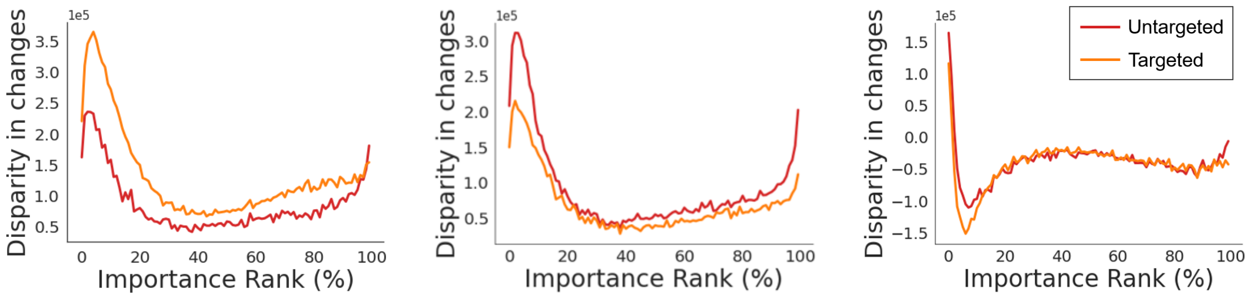


(a) CIFAR-10 (b) SVHN (c) TinyImageNet

Figure 4: Comparison of change patterns over three datasets under two different poisoning attack scenarios, untargeted and targeted attack, where the disparity is measured by the difference in changes of importance rank between benign and poisoned models after one training round.

A. Appendix

A.1. Release & Implementation details

We adopt ResNet18 as the default backbone architecture, building upon prior research in federated learning [14, 10]. In the case of the targeted attack, we follow the original literature [9] and generate a noise input pattern called a backdoor. The size of the backdoor is set to 5×5, and its location is in the bottom-right corner of the images. For the untargeted Gaussian attack, we set the standard deviation of the Gaussian noise to 0.05.

We follow the original works' implementations and hyper-parameter settings to reproduce all baselines. For Multi-Krum and Norm Bounding algorithms, we assume the central server already knows the upper bound of attacker numbers when deciding on hyper-parameters. The confidence interval and clipping threshold in the ResidualBase algorithm are set to 2.0 and 0.05, respectively. We calculate the geometric mean for RFA by setting the smoothing parameter to 1e-6 and the maximum number of Weiszfeld iterations to 100. More details on implementations are at https://github.com/Sungwon-Han/FEDCPA.

A.2. Time Complexity Analysis

For all experiments, we utilized four A100 GPUs. Table 7 compares the time costs in seconds of every defense strategy per each round of training. Note that FedCPA is not a huge burden and only took 10% more processing time than the classical FedAvg algorithm (i.e., No Defense).

A.3. Extra Results on Critical Parameter Analysis

In Section 4, we have shown that benign and poisoned local models exhibit distinct patterns in terms of parameter importance, with the poisoned model causing more significant disruptions to the top and bottom parameters. We conducted the same analysis across different datasets to validate our observation. The results of our analysis are presented in Figure 4, which compares the change patterns in importance rank between benign and poisoned models under two different attack scenarios, untargeted and targeted attacks. For the untargeted attack scenario, we used the label flipping attack method. After one training round, We measure the disparity in importance rank between benign and poisoned models. The results demonstrate that our observation remains consistent across the various datasets.

Method	Time costs in seconds
No Defense	87
Median	86
Trimmed Mean	87
Multi Krum	87
FoolsGold	90
Norm Bound	86
RFA	91
ResidualBase	185
FedCPA	96

Table 7: Comparison on time complexity among defense strategies against poisoning attacks. The CIFAR-10 dataset is used for the analysis.

