\ell_u &\leq \min_{\theta\in\mathbb{R}^d} \mathbb{E}_{(x,y)\in\mathcal{D}} \Big(\ell(f_\theta(x), y) + \lambda \mathbb{E}_{\|\delta_{\mathcal{F}}\|\geq\frac{1}{\lambda}} \frac{\|f_\theta(x) - f_\theta(x+\delta_{\mathcal{F}})\|_2}{\|\delta_{\mathcal{F}}\|_2}\Big) \\
&\leq \min_{\theta\in\Theta^*} \mathbb{E}_{(x,y)\in\mathcal{D}} \Big(\ell(f_\theta(x), y) + \lambda \mathbb{E}_{\|\delta_{\mathcal{F}}\|\geq\frac{1}{\lambda}} \frac{\|f_\theta(x) - f_\theta(x+\delta_{\mathcal{F}})\|_2}{\|\delta_{\mathcal{F}}\|_2}\Big) \\
&\leq \min_{\theta\in\Theta^*} \mathbb{E}_{(x,y)\in\mathcal{D}} \mathbb{E}_{\|\delta_{\mathcal{F}}\|\geq\frac{1}{\lambda}} \|y - f_{\theta^*}(x+\delta_{\mathcal{F}})\|_2 \\
&\leq \mathbb{E}_{(x,y)\in\mathcal{D}} \mathbb{E}_{\|\delta_{\mathcal{F}}\|\geq\frac{1}{\lambda}} \min_{\theta\in\Theta^*} \|y - f_{\theta^*}(x+\delta_{\mathcal{F}})\|_2 \\
&= \mathbb{E}_{\|\delta_{\mathcal{F}}\|\geq\frac{1}{\lambda}} \mathbb{E}_{(x,y)\in\mathcal{D}} \min_{\theta\in\Theta^*} \|y - f_{\theta^*}(x+\delta_{\mathcal{F}})\|_2 \\
&\leq \max_{\|\delta_{\mathcal{F}}\|\geq\frac{1}{\lambda}} \mathbb{E}_{(x,y)\in\mathcal{D}} \min_{\theta\in\mathbb{R}^d} \|y - f_{\theta^*}(x+\delta_{\mathcal{F}})\|_2 \\
&\leq \max_{\|\delta_{\mathcal{F}}\|\leq C} \mathbb{E}_{(x,y)\in\mathcal{D}} \min_{\theta\in\mathbb{R}^d} \|y - f_{\theta^*}(x+\delta_{\mathcal{F}})\|_2 \\
&= \ell_2,
\end{aligned}$$

Evaluation – Questions to be Answered

1. **Effectiveness** – How effective is the proposed Federated Feature Unlearning framework in removing the target feature?
    1. Sensitive Feature Unlearning
    2. Backdoor Feature Unlearning
    3. Biased Feature Unlearning
2. **Utility** – Can the unlearned model maintain its generalization capability on the test dataset?
3. **Efficiency** – How efficient is the unlearning process?

Result and Discussion

Effectiveness - Sensitive Feature Unlearning

Model Inversion Attack – Attack Success Rate

Scenario	Datasets	Unlearn Feature	Attack Success Rate(ASR) (%) ↓					
			Baseline	Retrain	Fine-tune	FedCDP	FedRecovery	Ours
Sensitive	CelebA	Mouth	84.36 ±3.22	47.52 ±1.04	77.43 ±10.98	75.36 ±9.31	71.52 ±6.07	**51.28 ±2.41**
	Adult	Marriage	87.54 ±13.89	49.28 ±2.13	83.45 ±8.44	72.83 ±5.18	80.39 ±10.68	**49.58 ±1.38**
	Diabetes	Pregnancies	92.31 ±7.55	38.89 ±2.52	88.46 ±5.01	81.91 ±8.17	78.27 ±2.47	**42.61 ±1.81**

