RAID 2024

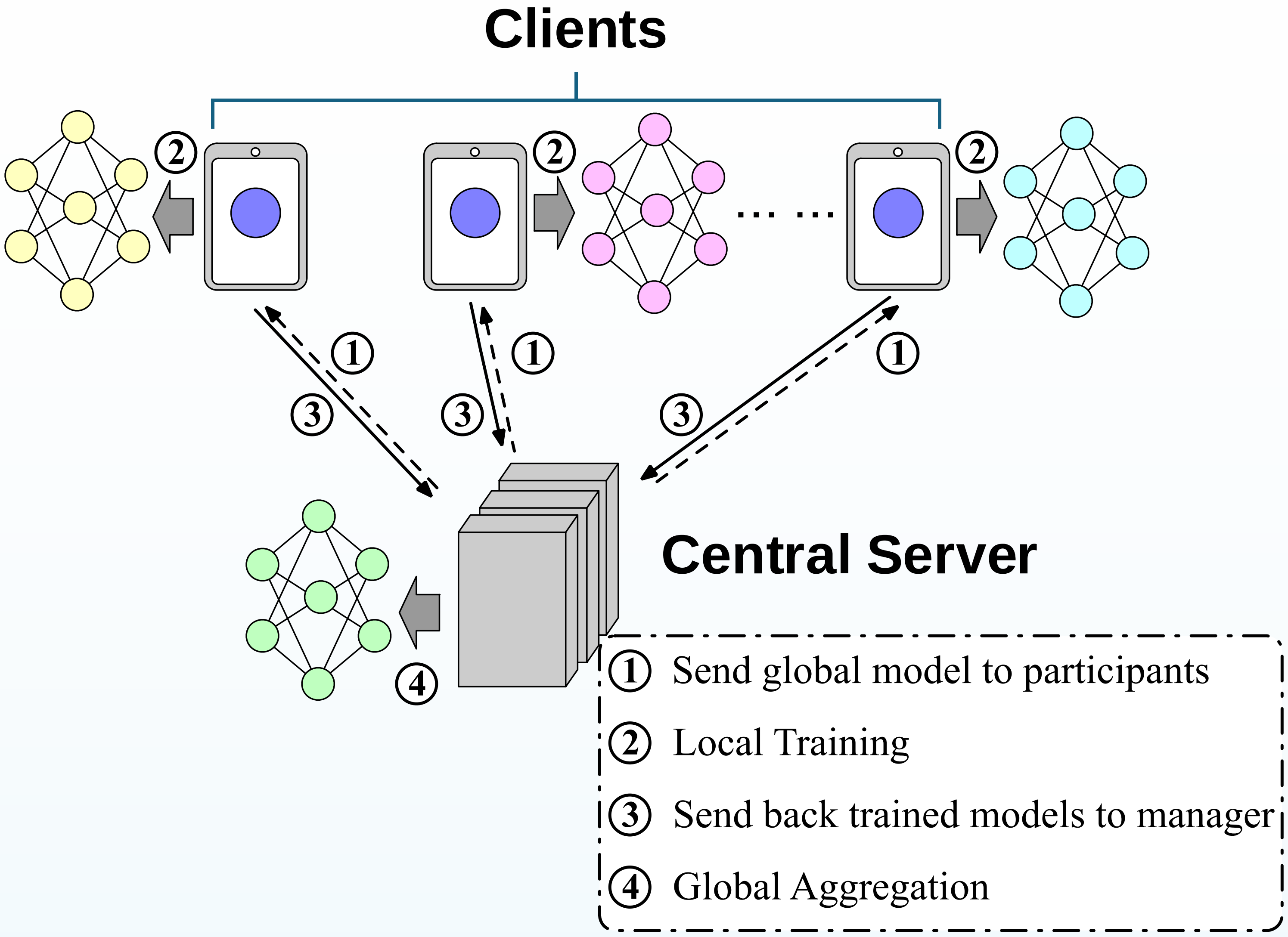
# Enhancing Model Poisoning Attacks to Byzantine-Robust Federated Learning via Critical Learning Periods


**Gang Yan** (UC Merced, USA)
**Hao Wang** (Stevens Institute of Technology, USA)
**Xu Yuan** (University of Delaware, USA)
**Jian Li** (Stony Brook University, USA)

# I. Basics of Federated Learning

# Federated Learning Workflow






- **Goal**

$$\min_{w \in \mathbb{R}^d} \mathcal{L}(\boldsymbol{w}, D) := \sum_{i \in \mathcal{N}} \frac{|D_i|}{|D|} \mathcal{L}_i(\boldsymbol{w}, D_i)$$

where local loss $\mathcal{L}_i(\boldsymbol{w}, D_i)$

- **Local Training**

$$\boldsymbol{w}_i^{(t)}(k) \leftarrow \boldsymbol{w}_i^{(t)}(k-1) - \eta \nabla \mathcal{L}_i$$

