Breaking Cross-View Associations: Byzantine Model Poisoning Attack against Vertical Federated Learning

Jarin Tasneem

Bangladesh University of Engineering and Technology
2005019@ugrad.cse.buet.ac.bd

Abstract

Federated learning (FL) has rapidly emerged as a pivotal framework for cross-silo collaborative training while keeping sensitive data localized, driven by growing data volumes and major privacy concerns. Within this paradigm, vertical federated learning (VFL) enables collaboration among parties holding different features of the same sample space, powering tasks like fraud detection, medical diagnosis, and credit scoring. However, the participation of multiple entities creates new vulnerabilities to malicious interference. One critical yet underexplored threat in VFL is the Byzantine poisoning attack, where an adversary intentionally corrupts training to degrade overall model performance. This work reveals a practical vulnerability showing how a single malicious participant can significantly reduce inference accuracy in a VFL system by breaking cross-view association through feature-space corruption. Our findings emphasize the urgent need for robust, VFL-specific defenses to ensure reliability in collaborative, cross-silo AI systems.

Introduction

Vertical federated learning (VFL) is one of the key frameworks for privacy-preserving AI, enabling organizations to jointly train models without exposing raw data. Unlike horizontal FL, which separates data by users, VFL allows institutions to share different features of the same records, such as hospitals that combine patient vitals with lab data or banks and telecoms that collaborate on fraud detection (Li et al. 2025). However, as reliance on VFL grows, ensuring robustness against adversarial behavior becomes critical. Most existing research focuses on targeted backdoor or privacy attacks, while Byzantine model poisoning, which aims to degrade overall model performance, remains largely unexplored (Cai et al. 2025a). This study proposes a practical untargeted poisoning framework that corrupts feature-space representations to expose how vertically partitioned systems fail even without shared labels. Understanding such attacks is vital; for example, a malicious participant in a joint credit scoring model could subtly corrupt its features, causing systemic misclassification and economic bias.

During training, VFL clients send intermediate embeddings of their feature subset to a central server that typically owns the labels and trains the global model (Cheng et al. 2021). A critical vulnerability arises when a single malicious client manipulates its embeddings and reshapes the joint decision surface. Additionally, the gradients returned to the bottom models may leak information or be altered to influence local training (Cai et al. 2025b; Jin et al. 2021). Previous work has studied both inference and training threats in VFL (Naseri, Han, and Cristofaro 2023; Fu et al. 2022; Yang et al. 2023). Most VFL poisoning research to date has focused on targeted backdoors that implant triggers for specific classes during training. Untargeted Byzantine corruption that degrades overall accuracy is uncommon in VFL. HFL defenses for Byzantine behavior rely on shared feature spaces and update similarity, assumptions that do not hold in VFL and can lead to false positives or missed attacks (Shi et al. 2022; So, Güler, and Avestimehr 2020). Existing VFL detectors that monitor smashed features or temporal gradients handle local deviations but struggle against globally consistent, slow-drift poisoning. Building on feature-space poisoning techniques used for targeted attacks in VFL, our main novelty lies in applying strategic cluster-based embedding swapping to break the cross-view association of feature subsets, causing catastrophic Byzantine model poisoning that is statistically plausible to existing detector, highlighting the need for new protocol-compatible VFL defenses.

Methodology

How cluster-swapping corrupts VFL training. The server maps intermediate local embeddings to server-held labels by leveraging consistent cross-view associations established during training. If a participant systematically alters the correspondence between local clusters and label groups (e.g., by consistently swapping embeddings of cluster A with cluster Z), the server encodes these spurious associations as legitimate patterns. During inference, when correctly aligned embeddings are provided, these learned dependencies are violated, causing the global model to collapse and resulting in severe performance degradation.

Stage 1: Recovering label-aligned clusters

- The attacker trains a semi-supervised local encoder and subsequently clusters the learned embeddings into K groups through SimCLR (Chen et al. 2020) and Gaussian Mixture Models (Cai and Akan 2025).

- Pairwise centroid distances are calculated to rank clusters by separation. The attacker then identifies the most distant cluster pairs for feature swapping, aiming to induce maximal cross-view mismatch at the server.

Stage 2: Strategic cluster-swap poisoning

- For each target cluster i, the adversary replaces its feature embeddings with samples drawn from the most distant cluster j, thereby maximizing the misalignment of cross-view associations.
- This substitution is executed locally within the adversary's feature space prior to transmitting the smashed embeddings, compelling the server to learn from a systematically corrupted cross-view representation.

Why this breaks the cross-view space. Since VFL servers only observe fused embeddings, systematic, label-consistent swaps by the adversary create persistent, spurious cross-view correlations that the server internalizes as legitimate patterns. Unlike isolated outliers, cluster-level swaps fundamentally distort the representation structure.

Experimental Setup

We construct a two-stage experimental pipeline that closely emulates a practical VFL deployment of a classical two-party configuration. The malicious client operates with access to only its own feature subset and a small auxiliary labeled dataset to classify its local data into clusters.

Datasets: Two image datasets (MNIST, FashionMNIST) and a tabular dataset (UCI-HAR) were used to evaluate the proposed pipeline.

Defense: The attack was evaluated against existing VFL defenses, including gradient clipping, smashed-feature reconstruction monitoring, and AE-based anomaly detection.

Baselines: We compared the cluster-guided Byzantine attack against two random poisoning baselines: random cluster-swap and random sample-swap. Additionally, a round-robin swap strategy was evaluated alongside the optimal most-distant swap.

Ablations: We systematically varied: (a) the auxiliary labeled fraction (e.g., 1%, 3%, 5%, 10%) to quantify effects on cluster purity and attack efficacy; and (b) the swap policy to analyze the trade-off between stealth and impact.