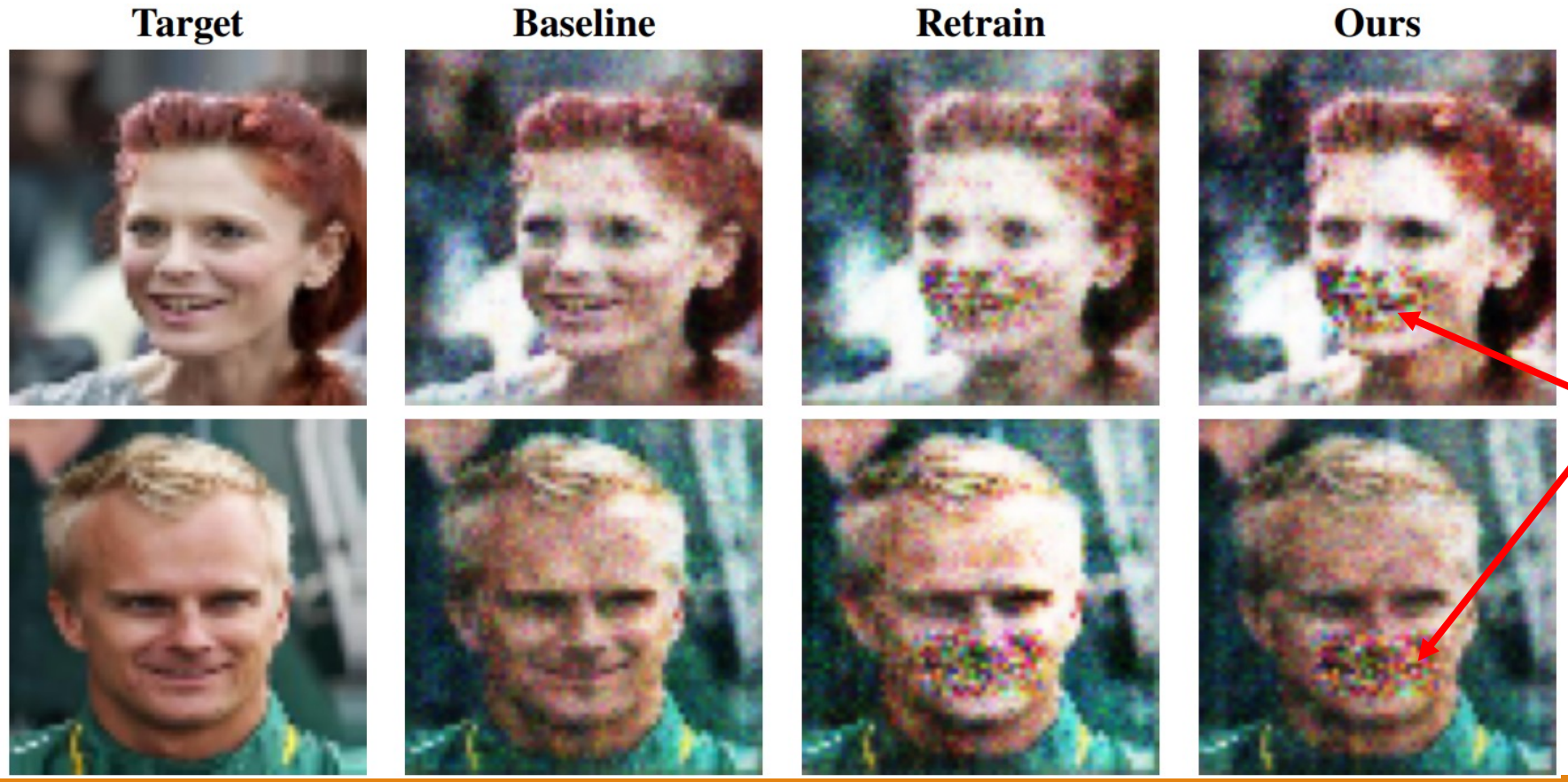
Feature Sensitivity

Scenario	Datasets	Unlearn Feature	Feature Sensitivity					
			Baseline	Retrain	Fine-tune	FedCDP	FedRecovery	Ours
Sensitive	CelebA	Mouth	$0.96 \pm 1.41 \times 10^{-2}$	$0.07 \pm 8.06 \times 10^{-4}$	$0.79 \pm 2.05 \times 10^{-2}$	$0.93 \pm 2.87 \times 10^{-2}$	$0.91 \pm 3.41 \times 10^{-2}$	$\mathbf{0.09 \pm 3.04} \times 10^{-4}$
	Adult	Marriage	$1.31 \pm 1.53 \times 10^{-2}$	$0.02 \pm 6.47 \times 10^{-4}$	$0.94 \pm 6.81 \times 10^{-2}$	$1.07 \pm 7.43 \times 10^{-2}$	$1.14 \pm 2.57 \times 10^{-2}$	$\mathbf{0.05 \pm 1.72} \times 10^{-4}$
	Diabetes	Pregnancies	$1.52 \pm 0.91 \times 10^{-2}$	$0.05 \pm 5.07 \times 10^{-4}$	$0.96 \pm 1.28 \times 10^{-2}$	$1.23 \pm 3.82 \times 10^{-2}$	$0.83 \pm 5.08 \times 10^{-2}$	$\mathbf{0.07 \pm 1.07} \times 10^{-4}$

