A. Detailed Experimental Settings

A.1. Baseline attack

BadNets [3]: An unconstrained backdoor attack with high attack strength and low stealthiness.
DBA [11]: DBA decomposes a global trigger pattern into separate local patterns and embeds them into the training sets of different attackers, resulting in a more negligible difference between benign and backdoor gradients. In our experiment, we split the 5×5 region into four sub-triggers with 2×2, 2×3, 2×3, and 3×3, respectively.
LP attack [11]: LP attack injects a backdoor into backdoor-critical (BC) layers, resulting in high stealthiness. In our experiment, we set the hyperparameters consistent with the original paper.

A.2. Defense

MultiKrum [1]: MK selects the $m = n - f$ clients with the smallest sum of pairwise $L2$ distances for aggregation, where n denotes the number of clients selected in each round and f denotes the number of tolerable attackers. In our experiment, we use $f = 2$ by default.
FLTrust [2]: Following the original paper, we select a small root dataset from clean training examples uniformly at random. Given that CIFAR100 has more classes, we increase the root dataset size from 100 to 200 for more effective detection.
FLAME [6]: In our experiment, we set the $min\_cluster\_size = n/2 + 1$, $min\_sample = 1$ for HDBSCAN clustering and $\sigma = 0.001$ for Gaussian noise, following the original paper.
Multi-Metrics [5]: In our experiments, the fraction of selected clients for aggregation is set to $p = 0.3$, which is the optimal value specified in the original paper.
DnC [9]: In accordance with the original paper, we set the filtering fraction $c = 1$ and the size of sub-samples $b = 10000$. We increase the number of iterations $niters$ from 1 to 5 for precise detection.
RLR [7]: The RLR method operates by monitoring sign-based patterns, with the learning rate flip threshold for individual parameters set to 4, following the original paper.
FLARE [10]: We select 10 clean samples from one class as the root dataset, following the original paper.
DeepSight [8]: In our experiment, we set the same hyperparameters consistent with the original paper.

B. Additional experiment result

B.1. The effectiveness of our attack in IID setting

We also evaluate our attack in the IID setting. As shown in Table 6, our attack successfully bypasses all defenses and achieves the highest BA in most settings.

Table 1. Main task accuracy on the unpoisoned global model.

Dataset (Model)	**MK**	**FLTrust**	**FLAME**	**MM**	**DnC**	**FLARE**	**RLR**
CIFAR10 (ResNets18)	77.69	81.27	74.84	72.11	79.16	77.96	75.55
CIFAR10 (VGG19)	73.55	83.99	62.04	73.92	82.72	84.76	77.62
CIFAR100 (ResNets18)	57.02	60.98	56.37	52.99	64.17	59.63	57.17

Table 2. The performance against DeepSight defense.

Dataset (Model)	Attack	MA	Best BA	Avg BA	MAR	BAR
CIFAR10 (VGG19)	Ours	77.94	**98.32**	**93.81**	**0.86**	0.63
	LP attack	77.96	94.61	93.55	0.81	0.34
CIFAR10 (ResNet18)	Ours	81.52	**97.53**	**95.94**	**0.80**	0.55
	LP attack	80.54	90.88	80.2	0.74	0.32
CIFAR100 (ResNet18)	Ours	66.5	**99.92**	**99.74**	**0.88**	0.73
	LP attack	66.81	3.70	1.19	0.39	0.85

B.2. The main task accuracy on the unpoisoned global model

We compare the main task accuracy between the backdoored model and the unpoisoned global model. As shown in Table 1, our attack achieves an MA close to that of the unpoisoned model, demonstrating its ability to maintain the performance while successfully injecting the backdoor.

B.3. The performance against DeepSight defense

We evaluate our attack performance against DeepSight defense. Table 2 shows that our attack achieves higher BA across all settings compared with the LP attack.

B.4. More evaluation on attack stealthiness

Figure 1 and Figure 2 show our crafted malicious update is close to the benign updates distribution, demonstrating that our attack is much stealthier than BadNets and DBA.

Table 3. The performance on More Networks.