where $\eta$ is learning rate

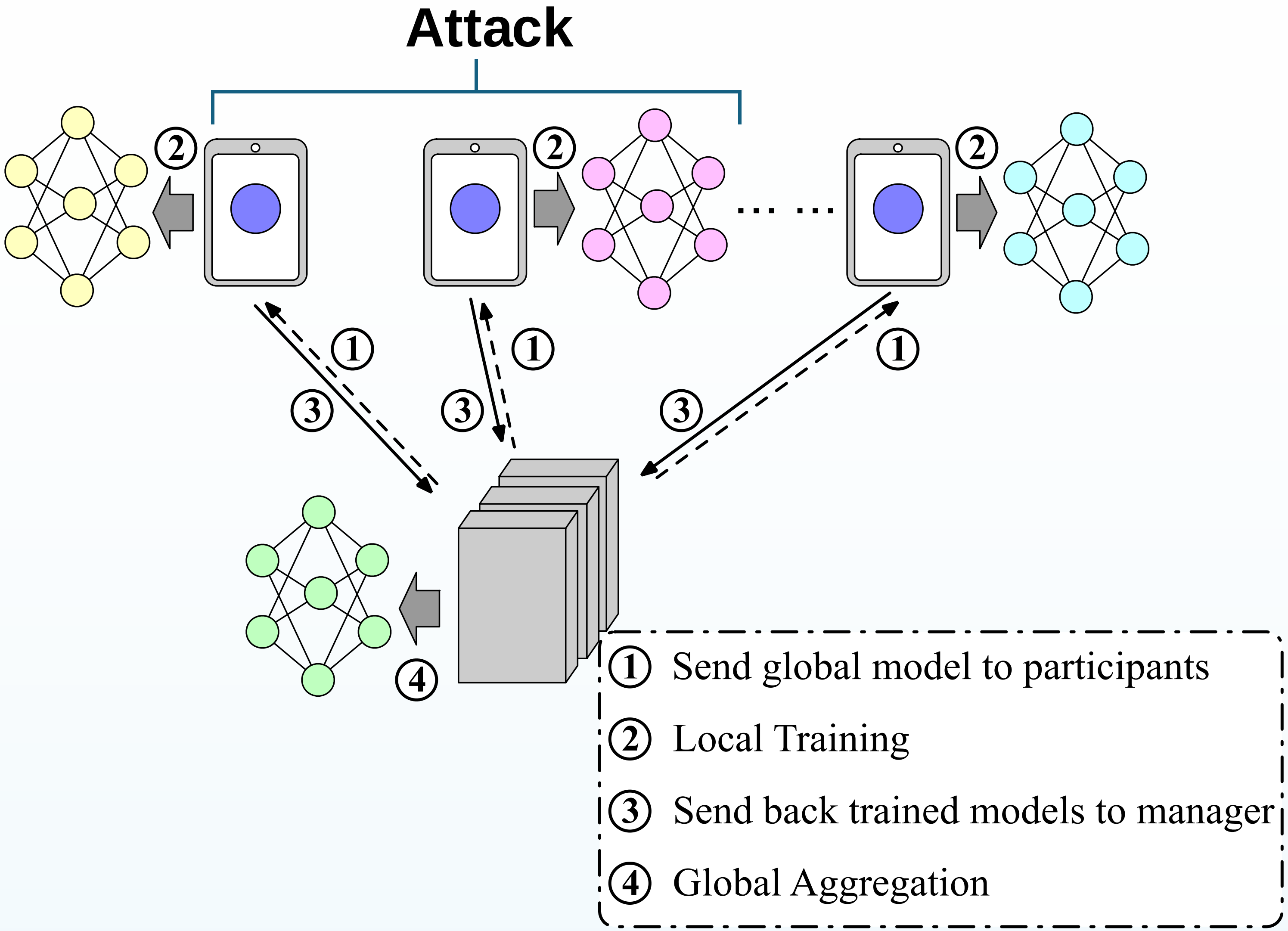
- **Global Aggregation**

$$\boldsymbol{w}^{(t)} \leftarrow \sum_{i \in \mathcal{N}^{(t)}} \frac{|D_i|}{|\cup_{i \in \mathcal{N}^{(t)}} D_i|} \boldsymbol{w}_i^{(t)}(K)$$

- Byzantine-robust Aggregation Rules on server

# Attack & Defense in FL

RAiD 2024




### ❑ Targeted Attack

- Minimize the accuracy on specific test inputs

### ❑ Untargeted Attack

- Minimize the global model accuracy on any test input

### ❑ Defense

Byzantine-robust methods on central server

- Detect and remove outliers
- Limit malicious updates' impacts

❖ **Attack:** degrade the global model accuracy by contributing malicious model updates

# Challenges in Existing Attacks



- **Fixed Attack Budget**: Utilizes a **constant number of malicious clients**, leading to a tradeoff between attack impact and budget

- **Uniform Attack Strategy**: Assumes **all training phases are equally important**, overlooking the significance of initial learning phases

- **Vulnerability to Defenses**: **Susceptible** to detection and mitigation by robust defenses (e.g., FLTrust, SparseFed)

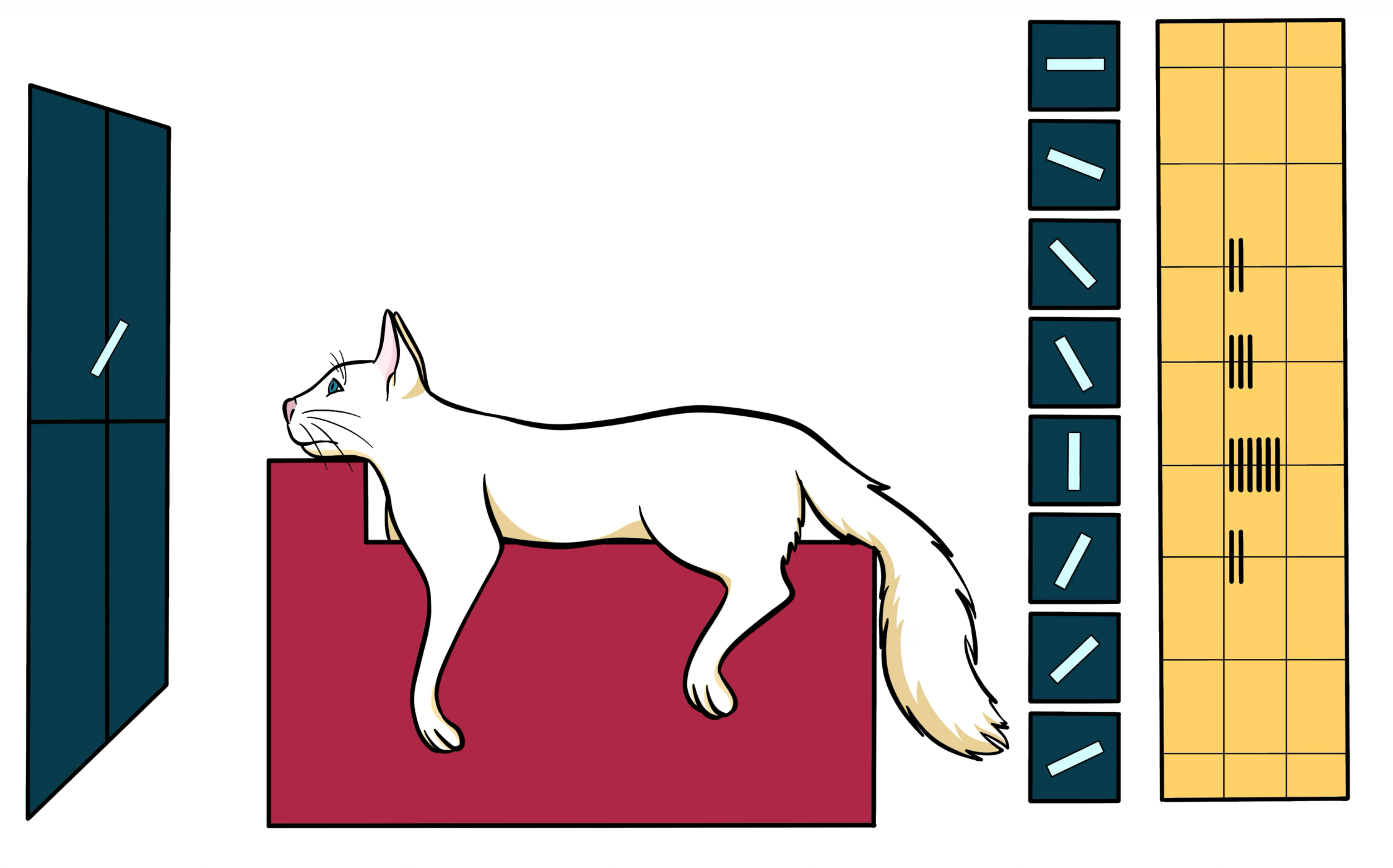
- **Lack of Adaptiveness**: Fails to adjust the attack strategy based on ***critical learning periods***, missing the opportunity for maximum impact