Result and Discussion

Effectiveness - Sensitive Feature Unlearning

Model Inversion Attack – Reconstructed Images



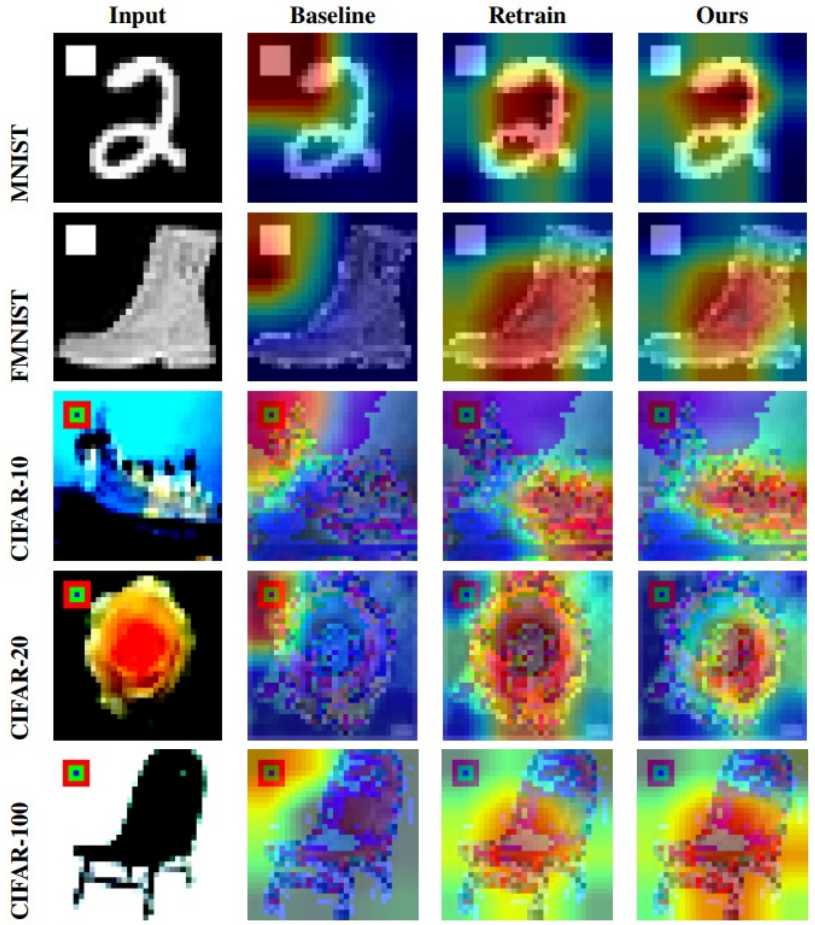
Result and Discussion

Effectiveness - Backdoor Feature Unlearning

Scenarios	Datasets	Unlearn Feature		Accuracy (%)					
				Baseline	Retrain	Fine-tune	FedCDP	RedRecovery	Ours
Backdoor	MNIST	Backdoor pattern	$\mathcal{D}_r$ ↑	95.65 ±1.39	97.19 ±2.49	**96.16 ±0.37**	65.82 ±6.85	40.81 ±4.31	95.93 ±0.45
			$\mathcal{D}_u$ ↓	97.43 ±3.69	0.00 ±0.00	72.64 ±0.24	69.37 ±0.83	53.72 ±3.14	**0.11 ±0.01**
	FMNIST		$\mathcal{D}_r$	91.07 ±0.54	93.85 ±1.08	**94.36 ±1.98**	68.46 ±3.39	42.93 ±2.50	92.83 ±0.61
			$\mathcal{D}_u$	94.51 ±6.29	0.00 ±0.00	43.91 ±0.28	72.19 ±0.49	48.15 ±4.37	**0.90 ±0.03**
	CIFAR-10		$\mathcal{D}_r$	87.63 ±1.16	91.12 ±1.60	**92.02 ±3.15**	54.91 ±6.91	27.49 ±4.96	89.91 ±0.95
			$\mathcal{D}_u$	95.05 ±2.30	0.00 ±0.00	88.44 ±0.92	62.75 ±5.07	49.26 ±2.23	**0.29 ±0.04**
	CIFAR-20		$\mathcal{D}_r$	75.06 ±6.41	81.91 ±4.68	**82.67 ±1.32**	55.67 ±6.35	23.76 ±2.17	78.29 ±3.12
			$\mathcal{D}_u$	94.21 ±4.11	0.00 ±0.00	86.53 ±1.47	50.17 ±9.11	50.38 ±4.25	**0.78 ±0.08**
	CIFAR-100		$\mathcal{D}_r$	54.14 ±3.96	73.54 ±5.70	**73.66 ±6.57**	34.62 ±2.24	15.62 ±7.78	69.57 ±3.81
			$\mathcal{D}_u$	88.98 ±6.63	0.00 ±0.00	65.38 ±4.76	57.29 ±3.62	46.17 ±9.25	**0.15 ±0.01**
	Adult		$\mathcal{D}_r$	75.12 ±1.09	81.55 ±2.31	**76.51 ±3.59**	38.17 ±3.05	45.19 ±5.75	74.95 ±1.54
			$\mathcal{D}_u$	95.88 ±0.59	0.00 ±0.00	89.07 ±1.38	41.93 ±2.75	31.94 ±6.79	**3.51 ±1.16**
	Diabetes		$\mathcal{D}_r$	75.67 ±1.73	79.57 ±1.25	**78.58 ±2.39**	43.76 ±4.91	37.14 ±2.74	73.38 ±1.93
			$\mathcal{D}_u$	97.29 ±0.91	0.00 ±0.00	82.19 ±1.87	54.48 ±6.71	59.32 ±5.29	**5.84 ±0.47**

