Split to Spot: Multi-Reference Byzantine Detection via Gradient Splitting in Federated Learning

Zihan Yu[1], Zukang Ai[2], Yilin Su[1], Dan Yu[2*], Yongle Chen[1], Jianhua Wang[2]
[1] *College of Artificial Intelligence, Taiyuan University of Technology*, Taiyuan, China
yuy63434@gmail.com; 2766733506@qq.com; chenyongle@tyut.edu.cn
[2] *College of Computer Science and Technology, Taiyuan University of Technology*, Taiyuan, China
2906084374@qq.com; yudan@tyut.edu.cn; wangjianhua02@tyut.edu.cn

***Abstract*—Federated learning under non-independent and identically distributed (non-IID) data settings faces severe Byzantine security challenges. To mitigate the performance degradation of robust AGgregation Rules(AGRs) caused by gradient heterogeneity and the curse of dimensionality, a GrAdient Splitting (GAS) mechanism has been proposed. However, existing GAS method typically relies on a single reference when identifying Byzantine clients, making their defense performance sensitive to reference selection. This sensitivity arises because different attack types often require different reference choices to achieve satisfactory robustness.**

Motivated by this observation, we propose GAS-MR, namely Gradient Splitting with Multi-Reference Consistency-Based Byzantine Detection, which leverages multiple references to quantify the deviation-based inconsistency of Byzantine clients after gradient splitting, thereby enabling robust Byzantine client identification. Experimental results demonstrate that the proposed method consistently outperforms GAS under various non-IID data distributions and five representative Byzantine attack scenarios. Even in cases where performance is comparable, our approach exhibits superior robustness and stability, thereby significantly reducing the dependence on reference selection and prior knowledge of attack strategies Code is available at https://github.com/SKLIIS-AIS/GAS-MR.git.

***Index Terms*—federated learning, byzantine robustness, non-iid data, multi-reference consistency**

I. Introduction

Federated Learning enables collaborative model training across multiple clients without sharing their local raw data, providing an effective paradigm for distributed learning in privacy-sensitive scenarios [1]. However, the decentralized nature of federated learning also exposes the training process to severe security threats, particularly Byzantine attacks. In such attacks, malicious clients can arbitrarily manipulate their uploaded gradients or model updates to steer the global model toward undesired directions, severely degrading overall performance [2] [3] [4]. Since the central server has neither access to local client data nor visibility into client-side training procedures, designing robust aggregation mechanisms under untrusted environments has become a central challenge in federated learning [5].

To defend against Byzantine attacks, numerous robust robust AGRS [6], such as Multi-Krum [7], Median [7],Bulyan [8] and RFA [9], have been proposed. These methods can suppress the influence of anomalous updates to some extent, but they remain limited in practice. First, most existing robust AGRs are sensitive to the curse of dimensionality [8] in high-dimensional gradient spaces , which allows malicious gradients to be hidden and evade detection. Second, in non-independent and identically distributed (non-IID) data settings, the heterogeneity of client data induces gradient heterogeneity, further increasing the difficulty of distinguishing malicious updates from benign ones [10] [11].

To address these challenges, GAS [12] was proposed. GAS first partitions high-dimensional gradients into multiple low-dimensional sub-vectors. Within each subspace, it applies aggregation result of a specific robust AGR as an honest reference to detect Byzantine updates.

Despite its promising adaptability to non-IID scenarios, Existing GAS implementations typically rely on a single robust AGR reference to identify Byzantine updates. This design introduces **inherent limitation** in practice,which we call reference sensitivity. Systematic experimental analyses reveal that single-reference GAS exhibits varying effectiveness across different Byzantine attack scenarios. A reference constructed from a specific AGR may perform well under one or two attack settings, but fails to maintain comparable robustness under other attack types. In particular, no single reference consistently achieves stable or optimal defense performance across all attack scenarios. This indicates that the defensive effectiveness of GAS is highly sensitive to the choice of reference, and implicitly coupled with the underlying attack strategy. Moreover, the aggregation results produced by robust AGRs do not necessarily correspond to reliable references for identifying anomalous updates, revealing a fundamental role mismatch between aggregation and detection.