**What are Critical Learning Periods?**

# What are CL Periods?


# What are CL Periods?

- **Two special cylinders**
  - Vertical/horizontal lines
- Kittens that were exposed to vertical lines for **the first few months since birth**
  - Only see vertical lines, but not horizontal ones—for the rest of their lives
  - And vice versa



https://www.futurelearn.com/info/courses/research-methods-psychology-animal-models-to-understand-human-behaviour/0/steps/265398

# Impact of Critical Learning Periods

RAiD 2024

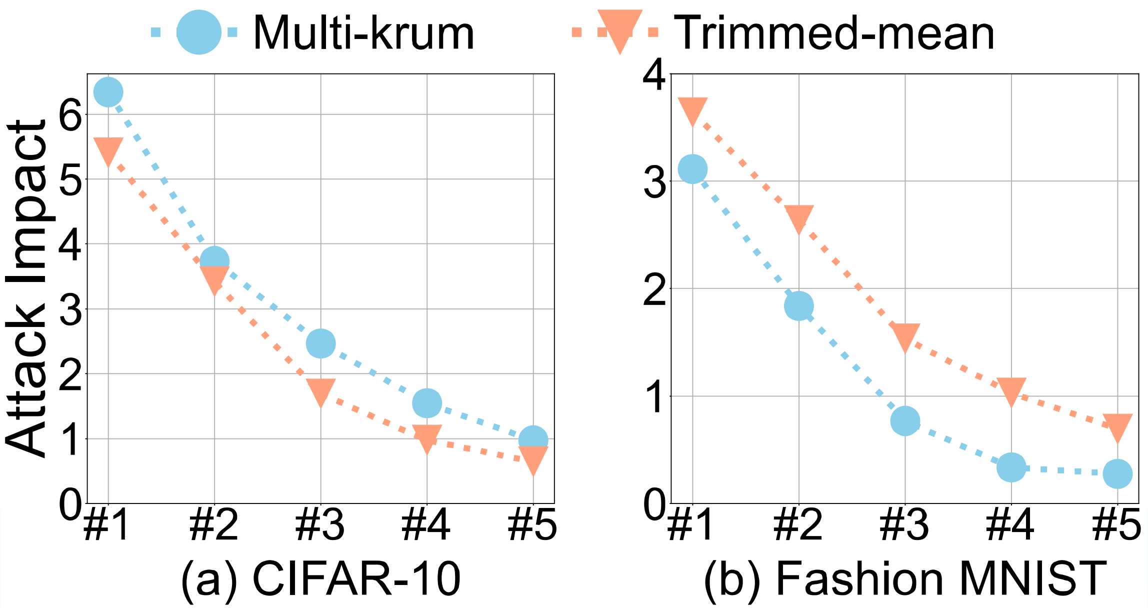



**Figure 1: FL under model poisoning attacks exhibits CLP, where the Min-Max attack occurs in (#1) rounds 0-20; (#2) rounds 20-40; (#3) rounds 40-60; (#4) rounds 60-80; and (#5) rounds 80-100, respectively.**

- **Critical Learning Periods (CLP)**: *Initial training phases* in deep neural networks that have an irreversible impact on the model's final quality

# Detection of Critical Learning Periods



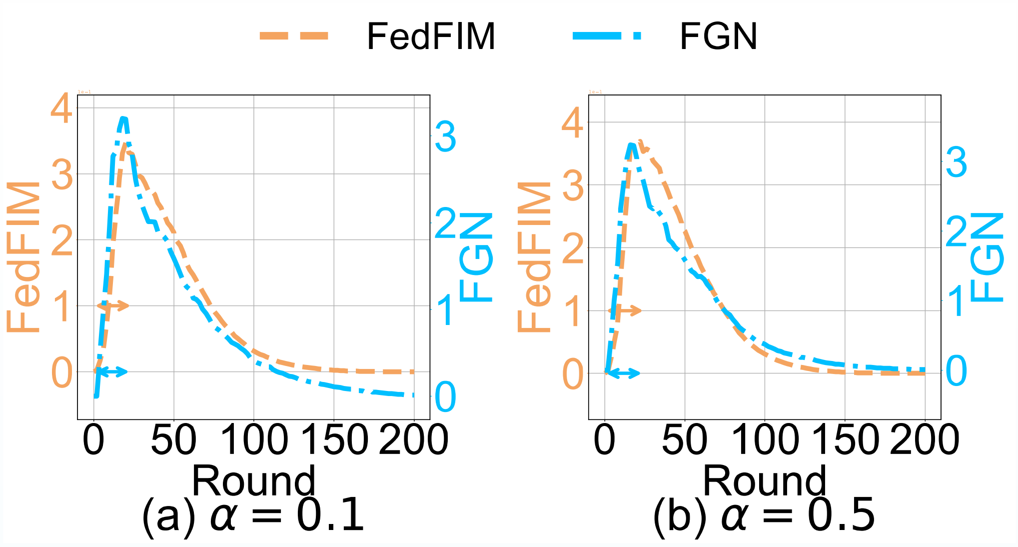



Figure 2: Detecting CLP via FGN and FedFIM, where the shade and double-arrows indicate identified CLP.

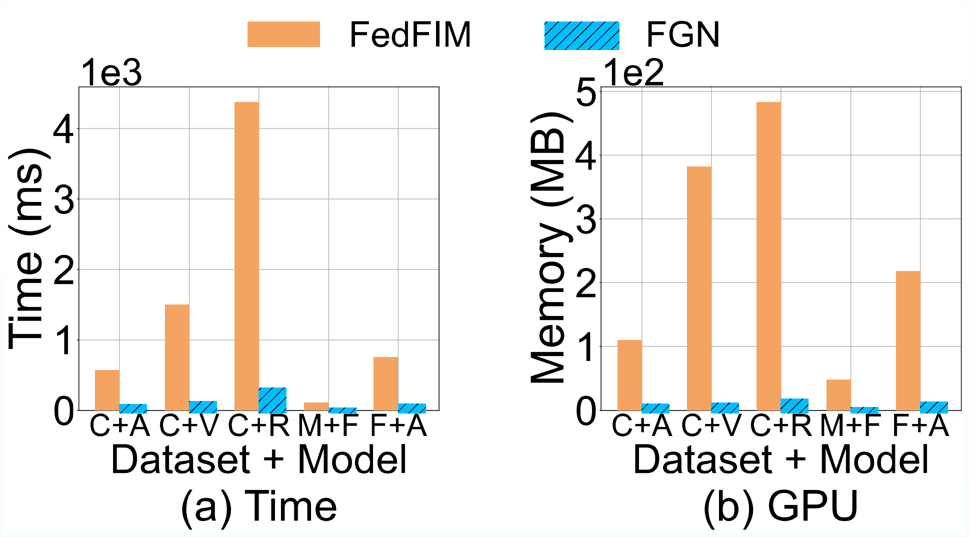



**Figure 3: Computation time and memory consumption of FGN and FedFIM approach to detect CLP.**

- **Federated Gradient Norm (FGN)**: CLP is identified using the changes in the Federated Gradient Norm during training