Feasibility: All experiments were conducted on a single NVIDIA GeForce RTX 4080 GPU (16 GB VRAM). Larger multi-party parameter sweeps may require small-scale clusters or cloud instances.

Results and Discussion

Clustering Quality: Across all datasets, the clustering achieves strong alignment with ground-truth labels, with accuracies of 86.55% on MNIST, 72.22% on Fashion-MNIST, and 87.12% on UCI-HAR. Leveraging these predicted clusters in the adversary's feature space, the cluster-guided optimal swapping yields a degraded MNIST accuracy of 42.3%, as illustrated in Table 1. A similar performance collapse is observed on FashionMNIST and UCI-HAR.

Defense	MNIST		FashionMNIST		UCI-HAR	
	Acc	Det	Acc	Det	Acc	Det
No Defense	42.3	0	45.7	0	65.6	0
Grad-Norm Clip	42.3	0	45.7	0	65.6	0
AE Anomaly	41.8	0.4	45.6	1.5	64.7	0.7

Table 1: Accuracy (Acc, %) and detection rate (Det, %) under the proposed attack.

Swap Strategy	MNIST Acc (%)
Round-Robin Swap	75.99
Random-Cluster Swap	71.69
Random-Sample Swap	88.71
Optimal Swap (Ours)	46.04

Table 2: MNIST accuracy under different cluster-swap and random-swap baselines.

Robustness Against Existing Defenses: Cluster swapping evades the classical VFL defenses, as shown in Table 1 because each substituted embedding is still a valid in-distribution feature, preserving normal magnitudes and statistics. Existing defense mechanisms, which typically monitor for numeric anomalies such as gradient explosions, reconstruction spikes, or abrupt per-label drift, fail to detect these manipulations as such signatures never manifest. Thus, the attack breaks cross-view alignment without producing the abnormal signatures most VFL defenses are designed to flag.

Baseline Comparisons: Round-robin, random-cluster swap and random-sample swaps remain relatively weak in attack success, as shown in Table 2. In contrast, our optimal cluster-swap strategy is far more damaging, driving accuracy down to 46.04%. Increasing auxiliary label access further refines cluster quality, thereby strengthening the attack. As a result, global accuracy degrades steadily as label access grows; for example, on MNIST, performance drops from 78.2% (at 0.5% access) to 31.1% (at 10% access).

Future Work: Future research directions include designing defenses that explicitly enforce cross-view consistency and extending the attack framework to multi-party settings (e.g., 4, 8, or 16 parties) to investigate coordinated disruption in large-scale VFL deployments. Additionally, evaluating the attack on noisier, real-world datasets will further validate its robustness under complex, imperfect data conditions.

Conclusion

This work uncovers a critically underexplored vulnerability in VFL: an untargeted Byzantine poisoning framework based on strategic cluster-swapping in the feature space. The adversary learns label-aligned clusters and swaps them between the most distant groups to effectively sever cross-view associations during training. By demonstrating how subtle feature-level inconsistencies can silently collapse model performance, this study underscores the urgent need for robust, VFL-specific defenses to ensure the reliability and accountability of federated AI in sensitive real-world deployments.

References

Cai, H.; and Akan, O. B. 2025. Semantic Learning for Molecular Communication in Internet of Bio-Nano Things. *arXiv preprint arXiv:2502.08426*.

Cai, H.; Dong, H.; Wang, H.; Li, K.; and Akan, O. B. 2025a. Graph Representation-based Model Poisoning on Federated LLMs in CyberEdge Networks. *arXiv preprint arXiv:2507.01694*.

Cai, H.; Wang, H.; Dong, H.; Li, K.; and Akan, O. B. 2025b. Graph Representation-based Model Poisoning on the Heterogeneous Internet of Agents. *arXiv preprint arXiv:2511.07176*.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, 1597–1607. PmLR.

Cheng, K.; Fan, T.; Jin, Y.; Liu, Y.; Chen, T.; Papadopoulos, D.; and Yang, Q. 2021. SecureBoost: A Lossless Federated Learning Framework. arXiv:1901.08755.

Fu, C.; Zhang, X.; Ji, S.; Chen, J.; Wu, J.; Guo, S.; Zhou, J.; Liu, A. X.; and Wang, T. 2022. Label inference attacks against vertical federated learning. In *31st USENIX security symposium (USENIX Security 22)*, 1397–1414.

Jin, X.; Chen, P.-Y.; Hsu, C.-Y.; Yu, C.-M.; and Chen, T. 2021. Cafe: Catastrophic data leakage in vertical federated learning. *Advances in neural information processing systems*, 34: 994–1006.

Li, K.; Liang, Y.; Yuan, X.; Ni, W.; Crowcroft, J.; Yuen, C.; and Akan, O. B. 2025. A novel framework of horizontal-vertical hybrid federated learning for edgeIoT. *IEEE Networking Letters*.

Naseri, M.; Han, Y.; and Cristofaro, E. D. 2023. BadVFL: Backdoor Attacks in Vertical Federated Learning. arXiv:2304.08847.

Shi, J.; Wan, W.; Hu, S.; Lu, J.; and Zhang, L. Y. 2022. Challenges and approaches for mitigating byzantine attacks in federated learning. In *2022 IEEE International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)*, 139–146. IEEE.

So, J.; Güler, B.; and Avestimehr, A. S. 2020. Byzantine-resilient secure federated learning. *IEEE Journal on Selected Areas in Communications*, 39(7): 2168–2181.

Yang, R.; Ma, J.; Zhang, J.; Kumari, S.; Kumar, S.; and Rodrigues, J. J. 2023. Practical feature inference attack in vertical federated learning during prediction in artificial Internet of Things. *IEEE Internet of Things Journal*, 11(1): 5–16.