VFedAD: A Defense Method Based on the Information Mechanism Behind the Vertical Federated Data Poisoning Attack

Jinrong Lai
Sun Yat-sen University
Guangzhou, China
laijr@mail2.sysu.edu.cn

Tong Wang
Sun Yat-sen University
Guangzhou, China
wangt328@mail2.sysu.edu.cn

Chuan Chen*
Sun Yat-sen University
Guangzhou, China
chenchuan@mail.sysu.edu.cn

Yihao Li
Sun Yat-sen University
Guangzhou, China
liyh328@mail2.sysu.edu.cn

Zibin Zheng
Sun Yat-sen University
Guangzhou, China
zhzibin@mail.sysu.edu.cn

ABSTRACT

In recent years, federated learning has achieved remarkable results in the medical and financial fields, but various attacks have always plagued federated learning. Data poisoning attack and defense research in horizontal federated learning are sufficient, yet vertical federated data poisoning attack and defense remains an open area due to two challenges: (1) Complex data distributions lead to immense attack possibilities, and (2) defense methods are insufficient for complex data distributions. We have discovered that from the perspective of information theory, the above challenges can be addressed elegantly and succinctly with a solution. We first reveal the information-theoretic mechanisms underlying vertical federated data poisoning attacks and then propose an unsupervised vertical federated data poisoning defense method (VFedAD) based on information theory. VFedAD learns semantic-rich client data representations through contrastive learning task and cross-client prediction task to identify anomalies. Experiments show VFedAD effectively detects vertical federated anomalies, protecting subsequent algorithms from vertical federated data poisoning attacks.

CCS CONCEPTS

• **Computing methodologies → Anomaly detection**; Neural networks.

KEYWORDS

Data Poisoning Attack and Defense, Vertical Federated Learning, Information Theory

ACM Reference Format:
Jinrong Lai, Tong Wang, Chuan Chen, Yihao Li, and Zibin Zheng. 2023. VFedAD: A Defense Method Based on the Information Mechanism Behind the Vertical Federated Data Poisoning Attack. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management (CIKM '23), October 21–25, 2023, Birmingham, United Kingdom.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3583780.3615106

1 INTRODUCTION

As distributed technologies have evolved and people's awareness of data privacy has increased [17], federated learning has garnered increasing attention. FL enables many participants to build a joint ML model without exposing their private training data [3, 14]. Considered a promising area of research, federated learning has achieved success in multiple fields such as finance, healthcare, and the Internet of Things [11, 13, 15, 21, 23].

However, federal learning has always faced the threat of data poisoning attacks [2, 10, 20]. Data poisoning attacks are a type of attack targeted at machine learning systems [2, 10, 20]. Recently, the impact of label flipping attack on federated learning is studied, and a method to defend against label flipping attack is proposed [18]. In order to conduct poisoning attacks more stealthily, some people propose that attackers can forge private samples of other participants by training GAN locally, and then perform label flip attacks on these generated samples to achieve stealthy attacks on the global model [22] .

Although robust federated learning studies have demonstrated promising results, they solely concentrated on horizontal federated learning. However, defending against vertical federated data poisoning attacks remains an open and challenging area due to the following two reasons: (1) The complex distribution of vertical federated data includes different features of the same sample, creating a vast space for potential attacks that are hard to detect. (2) Current defenses against data poisoning mainly employ shallow anomaly detection methods, which are inadequate to handle the intricate, high-dimensional data distributions in vertical federated learning.

Yet, our study has uncovered a novel approach based on information theory that may offer a simple yet effective solution to these enduring challenges. Firstly, this study conducts an extensive examination of the data flow process and identifies the three primary forms of vertical federated data poisoning attacks. Secondly, this study examines the impact of these attacks on the information structure of vertical federated data. Finally, this study proposes a new unsupervised anomaly detection method, in order to improve

[*] Corresponding author

CIKM '23, October 21–25, 2023, Birmingham, United Kingdom
© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 979-8-4007-0124-5/23/10…$15.00
https://doi.org/10.1145/3583780.3615106

the robustness of the federated system by effectively detecting poisoned samples introduced by data poisoning attacks. Our major contributions are outlined as:

- We conduct an extensive examination of the data flow process and identify the three primary forms of vertical federated data poisoning attacks. Furthermore, we reveal the impact of these attacks on the information structure of vertical federated data through both theoretical analysis and experimental evaluation.
- We propose the defense method VFedAD, an unsupervised vertical federated anomaly detection algorithm based on the information theory. VFedAD is built based on the semantic information structure of vertical federated data and does not depend on the shallow data distribution structure, so it can handle more complex data distributions.
- Sufficient experiments demonstrate that the vertical federated anomalies introduced by data poisoning attacks can be accurately detected by VFedAD even though the data distribution is very complex. By inserting VFedAD before any vertical federated learning algorithm to remove the poisoned samples, it can effectively defend against various vertical federated data poisoning attacks.

2 RELATED WORKS

2.1 Vertical Federated Learning

Vertical federated learning (VFL) is a strategy for training cross-silo datasets that share the same samples in different feature spaces. Unlike horizontal FL, which aggregates model parameters or gradient information, VFL uses encoded embeddings to aggregate all feature spaces for training and inference in a central server. The process first aligns datasets, then encodes local features and sends them to the server for supervised learning. Gradient descent can also be used to optimize models on both client and server, with the server sharing gradient information to help complete local updates.

Previous work has focused on improving the efficiency of VFL, such as using asynchronous updating algorithms and reducing communication times. Some studies proposed methods like VAFL [4], T-VFL [24], FDML [7], FedOnce [19], and CE-VFL [9], to speed up the training process. Other work has focused on improving the effectiveness of VFL, by addressing issues like sample unalignment in real-world datasets using strategies like self-supervised learning as in VFed-SSD [12] and FedHSSL [6].

Nonetheless, few studies have focused on the risk of poisoned data in VFL, which can disrupt its efficiency and effectiveness. Our work addresses this issue by utilizing abnormal detection to detect and resolve poisoned data, an unexplored area in VFL research.

2.2 Data Poisoning Attack

Data poisoning attacks are prevalent in machine learning models. They aim to inject false information into the training data, resulting in incorrect predictions during model inference. Label flipping attacks [18], a type of poisoning attack, corrupt the labels of samples and degrade the performance of the global model. To defend against such attacks, some people proposed using principal component analysis (PCA) to lower the dimensionality of local model parameters and detect malicious clients [18]. An anomaly detection method was introduced to filter potentially poisoned samples during training [5]. This approach generates sub-models and applies a voting scheme to identify poisoned samples. Some people proposed a covert poisoning attack method using generative adversarial networks to forge private samples of other participants [22]. The attacker can perform label flipping attacks on these generated samples, which can effectively tamper with the global model. This type of attack is more covert, as it does not require access to participant devices or bypass local intrusion detection mechanisms.

However, the above data poisoning attacks and defense methods are all aimed at horizontal federated learning and cannot be directly transferred to vertical federated learning scenarios.

3 ALGORITHM

In this section, we propose three primary forms of data poisoning attacks against vertical federated learning. Then, we reveal the impact of data poisoning attacks on the information structure of vertical federated data. Finally, we propose a defense method VFedAD that can accurately detect the anomalies introduced by the above vertical federated data poisoning attacks.

3.1 Data Poisoning Attacks Against Vertical Federated Learning

3.1.1 Characteristics of Vertical Federated Data. In the context of vertical federated learning, we observe that different clients possess features about distinct aspects of the same entity. Although these features contain diverse information, there is a certain internal correlation among them since they all pertain to the same entity. Hence, some of the information is shared among the features, while some are unique to each client (as Example 1).