- **Why FGN?**: FGN provides a computationally efficient and online method to detect CLP, allowing for adaptive adjustments in the attack strategy

# Proposed Attack Methods



## ❑ $\mathcal{A}$-CLP (CLP-Aware Model Poisoning)

- **Adaptive Budget**: Dynamically adjusts the number of malicious clients
- **Optimized Strategy**: Increases attack budget during CLP for maximum impact, reducing it afterward to enhance efficiency
- **Improved Resilience**: Strengthens resistance against defenses like FLTrust

## ❑ GraSP (CLP-Aware Similarity-Based Attack)

- **Lightweight**: Uses a cosine similarity approach to craft malicious gradients
- **Approximate Deviations**: Deviates gradients based on similarity, without strictly following the global model's inverse direction
- **Superior Impact**: Achieves better attack performance

# II. Design of $\mathcal{A}$-CLP

# Overview of $\mathcal{A}$-CLP



---

**Algorithm 1** $\mathcal{A}$-CLP: CLP Aware Model Poisoning Attacks

---

1: **for** $t = 0, 1, \cdots, T - 1$ **do**
2:     **if** $\frac{\mathrm{FGN}(t) - \mathrm{FGN}(t-1)}{\mathrm{FGN}(t-1)} \geq \delta$ **then**
3:         The adversary invokes a larger number of malicious clients to share malicious gradients (e.g., $2m$) with the central server //More malicious clients during CLP
4:     **else**
5:         A smaller number of malicious clients is invoked to share malicious gradients (e.g., $m/2$) with the central server //Fewer malicious clients after CLP
6:     **end if**
7: **end for**

---

- **Concept**: $\mathcal{A}$-CLP adapts the number of malicious clients during federated learning rounds based on the identification of CLP
- **Key Insight**: Larger attack budgets are only required during the initial critical learning periods for maximum impact

# Feasibility Guarantee for $\mathcal{A}$-CLP

RAiD 2024

$$M' = \left\lceil \frac{(N-M)m}{n-m} \right\rceil, \quad m \le \left\lceil \frac{nM}{N} \right\rceil.$$

$N$ denotes the total number of clients,

$n$ represents the clients selected in each round

$M$ is the total number of controlled clients

$M'$ is the number of **activated clients**

$m$ **is the corresponding number of selected malicious clients (we want to guarantee)**

| $M'$ | Method | $n = 16$ | **n=32** | $n = 48$ |
|---|---|---|---|---|
| $m = 0.0625n$ | Equation (1) | 7 | **7** | 7 |
| | Simulation | 7 | **7** | 7 |
| $m = 0.125n$ | Equation (1) | 14 | **14** | 14 |
| | Simulation | 14 | **14** | 14 |
| $m = 0.25n$ | Equation (1) | 32 | **32** | 32 |
| | Simulation | 32 | **32** | 32 |

**Table 1: The number of malicious clients $M'$ invoked by the adversary so as to guarantee that on average $m$ malicious clients are selected by the server.**

# Experimental Details



- **Datasets:** CIFAR-10, CIFAR-100, MNIST, Fashion-MNIST, Shakespeare
- **Models:** AlexNet, VGG-11, ResNet-18, LSTM
- **Baseline Attacks:** Fang, LIE, Min-Sum, Min-Max, MPHM
- **Settings:** Total number of clients $N = 128$ selected clients $n = 32$, controlled clients $M = 32$
- **Objective:** Evaluate the attack impact, budget, and resilience against different defense mechanisms

