Batman: Benign Knowledge Alignment Through Malicious Null Space in Federated Backdoor Attack

Supplementary Material

A. Derivation of the Closed-Form Solution for Equation 7

We provide the derivation of the closed-form solution to the following constrained optimization problem:

$$\min_{\Delta} \|(W_1 + \Delta) - W_b\|_F^2 + \lambda \|\Delta\|_F^2, \quad \text{s.t. } \Delta = \Delta P.$$

Let us define $D := W_1 - W_b$, then the objective becomes:

$$\min_{\Delta} \|\Delta + D\|_F^2 + \lambda \|\Delta\|_F^2, \quad \text{s.t. } \Delta = \Delta P.$$

We now solve the unconstrained version under the assumption that Δ lies in the row space of the projection matrix P, i.e., $\Delta = \Delta P$. Then, let us consider the objective function:

$$\begin{aligned}\mathcal{L}(\Delta) &= \|\Delta + D\|_F^2 + \lambda \|\Delta\|_F^2 \\ &= \mathrm{Tr}\big[(\Delta + D)(\Delta + D)^\top\big] + \lambda \, \mathrm{Tr}\big(\Delta\Delta^\top\big).\end{aligned}$$

Taking the gradient of $\mathcal{L}$ with respect to Δ and setting it to zero:

$$\nabla \mathcal{L} = 2(\Delta + D) + 2\lambda\Delta = 0.$$

Solving this yields the stationary point:

$$(1 + \lambda)\Delta = -D \quad \Rightarrow \quad \Delta = -\frac{1}{1+\lambda} D.$$

However, we must project this solution into the null space defined by P. Therefore, we enforce the constraint $\Delta = \Delta P$ by multiplying both sides by P:

$$\Delta = -\frac{1}{1+\lambda} DP.$$

Thus, the optimal solution that satisfies both the objective and the constraint is:

$$\Delta^* = -\frac{1}{1+\lambda}(W_1 - W_b)P.$$

Alternatively, writing it in a numerically stable matrix form using projection properties, we can express:

$$\Delta^* = -\left(I + \lambda^{-1} P\right)^{-1} P(W_1 - W_b),$$

which corresponds to Equation 6. This solution performs the optimal projection of the discrepancy $W_1 - W_b$ into the null space, balancing alignment accuracy and perturbation magnitude under the orthogonality constraint.

B. Algorithm of *Batman*

Algorithm 1 Batman: Benign Alignment Through Malicious Null Space

Input : Poisoned dataset $\mathcal{D}_k$, clean dataset $\mathcal{D}_k^{\emptyset}$, global model W_g^t, regularization λ, top-r SVD threshold

Output: Aligned malicious model $W_k^{t+1} + \Delta^*$

Train poisoned model W_k^{t+1} by $\mathcal{D}_k$ according to Eq. 1

$$W_k^{t+1} \leftarrow W_g^t - \eta \left[(1-\alpha)\nabla\mathcal{L}(x, y) + \alpha \nabla\mathcal{L}(x + \tau(x), \widetilde{y})\right]$$

Train clean model $W_k^{\emptyset}$ using $\mathcal{D}_k^{\emptyset}$

$$W_k^{\emptyset} \leftarrow W_g^t - \eta \nabla\mathcal{L}(x, y)$$

Construct benign set: $\mathcal{W}_b \leftarrow \{W_k^{\emptyset}, W_g^t\}$ according to Eq. 5

`// Get the benign knowledge` $\mathcal{W}_b$

Compute truncated SVD: $W_k^{t+1} = W_k^{\text{main}} + W_k^{\text{res}}$ according to Eq. 3

`// Get the malicious knowledge` W_k^{main}

Apply SVD on W_k^{main} to obtain null space projection: $P = \tilde{U}_0 \tilde{U}_0^\top$ according to Eq. 4

`// Construct the malicious Null Space`

Initialize $\Delta_{\text{sum}} \leftarrow 0$

foreach $W_k^{\text{b}} \in \mathcal{W}_k^{\text{b}}$ **do**

  $\Delta \leftarrow -(I + \lambda^{-1}P)^{-1} P(W_k^{t+1} - W_k^{\text{b}})$ according to Eq. 7

  $\Delta_{\text{sum}} \leftarrow \Delta_{\text{sum}} + \Delta$

end `// Compute the final perturbation`

$\Delta^* \leftarrow \frac{1}{2}\Delta_{\text{sum}}$

return $W_k^{t+1} + \Delta^*$

C. The detail of Experiment setup

C.1. The base setup

We evaluate our method on three standard image classification datasets: **CIFAR-10**, **Fashion-MNIST**, and **CINIC**.