Example 1 (Characteristics of vertical federated data distribution). *In the context of vertical federated learning, a hospital and a supermarket in the same area can be used as clients. The hospital's data features reflect the health status of users, while the supermarket's features reflect user consumption preferences. Although these features contain different information, they also have certain correlations and share some features. For instance, if Bob suffers from diabetes, his hospital features will show corresponding indicators such as abnormal blood sugar levels, while the supermarket's features will also be affected. For example, the possibility of sweets in the supermarket's consumption preferences will be reduced.*

3.1.2 Vertical Federated Data Poisoning Attack. Complex vertical federated data distribution means that the space of possible attacks is huge. This brings difficulties to the systematic research of vertical federated data poisoning attacks. Observing the entire life cycle of data, from data collection to data storage and application, we propose three primary forms vertical federated data poisoning attacks (shown in Figure 1):

Definition 1 (Random Failure Attack). *The samples in client c are replaced with random values with a certain probability, which means that client c suffers from a random failure attack.*

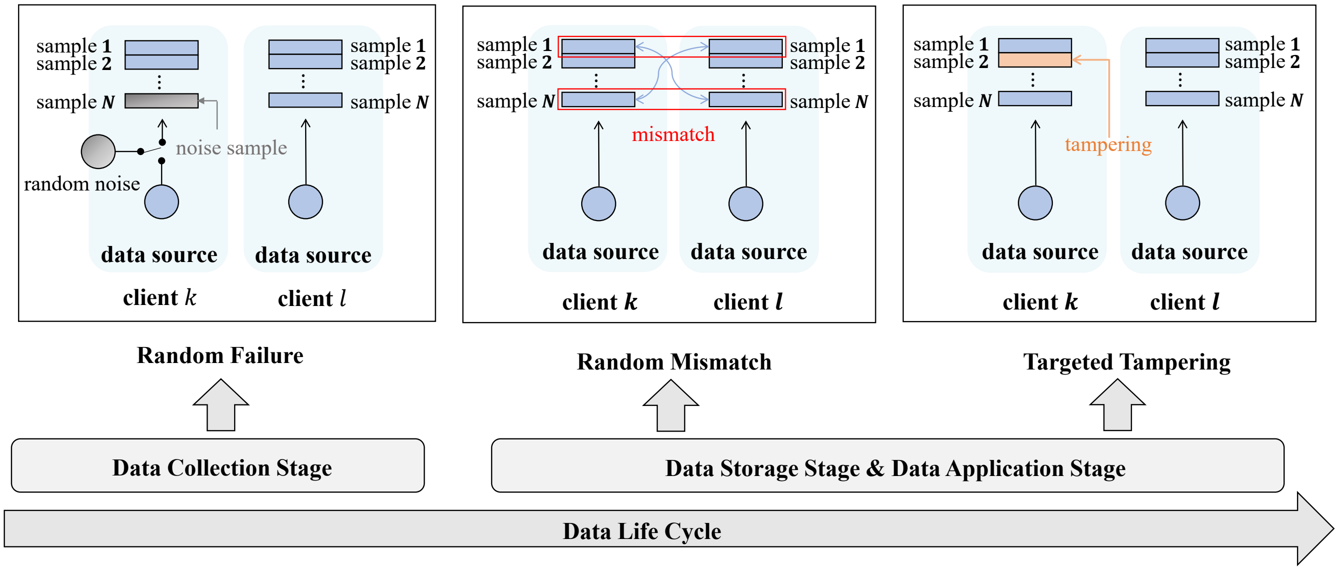

Figure 1: Three primary forms vertical federated data poisoning attacks

Definition 2 (Random Mismatch Attack). *A certain proportion of features in client c are randomly shuffled, resulting in the destruction of their correspondence with features in other clients, which means that client c suffers from a random mismatch attack.*

Definition 3 (Targeted Tampering Attack). *The data of client c is artificially tampered with so that the tampered data obeys the target data distribution, which means that client c suffers from a targeted tampering attack.*

The **random failure attack** models the unavoidable natural causes of noise in the data collection phase in the real world, for example, sensor faults or data transmission faults. Also, some low-level attackers may tamper with sample data into random noise, and random failure attack also covers this situation.

The **random mismatch attack** models the mismatch between samples in different clients. For example, in the data storage and application stage, due to the lack of communications among clients, all entities in all clients are sometimes not perfectly aligned. Or an attacker may also select a part of the dataset and shuffle them.

The **targeted tampering attack** models those anomalous data patterns caused by intended attacking. For example, an attacker may tamper with the features of some samples as the target features, for example, tamper with the facial features of some users as the target facial features.

3.2 Vertical Federated Data from the Perspective of Information Theory

It is not difficult to find that these proposed vertical federated data poisoning attacks will destroy the information structure of vertical federated data because the correlation of features in different clients of the attacked samples is weakened.

3.2.1 Theoretical Analysis. We summarize the above conclusion as Theorem 1 and prove it in the appendix.

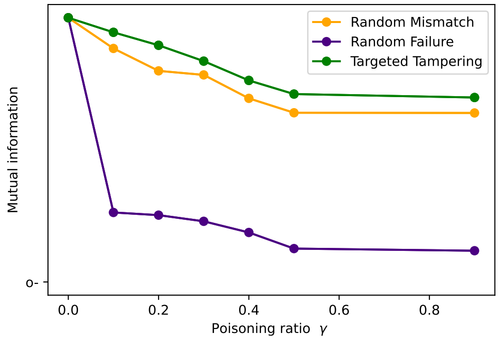

Figure 2: Sum of mutual information values for all client pairs of the Caltech-7 dataset at different poisoning ratios γ.

定义了三种在纵向联邦学习的攻击方法，提出一种纵向联邦学习防御方法。

Theorem 1. *The occurrence of any one of the vertical federated data poisoning attacks (random mismatch, targeted tampering, or random failure), will lead to the reduction of the mutual information of data from different clients, that is, $MI^{poisoned} < MI^{ideal}$.*

3.2.2 Experimental Evaluation. To verify the Theorem 1, we construct a series of datasets containing different types of poisoning attacks and different poisoning ratios based on the Caltech-7 dataset. With MINE [1], we estimate the sum of the mutual information between all client pairs. As shown in Figure 2, the mutual information between clients will decrease as the poisoning ratio grows. Thus we validate the Theorem 1 that these three vertical federated data poisoning attacks will disrupt shared semantic information between clients.