Result and Discussion

Effectiveness - Backdoor Feature Unlearning



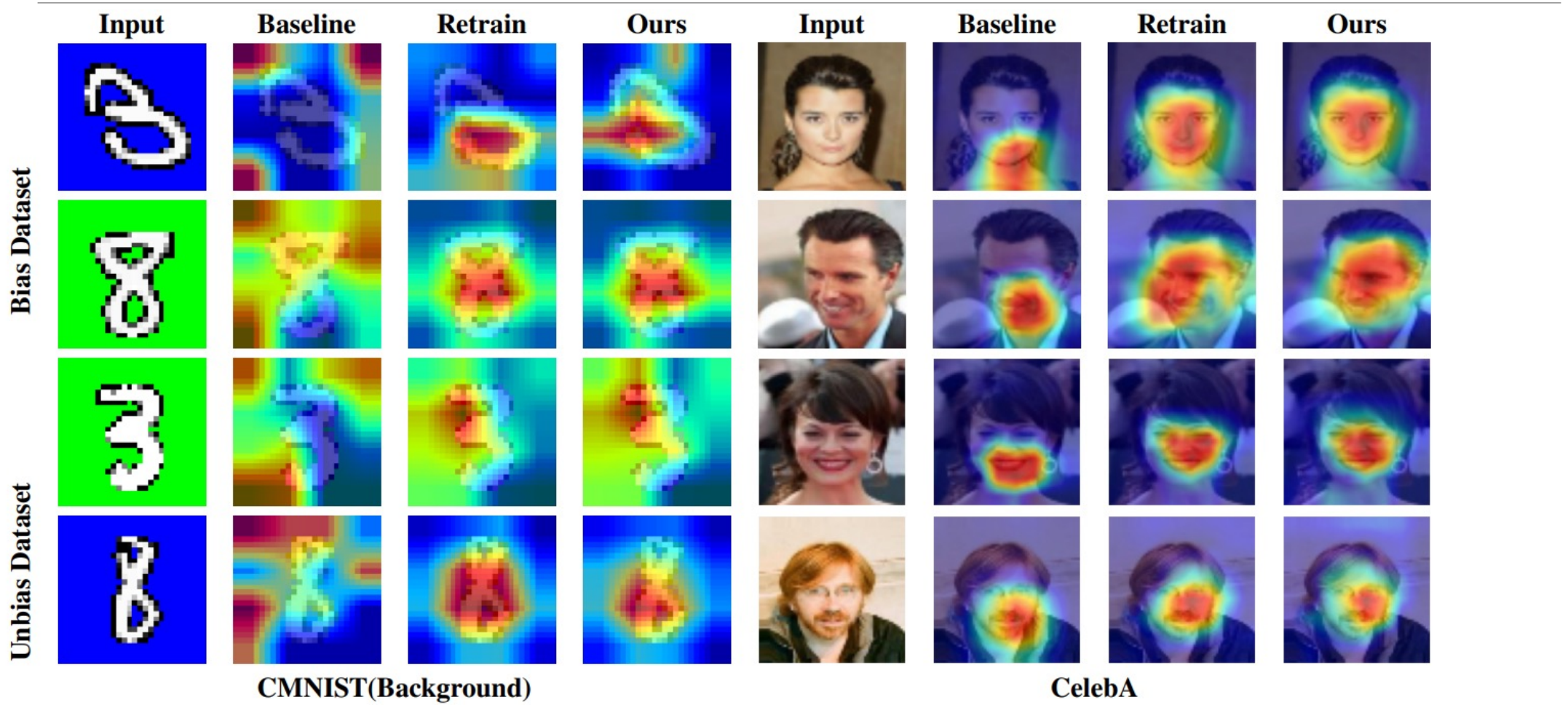
Result and Discussion

Effectiveness - Biased Feature Unlearning

Scenarios	Datasets	Unlearn Feature		Accuracy (%) ≈					
				Baseline	Retrain	Fine-tune	FedCDP	RedRecovery	Ours
Biased	CMNIST	Digit	$\mathcal{D}_r$	64.94 ±7.88	98.76 ±3.65	67.15 ±2.60	25.85 ±1.58	23.92 ±1.08	**84.31 ±2.63**
			$\mathcal{D}_u$	98.88 ±4.90	98.44 ±1.90	97.95 ±1.13	30.17 ±4.69	27.64 ±9.37	**84.62 ±3.59**
	CMNIST	Background	$\mathcal{D}_r$	61.76 ±5.31	99.05 ±4.97	70.57 ±0.92	19.24 ±1.87	24.71 ±2.93	**87.98 ±1.85**
			$\mathcal{D}_u$	98.27 ±2.85	98.39 ±1.83	96.06 ±2.08	32.67 ±5.72	35.59 ±5.08	**87.21 ±0.84**
	CelebA	Mouth	$\mathcal{D}_r$	79.46 ±2.09	96.47 ±6.15	84.45 ±1.48	14.29 ±0.81	16.34 ±3.43	**94.18 ±3.08**
			$\mathcal{D}_u$	96.38 ±3.87	96.11 ±2.17	94.23 ±0.66	21.58 ±3.48	25.72 ±8.02	**94.79 ±1.48**
	Adult	Marriage	$\mathcal{D}_r$	64.68 ±3.73	80.02 ±1.49	68.28 ±3.63	36.19 ±5.69	42.86 ±4.28	**79.68 ±1.26**
			$\mathcal{D}_u$	87.48 ±1.93	80.57 ±2.08	87.06 ±2.85	56.28 ±3.75	28.73 ±1.85	**79.76 ±0.63**
	Diabetes	Pregnancies	$\mathcal{D}_r$	57.46 ±3.36	78.35 ±3.53	63.76 ±2.07	25.77 ±1.58	48.93 ±5.64	**71.25 ±1.33**
			$\mathcal{D}_u$	73.42 ±1.68	77.57 ±2.51	70.56 ±3.43	40.73 ±2.95	35.28 ±4.71	**72.84 ±2.05**