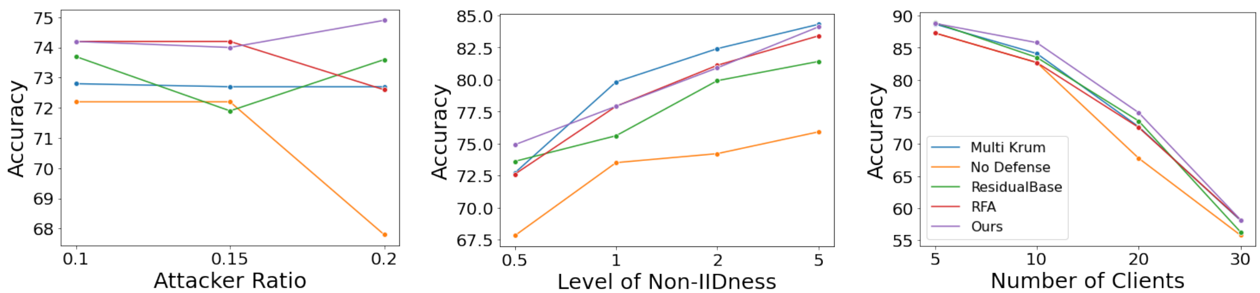
A.4. Extra Results on Robustness Tests

We evaluate the robustness of FedCPA through experiments conducted under different settings, varying key simulation parameters such as (a) the number of malicious clients $|\mathcal{C}_m|$, (b) the total number of participating clients N, and (c) the degree of non-IIDness, controlled by the β parameter in the Dirichlet distribution. A lower β value results in a higher level of non-IIDness.

This section presents additional comparison results among different defense strategies under an untargeted attack scenario (i.e., label flipping attack) on the CIFAR-10 dataset. Results presented in Figure 5 show that FedCPA consistently performs comparably well despite variations in simulation parameters.

A.5. Full Results on Performance Evaluation

Table 8-12 shows the complete evaluation results on defense performance over three datasets under various poisoning attack scenarios: targeted attack with $\gamma_p = 0.5, 0.8$, untargeted label flipping attack with $\gamma_p = 0.8, 1.0$, and untargeted Gaussian attack. The results are obtained by averaging over the last ten rounds and are reported with mean and standard deviation values.



(a) Effect of the ratio of attackers (b) Effect of the level of non-IIDness (c) Effect of the number of clients

Figure 5: Robustness test results under label flipping attack across different simulation hyper-parameters: (a) the attacker ratio, (b) the level of non-IIDness, and (c) the number of clients over the CIFAR-10 dataset.

Method	CIFAR-10		SVHN		TinyImageNet	
($\gamma_p = 0.5$)	ACC	ASR	ACC	ASR	ACC	ASR
No Defense	72.1±3.07	71.0±0.61	93.0±0.48	22.2±12.02	39.5±2.71	96.6±0.41
Median	65.6±3.54	77.8±1.09	90.7±0.44	23.0±7.02	32.5±3.43	96.1±0.58
Trimmed Mean	70.1±3.15	51.4±0.82	92.2±0.57	20.9±20.66	39.3±1.13	97.2±0.26
Multi Krum	69.9±0.89	63.8±1.24	92.1±0.82	21.4±10.88	37.1±2.85	74.6±6.93
FoolsGold	45.5±12.24	54.3±18.25	79.6±5.32	23.5±36.78	24.3±7.37	92.4±14.82
Norm Bound	68.2±4.12	61.2±25.00	93.1±0.69	20.8±0.91	36.6±0.38	96.7±0.69
RFA	72.8±3.09	56.4±13.52	92.3±1.09	20.8±1.57	37.1±0.48	93.9±0.63
ResidualBase	70.6±3.12	59.9±0.61	93.1±0.34	21.1±15.45	39.6±1.27	96.9±0.19
FedCPA	68.8±3.74	21.9±0.73	93.3±9.36	20.6±2.69	30.1±1.51	43.2±44.66

Table 8: Comparison of defense performance over three datasets under targeted attack scenarios with pollution ratio $\gamma_p = 0.5$. Mean and standard deviation over ten last rounds are reported.