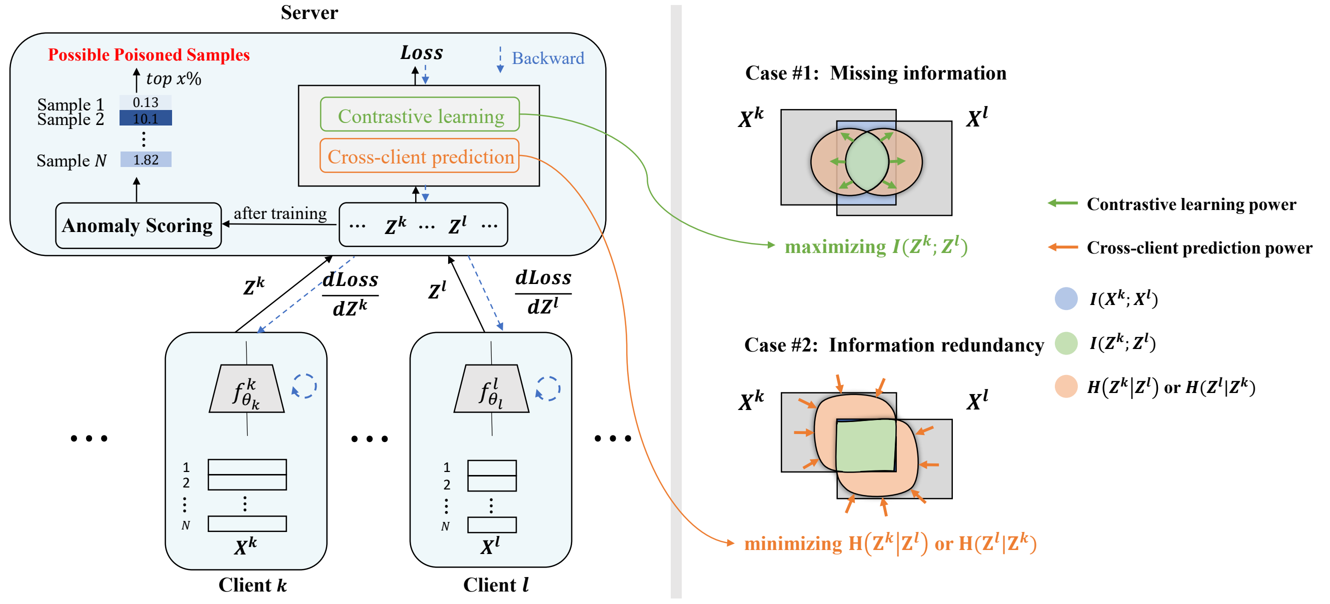

Figure 3: Overview of the VFedAD. To ensure that the representations Z contain as much semantic information as possible, we use contrastive learning to maximize the mutual information of representations in different clients. Meanwhile, to ensure that the representations Z do not contain semantically irrelevant information, a cross-client prediction task is used to minimize the conditional entropy of representations in different clients to discard semantically irrelevant information. After training, we learn a compact representation rich in consistent semantic information and free of irrelevant information. Then we perform an anomaly scoring, the higher the score the sample is more likely to be a poisoned sample.

3.3 Proposed Defense Model

Motivated by the relationship between poisoning attacks and the information structure in vertical federated learning, we present a defense method, VFedAD, illustrated in Figure 3, to detect anomalies generated by poisoning attacks. For privacy preservation, we utilize the approach used in VAFL [4] where clients upload representations of their private data to the server and apply local perturbation to enhance differential privacy. Assuming that the samples in different clients have been aligned, VFedAD learns hidden representations containing precise semantic information to detect anomalies with abnormal semantic information based on these representations. To ensure that the learned semantic representations contain maximum semantic information and no irrelevant information, we address two issues: (1) how to maximize the amount of captured semantic information in the representations, and (2) how to discard semantically irrelevant information in the representations. To address these problems, two tasks are designed: the contrastive learning task and the cross-client prediction task.

3.3.1 *Maximize semantic information shared by clients.*

When the learned representation Z lacks semantic information, the contrastive learning task drives the representation Z to capture more shared semantic information $I(Z^k; Z^l)$ (as shown in Figure 3) . Consider a case containing two clients as an example: we maximize the shared information of the representation Z^k in client k and the representation Z^l in client l by optimizing the following contrastive loss:

$$\mathcal{L}_{cl} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{\mathrm{sim}(\mathrm{Z}_{\mathrm{i}}^{k},\mathrm{Z}_{\mathrm{i}}^{l})}}{\sum_{j=1}^{N} e^{\mathrm{sim}\left(\mathrm{Z}_{\mathrm{i}}^{k},\mathrm{Z}_{\mathrm{j}}^{l}\right)}} \tag{1}$$

where N is the number of samples, $\mathrm{sim}(\cdot,\cdot)$ is euclidean distance, $\mathrm{Z}_{\mathrm{i}}^{k}$ is the representation of sample i corresponding to client k, and so on. According to previous research [16], minimizing such contrastive loss is equivalent to maximizing the lower bound of the mutual information of the representations between the clients: $I(Z^k; Z^l) \geq \log(N) - \mathcal{L}_{cl}$. Therefore, by minimizing the contrastive loss, we can force Z^k and Z^l to capture as much shared semantic information as possible as their mutual information grows. For multiple clients, the optimization objective is as follows:

$$\mathcal{L}_{CL} = \sum_{k\neq l} -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{\mathrm{sim}(\mathrm{Z}_{\mathrm{i}}^{k},\mathrm{Z}_{\mathrm{i}}^{l})}}{\sum_{j=1}^{N} e^{\mathrm{sim}\left(\mathrm{Z}_{\mathrm{i}}^{k},\mathrm{Z}_{\mathrm{j}}^{l}\right)}} \tag{2}$$

3.3.2 *Discard semantically irrelevant information.*

When there are some redundant semantically irrelevant information contained in Z^k and Z^l, the cross-client prediction task will force the representation to discard them as reducing conditional entropy $H(Z^k|Z^l)$ and $H(Z^l|Z^k)$ (as shown in the Figure 3).

Take minimizing $H(Z^k|Z^l)$, which measures the redundant information in Z^k not relevant to Z^l, as an example. Minimizing $H(Z^k|Z^l)$ is equivalent to maximizing $\mathbb{E}_{P_{Z^k,Z^l}}\left[\log P\left(Z^k \mid Z^l\right)\right] = -H\left(Z^k \mid Z^l\right)$. Since it is intractable to optimize $\mathbb{E}_{P_{Z^k,Z^l}}\left[\log P\left(Z^k \mid Z^l\right)\right]$

directly, we introduce a variational distribution $Q_\phi\left(Z^k \mid Z^l\right)$ with parameter ϕ . We can show that

$$\begin{aligned}&\mathbb{E}_{P_{Z^k,Z^l}}\left[\log P\left(Z^k \mid Z^l\right)\right]\\ &=\mathbb{E}_{P_{Z^k,Z^l}}\left[\log Q_\phi\left(Z^k \mid Z^l\right)\right]\\ &\quad+D_{\mathrm{KL}}\left(P\left(Z^k \mid Z^l\right)\|Q_\phi\left(Z^k \mid Z^l\right)\right)\\ &\geq\mathbb{E}_{P_{Z^k,Z^l}}\left[\log Q_\phi\left(Z^k \mid Z^l\right)\right]\end{aligned} \tag{3}$$

Therefore, we find $\mathbb{E}_{P_{Z^k,Z^l}}\left[\log Q_\phi\left(Z^k \mid Z^l\right)\right]$ is a lower bound of $\mathbb{E}_{P_{Z^k,Z^l}}\left[\log P\left(Z^k \mid Z^l\right)\right]$. We can let $Q_\phi\left(Z^k \mid Z^l\right)$ be the Gaussian $\mathcal{N}\left(Z^k \mid g^{l\to k}(Z^l), \sigma\mathbf{I}\right)$ then minimizing $H(Z^k|Z^l)$ is equivalent to minimizing:

$$\begin{aligned}\mathcal{L}_{CP}^{k|l} &=\mathbb{E}_{Z^k,g^{l\to k}(Z^l)\sim P_{Z^k,g^{l\to k}(Z^l)}}\left[\left\|Z^k - g^{l\to k}(Z^l)\right\|_2^2\right]\\ &=\frac{1}{N}\sum_{i=1}^{N}\left\|Z_i^k - g^{l\to k}(Z_i^l)\right\|_2^2\end{aligned} \tag{4}$$

where $g^{l\to k}(\cdot)$ is the predictor that predicts client k from client l. Similarly, the loss function to achieve minimizing $H(Z^l|Z^k)$ is $\mathcal{L}_{CP}^{l|k}$. Therefore, by minimizing our cross-client prediction loss function:

$$\begin{aligned}\mathcal{L}_{CP} &=\mathcal{L}_{CP}^{k|l}+\mathcal{L}_{CP}^{l|k}\\ &=\frac{1}{N}\sum_{i=1}^{N}\left[\left\|Z_i^k - g^{l\to k}(Z_i^l)\right\|_2^2+\left\|Z_i^l - g^{k\to l}(Z_i^k)\right\|_2^2\right]\end{aligned} \tag{5}$$

the redundant information in Z^k and Z^l will get discarded.