Result and Discussion

Effectiveness - Biased Feature Unlearning



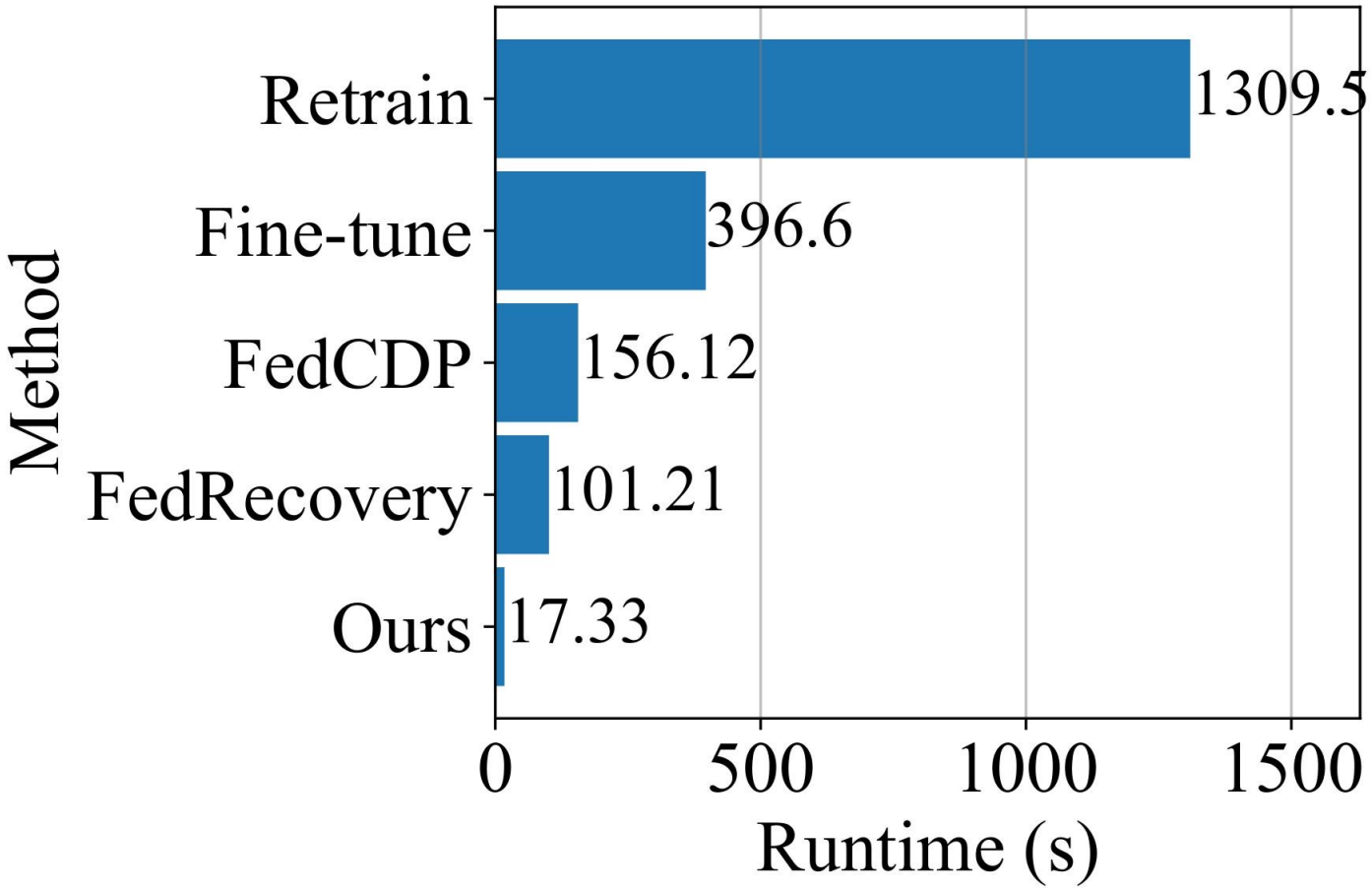
Result and Discussion

Utility

Scenarios	Datasets	Unlearn Feature	Accuracy(%) ↑					
			Baseline	Retrain	Fine-tune	FedCDP	RedRecovery	Ours
Sensitive	CelebA	Mouth	94.87 ±1.38	79.46 ±2.32	62.79 ±1.62	34.03 ±4.20	29.78 ±6.69	**92.26 ±1.73**
	Adult	Marriage	82.45 ±2.59	65.27 ±0.58	61.02 ±1.05	30.19 ±1.62	27.89 ±3.71	**81.02 ±0.58**
	Diabetes	Pregnancies	82.11 ±0.49	64.19 ±0.72	59.57 ±0.68	36.71 ±4.56	17.56 ±2.32	**79.53 ±0.79**
Backdoor	MNIST	Backdoor Pixel Pattern	94.75 ±4.88	96.23 ±0.16	**96.85 ±0.91**	65.31 ±4.39	40.52 ±7.38	95.83 ±1.14
	FMNIST		90.68 ±2.19	92.98 ±0.75	**93.52 ±1.63**	67.62 ±0.81	42.24 ±4.45	92.61 ±1.57
	CIFAR-10		87.55 ±3.71	90.92 ±1.83	**91.23 ±0.44**	53.98 ±2.17	27.16 ±9.68	89.52 ±2.18
	CIFAR-20		74.47 ±2.38	81.61 ±1.75	**82.52 ±0.69**	54.76 ±0.98	23.02 ±3.11	78.34 ±2.35
	CIFAR-100		54.13 ±7.62	73.12 ±1.54	**73.59 ±1.66**	34.30 ±0.42	15.21 ±5.83	69.30 ±2.27
	Adult		77.51 ±0.94	80.38 ±1.92	**81.75 ±3.16**	42.57 ±2.38	43.86 ±4.55	79.73 ±1.46
	Diabetes		75.13 ±1.69	79.04 ±0.73	**80.53 ±1.59**	48.29 ±5.35	40.83 ±3.65	79.57 ±0.82