Method	CIFAR-10		SVHN		TinyImageNet	
($\gamma_p = 0.8$)	ACC	ASR	ACC	ASR	ACC	ASR
No Defense	69.3±3.74	50.9±25.09	92.5±0.93	22.0±2.21	38.8±1.12	96.1±1.34
Median	62.4±3.32	70.6±16.51	90.0±1.53	23.6±3.39	31.5±0.99	96.2±0.59
Trimmed Mean	71.4±2.77	19.0±10.29	91.7±1.25	21.4±1.78	37.9±1.12	97.0±0.81
Multi Krum	69.0±2.21	40.4±21.85	90.7±2.33	23.4±4.06	36.3±1.78	19.0±13.91
FoolsGold	49.1±9.46	46.8±34.83	69.8±24.56	32.3±27.72	28.5±4.27	69.1±43.20
Norm Bound	64.9±4.28	53.1±30.29	92.7±1.31	20.9±1.42	35.7±1.00	97.1±0.83
RFA	70.1±3.37	44.8±21.58	91.8±1.44	22.1±2.01	36.3±1.05	11.4±5.80
ResidualBase	69.9±3.59	54.0±27.50	92.5±0.81	21.9±2.34	38.6±0.47	96.2±0.81
FedCPA	72.3±0.88	12.5±1.02	93.1±1.02	20.8±1.35	38.7±0.63	4.8±1.40

Table 9: Comparison of defense performance over three datasets under targeted attack scenarios with pollution ratio $\gamma_p = 0.8$. Mean and standard deviation over ten last rounds are reported.

Method ($\gamma_p = 0.8$)	CIFAR-10	SVHN	TinyImageNet
No Defense	69.8±3.49	90.6±1.80	33.0±4.76
Median	59.8±3.16	89.9±1.55	28.7±4.73
Trimmed Mean	72.9±3.47	91.0±1.49	34.1±3.73
Multi Krum	72.7±3.61	92.6±0.99	35.9±2.22
FoolsGold	18.6±7.53	47.6±19.76	4.6±3.36
Norm Bound	64.9±4.19	90.8±2.06	29.3±5.18
RFA	72.6±2.31	92.7±0.96	36.5±0.78
ResidualBase	73.6±3.40	92.1±1.03	36.0±3.38
FedCPA	74.9±3.30	93.2±0.72	36.8±1.53

Table 10: Comparison of defense performance over three datasets under label flipping attack scenarios with pollution ratio $\gamma_p = 0.8$. Mean and standard deviation over ten last rounds are reported.

Method ($\gamma_p = 1.0$)	CIFAR-10	SVHN	TinyImageNet
No Defense	63.8±5.85	86.1±5.21	24.4±8.94
Median	56.8±7.23	89.6±2.49	21.2±8.71
Trimmed Mean	66.2±5.12	87.9±3.97	27.2±8.25
Multi Krum	73.0±3.78	92.6±1.42	35.9±3.10
FoolsGold	24.9±10.72	41.9±17.53	1.3±1.60
Norm Bound	63.5±4.45	86.6±7.05	24.1±8.86
RFA	71.5±2.66	92.4±1.06	36.3±1.12
ResidualBase	70.3±3.95	91.8±1.38	30.5±8.23
FedCPA	74.4±2.85	93.2±0.57	34.9±2.18

Table 11: Comparison of defense performance over three datasets under label flipping attack scenarios with pollution ratio $\gamma_p = 1.0$. Mean and standard deviation over ten last rounds are reported.

Method	CIFAR-10	SVHN	TinyImageNet
No Defense	32.7±4.18	47.8±8.72	2.1±1.09
Median	67.8±4.30	91.5±1.21	28.8±3.44
Trimmed Mean	55.6±4.38	72.5±9.72	12.1±5.63
Multi Krum	52.8±5.86	68.4±13.72	15.0±4.55
FoolsGold	13.9±4.13	6.7±0.00	0.5±0.08
Norm Bound	28.2±2.49	42.9±10.39	1.2±0.67
RFA	72.0±2.85	92.2±0.49	35.8±0.80
ResidualBase	74.6±2.11	93.7±0.39	37.0±1.05
FedCPA	74.8±2.42	93.6±0.58	36.1±1.37

Table 12: Comparison of defense performance over three datasets under Gaussian noise attack scenarios. Mean and standard deviation over ten last rounds are reported.