In the case of multiple clients, the total cross-client prediction loss can be written as

$$\mathcal{L}_{CP-dual}=\sum_{k\neq l}\mathcal{L}_{CP}^{k|l}=\frac{1}{N}\sum_{i=1}^{N}\sum_{k\neq l}\left\|Z_i^k - g^{l\to k}(Z_i^l)\right\|_2^2 \tag{6}$$

3.3.3 The Ring Prediction Loss with Good Scalability. However, to calculate the loss function above (the dual prediction loss), $M(M-1)$ predictor networks are needed if there are M clients in the federated system, which is not affordable in practice. Hence, we relax the dual prediction loss above to the ring prediction loss:

$$\begin{aligned}\mathcal{L}_{CP-ring} &=\sum_{k=1}^{M-1}\mathcal{L}_{CP}^{k+1|k}+\mathcal{L}_{CP}^{1|M}\\ &=\frac{1}{N}\sum_{i=1}^{N}\left[\sum_{k=1}^{M-1}\left\|Z_i^{k+1} - g^{k\to k+1}(Z_i^k)\right\|_2^2\right.\\ &\quad\left.+\left\|Z_i^1 - g^{M\to 1}(Z_i^M)\right\|_2^2\right]\end{aligned} \tag{7}$$

in which we need only M predictors.

However, we found that the ring prediction loss can achieve the same effect even though it uses much fewer predictors than the dual prediction loss. Next, we prove that they can achieve the same goal of reducing the irrelevant information with the optimal representations, as stated in Theorem 2.

Theorem 2. *It is possible to achieve the following goals through the optimization of either the ring cross-client prediction loss $\mathcal{L}_{CP-ring}$ or the dual cross-client prediction loss $\mathcal{L}_{CP-dual}$:*

$$H(Z^i|Z^j)=0,\quad i,j=1,2,\ldots,M,\ i\neq j$$

where Z^i is all sample representations of client i and M is the number of clients.

We prove the theorem 2 in the appendix. If not specified, the cross-client prediction loss $\mathcal{L}_{CP}$ used in VFedAD is the ring prediction loss $\mathcal{L}_{CP-ring}$.

3.3.4 Overall Objective Function. The overall objective function of VFedAD is:

$$\mathcal{L}=\mathcal{L}_{CL}+\lambda\mathcal{L}_{CP} \tag{8}$$

where λ is a trade-off parameter. Optimizing this loss can ensure that the representation captures as much semantic information as possible while discarding semantically irrelevant information. After the loss converges, we perform anomaly scoring, and samples with higher scores are more likely to be poisoned samples.

3.4 Anomaly Score Measurement

With the learned latent representations, we propose an anomaly scoring function:

$$S(i)=S_{NC}(i)+S_{CC}(i) \tag{9}$$

where

$$\begin{gathered}S_{NC}(i)=\sum_{Z_j\in\mathrm{knn}(Z_i)}\left[\left\|Z_i - Z_j\right\|_2^2\right],\\ \left(Z_i=Z_i^1\oplus Z_i^2\ldots\oplus Z_i^M\right)\end{gathered} \tag{10}$$

$$S_{CC}(i)=\sum_{k=1}^{M-1}\left\|Z_i^{k+1} - g^{k\to k+1}(Z_i^k)\right\|_2^2+\left\|Z_i^1 - g^{M\to 1}(Z_i^M)\right\|_2^2 \tag{11}$$

$S_{NC}(i)$ is the neighbor consistency score (concatenate the representations Z_i^k of all clients of sample i to get Z_i). $S_{CC}(i)$ is the client consistency score (cross-client prediction loss for the sample i). We define the set $Clients=\{1,2,\ldots M\}$ and $Samples=\{1,2,\ldots N\}$. The anomaly scoring function can detect anomalies injected by three types of poisoning attacks: (1) **For random failure**: If sample i in client k is attacked by this attack, it means that its feature becomes a significant anomaly, and its original semantic information is completely destroyed. Therefore, its semantic representation Z_i^k will be far away from $Z^k_{Samples\setminus\{i\}}$, the semantic representation of other samples, and will also be far away from its representation on other clients ($Z_i^{Clients\setminus\{k\}}$), so the poisoned sample i will have a larger neighbor consistency score and client consistency score. (2) **For random mismatch**: If the sample i in the client k is subjected to a random mismatch attack, it means that the semantic consistency of the features of the poisoned sample i in different clients is destroyed, so the corresponding semantic representations Z_i^k is away from $Z_i^{Clients\setminus\{k\}}$, the poisoned sample i will have a larger client consistency score. (3) **For targeted tampering**: If sample i in client k is subjected to this attack, it means that X_i^k is changed to the feature of the target category, resulting in X_i^k, the feature of the poisoned sample i in the client k, contain different semantic

Table 1: Details of several datasets. The numbers in the table are the dimensions of the data in the client (with feature name in parenthesis)

Client	**Synthetic dataset**	**Real multi-view datasets**		
		MSRC-v1	**AWA-10**	**Caltech-7**
1	2	24 (CM)	2688 (CQ)	48 (Gabor)
2	2	576 (HOG)	2000 (LSS)	40 (WM)
3	-	512 (GIST)	252 (PHOG)	254 (CENTRIST)
4	-	256 (LBP)	2000 (SIFT)	1984 (HOG)
5	-	254 (CENT)	2000 (RGSIFT)	512 (GIST)
6	-	-	2000 (SURF)	928 (LBP)
Number of samples	400	210	800	1474
Number of categories	2	7	10	7

information with the features $X_i^{Clients\backslash\{k\}}$ in other clients. Therefore, the corresponding semantic representation $X_i^{Clients\backslash\{k\}}$ and $Z_i^{Clients\backslash\{k\}}$ are far apart, so the client consistency score of the poisoned sample i will be larger.

After model convergence, poisoned samples will have larger anomaly scores than normal samples. Therefore, VFedAD can detect and remove poisoned samples, defending against vertical federated data poisoning attacks.

4 EXPERIMENTS

4.1 Experimental Setup

4.1.1 Datasets. This section uses a synthetic dataset and three real datasets (MSRC-v1, AWA-10, Caltech-7) that are commonly employed in multi-view learning. Each view's data corresponds to a client in subsequent experiments. Table 1 summarizes the datasets' details. A synthetic dataset is created to showcase VFedAD's superiority in processing datasets with complex data distribution that does not adhere to cluster structure assumptions. This synthetic dataset is a two-client dataset with two categories in a single cluster structure that fails the cluster structure assumption. Client 1 has 400 data points sampled from a two-dimensional Gaussian distribution $N(0, 0, 0.1, 0.1, 0)$, segmented into two categories with the line y=x as a boundary. We apply linear transformations and random perturbations to client 1 to create client 2.