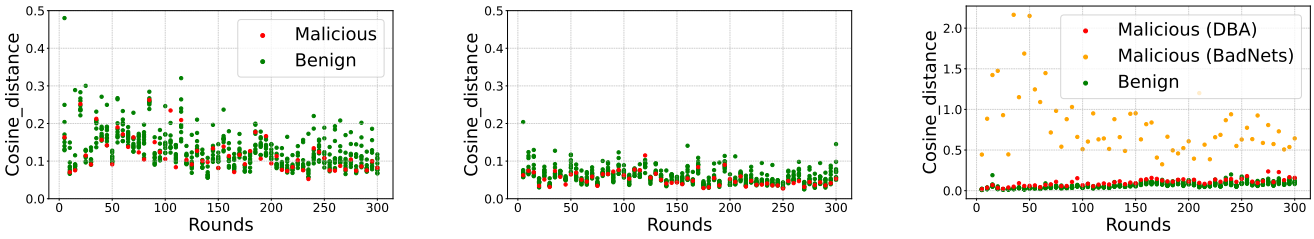
Model	Defense	MA	Best BA	Avg BA
ResNet50	FLAME	69.30	98.08	98.5
	MM	71.39	97.14	90.52
ResNet101	FLAME	56.02	97.54	96.8
	MM	65.25	98.71	95.52

Table 4. The evaluation of our attack trained on ViT model.

Defense	MK	FLTrust	FLAME	MM	DnC	RLR	FLARE
MA/BA	**97.21**	**94.28**	**99.14**	**97.14**	**98.89**	**97.75**	**97.39**

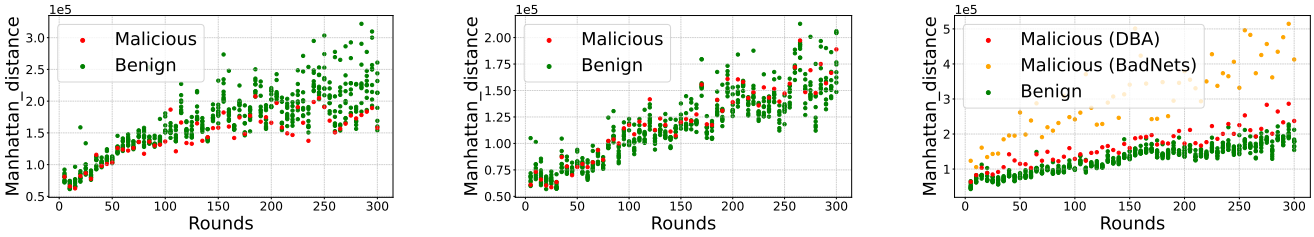
B.5. The performance on more networks

We further evaluate our attack on additional networks under the FLAME and MM settings. Specifically, we test two



(a) Ours (b) LP attack (c) BadNets and DBA

Figure 1. Cosine distance of malicious and benign model updates. A larger distance indicates that the update deviates significantly from benign updates. The result shows the stealthiness of our attack.



(a) Ours (b) LP attack (c) BadNets and DBA

Figure 2. Manhattan distance of malicious and benign model updates. A larger distance indicates that the update deviates significantly from benign updates. The result shows the stealthiness of our attack.

larger architectures, ResNet50 [4] and ResNet101 [4], and a transformer-based network ViT-tiny on the CIFAR10 dataset. As shown in Table 3 and Table 4, our attack consistently achieves high BA across both networks, highlighting its strong generalizability.

B.6. Impact of different degrees of non-IID

We evaluate the performance of our attack under varying degrees of non-IID data distributions, with $q = 0.5$ for CIFAR10 and $q = 0.2$ for CIFAR100 by default. We then assess the impact of higher non-IID degrees on both datasets to explore the challenges posed by data heterogeneity. Figure 3 shows that as the degree of non-IID data distribution increases, the BA declines, suggesting that higher non-IID levels present greater challenges in successfully injecting a backdoor. Nevertheless, our attack maintains robust performance even under high non-IID conditions, consistently outperforming the LP attack across all scenarios.

B.7. Impact of alignment interval

We further compare the impact of batch-wise alignment versus epoch-wise alignment on the attack's effectiveness. Batch-wise alignment updates the model after each batch, while epoch-wise alignment updates occur after each full epoch. Table 5 shows that batch-wise alignment results in lower BA on certain defenses, such as FLAME, compared to epoch-wise alignment. Even under FedAvg, batch-wise alignment performs worse than our attack. The reason for this drop in effectiveness lies in the nature of batch-wise updates: with each batch only considering a small subset of samples, the attack does not effectively integrate information from the entire global model. This can lead to inconsistent alignment and makes the backdoor less effective.

Table 5. Ablation study on scaling interval.