Motivation: In practical federated learning systems, the server typically lacks prior knowledge of the attack type, while adversarial behaviors may evolve dynamically during training. Under such conditions, a fixed reference is often inadequate, as its effectiveness is highly sensitive to the underlying attack

*Corresponding author: Dan Yu(yudan@tyut.edu.cn). All authors are affiliated with the Shanxi Key Laboratory of Industrial Internet Security, Taiyuan University of Technology.

This work was supported in part by the Central Government Guiding Local Scientific and Technological Development Fund of Shanxi Province under Grant YDZJSX2024C003 and YDZJSX2025D029.

patterns. Consequently, to enhance the stability and robustness of defense performance in the presence of evolving attacks, it is necessary to revisit the design of the reference mechanism.

Based on this observation, we propose GAS-MR. GAS-MR introduces multiple complementary reference updates and identifies malicious clients by modeling the consistency of their deviations across different references. The key insight is that benign updates tend to exhibit consistent deviation patterns across multiple references, whereas Byzantine updates are unlikely to align with all references simultaneously. By design, GAS-MR is attack-agnostic and targets realistic federated learning scenarios where attack priors are unavailable or dynamically evolving, aiming to achieve both robustness and generalization.

In summary, the main contributions of this work are as follows:

- We propose **GAS-MR**, a novel *multi-reference consistency-based* Byzantine detection mechanism built upon GAS. GAS-MR explicitly reformulates reference construction from a single fixed aggregation result to a set of complementary references, and detects Byzantine updates by exploiting the consistency of client deviations across references, thereby fundamentally alleviating the instability inherent in single-reference GAS.
- We conduct extensive experiments under multiple non-IID data distributions and five representative Byzantine attack scenarios. The results demonstrate that GAS-MR consistently outperforms single-reference GAS, and even when achieving comparable accuracy, exhibits markedly improved robustness and stability, substantially reducing sensitivity to reference selection and reliance on prior knowledge of attack strategies.

II. Notations and Preliminaries

Notations. Let $[n] = \{1, \dots, n\}$ denote the set of the first n positive integers, and let $|S|$ denote the cardinality of a set S. For a model vector $w \in \mathbb{R}^d$, let $\|w\|$ denote its ℓ_2 norm, and let $w[j]$ denote its j-th component. For an index set $J \subseteq [d]$, $w[J]$ denotes the sub-vector of w containing components indexed by J. For a scalar $x \in \mathbb{R}$, $|x|$ denotes its absolute value. For a random variable X, $\mathbb{E}[X]$ denotes its expectation.

Federated Learning Setup. We consider a standard federated learning system with a central server S and n clients $\{c_i\}_{i\in[n]}$ following [3], [13]. The objective is to minimize the global loss function

$$L(w) = \frac{1}{n}\sum_{i=1}^{n} L_i(w), \tag{1}$$

where

$$L_i(w) = \mathbb{E}_{\xi_i}[L(w;\xi_i)], \quad i \in [n]. \tag{2}$$

Here, w denotes the model parameters, L_i is the local loss function of client c_i, and ξ_i represents its local data distribution, which is non-IID across clients.

At communication round t, the server broadcasts the global model w_t to all clients. Each client c_i performs local training on local data to obtain updated parameters w_i^t and computes the local gradient

$$g_i^t = w^t - w_i^t. \tag{3}$$

The server then aggregates the received gradients to update the global model as

$$w^{t+1} = w^t - g^t, \quad g^t = \frac{1}{n}\sum_{i=1}^{n} g_i^t. \tag{4}$$

This process is repeated for T communication rounds, yielding the final global model w^T.

Byzantine Threat Model. Among n clients, f clients are Byzantine. Let $B \subseteq [n]$ denote the set of Byzantine clients and $H = [n] \setminus B$ denote the honest clients. In the t-th round, the update uploaded from client i is

$$g_i^t = \begin{cases} w^t - w_i^{t+1}, & i \in H, \\ *, & i \in B, \end{cases} \tag{5}$$

where $*$ represents an arbitrary value.

Robust AGRs. To defend against Byzantine attacks, most existing methods replace the averaging step with robust AGRs :

$$w^{t+1} = w^t - g_{\text{agg}}^t, \quad g_{\text{agg}}^t = A(g_1^t, \dots, g_n^t) \tag{6}$$

where g_{agg}^t is the aggregated gradient, and A can be any robust AGR such as Minmax or Minsum [6]. For simplicity, we omit the superscript t when the round index is clear from context.