4.1.2 Implementation Details. We conduct our experiments with a batch size of 256 on RTX 3090. The number of nearest neighbors in the outlier score function is set to 5. The trade-off parameter λ in the loss function is set to 10 (Synthetic), 100 (Caltech-7), 1 (AWA-10), and 10 (MSRC-v1) on different datasets. The intra-client encoder and cross-client predictor (in server) are both implemented with MLP (Multi-Layer Perceptron). We use the Adam optimizer to optimize all encoders and predictors. The learning rate is set to 0.001 in all experiments. We use AUC (area under ROC curve) [8] and F1-score as evaluation metrics for the anomaly detection ability of the model. We will publish our code on GitHub after the paper is accepted.

4.2 Effectiveness of Vertical Federated Data Poisoning Attacks

We conducted varying degrees of our proposed data poisoning attacks on three datasets and observed the classification performance of the downstream vertical federated learning algorithms (FedAVG [14] and VAFL [4]) on the poisoned datasets. For example, "vafl@msrc-v1" in Figure 4 corresponds to the experimental results of the VAFL algorithm on MSRC-V1 dataset. As shown in Figure 4, as the poisoning samples ratio increased, the performance of the downstream model significantly decreased. Furthermore, a relatively small rate of poisoning samples (3%) was sufficient to significantly damage the model. This demonstrates that our proposed data poisoning attack can effectively damage vertical federated learning algorithms.

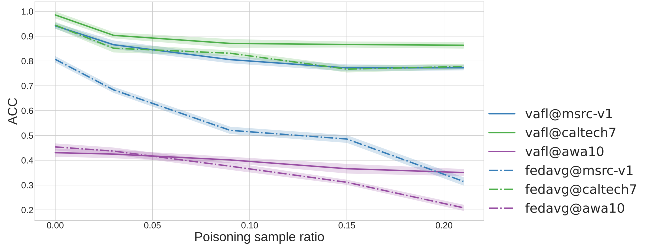

Figure 4: Performance of downstream algorithms on datasets with different poisoning attack ratios

Table 2: Experimental evaluation of anomaly detection capability of VFedAD. The average AUC and F1-score of 5 replicate experiments were recorded.

	Synthetic dataset	MSRC-v1	AWA-10	Caltech-7
AUC	0.984	0.967	0.912	0.957
F1-score	0.917	0.862	0.792	0.831

4.3 Evaluate the ability of VFedAD to detect poisoned samples

4.3.1 Synthetic Datasets. To demonstrate VFedAD's ability to handle complex data distribution, we evaluate VFedAD using a synthetic dataset with odd data distribution. The experimental results are recorded in Table 2. Experimental results show that VFedAD can accurately detect anomalies from synthetic dataset even if the dataset does not obey common cluster structure assumptions.

4.3.2 Real Datasets. In order to further verify the ability of VFedAD to handle real high-dimensional and complex data sets, we conduct experiments on real datasets with multiple clients, high dimensions, and large sample numbers. The experimental results are recorded in Table 2. The experimental results show that even on real data sets with complex data distribution, VFedAD can still accurately identify the anomalies introduced by the three data poisoning attacks.

4.4 VFedAD is A Good Bodyguard

In this section, we conduct experiments on real datasets to check whether VFedAD can effectively protect downstream vertical federated algorithms from vertical federated data poisoning attacks. To this end, we insert the VFedAD algorithm before other vertical federated learning algorithms (FedAVG [14] and VAFL [4]) to eliminate possible poisoned samples and then observe the performance of the downstream vertical federated algorithm on the remaining

Table 3: Insert VFedAD before other vertical federation algorithms to discard possible poisoned samples, and record the classification task performance of downstream vertical federated algorithms on the remaining samples. The average accuracy of 10 replicate experiments was recorded.

	MSRC-v1	AWA-10	Caltech-7
PSR-15% →FedAVG	46.03	37.50	77.86
PSR-15% →VFedAD-5% →FedAVG	68.01	39.55	87.97
PSR-15% →VFedAD-10% →FedAVG	72.32	42.19	90.85
PSR-15% →VFedAD-15% →FedAVG	78.46	45.00	91.39
PSR-0% →FedAVG	80.95	45.25	94.37
PSR-15% →VAFL	80.71	36.66	92.00
PSR-15% →VFedAD-5% →VAFL	82.50	39.57	92.26
PSR-15% →VFedAD-10% →VAFL	84.54	43.25	95.40
PSR-15% →VFedAD-15% →VAFL	90.00	43.67	97.50
PSR-0% →VAFL	94.33	43.95	97.96

data. Specifically, for each experimental data set, we first perform the VFedAD algorithm to obtain the anomaly score of each sample. Next, we sort the anomaly scores of all samples in descending order and remove the top x%(x=5,10,15) samples. Finally, the downstream federated learning algorithm performs classification task learning on the remaining samples and records the final ACC. The experimental results are recorded in Table 3. "PSR-15% →VFedAD-5% →VAFL "means that the dataset has a poisoning sample rate of 15%. VFedAD is employed to remove the top 5% of samples with the highest anomaly scores, and the downstream federated algorithm used is VAFL. From all the experimental results, as the number of poisoned samples removed by VFedAD increases, the performance of downstream federated learning algorithms improves significantly. In summary, VFedAD can effectively protect downstream vertical federated learning from harm caused by upstream vertical federated poisoning attacks.

4.5 Ablation Study

In this part, we conduct the ablation study to demonstrate the effectiveness of various parts of VFedAD, including the contrastive loss L_{CL}, the ring cross-client prediction loss $\mathcal{L}_{CP-ring}$ and the dual cross-client prediction loss $\mathcal{L}_{CP-dual}$. The results of the ablation study are shown in Table 4. It can be seen that applying the contrastive learning task and the cross-client prediction task together outperforms applying one task alone, which demonstrates the effectiveness of our learning task design. And according to the last two rows of Table 4, the ring cross-client prediction loss has almost the same effect as the dual cross-client prediction loss, even though the ring loss needs to introduce far fewer cross-client predictors than the dual loss. As Theorem 2 states, the ring cross-client prediction and the dual prediction have the same effect.

Table 4: The results of ablation experiments on the real datasets. The average AUC of 5 replicate experiments was recorded. The best and the second-best results are in bold and underline, respectively.

	MSRC-v1	AWA-10	Caltech-7
L_{CL}	0.906	0.866	0.934
$L_{CP-dual}$	0.891	0.813	0.876
$L_{CP-ring}$	0.886	0.799	0.881
$L_{CL} + L_{CP-dual}$	<u>0.964</u>	**0.928**	<u>0.953</u>
$L_{CL} + L_{CP-ring}$	**0.967**	<u>0.912</u>	**0.957**

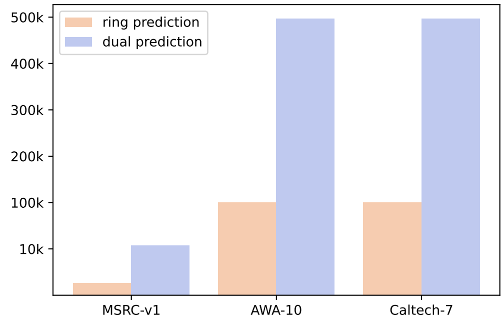

Figure 5: A comparison of the number of model parameters in the ring prediction loss and the dual prediction loss

5 CONCLUSION

We first propose three basic types of vertical federated poisoning attacks and elucidate their underlying information-theoretic mechanisms. We then propose an unsupervised defense method VFedAD which can accurately detect poisoned samples. Experiments have proved that VFedAD can accurately identify poisoned samples even in the face of high-dimensional and complex data distribution, thus effectively protecting downstream vertical federated learning algorithms from data poisoning attacks. To the best of our knowledge, this study is the first work that investigates the information-theoretic mechanisms underlying vertical federated poisoning attacks and defenses. We envisage that our work could serve as a basis for future studies on poisoning attack and defense in vertical federated learning.