Table 3. Federated learning setup across datasets.

Symbol	Description	Fashion-MNIST	CIFAR-10/100	CINIC
N	# of clients	100	100	100
C	Selected clients proportion	10%	10%	10%
E	Local epochs	5	5	5
B	Local batch size	64	64	64
R	Global training rounds	100	100	100
M/N	Malicious client proportion	20%	20%	20%
α	Malicious data proportion	50%	50%	50%
η	Local learning rate	0.01	0.1	0.1
r	Rank r in malicious knowledge extraction	4	4	4

Table 4. Performance comparison of different backdoor attacks under Fashion-MNIST, three defense mechanisms (DnC, FLTrust, TrimmedMean), and one no-defense aggregation (FedAvg).

Method	FedAvg			DnC			FLTrust			TrimmedMean		
	ASR(↑)	ACC(↑)	AVG(↑)	ASR(↑)	ACC(↑)	AVG(↑)	ASR(↑)	ACC(↑)	AVG(↑)	ASR(↑)	ACC(↑)	AVG(↑)
BadPFL	58.50	90.05	74.27	42.85	89.96	66.40	11.54	89.17	50.36	13.06	89.90	51.48
Lp-attack	94.67	90.20	92.44	58.60	90.29	74.45	20.60	89.68	55.14	8.46	89.95	49.21
BadNet	99.85	90.19	95.02	94.19	90.41	92.30	99.78	89.44	89.44	90.56	90.04	90.30
DBA	99.98	90.14	95.06	99.97	90.02	95.00	99.95	89.37	94.66	99.99	89.91	**94.95**
Neurotoxin	99.89	90.08	94.99	98.64	90.10	94.37	6.65	90.01	48.33	26.72	89.75	58.24
Batman (Ours)	99.77	90.80	**95.29**	99.59	90.71	**95.15**	99.74	89.81	**94.78**	88.00	90.27	89.13

In each communication round, only a subset of clients is randomly selected for participation. The federated learning setup is summarized in Table 3. Specifically, the total number of clients is set to 100, with 10% selected per round. Each client performs 5 local training epochs with a batch size of 64. The global training process lasts for 100 rounds. To simulate adversarial behavior, 20% of the clients are designated as malicious, each injecting 50% poisoned data. The local learning rate is set to 0.01 for Fashion-MNIST and 0.1 for CIFAR-10 and CINIC. For malicious knowledge extraction, the rank r in truncated SVD is fixed at 4 across all datasets. These settings aim to simulate a realistic non-IID scenario and allow fair comparison of different attack and defense strategies.

C.2. Baseline Attack

We compare our proposed method with several representative baseline backdoor attacks in federated learning:

- **BadNet** [11] [arXiv'17]: A classic data poisoning backdoor attack where a fixed visual pattern is embedded into inputs, associated with a target label.
- **DBA** [41] [ICLR'20]: Distributed Backdoor Attack, which spreads different trigger components across multiple compromised clients to collaboratively form a global trigger.
- **BadPFL** [7] [ICLR'25]: An attack embedding natural feature triggers to both global and personalized models, aiming to improve stealth and durability.
- **LP-Attack** [48] [ICLR'24]: A model poisoning strategy that selectively modifies a few critical layers to minimize perturbation magnitude while maintaining backdoor effectiveness.
- **Neurotoxin** [45] [ICML'22]: A gradient manipulation attack that avoids tampering with parameters heavily updated by benign clients, preserving backdoor behavior during aggregation.