Reference (r): a gradient used to distinguish honest and Byzantine client updates. It is computed by splitting the gradients of multiple clients $\{g_1, \dots, g_n\}$ into p segments each:

$$g_i \to \{g_i^{(1)}, \dots, g_i^{(p)}\}, \quad i = 1, \dots, n. \tag{7}$$

where p denotes the number of segments . Then, a robust AGR A is applied to the corresponding segments across clients to obtain

$$r_j = A\big(\{g_1^{(j)}, \dots, g_n^{(j)}\}\big), \quad j = 1, \dots, p. \tag{8}$$

which serves as the *honest reference* for evaluating client corresponding segment.

III. Method

Our method is motivated by the intuition that references should expose anomalies introduced by Byzantine clients, while remaining stable and effective under diverse attack scenarios. However, a robust aggregation result does not necessarily constitute a reliable reference, as robustness in aggregation does not imply effectiveness in identifying Byzantine anomalies.

Based on this insight, we construct three references using the **mean**, **median**, and **trimmed mean**, respectively.These references capture distinct characteristics of the client update distribution: the mean reflects the global trend, the median provides strong resistance to extreme outliers, and the trimmed

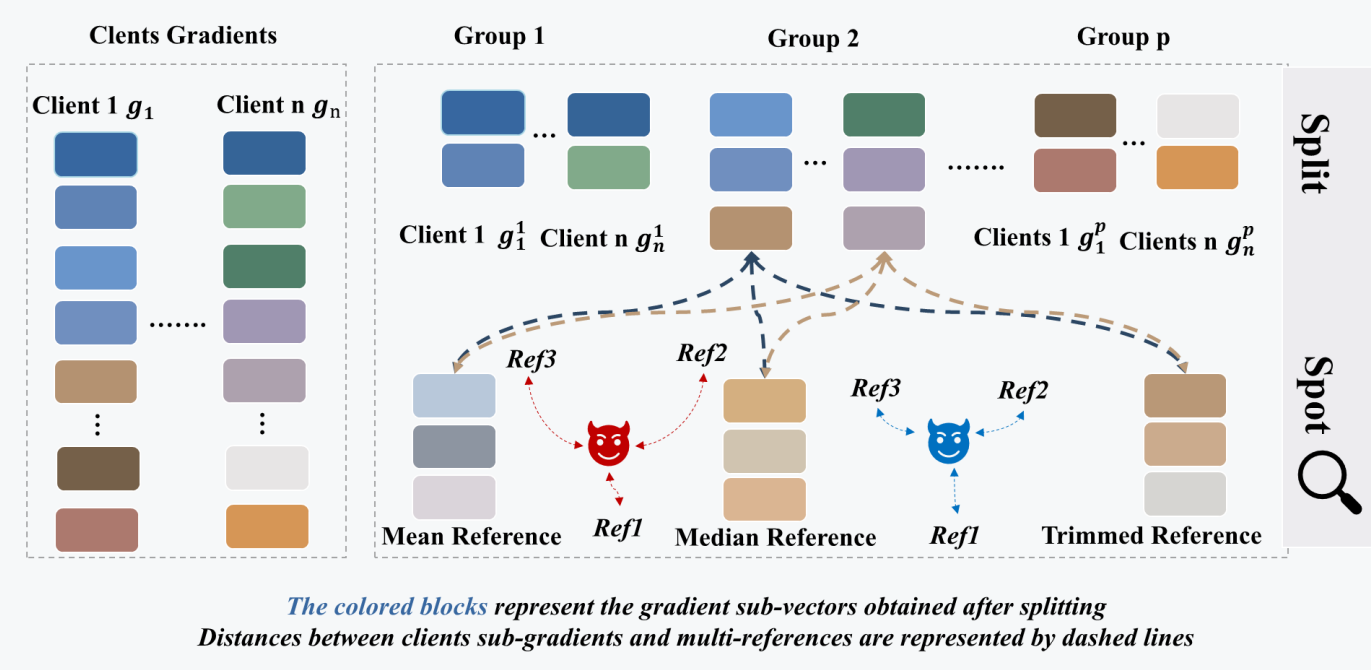

Fig. 1. Framework of GAS-MR

mean suppresses moderately corrupted updates. By jointly leveraging these three complementary references, GAS-MR obtains multiple robustness-oriented views of client deviations, enabling more reliable and stable Byzantine identification through cross-reference consistency.