ACKNOWLEDGMENTS

The research is supported by the National Key R&D Program of China (No. 2020YFB1006001), the Key-Area Research and Development Program of Guangdong Province (No. 2020B1111370001), the National Natural Science Foundation of China (62176269), the Guangzhou Science and Technology Program (2023A04J0314), and the Sun Yat-Sen University Young Faculty Development Program (No. 23ptpy109).

REFERENCES

[1] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. 2018. Mutual information neural estimation. In *International conference on machine learning*. PMLR, 531–540.

[2] Battista Biggio, Blaine Nelson, and Pavel Laskov. 2012. Poisoning attacks against support vector machines. *arXiv preprint arXiv:1206.6389* (2012).

[3] Keith Bonawitz, Vladimir Ivanov, Ben Kreuter, Antonio Marcedone, H Brendan McMahan, Sarvar Patel, Daniel Ramage, Aaron Segal, and Karn Seth. 2017. Practical secure aggregation for privacy-preserving machine learning. In *proceedings*

of the 2017 ACM SIGSAC Conference on Computer and Communications Security. 1175–1191.

[4] Tianyi Chen, Xiao Jin, Yuejiao Sun, and Wotao Yin. 2020. Vafl: a method of vertical asynchronous federated learning. *arXiv preprint arXiv:2007.06081* (2020).

[5] Gabriela F Cretu, Angelos Stavrou, Michael E Locasto, Salvatore J Stolfo, and Angelos D Keromytis. 2008. Casting out demons: Sanitizing training data for anomaly sensors. In *2008 IEEE Symposium on Security and Privacy (sp 2008)*. IEEE, 81–95.

[6] Yuanqin He, Yan Kang, Jiahuan Luo, Lixin Fan, and Qiang Yang. 2022. A hybrid self-supervised learning framework for vertical federated learning. *arXiv preprint arXiv:2208.08934* (2022).

[7] Yaochen Hu, Di Niu, Jianming Yang, and Shengping Zhou. 2019. FDML: A collaborative machine learning framework for distributed features. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2232–2240.

[8] Jin Huang and Charles X Ling. 2005. Using AUC and accuracy in evaluating learning algorithms. *IEEE Transactions on knowledge and Data Engineering* 17, 3 (2005), 299–310.

[9] Afsana Khan, Marijn ten Thij, and Anna Wilbik. 2022. Communication-Efficient Vertical Federated Learning. *Algorithms* 15, 8 (2022), 273.

[10] Pang Wei Koh and Percy Liang. 2017. Understanding black-box predictions via influence functions. In *International conference on machine learning*. PMLR, 1885–1894.

[11] Yogesh Kumar and Ruchi Singla. 2021. Federated learning systems for healthcare: perspective and recent progress. *Federated Learning Systems* (2021), 141–156.

[12] Wenjie Li, Qiaolin Xia, Junfeng Deng, Hao Cheng, Jiangming Liu, Kouying Xue, Yong Cheng, and Shu-Tao Xia. 2022. Semi-Supervised Cross-Silo Advertising with Partial Knowledge Transfer. *arXiv preprint arXiv:2205.15987* (2022).

[13] Guodong Long, Yue Tan, Jing Jiang, and Chengqi Zhang. 2020. Federated learning for open banking. In *Federated learning*. Springer, 240–254.

[14] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. 2017. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*. PMLR, 1273–1282.

[15] Dinh C Nguyen, Ming Ding, Pubudu N Pathirana, Aruna Seneviratne, Jun Li, and H Vincent Poor. 2021. Federated learning for internet of things: A comprehensive survey. *IEEE Communications Surveys & Tutorials* 23, 3 (2021), 1622–1658.

[16] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748* (2018).

[17] Protection Regulation. 2016. Regulation (EU) 2016/679 of the European Parliament and of the Council. *Regulation (eu)* 679 (2016), 2016.

[18] Vale Tolpegin, Stacey Truex, Mehmet Emre Gursoy, and Ling Liu. 2020. Data poisoning attacks against federated learning systems. In *European Symposium on Research in Computer Security*. Springer, 480–501.

[19] Zhaomin Wu, Qinbin Li, and Bingsheng He. 2022. Practical Vertical Federated Learning with Unsupervised Representation Learning. *IEEE Transactions on Big Data* (2022).

[20] Han Xiao, Huang Xiao, and Claudia Eckert. 2012. Adversarial label flips attack on support vector machines. In *ECAI 2012*. IOS Press, 870–875.

[21] Qiang Yang, Yang Liu, Tianjian Chen, and Yongxin Tong. 2019. Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)* 10, 2 (2019), 1–19.

[22] Jiale Zhang, Junjun Chen, Di Wu, Bing Chen, and Shui Yu. 2019. Poisoning attack in federated learning using generative adversarial nets. In *2019 18th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/13th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE)*. IEEE, 374–380.

[23] Weishan Zhang, Tao Zhou, Qinghua Lu, Xiao Wang, Chunsheng Zhu, Haoyun Sun, Zhipeng Wang, Sin Kit Lo, and Fei-Yue Wang. 2021. Dynamic-fusion-based federated learning for COVID-19 detection. *IEEE Internet of Things Journal* 8, 21 (2021), 15884–15891.

[24] Zezhong Zhang, Guangxu Zhu, and Shuguang Cui. 2022. Low-Latency Cooperative Spectrum Sensing via Truncated Vertical Federated Learning. *arXiv preprint arXiv:2208.03694* (2022).

A PROOF OF THEOREM 1

Proof. Here we theoretically prove the Theorem ??. We first consider the data distribution under ideal conditions without attacks. For brevity, we take the data from two clients as an example. It is assumed that there are c underlying semantic classes in the data. When no attack occurs, we can model the joint distribution of the semantic class related to the features from two clients as:

$$P_{\mathbb{C}_1\mathbb{C}_2}(i,j) = \begin{cases} p_i, & i = j \text{ and } i = 1, 2, ..., c \\ 0, & \text{otherwise.} \end{cases} \tag{12}$$

where $P_{\mathbb{C}_1\mathbb{C}_2}(i,j)$ represents the probability of a sample that belongs to class i in client 1 and belongs to class j in client 2. For clarity, the joint distribution is recorded as Table 5. Normal samples correspond to the diagonal elements (excluding $P_{\mathbb{C}_1\mathbb{C}_2}(i = \text{anomaly}, j = \text{anomaly})$) of Table 5.

The inter-client mutual information MI for an ideal data distribution without poisoning attacks is:

$$\begin{aligned} \mathrm{MI}^{\text{ideal}} &= \sum_{i,j} P_{\mathbb{C}_1\mathbb{C}_2}(i,j) \log\left(\frac{P_{\mathbb{C}_1\mathbb{C}_2}(i,j)}{P_{\mathbb{C}_1}(i)P_{\mathbb{C}_2}(j)}\right) \\ &= -\sum_{i=1}^{c} p_i \log p_i \end{aligned} \tag{13}$$

Next, we show that all three types of poisoning attacks will lead to a reduction in mutual information between clients.