C.3. Defense Mechanisms

To evaluate the stealth and robustness of the attack methods, we include multiple widely adopted federated learning defense mechanisms:

- **FedAvg** [27] [AISTATS'17]: The standard federated averaging protocol without specific defense mechanisms.
- **DnC** [35] [NDSS'21]: Divide-and-Conquer defense, grouping clients and applying group-wise aggregation to detect anomalous updates.
- **FLTrust** [5] [NDSS'21]: A defense framework that anchors trust in a small set of trusted data to calibrate client model updates before aggregation.
- **TrimmedMean** [43] [ICML'18]: It discusses the optimal statistical performance even in the presence of adversarial behavior from worker machines, using techniques such as median and trimmed mean operations.
- **FLuardian** [47] [TIFS'25]: A novel defense against layer-space backdoor attacks in federated learning, using a

layer-wise scoring mechanism that weights layers by importance to detect malicious clients.

- **AlignIns** [42] [CVPR'25]: AlignIns detects malicious clients by inspecting the directional alignment of local updates. It filters anomalous updates based on coarse- and fine-grained alignment patterns.

C.4. Datasets and model

We conduct experiments on four widely used image classification datasets: **CIFAR-10**, **CIFAR-100**, **Fashion-MNIST**, and **CINIC**. The datasets vary in image modality and difficulty level, offering a comprehensive evaluation of the proposed method across different scenarios. For CIFAR-10, CIFAR-100, and CINIC—which consist of RGB images with a resolution of 32×32—we adopt the ResNet-18 architecture as the backbone model due to its strong performance and common usage in federated learning benchmarks. For Fashion-MNIST, which contains grayscale images of size 28×28, we use a lightweight convolutional neural network (CNN) to better suit the lower image complexity and resolution. The details are summarized in Table 5.

Table 5. Datasets and model architectures.

Dataset	Image	Model
CIFAR-100	RGB, 32×32	ResNet-18
CIFAR-10	RGB, 32×32	ResNet-18
Fashion-MNIST	Gray, 28×28	CNN
CINIC	RGB, 32×32	ResNet-18

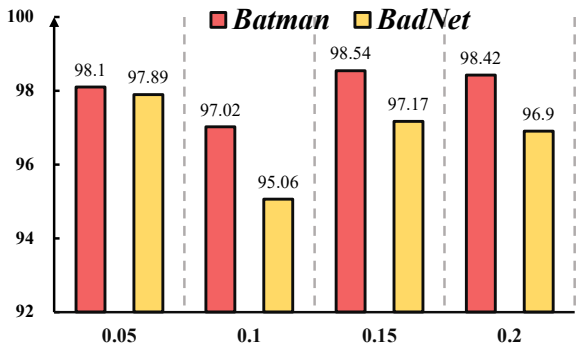

Figure 8. ASR comparison between Batman and BadNet under different client online ratios.

Table 6. Ablation on aligning objective tested under FedAvg and attack time tested under TrimmedMean.

Metric	Aligning objective			Attack time	
	Only W_k^0	Only W_g^t	Batman	Start epoch	End epoch
ASR	96.73	95.56	**97.02**	81.75	**95.59**
ACC	71.45	69.09	**72.54**	**75.64**	73.69
AVG	84.09	82.33	**84.78**	78.70	**84.64**

C.5. The Additional Experiments

Additional results on multiple datasets. Table 4 reports the performance of different backdoor attacks on Fashion-MNIST under four aggregation schemes, while Table 2 presents additional results on CIFAR-100, CIFAR-10, and CINIC under multiple defenses. Overall, Batman consistently achieves competitive AVG scores across different datasets and defense settings, demonstrating a favorable balance between attack success rate and clean accuracy.

Impact of client online ratio. We further evaluate the effect of different client online ratios, as shown in Fig. 8. As the online ratio increases from 0.05 to 0.2, Batman consistently maintains higher ASR than BadNet. These results indicate that Batman remains effective under varying levels of client participation and exhibits stronger robustness than the baseline attack in dynamic federated settings.

Aligning objective and attack time. We align the attacker to both the clean local model W_k^0 and the current global model W_g^t. Using only W_k^0 may deviate from the global benign trend, while using only W_g^t may overlook client-specific benign dynamics. As shown in Table 6, combining both leads to the best overall performance. We further apply Batman only in the final training round. Injecting the attack earlier can be gradually overridden by subsequent malicious updates, whereas applying it at the end makes the correction persist after full optimization.