GAS-MR consists of three steps: **Splitting**, **Identification**, and **Aggregation**.

Step 1: Splitting. Each client gradient $g_i \in \mathbb{R}^d$ is split into p groups according to a random partition $\{J_1, \ldots, J_p\}$ of the gradient dimensions:

$$g_i = \left(g_i^1, g_i^2, \ldots, g_i^p\right), \quad g_i^q = [\, g_i \,]_{J_q}, \quad i \in [n],\ q \in [p] \tag{9}$$

Step 2: Identification with Three References. For each sub-vector index $q \in [p]$, corresponding to the dimension set J_q, we compute three reference gradients based on $\{g_i^q\}_{i=1}^n$: the mean, median, and trimmed mean:

$$\hat{g}_{\text{mean}}^q = \frac{1}{n}\sum_{i=1}^{n} g_i^q, \tag{10}$$

$$\hat{g}_{\text{median}}^q = \text{median}\left(\{g_i^q\}_{i=1}^n\right), \tag{11}$$

$$\hat{g}_{\text{trim}}^q = \frac{1}{n-2k}\sum_{i=k+1}^{n-k} g_i^q. \tag{12}$$

where g_i^q denotes the gradient element of the i-th client in the q-th group; for each dimension, the elements at the same position across clients are sorted prior to computing the median and trimmed mean. The trimming parameter k is constrained as

$$k = \min\left(\max(1, \lfloor 0.1n \rfloor), \left\lfloor \tfrac{n-1}{2} \right\rfloor\right), \tag{13}$$

which ensures that at least 10% of extreme values are trimmed (*lower bound*) while not removing more than half of the clients (*upper bound*), providing both robustness and theoretical safety under the Byzantine threat model.

Next, for each client i, we compute the distances to the three references:

$$d_{\text{mean},i}^q = \|g_i^q - \hat{g}_{\text{mean}}^q\|, \tag{14}$$

$$d_{\text{median},i}^q = \|g_i^q - \hat{g}_{\text{median}}^q\|, \tag{15}$$

$$d_{\text{trim},i}^q = \|g_i^q - \hat{g}_{\text{trim}}^q\|. \tag{16}$$

The *identification score* of client i for group q is defined as the maximum absolute pairwise disagreement between references:

$$s_i^q = \max\left(|d_{\text{mean,i}}^q - d_{\text{median,i}}^q|, |d_{\text{mean,i}}^q - d_{\text{trim,i}}^q|, |d_{\text{median,i}}^q - d_{\text{trim,i}}^q|\right), \tag{17}$$

The maximum operator is used to capture the most pronounced inconsistency among different references. If a client update deviates significantly from any one of the reference perspectives, the resulting large discrepancy is sufficient to indicate potential Byzantine behavior, making the identification sensitive to worst-case abnormality while remaining robust to minor statistical fluctuations.

and the overall identification score is the sum over all groups:

$$s_i = \sum_{q=1}^{p} s_i^q. \tag{18}$$

The top $n - f$ clients with the lowest scores are selected as honest for aggregation.

Step 3: Aggregation. The final aggregation is computed as the average of the selected honest gradients:

$$\hat{g} = \frac{1}{n-f} \sum_{i \in \mathcal{I}} g_i, \tag{19}$$

where $\mathcal{I}$ denotes the index set of selected clients.

Algorithm 1: Gradient Splitting with Multi-Reference Consistency-Based Byzantine Detection

Input: Client gradients $\{g_i \in \mathbb{R}^d\}_{i=1}^n$, number of splits p
Output: Aggregated gradient $\hat{g}$