(1) For random mismatch: Without loss of generality, we assume that when a random mismatch occurs, the features in client 2 will be randomly shuffled. Assuming that the attack occurs at a probability of $0 < \alpha < 1$, then each sample has a probability of α to be randomly shuffled in client 2. It is not difficult to calculate that, when considering random mismatch, the joint distribution is:

$$P^{rm}_{\mathbb{C}_1\mathbb{C}_2}(i,j) = (1-\alpha)P_{\mathbb{C}_1\mathbb{C}_2}(i,j) + \alpha P_{\mathbb{C}_2}(j)P_{\mathbb{C}_1\mathbb{C}_2}(i,i) \tag{14}$$

And the marginal distribution of each client under random mismatch attacks is:

$$\begin{aligned} P^{rm}_{\mathbb{C}_1}(i) &= \sum_{j=1}^{c} P^{rm}_{\mathbb{C}_1\mathbb{C}_2}(i,j) = P_{\mathbb{C}_1}(i) \\ P^{rm}_{\mathbb{C}_2}(j) &= \sum_{i=1}^{c} P^{rm}_{\mathbb{C}_1\mathbb{C}_2}(i,j) = P_{\mathbb{C}_2}(j) \end{aligned} \tag{15}$$

The random mismatch attack does not change the marginal distribution of each client, since such attack only destroy the association between clients. Considering random mismatch attacks, the mutual information between clients MI^{rm} is:

Table 5: Ideal multi-client joint distribution without considering poisoning attacks

$\mathbb{C}_2$ $\mathbb{C}_1$	1	2	3	...	c	anomaly
1	p_1	0	0	$\cdots$	0	0
2	0	p_2	0	$\cdots$	0	0
3	0	0	p_3	$\cdots$	0	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
c	0	0	0	$\cdots$	p_c	0
anomaly	0	0	0	$\cdots$	0	0

$$\begin{aligned}
\mathrm{MI}^{rm} &= \sum_{i,j} P^{rm}_{\mathbb{C}_1\mathbb{C}_2}(i,j)\log\left(\frac{P^{rm}_{\mathbb{V}_1\mathbb{V}_2}(i,j)}{P^{rm}_{\mathbb{C}_1}(i)P^{rm}_{\mathbb{C}_2}(j)}\right)\\
&= \sum_{i}[(1-\alpha)p_i+\alpha p_i^2]\log\left(\frac{(1-\alpha)p_i+\alpha p_i^2}{p_i^2}\right)\\
&\quad + \alpha\log\alpha\sum_{i\neq j}p_ip_j\\
&< \sum_{i}[(1-\alpha)p_i+\alpha p_i^2]\log\left(\frac{(1-\alpha)p_i+\alpha p_i^2}{p_i^2}\right)\\
&< \sum_{i}[(1-\alpha)p_i+\alpha p_i]\log\left(\frac{(1-\alpha)p_i+\alpha p_i}{p_i^2}\right)\\
&= -\sum_{i=1}^{c}p_i\log(p_i) = \mathrm{MI}^{ideal}
\end{aligned} \tag{16}$$

So far, we have demonstrated that random mismatches lead to reduced mutual information shared across clients.

(2) For targeted tampering: Since different malicious attackers want different target data distributions, targeted tampering attacks are complex and diverse. To simplify the problem, we consider a commonly targeted tampering attack: the sample features of a certain class in the malicious client are modified to the features from another semantic class. For this case, without loss of generality, we assume that the attacker holds client 1 and changes those data features of class 1 to features from class 2. Under such a situation, we can write the joint distribution as:

$$P^{tt}_{\mathbb{C}_1\mathbb{C}_2}(i,j) = \begin{cases} P_{\mathbb{C}_1\mathbb{C}_2}(i,j), & i=j \text{ and } i=2,...,c\\ P_{\mathbb{C}_1\mathbb{C}_2}(1,1), & i=2 \text{ and } j\text{=}1\\ 0, & \text{otherwise.}\end{cases} \tag{17}$$

Thus, the mutual information between views is:

$$\begin{aligned}
\mathrm{MI}^{tt} &= \sum_{i,j} P^{tt}_{\mathbb{C}_1\mathbb{C}_2}(i,j)\log\left(\frac{P^{tt}_{\mathbb{C}_1\mathbb{C}_2}(i,j)}{P^{tt}_{\mathbb{C}_1}(i)P^{tt}_{\mathbb{C}_2}(j)}\right)\\
&=p_2\log\left(\frac{p_2}{(p_1+p_2)p_2}\right)+\sum_{i=3}^{k}p_i\log\left(\frac{p_i}{p_ip_i}\right)\\
&\quad + p_1\log\left(\frac{p_1}{(p_1+p_2)p_1}\right)\\
&=p_1\log\left(\frac{p_1}{p_1+p_2}\right)+p_2\log\left(\frac{p_2}{p_1+p_2}\right)+\mathrm{MI}^{ideal}\\
&<\mathrm{MI}^{ideal}
\end{aligned} \tag{18}$$

Therefore it is proved that targeted tampering leads to less semantic information shared between clients.

(3) For random failure: Assuming that the features in client 1 and client 2 become noise with failure rate α_1 and α_2 ($0 < \alpha_1, \alpha_2 < 1$), respectively. Considering the random failure, the joint distribution is:

$$P^{rf}_{\mathbb{C}_1\mathbb{C}_2}(i,j) = \begin{cases} (1-\alpha_1)(1-\alpha_2)P_{\mathbb{C}_1\mathbb{C}_2}(i,j); \quad i,j=1,2...c & .\\ \alpha_1(1-\alpha_2)P_{\mathbb{C}_2}(j); i\text{=anomaly and } j=1,2...c & .\\ \alpha_2(1-\alpha_1)P_{\mathbb{C}_1}(i); j\text{=anomaly and } i=1,2...c & .\\ \alpha_1\alpha_2; \qquad i\text{=anomaly and } j\text{=anomaly} & \end{cases} \tag{19}$$

And according the joint distribution, it can be proved that:

$$\begin{aligned}
\mathrm{MI}^{rf} &= \sum_{i,j} P^{rf}_{\mathbb{C}_1\mathbb{C}_2}(i,j)\log\left(\frac{P^{rf}_{\mathbb{C}_1\mathbb{C}_2}(i,j)}{P^{rf}_{\mathbb{C}_1}(i)P^{rf}_{\mathbb{C}_2}(j)}\right)\\
&=\sum_{i=1}^{c}(1-\alpha_1)(1-\alpha_2)p_i\log\left(\frac{(1-\alpha_1)(1-\alpha_2)p_i}{(1-\alpha_1)p_i(1-\alpha_2)p_i}\right)\\
&\quad+\sum_{j=1}^{c}(1-\alpha_2)\alpha_1p_j\log\left(\frac{(1-\alpha_2)\alpha_1p_j}{\alpha_1(1-\alpha_2)p_j}\right)\\
&\quad+\sum_{i=1}^{c}(1-\alpha_1)\alpha_2p_i\log\left(\frac{(1-\alpha_1)\alpha_2p_i}{\alpha_2(1-\alpha_1)p_i}\right)\\
&\quad+\alpha_1\alpha_2\log\left(\frac{\alpha_1\alpha_2}{\alpha_1\alpha_2}\right)\\
&=(1-\alpha_1)(1-\alpha_2)\mathrm{MI}^{ideal} < \mathrm{MI}^{ideal}
\end{aligned} \tag{20}$$

Thus, random failures will lead to less mutual information shared between views.