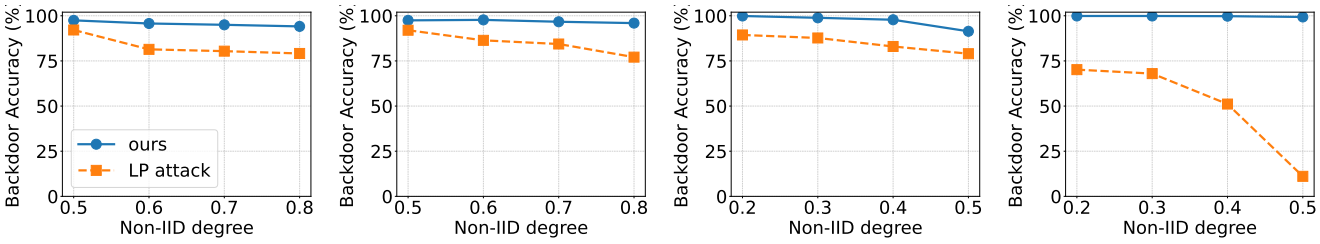
Model	**attack**	No Defense	FLAME	MM
CIFAR10 (VGG19)	batch-wise	93.26	92.63	**93.97**
	epoch-wise (ours)	**96.16**	97.46	93.40
CIFAR10 (ResNet18)	batch-wise	96.66	98.03	96.81
	epoch-wise (ours)	**97.21**	**98.16**	**97.52**
CIFAR100 (ResNet18)	batch-wise	79.67	98.03	96.93
	epoch-wise (ours)	**96.27**	**99.88**	**97.05**

B.8. The impact of attack interval

We further evaluate the impact of various attack intervals under all defenses trained on CIFAR10 (ResNet18). Figure 4 shows that our attack achieves high BA even when $F = 5$.

B.9. The impact of the compromised client proportion

We further evaluate the impact of various attack intervals under all defenses trained on CIFAR10 (ResNet18). Figure 5 shows that our attack achieves high BA even when $C = 0.02$.



(a) MK (CIFAR10) (b) FLAME (CIFAR10) (c) MK (CIFAR100) (d) FLAME (CIFAR100)

Figure 3. Impact of Non-IID degree in MK and FLAME.

Table 6. The effectiveness of our approach against the baseline attack across various state-of-the-art (SOTA) defenses under a fixed-frequency attack setting on IID datasets. BA values below 10% in CIFAR10 (10 classes) and below 1% in CIFAR100 (100 classes) are highlighted in red, indicating a failed attack. The highest BA achieved in each setting is presented in bold. Our attack results are reported as $a \pm b$, where a denotes the mean and b represents the standard deviation. MA and BA unit: %. Avg indicates the average. Each result is obtained as the average over five independent runs.