Splitting:;
Randomly partition $[d]$ into p disjoint subsets $\{J_1, \ldots, J_p\}$ with $|J_q| \leq \lceil d/p \rceil$;
for $i = 1$ **to** n **do**
  for $q = 1$ **to** p **do**
    $g_i^q \leftarrow [g_i]_{J_q}$;
Identification (Multi-Reference):;
Initialize $s_i \leftarrow 0$ for $i = 1, \ldots, n$;
for $q = 1$ **to** p **do**
  Group construction:;
  $G^q \leftarrow \{g_i^q \mid i = 1, \ldots, n\}$;
  Compute references on G^q:;
    $ref\_mean \leftarrow \text{mean}(G^q)$;
    $ref\_median \leftarrow \text{median}(G^q)$;
    $ref\_trim \leftarrow \text{trimmed-mean}(G^q)$;
  for *each client* c_i *in* $G^{(q)}$ **do**
    Compute distances:;
      $d\_mean \leftarrow \|g_i^q - ref\_mean\|$;
      $d\_median \leftarrow \|g_i^q - ref\_median\|$;
      $d\_trim \leftarrow \|g_i^q - ref\_trim\|$;
    Cross-reference disagreement score:;
      $s_i \leftarrow s_i + \max\Big(|d\_mean - d\_median|, |d\_mean - d\_trim|, |d\_trim - d\_median|\Big)$;
Aggregation:;
Select index set $\mathcal{I}$ of $n - f$ clients with smallest s_i;
$\hat{g} \leftarrow \frac{1}{n-f} \sum_{i \in \mathcal{I}} g_i$;

IV. Experiments

A. Experiment Settings

Datasets. Our experiments are conducted on three widely-used benchmark datasets: CIFAR-10 [3], CIFAR-100 [14], and FEMNIST [15].

Data distribution. For CIFAR-10 and CIFAR-100, we follow prior works [16] and use the Dirichlet distribution to generate non-IID data partitions among clients. Unless otherwise specified, we set the number of clients to $n = 50$ and the concentration parameter of the Dirichlet distribution to $\beta = 0.5$. FEMNIST is a dataset with a natural non-IID partition. Specifically, the data is partitioned into 3,597 clients according to the writer of each digit or character. For each client, we randomly split its local data into training and testing sets, using 90% of the samples for training and the remaining 10% for testing, following [15].

Attacks. We evaluate the robustness of our method against five representative Byzantine attacks: BitFlip [17], LIE [18], Minmax and Minsum [6], and IPM [19].

Baselines. We consider six representative robust AGRs: Multi-Krum [7], Bulyan [8], Coordinate-wise Median [7], RFA [9], DnC [6], and RBTM [20]. Each AGR is combined with the GAS framework, resulting in GAS (Multi-Krum), GAS (Bulyan), GAS (Median), GAS (RFA), GAS (DnC), and GAS (RBTM), respectively. We compare our proposed GAS-MR with them to evaluate its performance under different Byzantine attack scenarios.

Evaluation. We use top-1 accuracy, i.e., the proportion of correctly predicted testing samples to total testing samples, to evaluate the performance of global models.

Other settings. We utilize AlexNet [21], SqueezeNet [22], and a four-layer CNN [15] for CIFAR-10, CIFAR-100, and FEMNIST, respectively. The number of Byzantine clients for all datasets is set to $f = 0.2 \cdot n$. We further consider a higher Byzantine ratio of up to $f = 0.3 \cdot n$ in the ablation study. Experimental results are presented in Tables I–V.

B. Results Analysis

Robustness. GAS-MR demonstrates strong robustness against various Byzantine attacks. On the CIFAR-10 dataset (Table I), GAS-MR outperforms the original GAS under four attack types, exceeding even the best single-reference performance, while slightly underperforming under the Min-Max attack. Notably, under the IPM attack, GAS-MR achieves an accuracy improvement of approximately 12% compared to the optimal single-reference GAS. Furthermore, fixing the total number of clients at $n = 50$, GAS-MR's performance remains largely unaffected by the number of Byzantine clients (Table II): for $f = 5$, the accuracy is 65.73%, and for $f = 15$, it slightly drops to 65.11%, a decrease of less than 0.62%, indicating high robustness even as attack intensity increases.

Stability. GAS-MR also exhibits high stability under different experimental conditions. First, under varying non-IID levels ($\beta = 0.3$ and 0.7) for the LIE attack on CIFAR-10 (Table III), GAS-MR consistently achieves high accuracy and outperforms the best single-reference GAS. The improvement is more pronounced under higher non-IID conditions ($\beta = 0.3$), demonstrating low sensitivity to data heterogeneity. Second, varying the total number of clients ($n = 75, 100$) (Table IV) does not significantly affect GAS-MR's performance, indicating reliable performance under scale changes. Finally, ablation studies (Table V) confirm that single statistical references (mean, median, or trimmed mean) perform poorly under most attacks, whereas GAS-MR leverages complementary information from all three references to accurately identify Byzantine gradients and maintain consistent performance across attack scenarios.