In summary, we demonstrate that the occurrence of all proposed types of poisoning attacks will decrease the mutual information shared across clients, that is, $MI^{poisoned} < MI^{ideal}$. □

B PROOF OF THEOREM 2

We finish the proof of Theorem 2 by four steps: (1) We define the *information equivalence* relation of the random variables and verify that it is an equivalence relation. (2) We define the *determined by* relation between the information non-equivalent random variables and prove it to be a partial order. (3) We reveal the relationship between conditional entropy and the *determined by* relation. (4) Based on the above discussion, we prove the theorem 2.

B.1 The *Information Equivalence* Relation

Suppose $\hat{\mathcal{X}}$ represents the family of random variables on some sample space Ω, we define the *information equivalence* as follows:

Definition 4. *Two random variables are information equivalence if there exist two functions f and $\bar{f}$ that make $Y = f(X), X = \bar{f}(Y)$ holds, denoted by $X \sim Y$.*

In intuition, the information equivalence of two random variables means the information in the two variables is the same, and the variables can predict each other using a pair of predictor functions. According to the Definition 4, it is not difficult to find that it is an equivalence relation.

B.2 The *Determined by* Relation

After defining the information equivalence relantion $\sim$, we consider the quotient space $\mathcal{X} = \hat{\mathcal{X}}/\sim$. The $\mathcal{X}$ represents the family of all information equivalence classes. For brevity, we still use the representative X to represent the its equivalence class $[X] = \{X' : X' \sim X\}$ in $\mathcal{X}$ when there is no ambiguity. Under such notion, two random variables X and Y represent the same element in $\mathcal{X}$ if $X \sim Y$.

In the space $\mathcal{X}$, we define the "*determined by*" relation as follow:

Definition 5. *We call a random variable X to be determined by Y if there exits a function f such that $X = f(Y)$ holds almost surely, denoted by $X \preceq Y$.*

Since we use the representatives instead of equivalence classes in the definition, we can first prove that the definition of $X \preceq Y$ is irrelevant to the choice of representatives $X \in [X], Y \in [Y]$, i.e. if $X \preceq Y$, then $X' \preceq Y'$ for any $X' \sim X, Y' \sim Y$.

We can easily check that the *determined by* relation is a partial order on $\mathcal{X}$

B.3 Information Properties of "determined by"

To relate the "determined by" relation based on the predictor function with information theory, we proposed the proposition below:

Proposition 1. *Suppose X, Y are two random variables taking values in finite sets $\{x_1, x_2, \ldots, x_n\}$ and $\{y_1, y_2, \ldots, y_m\}$ respectively, then the following four statements are equivalent:*
(1) $X \preceq Y$, i.e. X is determined by Y
(2) $H(X, Y) = H(Y)$
(3) $H(X|Y) = 0$
(4) $I(X; Y) = H(X)$

Proof. The equivalence of statement (2), (3), (4) can be derived directly just by their definitions, here we mainly focus on the equivalence of (1) and (3).

(1) $\Longrightarrow$ (3): $X \preceq Y$ means that there is some function f such that $X = f(Y)$ and $P(x_i|y_j) = \delta_{x_i, f(y_j)}$. Substituting this into the definition of conditional entropy, we can get

$$\begin{aligned} H(X|Y) &= -\sum_{j=1}^{m} P(y_j) \sum_{i=1}^{n} P(x_i|y_j) \log P(x_i|y_j) \\ &= -\sum_{j=1}^{m} P(y_j) \sum_{i=1}^{n} \delta_{x_i, f(y_j)} \log \delta_{x_i, f(y_j)} = 0 \end{aligned}$$

since that $\delta_{x_i, f(y_j)}$ can only be 1 and 0, where $\delta_{x_i, f(y_j)} \log \delta_{x_i, f(y_j)}$ is always 0.

(3) $\Longrightarrow$ (1): First, by definition, we have

$$H(X|Y) = -\sum_{j=1}^{m} P(y_j) \sum_{i=1}^{n} P(x_i|y_j) \log P(x_i|y_j)$$

$$P(x_i|y_j) \geq 0, \forall (i, j); \quad \sum_{i} P(x_i|y_j) = 1, \forall j.$$

We denote that

$$H(X|Y = y_j) = -\sum_{i=1}^{n} P(x_i|y_j) \log P(x_i|y_j) \geq 0$$

$$H(X|Y) = \sum_{j=1}^{m} P(y_j) H(X|Y = y_j).$$

$H(X|Y) = 0$ means that for those $j : P(y_j) \neq 0$, we must have $H(X|Y = y_j) = 0$. And for a fixed y_j, $H(X|Y = y_j)$ can be zero only when it is minimized under constraints.

It is noticed that the entropy above is a strictly concave function of $\{P(x_i|y_j), i = 1, 2, \ldots, n\}$. Thus the minimal point of that strictly concave function must lying on a corner point of its feasible domain, the probability simplex .

Denote the corner point, which is the only choice of $P(x_i|y_j)$ given $H(X|Y) = 0$, by

$$P(x_i|y_j) = \begin{cases} 1, & i = \tilde{i}_j, \text{ for some } \tilde{i}_j, \\ 0, & \text{otherwise.} \end{cases}$$

It is easy to check that $H(X|Y = y_j)$ does take zero at this point. Hence, We can construct the predictor function f as

$$f(y_j) = \begin{cases} x_{\tilde{i}_j}, & P(y_j) \neq 0, \\ x_1, & \text{otherwise.} \end{cases}$$

and the joint distribution

$$P(x_i, y_j) = P(y_j) P(x_i|y_j) \begin{cases} 0, & x_i \neq f(y_j) \text{ or } P(y_j) = 0 \\ P(y), & x_i = f(y_j) \text{ and } P(y_j) = 0. \end{cases}$$

Therefore, It is derived that $P[x_i \neq f(y_j)] = 0$, meaning that $X = f(Y)$ holds almost surely and $X \preceq Y$ gets proved. □

By the anti-symmetry of the "determined by" relation, we can directly get the corollary 3.

Corollary 3. *Suppose X, Y are random variables taking values in finite sets $\{x_1, x_2, \ldots, x_n\}$ and $\{y_1, y_2, \ldots, y_m\}$ respectively, then the following for statements are equivalent:*
(1) $X \sim Y$
(2) $H(X, Y) = H(Y) = H(X)$
(3) $H(X|Y) = H(Y|X) = 0$
(4) $I(X; Y) = H(X) = H(Y)$

B.4 Final Proof of the Theorem 2

By minimizing the ring prediction loss, we can drive the representations to satisfy:

$$H(Z_2 \mid Z_1) = H(Z_1 \mid Z_M) = 0 \tag{21}$$

By the proposition above, we have

$$Z_1 \preceq Z_M \preceq Z_{M-1} \preceq \cdots \preceq Z_2 \preceq Z_1 \tag{22}$$

Since "$\preceq$" is a partial order, by which we can never form a ring, the only possibility is that

$$Z_1 \sim Z_2 \sim \cdots \sim Z_M. \tag{23}$$

This means that the representations of all clients are information equivalent. According to Corollary 3, we have:

$$H(Z^i|Z^j) = 0, \quad \forall i, j = 1, 2, \ldots, M, \ i \neq j \tag{24}$$

Therefore, like the dual prediction loss, ideally optimizing the ring prediction loss can also make the conditional entropy to be zero.