# Different CLP Augmented Schemes



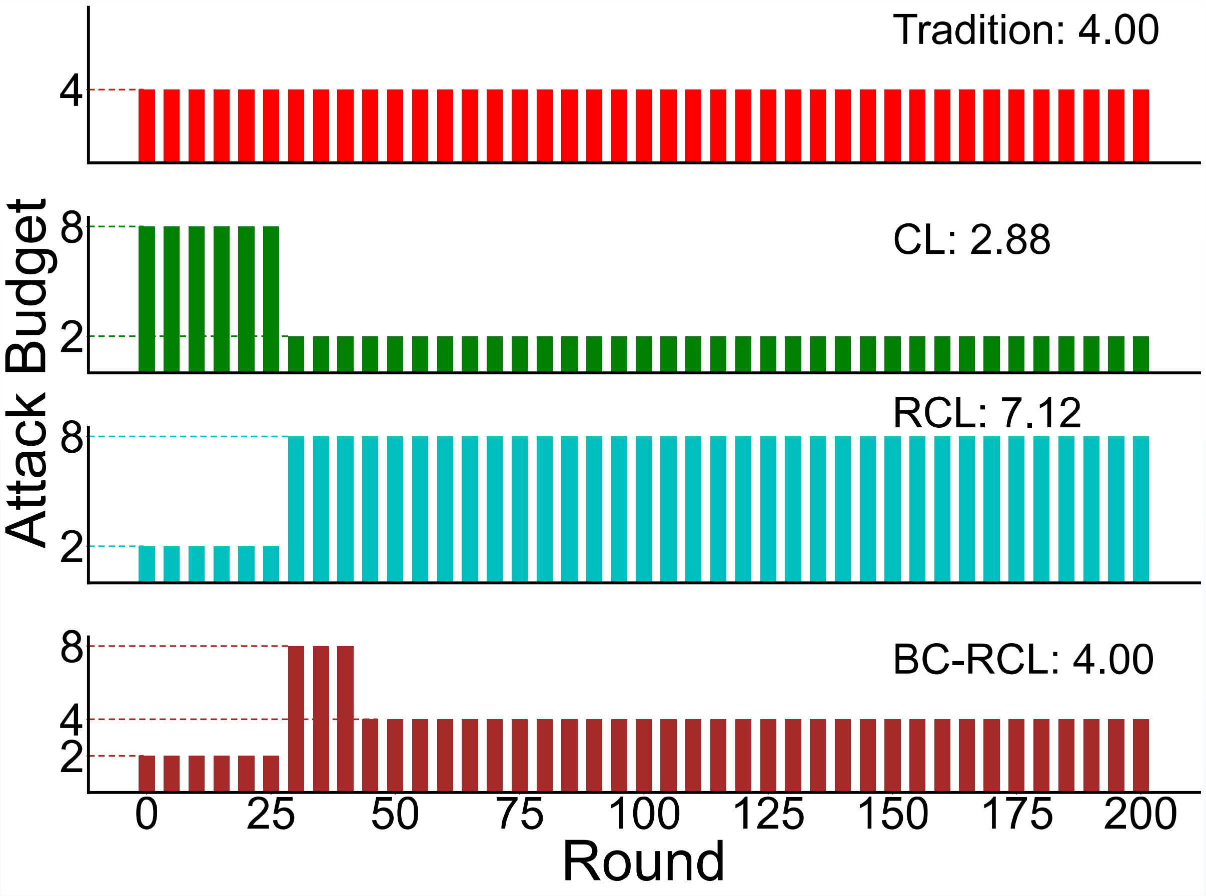



**Figure 4: The attack budget: A fixed average attack budget of 4 per round.**

- **Traditional**: Uses a fixed number of malicious clients throughout all training rounds
- **CL (CLP-Aware)**: Increases the number of malicious clients during CLP and reduces it after
- **RCL (Reverse CLP)**: Reduces malicious clients during CLP and increases them afterward
- **BC-RCL (Budget-Constrained RCL)**: Similar to RCL but with a fixed overall attack budget

# Different CLP Augmented Schemes

RAiD 2024

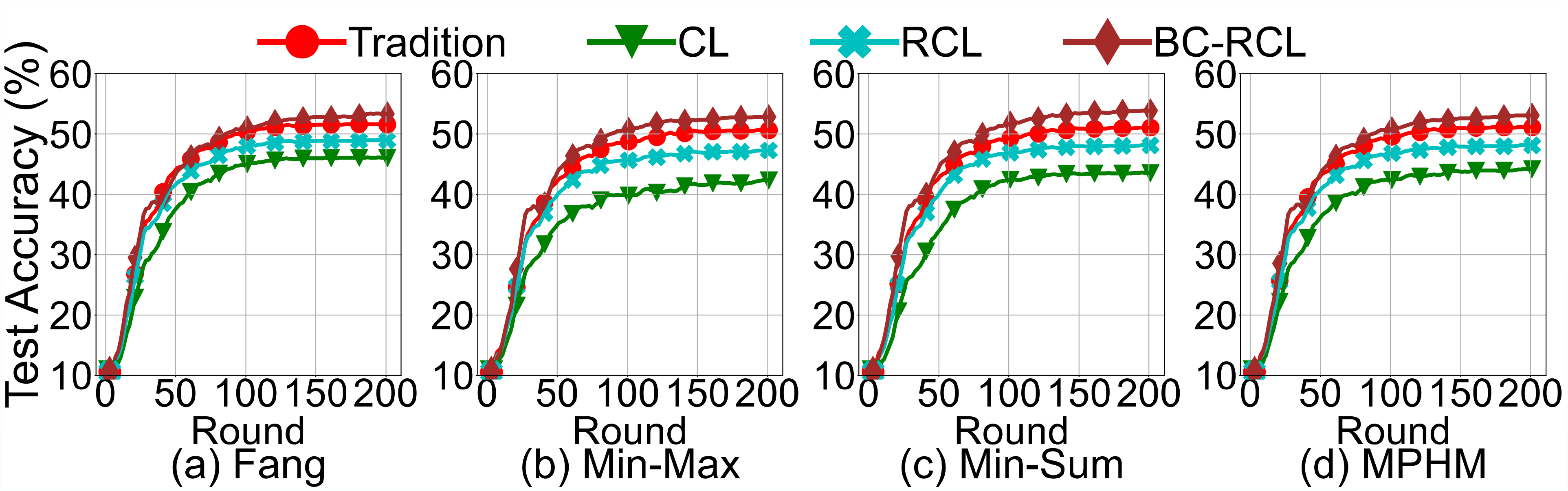



**Figure 5: Comparisons of different CLP aware attacks to FL. All attacks *do not know the gradients on benign clients.***

- **$\mathcal{A}$-CLP (CL)** significantly **outperforms** traditional attacks, achieving **higher accuracy reduction** by dynamically targeting CLP
- **RCL** and **BC-RCL** show **limited impact** as they fail to fully leverage the CLP