TABLE I
ACCURACY (%) UNDER DIFFERENT BYZANTINE ATTACKS

Dataset	Attack	BitFlip	LIE	Min-Max	Min-Sum	IPM
CIFAR-10	GAS (Multi-Krum)	63.13 ± 0.67	51.47 ± 0.81	51.31 ± 0.93	52.31 ± 0.74	**51.45 ± 0.88**
	GAS (Bulyan)	42.98 ± 0.84	54.30 ± 0.77	54.00 ± 0.95	53.38 ± 1.06	29.96 ± 1.21
	GAS (Median)	62.66 ± 0.58	49.21 ± 0.86	49.77 ± 0.69	49.77 ± 0.91	50.56 ± 0.72
	GAS (RFA)	63.28 ± 0.61	52.74 ± 0.79	49.13 ± 0.97	**56.79 ± 0.88**	18.37 ± 1.34
	GAS (DnC)	63.39 ± 0.73	**61.24 ± 0.62**	**61.51 ± 0.54**	55.92 ± 0.83	50.66 ± 0.76
	GAS (RBTM)	**63.65 ± 0.64**	47.73 ± 0.92	48.56 ± 0.81	52.15 ± 0.68	51.05 ± 0.85
	GAS-MR	**65.20 ± 0.39**	**65.87 ± 0.44**	**58.63 ± 0.52**	**61.07 ± 0.47**	**63.90 ± 0.41**
CIFAR-100	GAS (Multi-Krum)	28.81 ± 0.92	26.26 ± 0.83	22.37 ± 0.71	25.19 ± 0.88	17.66 ± 0.79
	GAS (Bulyan)	25.52 ± 1.04	23.74 ± 0.91	21.76 ± 0.86	21.13 ± 0.97	25.27 ± 1.32
	GAS (Median)	40.15 ± 1.18	20.62 ± 0.89	22.54 ± 0.77	22.13 ± 0.93	41.39 ± 1.41
	GAS (RFA)	27.56 ± 0.87	30.13 ± 0.95	24.88 ± 0.92	25.03 ± 0.84	13.36 ± 0.81
	GAS (DnC)	34.27 ± 1.01	46.38 ± 1.12	36.16 ± 0.98	35.57 ± 0.94	40.27 ± 1.09
	GAS (RBTM)	39.46 ± 0.96	20.94 ± 0.82	22.69 ± 0.88	21.76 ± 0.79	40.18 ± 1.15
	GAS-MR	**44.46 ± 0.56**	**49.25 ± 0.49**	**43.31 ± 0.58**	**42.24 ± 0.53**	**43.41 ± 0.51**
FEMNIST	GAS (Multi-Krum)	80.36 ± 0.61	73.08 ± 0.58	66.12 ± 0.63	67.26 ± 0.54	80.55 ± 0.60
	GAS (Bulyan)	77.64 ± 0.72	70.67 ± 0.69	62.74 ± 0.66	68.24 ± 0.61	79.18 ± 0.68
	GAS (Median)	77.55 ± 0.64	74.89 ± 0.59	67.22 ± 0.62	70.82 ± 0.57	80.12 ± 0.60
	GAS (RFA)	74.63 ± 0.75	68.28 ± 0.71	68.67 ± 0.69	76.44 ± 0.63	78.19 ± 0.72
	GAS (DnC)	74.70 ± 0.66	80.64 ± 0.58	72.35 ± 0.61	73.53 ± 0.64	79.45 ± 0.59
	GAS (RBTM)	72.75 ± 0.73	72.87 ± 0.76	64.13 ± 0.71	74.41 ± 0.65	78.64 ± 0.74
	GAS-MR	**86.98 ± 0.44**	**85.76 ± 0.41**	**87.26 ± 0.39**	**80.78 ± 0.46**	**87.66 ± 0.42**

TABLE II
ACCURACY (%) WITH DIFFERENT NUMBERS OF BYZANTINE CLIENTS f UNDER LIE ATTACK ON CIFAR-10

f	GAS (Multi-Krum)	GAS (Bulyan)	GAS (Median)	GAS (RFA)	GAS (DnC)	GAS (RBTM)	**GAS-MR**
5	63.71 ± 1.42	61.24 ± 1.13	59.36 ± 1.18	62.77 ± 1.31	65.34 ± 0.94	58.75 ± 1.26	**65.73 ± 0.62**
15	48.75 ± 1.87	47.66 ± 1.35	42.66 ± 1.55	49.23 ± 1.69	61.74 ± 1.08	42.21 ± 1.74	**65.11 ± 0.71**