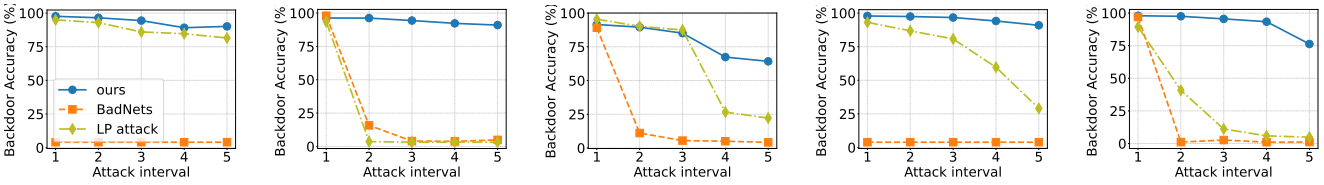
Defense		VGG19 (CIFAR-10)				ResNet18 (CIFAR-10)				ResNet18 (CIFAR-100)			
		BadNet	**DBA**	**LP**	**Ours**	**BadNet**	**DBA**	**LP**	**Ours**	**BadNet**	**DBA**	**LP**	**Ours**
FedAvg	MA	**84.15**	84.03	82.57	82.41±0.59	79.47	78.44	79.43	**79.92**±0.47	66.91	62.4	**68.47**	67.48±0.57
	Best BA	**98.31**	59.66	95.61	97.15±0.31	**98.33**	21.20	97.07	97.38±0.24	**100.0**	0.99	78.66	99.51±0.11
	Avg BA	**97.95**	54.91	95.36	96.85±0.29	**97.84**	14.87	95.73	97.27±0.21	**99.99**	0.68	72.62	99.31±0.04
MK	MA	**82.47**	82.10	81.93	82.37±1.20	76.84	76.79	76.62	**79.10**±1.19	62.01	60.05	**63.09**	62.13±0.54
	Best BA	7.97	11.21	97.42	**98.20**±0.29	4.26	11.24	96.54	**98.44**±0.07	0.58	1.28	95.90	**99.99**±0.01
	Avg BA	3.07	2.91	94.10	**97.66**±0.38	2.41	3.91	93.05	**97.98**±0.17	0.27	0.21	78.29	**99.93**±0.08
FLTrust	MA	**83.25**	83.18	82.86	82.95±0.41	**81.67**	80.99	81.61	81.25±0.98	61.93	63.21	63.75	**64.10**±0.51
	Best BA	79.6	16.21	95.99	**96.22**±2.14	95.72	21.19	96.79	**96.96**±0.18	46.97	0.75	74.88	**99.71**±0.26
	Avg BA	70.96	10.68	**95.15**	90.53±5.39	90.6	9.33	**96.64**	92.34±4.10	38.52	0.42	53.10	**95.58**±6.02
FLAME	MA	78.72	**79.16**	78.90	79.13±0.98	**77.71**	74.01	75.14	76.09±1.87	**58.67**	58.32	58.09	57.67±1.22
	Best BA	31.71	10.22	95.19	**97.76**±0.19	10.23	11.80	97.53	**98.89**±0.10	1.06	0.80	74.87	**99.99**±0.01
	Avg BA	6.15	2.52	85.33	**97.49**±0.93	3.33	4.57	95.69	**98.57**±0.15	0.51	0.42	69.70	**99.98**±0.01
MM	MA	**82.56**	81.97	81.78	82.30±0.65	**78.56**	73.06	78.03	78.26±0.44	58.32	58.16	**59.19**	58.34±0.16
	Best BA	4.66	12.73	96.64	**98.56**±0.55	6.74	5.38	95.39	**98.69**±0.30	0.93	0.93	92.41	**99.98**±0.03
	Avg BA	3.04	5.18	93.04	**97.20**±0.38	3.41	1.47	93.98	**98.25**±0.17	0.41	0.48	82.17	**90.57**±8.06
DnC	MA	84.2	83.89	84.46	**84.78**±0.11	79.72	79.27	78.87	**80.49**±0.34	66.69	66.99	**67.12**	66.71±0.23
	Best BA	5.22	7.97	95.53	**97.38**±0.11	13.09	8.90	97.37	**98.07**±0.39	0.32	0.62	76.07	**99.95**±0.04
	Avg BA	2.69	3.19	93.62	**96.77**±0.53	5.79	4.02	97.12	**97.89**±0.10	0.71	0.39	65.09	**99.65**±0.04
RLR	MA	**82.41**	81.21	82.34	81.76±0.49	**78.01**	75.22	76.85	77.09±0.49	57.77	60.13	60.88	**61.88**±0.79
	Best BA	95.71	15.02	95.16	**98.60**±0.30	64.46	13.72	78.21	**98.50**±0.13	0.02	0.05	81.49	**99.99**±0.01
	Avg BA	95.34	10.56	92.44	**98.23**±0.15	32.56	7.91	71.42	**97.80**±1.09	0.01	0.03	60.88	**99.98**±0.01
FLARE	MA	**85.24**	84.84	85.16	84.81±1.63	76.13	**76.30**	76.24	75.90±1.31	62.96	62.17	62.08	**63.43**±0.85
	Best BA	**98.14**	48.43	96.02	97.73±0.94	**97.94**	6.51	4.09	97.11±1.17	**100.0**	1.54	89.61	99.34±1.11
	Avg BA	**98.02**	30.03	95.24	97.28±1.23	**97.72**	5.76	3.77	96.46±1.30	**100.0**	0.72	79.6	98.34±2.92

B.10. Various trigger shapes

Figure 6 shows the three different trigger shapes applied to backdoor a ResNet18 model, trained on the CIFAR10 dataset.

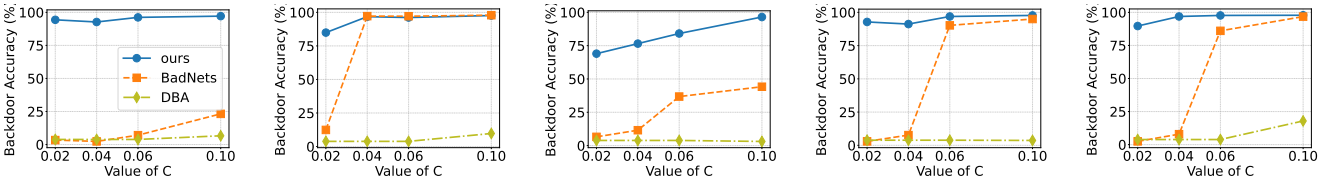
B.11. The performance on TinyImageNet

We further evaluate our attack on TinyImageNet. We follow prior work [12, 13] and assume a non-i.i.d. data distribution with a Dirichlet concentration parameter $h = 0.9$. As shown in Table 7, our attack achieves a BA of over 99% on MK, FLAME and MM, demonstrating its effectiveness on large-



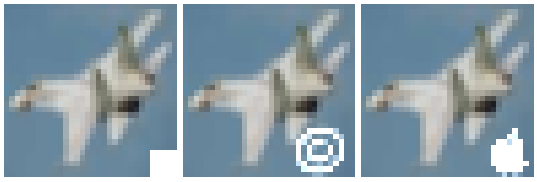
(a) DnC (b) FLARE (c) FLTrust (d) MM (e) RLR

Figure 4. Impacts of different attack frequencies on the attack performance on CIFAR10.