# Evaluations of Attack Impact



| Dataset (Model) | Aggregation Rule | No Attack (Accuracy) | Fang Trad. | Fang CL | LIE Trad. | LIE CL | Min-Max Trad. | Min-Max CL | Min-Sum Trad. | Min-Sum CL | MPHM Trad. | MPHM CL | Label Flipping Trad. | Label Flipping CL |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| CIFAR-10 (AlexNet) | Multi-krum [10] | 57.57 | 10.40 | **20.02** | 5.73 | **11.86** | 12.03 | **26.47** | 11.32 | **24.37** | 11.19 | **23.27** | 3.23 | **7.94** |
| | Bulyan [17] | 56.34 | 9.40 | **20.69** | 7.50 | **12.99** | 7.98 | **20.90** | 6.53 | **16.95** | 8.15 | **20.73** | 4.89 | **10.54** |
| | Trimmed-mean [59, 66] | 57.33 | 10.32 | **22.44** | 7.36 | **17.23** | 9.50 | **22.85** | 8.35 | **19.44** | 8.91 | **21.37** | 6.19 | **12.05** |
| | Median [59, 66] | 55.46 | 11.73 | **22.62** | 10.89 | **18.44** | 9.10 | **20.48** | 7.91 | **18.44** | 9.03 | **20.14** | 6.95 | **13.04** |
| | AFA [41] | 57.89 | 6.99 | **11.81** | 2.98 | **7.41** | 9.27 | **19.05** | 7.73 | **14.83** | 8.81 | **17.32** | 2.01 | **5.30** |
| CIFAR-10 (VGG-11) | Multi-krum [10] | 62.63 | 9.13 | **16.03** | 6.24 | **12.82** | 9.94 | **17.94** | 9.50 | **18.07** | 9.72 | **18.03** | 3.12 | **4.36** |
| | Bulyan [17] | 63.37 | 15.16 | **22.53** | 13.46 | **19.56** | 14.91 | **21.85** | 14.54 | **21.52** | 14.88 | **21.69** | 7.13 | **11.60** |
| | Trimmed-mean [59, 66] | 62.90 | 11.62 | **18.88** | 11.20 | **17.02** | 13.14 | **20.89** | 10.09 | **20.95** | 12.53 | **20.34** | 5.54 | **11.08** |
| | Median [59, 66] | 60.13 | 15.23 | **23.58** | 12.80 | **15.98** | 15.05 | **23.00** | 14.38 | **23.34** | 14.49 | **23.56** | 5.44 | **7.67** |
| | AFA [41] | 62.75 | 7.21 | **10.58** | 6.26 | **8.55** | 8.54 | **11.55** | 7.87 | **11.09** | 8.19 | **11.41** | 4.43 | **6.18** |
| CIFAR-100 (ResNet-18) | Multi-krum [10] | 34.89 | 17.68 | **25.62** | 5.33 | **11.09** | 16.62 | **25.53** | 10.69 | **20.23** | 18.49 | **25.22** | 3.29 | **6.02** |
| | Bulyan [17] | 35.21 | 14.28 | **16.61** | 8.15 | **11.67** | 12.58 | **19.11** | 10.36 | **14.93** | 13.72 | **18.62** | 4.65 | **8.75** |
| | Trimmed-mean [59, 66] | 35.26 | 10.01 | **18.49** | 7.85 | **9.41** | 10.60 | **18.20** | 11.17 | **19.62** | 10.93 | **19.34** | 5.70 | **8.19** |
| | Median [59, 66] | 34.79 | 12.41 | **23.59** | 4.97 | **9.71** | 9.83 | **21.18** | 9.68 | **17.93** | 12.37 | **23.90** | 3.10 | **6.84** |
| | AFA [41] | 34.59 | 9.94 | **11.85** | 2.05 | **6.33** | 9.33 | **13.70** | 8.12 | **13.38** | 10.13 | **13.93** | 1.59 | **3.67** |
| MNIST (FC) | Multi-krum [10] | 97.02 | 1.59 | **2.06** | 0.26 | **0.96** | 1.51 | **2.32** | 1.47 | **2.25** | 1.49 | **2.30** | 0.04 | **0.72** |
| | Bulyan [17] | 97.21 | 1.36 | **1.88** | 0.84 | **1.18** | 1.32 | **2.14** | 1.23 | **2.06** | 1.28 | **2.09** | 0.34 | **1.02** |
| | Trimmed-mean [59, 66] | 97.24 | 1.49 | **2.05** | 0.24 | **0.93** | 1.35 | **2.28** | 1.35 | **2.23** | 1.32 | **2.27** | 0.08 | **0.62** |
| | Median [59, 66] | 96.93 | 1.51 | **2.03** | 0.31 | **1.00** | 1.31 | **2.15** | 1.25 | **2.12** | 1.27 | **2.16** | 0.08 | **0.57** |
| | AFA [41] | 97.20 | 1.27 | **1.70** | 0.13 | **0.89** | 1.28 | **2.06** | 1.28 | **2.08** | 1.29 | **2.10** | 0.02 | **0.52** |
| Fashion MNIST (AlexNet) | Multi-krum [10] | 83.24 | 5.97 | **11.05** | 3.51 | **6.30** | 5.06 | **15.05** | 4.64 | **12.10** | 5.80 | **15.37** | 2.08 | **2.69** |
| | Bulyan [17] | 83.12 | 7.79 | **20.58** | 3.95 | **7.42** | 6.80 | **13.24** | 5.51 | **12.88** | 7.95 | **20.34** | 1.62 | **3.97** |
| | Trimmed-mean [59, 66] | 83.53 | 6.10 | **9.39** | 4.46 | **11.62** | 5.21 | **8.75** | 4.93 | **8.57** | 6.02 | **11.77** | 2.66 | **3.42** |
| | Median [59, 66] | 81.81 | 5.34 | **8.88** | 5.84 | **10.65** | 4.27 | **8.25** | 4.14 | **8.72** | 5.49 | **9.21** | 1.23 | **2.66** |
| | AFA [41] | 83.97 | 4.04 | **6.46** | 2.96 | **5.09** | 4.91 | **9.49** | 3.62 | **7.57** | 4.86 | **9.30** | 2.26 | **3.91** |
| Shakespeare (LSTM) | Multi-krum [10] | 47.14 | 9.65 | **11.94** | 2.65 | **4.73** | 8.80 | **11.75** | 8.08 | **11.07** | 8.23 | **11.29** | 1.68 | **3.34** |
| | Bulyan [17] | 46.52 | 10.38 | **13.71** | 1.63 | **3.48** | 8.25 | **12.14** | 7.71 | **11.50** | 7.99 | **11.60** | 1.22 | **2.69** |
| | Trimmed-mean [59, 66] | 46.93 | 9.03 | **12.18** | 2.23 | **3.98** | 8.26 | **11.12** | 7.92 | **10.76** | 8.04 | **10.98** | 1.53 | **3.26** |
| | Median [59, 66] | 45.76 | 9.09 | **11.53** | 1.37 | **3.16** | 7.45 | **10.38** | 7.05 | **9.96** | 7.25 | **9.52** | 1.05 | **2.44** |
| | AFA [41] | 47.41 | 7.19 | **10.14** | 4.09 | **5.50** | 8.58 | **10.98** | 8.47 | **9.91** | 8.36 | **9.68** | 1.43 | **2.97** |

**Table 2: The attack impact for state-of-the-art model poisoning attack $\mathcal{A}$ and the corresponding CLP aware attack $\mathcal{A}$-CLP under various threats using non-IID partitioned datasets when *benign gradients are unknown* to attack $\mathcal{A}$.**

- ***Attack Impact:***

$$\frac{\textit{pure accuracy} - \textit{attack accuracy}}{\textit{pure accuracy}} \times 100\%$$

- $\mathcal{A}$-CLP improves effectiveness by up to **6.85x** compared to traditional attacks