TABLE III
ACCURACY (%) WITH DIFFERENT non-IID LEVELS β UNDER LIE ATTACK ON CIFAR-10

β	GAS (Multi-Krum)	GAS (Bulyan)	GAS (Median)	GAS (RFA)	GAS (DnC)	GAS (RBTM)	**GAS-MR**
0.3	50.18 ± 0.96	53.15 ± 0.82	45.76 ± 0.88	51.75 ± 0.91	61.01 ± 1.04	45.96 ± 0.79	**63.36 ± 0.54**
0.7	56.84 ± 0.61	54.97 ± 0.48	51.55 ± 0.52	58.35 ± 0.57	63.47 ± 0.66	52.26 ± 0.49	**64.79 ± 0.36**

TABLE IV
ACCURACY(%) WITH DIFFERENT NUMBERS OF CLIENTS n UNDER LIE ATTACK ON CIFAR-10

n	GAS (Multi-Krum)	GAS (Bulyan)	GAS (Median)	GAS (RFA)	GAS (DnC)	GAS (RBTM)	**GAS-MR**
75	51.16 ± 0.72	48.87 ± 0.18	46.83 ± 1.15	53.38 ± 1.36	65.55 ± 0.43	48.36 ± 0.51	**65.29 ± 0.39**
100	48.80 ± 1.05	48.11 ± 0.12	43.46 ± 0.68	56.35 ± 1.28	63.43 ± 0.47	48.14 ± 0.36	**63.86 ± 0.42**

TABLE V
IMPACT OF SINGLE STATISTICAL REFERENCE ON ACCURACY (%) UNDER DIFFERENT BYZANTINE ATTACKS ON CIFAR-10

Attack	BitFlip	LIE	Min-Max	Min-Sum	IPM
GAS (Mean)	62.84 ± 0.85	55.63 ± 0.90	51.61 ± 0.80	51.63 ± 0.88	49.88 ± 1.65
GAS (Median)	62.66 ± 1.63	49.21 ± 0.85	49.77 ± 0.80	49.77 ± 0.88	50.56 ± 0.92
GAS (Trimmed-mean)	63.96 ± 0.88	56.49 ± 0.90	51.26 ± 1.22	51.44 ± 0.87	50.57 ± 0.90
GAS-MR	**65.20 ± 0.39**	**65.87 ± 0.44**	**58.63 ± 0.52**	**61.07 ± 0.47**	**63.90 ± 0.41**

Overall. These results demonstrate that GAS-MR not only exhibits strong **robustness** against various Byzantine attacks but also maintains high **stability** across different data distributions, client numbers, and design variations. This validates the effectiveness and broad applicability of the multi-reference identification strategy.

V. Conclusion

In this paper, we revisited reference-based Byzantine detection in federated learning and identified *reference sensitivity* as an inherent limitation of existing single-reference GAS frameworks. We showed that GAS's effectiveness is tightly coupled with reference choice and no single robust AGR-based reference reliably defends against diverse Byzantine attacks under non-IID data.

To address this, we proposed GAS-MR, By introducing multiple complementary references and evaluating the consistency of client deviations across them, GAS-MR effectively decouples Byzantine detection from any single reference choice. This design enables attack-agnostic and stable identification of malicious clients, even when attack patterns are unknown or dynamically evolving.

Experiments under non-IID settings and representative Byzantine attacks show that GAS-MR consistently outperforms single-reference GAS, with superior robustness and substantially reduced sensitivity to reference selection and prior attack knowledge.

References

[1] M. Hu, Y. Cao, A. Li, Z. Li, C. Liu, T. Li, and Y. Liu, "Fedmut: Generalized federated learning via stochastic mutation," in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, vol. 38, no. 11, 2024, pp. 12 528–12 537.

[2] C. Chen, J. Zhang, A. K. Tung, M. Kankanhalli, and G. Chen, "Robust federated recommendation system," *arXiv preprint*, 2020.

[3] P. Blanchard, E. M. E. Mhamdi, R. Guerraoui, and J. Stainer, "Machine learning with adversaries: Byzantine tolerant gradient descent," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017.