Biased	CMNIST	Digit	81.72 ±3.41	98.49 ±1.46	82.54 ±0.78	27.56 ±1.71	25.05 ±5.09	**83.85 ±1.63**
	CMNIST	Background	80.12 ±2.18	98.70 ±1.81	83.35 ±1.53	25.96 ±2.29	28.15 ±3.05	**86.03 ±1.36**
	CelebA	Mouth	87.35 ±4.07	95.87 ±1.52	88.93 ±2.65	16.98 ±0.23	20.19 ±7.21	**94.62 ±2.49**
	Adult	Marriage	76.08 ±2.79	80.47 ±1.73	77.24 ±2.24	46.35 ±3.72	35.69 ±2.56	**81.22 ±1.45**
	Diabetes	Pregnancies	65.48 ±3.07	77.93 ±2.51	67.13 ±2.78	38.25 ±2.28	45.11 ±3.18	**72.04 ±1.39**

Result and Discussion

Time Efficiency



Conclusion

- To best of our knowledge, this is the first work to achieve feature unlearning within Federated Learning settings (Federated Feature Unlearning)
- The proposed Federated Feature Unlearning framework effectively achieves feature unlearning via the proposed Sensitivity-Guided Optimization algorithm.
- Theoretical analysis and experimental results, both quantitative and qualitatively.
- Proposed Federated Feature Unlearning framework proven to be effective in unlearning:
  - Sensitive Feature
  - Backdoor Feature
  - Biased Feature
- Practical Federated Feature Unlearning Framework without participation of all clients, only participation of unlearn client is needed.