- Achieves a greater impact while using a **smaller** attack budget

# Resilience Against Defenses

RAiD 2024

| Dataset (Model) | Defense | Fang Trad. | Fang CL | Min-Max Trad. | Min-Max CL | Min-Sum Trad. | Min-Sum CL | MPHM Trad. | MPHM CL |
|---|---|---|---|---|---|---|---|---|---|
| CIFAR-10 (AlexNet) | FLTrust | 4.74 | **10.35** | 5.21 | **12.79** | 5.57 | **11.76** | 6.13 | **13.26** |
| | SparseFed | 6.84 | **11.73** | 6.32 | **12.61** | 6.22 | **12.46** | 7.03 | **12.68** |
| | cosDefense | 5.71 | **11.40** | 6.35 | **13.47** | 5.96 | **12.56** | 6.15 | **13.22** |
| | FLAIR | 6.57 | **12.53** | 7.27 | **13.81** | 8.03 | **14.13** | 7.53 | **14.01** |
| | LeadFL | 6.31 | **10.06** | 5.12 | **9.96** | 6.20 | **11.49** | 6.36 | **11.60** |
| CIFAR-10 (VGG-11) | FLTrust | 1.57 | **2.85** | 1.30 | **3.23** | 1.74 | **2.25** | 1.57 | **2.77** |
| | SparseFed | 2.71 | **4.63** | 2.60 | **4.42** | 2.57 | **4.02** | 3.01 | **4.88** |
| | cosDefense | 3.07 | **4.52** | 2.42 | **4.64** | 3.41 | **4.55** | 3.43 | **4.74** |
| | FLAIR | 2.86 | **4.25** | 3.05 | **4.96** | 4.01 | **5.04** | 3.76 | **4.79** |
| | LeadFL | 1.31 | **2.43** | 0.88 | **2.65** | 0.73 | **2.15** | 1.18 | **2.86** |
| CIFAR-100 (ResNet-18) | FLTrust | 3.10 | **4.49** | 2.45 | **5.19** | 2.43 | **5.47** | 2.99 | **5.73** |
| | SparseFed | 3.32 | **6.25** | 2.64 | **5.01** | 2.95 | **5.39** | 3.21 | **5.57** |
| | cosDefense | 3.39 | **5.42** | 4.51 | **5.56** | 3.62 | **5.85** | 5.39 | **6.45** |
| | FLAIR | 3.38 | **4.97** | 4.94 | **6.25** | 5.20 | **5.96** | 6.25 | **7.10** |
| | LeadFL | 1.85 | **3.55** | 0.97 | **3.95** | 0.82 | **3.72** | 2.19 | **4.34** |
| MNIST (FC) | FLTrust | 1.33 | **1.66** | 1.37 | **2.10** | 1.31 | **2.03** | 1.39 | **2.16** |
| | SparseFed | 1.22 | **1.65** | 1.12 | **1.81** | 1.53 | **1.79** | 1.60 | **1.84** |
| | cosDefense | 1.09 | **1.78** | 1.48 | **2.05** | 1.20 | **1.94** | 1.37 | **2.00** |
| | FLAIR | 1.28 | **1.58** | 1.01 | **1.93** | 1.24 | **2.07** | 1.18 | **2.05** |
| | LeadFL | 1.23 | **1.63** | 1.07 | **1.99** | 1.19 | **2.08** | 1.30 | **2.03** |
| Fashion MNIST (AlexNet) | FLTrust | 2.93 | **5.80** | 3.28 | **6.82** | 3.82 | **7.63** | 3.60 | **7.21** |
| | SparseFed | 3.25 | **5.35** | 2.56 | **4.41** | 2.51 | **5.39** | 3.06 | **4.76** |
| | cosDefense | 3.12 | **7.45** | 3.36 | **7.85** | 2.99 | **6.77** | 3.61 | **8.09** |
| | FLAIR | 3.48 | **7.32** | 3.72 | **7.58** | 3.92 | **7.86** | 4.17 | **8.13** |
| | LeadFL | 3.34 | **5.50** | 3.41 | **5.85** | 2.89 | **5.12** | 3.34 | **5.91** |
| Shakespeare (LSTM) | FLTrust | 4.43 | **5.58** | 5.41 | **7.24** | 5.74 | **7.05** | 5.64 | **7.38** |
| | SparseFed | 5.20 | **7.25** | 6.04 | **8.22** | 6.38 | **8.54** | 6.34 | **8.81** |
| | cosDefense | 5.31 | **6.94** | 5.35 | **7.78** | 6.05 | **7.58** | 5.78 | **7.74** |
| | FLAIR | 6.08 | **7.20** | 5.97 | **6.96** | 6.94 | **7.75** | 5.87 | **7.09** |
| | LeadFL | 4.05 | **6.57** | 5.12 | **7.89** | 4.39 | **7.41** | 4.78 | **8.16** |

**Table 4: Attack impacts of $\mathcal{A}$ and $\mathcal{A}$-CLP defended by FLTrust, SparseFed, cosDefense, FLAIR and LeadFL.**

- ***Attack Impact*:**

$$\frac{\textit{pure accuracy} - \textit{attack accuracy}}{\textit{pure accuracy}} \times 100\%$$

- Demonstrates **stronger resistance** against defenses (e.g., FLTrust, SparseFed), enhancing attack success by up to 2x

- Properly leveraging CLP with adaptive client selection significantly **boosts** the attack's **performance**

# III. Design of GraSP

# Design Motivation



- **High Complexity of Existing Attacks**: Traditional model poisoning attacks are computationally intensive and complex

- **CLP Vulnerabilities**: Small gradient errors during CLP have a lasting impact, providing a window for more effective attacks

- **Need for Difference**: Current methods + $\mathcal{A}$-CLP produce very similar malicious updates, easily to be detected

- **GraSP's Goal**: Introduce a lightweight, similarity-based attack that maximizes impact with minimal computational effort, targeting the most vulnerable training phases

# Design Details



- **Malicious local model** is calculated as:

$$\tilde{\mathbf{w}}_i(t) := \mathbf{w}_i(t) - \eta\lambda_i \mathbf{s}_t,$$

$\mathbf{w}$ represents the model parameters

$\eta$ is the learning rate

$\mathbf{s}_t$ is the update direction at the $t$-th training round, which is estimated by using received local updates.

**Model poisoning makes the model in the opposite direction of current update.**

# Design Details



LEMMA 1. *Suppose that $\lambda_i$ is the changing direction to craft malicious gradient of the malicious client $i$, $\forall i = 1, \cdots, m^{CLP}$. Then for any given attack threshold $\tau$, the value of $\lambda_i$ satisfies*

$$\lambda_i = \frac{\langle \mathbf{g}(t), \mathbf{g}_i(t) \rangle - \tau \|\mathbf{g}(t)\| \|\tilde{\mathbf{g}}(t)\|}{\mathbf{g}(t)^\top \mathbf{s}(t)}, \ \forall i = 1, \cdots, m^{CLP}. \tag{9}$$

$\mathbf{g}(t)$ is the estimated global update

$\tilde{\mathbf{g}}(t)$ is the **calculated targeted malicious global update**

$\mathbf{g}_i(t)$ represents the update of client $i$.

- Each client uses a **unique $\lambda_i$** to manipulate its local update. By **coordinating** their efforts, malicious clients make the attack harder to detect by defenses like FLTrust

# Design Details

RAiD 2024

- The **targeted malicious global update** is calculated as:

$$\tilde{\mathbf{g}}(t) = \mathbf{g}(t) + \lambda \mathbf{s}(t)$$

where $\mathbf{g}(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{g}_i(t)$, $\lambda$ is solved by using the below proposition.