(a) DnC (b) FLARE (c) FLTrust (d) MM (e) RLR

Figure 5. Effectiveness of our attack under different proportions of compromised clients ($C = 0.02, 0.04, 0.06, 0.1$) in fixed-pool attack setting trained on CIFAR10.



(a) Square (b) Watermark (c) Apple

Figure 6. Illustration of various trigger shapes ("square," "apple," and "watermark").

Table 7. BA on TinyImageNet on IID(top)/Non-IID(bottom) datasets.

Defense	MK	FLTrust	FALME	MM	DnC	RLR	FLARE
BadNets	0.35	53.11	0.97	15.63	0.26	0.24	98.1
LP attack	78.29	84.10	69.70	82.17	65.09	60.88	79.6
Ours	**99.01**	**91.85**	**99.88**	**99.36**	**96.63**	**98.93**	**98.23**
BadNets	0.12	28.07	0.51	0.93	0.53	0.02	**97.90**
LP attack	78.29	57.89	61.59	77.55	64.24	73.79	67.68
Ours	**98.87**	**93.41**	**99.12**	**99.89**	**95.35**	**96.63**	97.13

scale datasets.

References

[1] Peva Blanchard, El Mahdi El Mhamdi, Rachid Guerraoui, and Julien Stainer. Machine Learning with Adversaries: Byzantine Tolerant Gradient Descent. In *NeurIPS*, 2017. 1

[2] Xiaoyu Cao, Minghong Fang, Jia Liu, and Neil Zhenqiang Gong. FLTrust: Byzantine-robust Federated Learning via Trust Bootstrapping. In *NDSS*, 2021. 1

[3] Tianyu Gu, Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. BadNets: Evaluating Backdooring Attacks on Deep Neural Networks. *IEEE Access*, 7:47230–47244, 2019. 1

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *CVPR*, 2016. 2

[5] Siquan Huang, Yijiang Li, Chong Chen, Leyu Shi, and Ying Gao. Multi-Metrics Adaptively Identifies Backdoors in Federated Learning. In *ICCV*, 2023. 1

[6] Thien Duc Nguyen, Phillip Rieger, Huili Chen, Hossein Yalame, Helen Möllering, Hossein Fereidooni, Samuel Marchal, Markus Miettinen, Azalia Mirhoseini, Shaza Zeitouni, Farinaz Koushanfar, Ahmad-Reza Sadeghi, and Thomas Schneider. FLAME: Taming Backdoors in Federated Learning. In *USENIX Security 22*, 2022. 1

[7] Mustafa Safa Ozdayi, Murat Kantarcioglu, and Yulia R Gel. Defending against backdoors in federated learning with robust learning rate. In *AAAI*, 2021. 1

[8] Phillip Rieger, Thien Duc Nguyen, Markus Miettinen, and Ahmad-Reza Sadeghi. DeepSight: Mitigating Backdoor Attacks in Federated Learning Through Deep Model Inspection. In *NDSS*, 2022. 1

[9] Virat Shejwalkar and Amir Houmansadr. Manipulating the Byzantine: Optimizing Model Poisoning Attacks and Defenses for Federated Learning. In *NDSS*, 2021. 1

[10] Ning Wang, Yang Xiao, Yimin Chen, Yang Hu, Wenjing Lou, and Y. Thomas Hou. FLARE: Defending Federated Learning against Model Poisoning Attacks via Latent Space Representations. In *AsiaCCS*, 2022. 1

[11] Chulin Xie, Keli Huang, Pin-Yu Chen, and Bo Li. DBA: Distributed Backdoor Attacks against Federated Learning. In *ICLR*, 2020. 1

[12] Hangfan Zhang, Jinyuan Jia, Jinghui Chen, Lu Lin, and Dinghao Wu. A3FL: Adversarially Adaptive Backdoor Attacks to Federated Learning. In *NeurIPS*, 2023. 3

[13] Zhengming Zhang, Ashwinee Panda, Linyue Song, Yaoqing Yang, Michael W. Mahoney, Prateek Mittal, Kannan Ramchandran, and Joseph Gonzalez. Neurotoxin: Durable Backdoors in Federated Learning. In *ICML*, 2022. 3