[4] Z. Zhang, L. Wu, C. Ma, J. Li, J. Wang, Q. Wang, and S. Yu, "Lsfl: A lightweight and secure federated learning scheme for edge computing," *IEEE Transactions on Information Forensics and Security*, vol. 18, pp. 365–379, 2022.

[5] Z. Zhang, L. Wu, J. Jin, E. Wang, B. Liu, and Q.-L. Han, "Secure federated learning for cloud-fog automation: Vulnerabilities, challenges, solutions, and future directions," *IEEE Transactions on Industrial Informatics*, 2025.

[6] V. Shejwalkar and A. Houmansadr, "Manipulating the byzantine: Optimizing model poisoning attacks and defenses for federated learning," in *Network and Distributed System Security Symposium (NDSS)*, 2021.

[7] D. Yin, Y. Chen, R. Kannan, and P. Bartlett, "Byzantine-robust distributed learning: Towards optimal statistical rates," in *Proceedings of the 35th International Conference on Machine Learning (ICML)*, ser. Proceedings of Machine Learning Research, vol. 80. PMLR, 2018, pp. 5650–5659.

[8] R. Guerraoui, S. Rouault, S. Farhadkhani, A. Guirguis, and L.-N. Hoang, "The hidden vulnerability of distributed learning in byzantium," in *Proceedings of the 35th International Conference on Machine Learning (ICML)*, ser. Proceedings of Machine Learning Research, vol. 80. PMLR, 2018, pp. 3521–3530.

[9] K. Pillutla, S. M. Kakade, and Z. Harchaoui, "Robust aggregation for federated learning," *arXiv preprint*, 2019.

[10] Z. Su, Z. Lu, Y. Wu, R. Shen, and S. Lu, "Flseg: Enhancing privacy and robustness in federated learning under heterogeneous data via model segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2025, pp. 3916–3925.

[11] Z. Zhang, L. Wu, C. Ma, J. Li, J. Wang, Q. Wang, and S. Yu, "Using third-party auditor to help federated learning: An efficient byzantine-robust federated learning," *IEEE Transactions on Information Forensics and Security*, vol. 19, pp. 1–15, 2024.

[12] Y. Liu, C. Chen, L. Lyu, F. Wu, S. Wu, and G. Chen, "Byzantine-robust learning on heterogeneous data via gradient splitting," in *Proceedings of the 40th International Conference on Machine Learning (ICML)*, ser. Proceedings of Machine Learning Research, vol. 202. PMLR, 2023, pp. 21 404–21 425.

[13] C. Chen, L. Lyu, H. Yu, and G. Chen, "Practical attribute reconstruction attack against federated learning," *IEEE Transactions on Big Data*, 2022.

[14] A. Krizhevsky, V. Nair, and G. Hinton, "Learning multiple layers of features from tiny images," University of Toronto, Tech. Rep., 2009.

[15] S. Caldas, P. Wu, T. Li, J. Konečný, H. B. McMahan, V. Smith, and A. Talwalkar, "Leaf: A benchmark for federated settings," *arXiv preprint*, 2018.

[16] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," in *Proceedings of Machine Learning and Systems (MLSys)*, vol. 2, 2020, pp. 429–450.

[17] Z. Allen-Zhu, F. Ebrahimianghazani, J. Li, and D. Alistarh, "Byzantine-resilient non-convex stochastic gradient descent," in *International Conference on Learning Representations (ICLR)*, 2020.

[18] G. Baruch, M. Baruch, and Y. Goldberg, "A little is enough: Circumventing defenses for distributed learning," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 32, 2019.

[19] C. Xie, O. Koyejo, and I. Gupta, "Fall of empires: Breaking byzantine-tolerant sgd by inner product manipulation," in *Proceedings of the 37th International Conference on Machine Learning (ICML)*, ser. Proceedings of Machine Learning Research. PMLR, 2020.

[20] E. M. El-Mhamdi, S. Farhadkhani, R. Guerraoui, A. Guirguis, L.-N. Hoang, and S. Rouault, "Collaborative learning in the jungle (decentralized, byzantine, heterogeneous, asynchronous and nonconvex learning)," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 34, 2021, pp. 25 044–25 057.

[21] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[22] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "Squeezenet: Alexnet-level accuracy with 50x fewer parameters and ¡0.5mb model size," *arXiv preprint*, 2016.