Proposition 1. *Suppose that $\lambda$ is the changing direction to craft gradients of $m^{CLP}$ malicious clients based on the cosine similarity. For any given attack threshold $\tau$, the value of $\lambda$ is*

$$\lambda = \frac{-z - \sqrt{z^2 - 4xy}}{2x}, \tag{8}$$

*where $x = (\mathbf{g}(t)^\top \mathbf{s}(t))^2 - \tau^2 \|\mathbf{g}(t)\|^2 \cdot \|\mathbf{s}(t)\|^2$, $y = (1 - \tau^2) \cdot \|\mathbf{g}(t)\|^4$, and $z = 2(\tau^2 - 1)\|\mathbf{g}(t)\|^2 \cdot \mathbf{g}(t)^\top \mathbf{s}(t)$.*

- $\tau$ is the attack degree predefined by the attacker; in this work, it is set to 0.1.

# Compared With $\mathcal{A}$-CLP



| Dataset (Model) | Attack | FLTrust | SparseFed | cosDefense | FLAIR | LeadFL |
|---|---|---|---|---|---|---|
| CIFAR-10 (AlexNet) | Best $\mathcal{A}^*$-CLP | 13.26 | 12.68 | 13.47 | 14.13 | 11.60 |
| | GraSP | **13.98** | **14.21** | **14.98** | **15.14** | **13.33** |
| CIFAR-10 (VGG-11) | Best $\mathcal{A}^*$-CLP | 3.23 | 4.88 | 4.74 | 5.04 | 2.86 |
| | GraSP | **4.68** | **5.60** | **6.42** | **5.89** | **3.98** |
| CIFAR-100 (ResNet-18) | Best $\mathcal{A}^*$-CLP | 5.73 | 6.25 | 6.45 | 7.10 | 4.34 |
| | GraSP | **6.83** | **7.33** | **8.67** | **7.10** | **6.61** |
| MNIST (FC) | Best $\mathcal{A}^*$-CLP | 2.16 | 1.84 | 2.05 | 2.07 | 2.08 |
| | GraSP | **2.50** | **2.35** | **3.34** | **2.88** | **2.66** |
| F. MNIST (AlexNet) | Best $\mathcal{A}^*$-CLP | 7.63 | 5.39 | 8.09 | 8.13 | 5.91 |
| | GraSP | **7.95** | **6.74** | **9.42** | **8.57** | **7.19** |
| Shakespeare (LSTM) | Best $\mathcal{A}^*$-CLP | 7.38 | 8.81 | 7.78 | 7.75 | 8.16 |
| | GraSP | **8.23** | **10.23** | **9.84** | **8.59** | **9.40** |

**Table 6: Attack impacts of GraSP and $\mathcal{A}^*$-CLP when defended by FLTrust, SparseFed, cosDefense, FLAIR and LeadFL under various threat models using non-IID partitioned datasets.**

- Both demonstrate significantly **improved resilience** against advanced defenses (e.g., FLTrust, SparseFed, FLAIR).
- **GraSP** outperforms $\mathcal{A}$-CLP in most of scenarios, indicating its **strong adaptability** to defensive measures
- The similarity-based approach in GraSP leads to **sustained attack success**, especially during critical learning periods.

# Analysis of GraSP



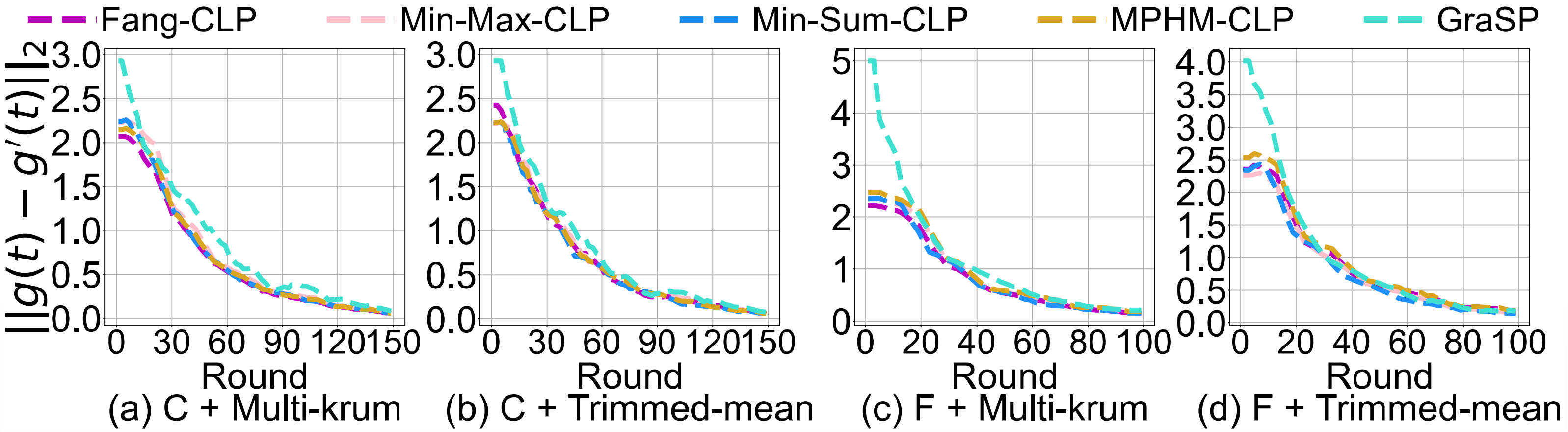



**Figure 16: The $\ell_2$-norm of gradient magnitude different of CLP augmented attacks when *benign gradients are unknown* to the adversary, where "C" and "F" stands for CIFAR-10 and Fashion-MNIST datasets, respectively.**

- ❑ GraSP shows **higher gradient magnitudes** during Critical Learning Periods (CLP), resulting in a greater impact on the global model

- ❑ Uses cosine similarity to target directions that most disrupt model updates, enabling **faster** and more **effective** gradient adjustments

# IV. Conclusions

# Defense Against Proposed Methods



## Defending Against $\mathcal{A}$-CLP:

- **Simple Approach**: Increase the number of participating clients during CLP to dilute the impact of malicious updates
- **Limitation**: Results in high communication costs and an increased attack budget

## Defending Against GraSP:

- **Layer-Based Similarity**: Calculate gradient similarity across clients at specific layers to identify anomalies
- **Anomaly Detection**: Inspired by methods like AFA and cosDefense, potential malicious clients are excluded from model aggregation during CLP for stricter protection

# Conclusions



## Key Contributions

- Proposed $\mathcal{A}$ **-CLP** and **GraSP**, adaptive attacks leveraging Critical Learning Periods (CLP) for greater impact
- Introduced the **FGN metric** for efficient, privacy-preserving CLP detection

## Main Findings

- $\mathcal{A}$**-CLP** enhances attack success, achieving up to 6.85x more impact than traditional methods by adjusting malicious client numbers
- **GraSP** uses similarity-based strategies for effective gradient deviations with lower computational costs, outperforming current attacks

**GitHub Repo:**
https://github.com/GYan58/RAID-2024-CLP

**IntelliSys Lab**
https://intellisys.haow.us