A Unified Defense Framework Against Membership Inference in Federated Learning via Distillation and Contribution-Aware Aggregation

Liwei Zhang*, Linghui Li*✉, Xiaotian Si*, Ziduo Guo*, Xingwu Wang*, Kaiguo Yuan*, Bingyu Li†
*Key Laboratory of Trustworthy Distributed Computing and Service (MoE),
Beijing University of Posts and Telecommunications
†School of Cyber Science and Technology, Beihang University
*{lw_z, lilinghui, sixt, duoduoboom, xingwuwang, flyingdreaming}@bupt.edu.cn,
†libingyu@buaa.edu.cn

***Abstract*—Federated learning enables decentralized model training without exposing raw data, making it a promising paradigm for privacy-preserving machine learning. However, it remains vulnerable to membership inference attacks (MIAs), where adversaries infer whether a specific data point is included in the training set, posing serious privacy risks and compromising data locality. Existing defenses against MIAs suffer from significant limitations: some incur substantial performance degradation, while others fail to provide protection against both passive and active attack vectors. To address these challenges, in this paper, we propose a unified defense framework that simultaneously mitigates both passive and active MIAs in federated learning, while preserving the utility of the target model. First, we incorporate a modified entropy regularization during teacher model training to enhance uncertainty on member data, offering stronger resistance to inference attacks than standard regularization. Second, we utilize a Conditional Variational Autoencoder (CVAE) to generate class-conditional synthetic data for supervised student training, which avoids direct exposure of sensitive data and provides better utility than unlabeled alternatives. Finally, we design a contribution-aware aggregation strategy that adjusts the influence of local models based on their utility, mitigating the impact of malicious clients during model aggregation. Experimental results on four benchmark datasets show that the proposed method significantly reduces the success rate of various membership inference attacks, outperforming existing state-of-the-art defenses. Moreover, it consistently maintains high model accuracy, demonstrating its practicality for real-world federated learning deployments [1].**

I. Introduction

Federated Learning (FL) is a prominent distributed machine learning framework that has been widely adopted in various domains, such as personalized recommendation on smartphones [16], [41], healthcare analytics [40], [32], and financial fraud detection [23], [4], [39]. Its core advantage lies in enabling decentralized model training by sharing only model updates, thereby avoiding direct access to user data and enhancing privacy protection [21]. However, despite its privacy-preserving design, FL is still susceptible to privacy threats. In particular, the model parameters exchanged during training may inadvertently leak sensitive information, exposing the system to membership inference attacks (MIAs). In such attacks, adversaries attempt to infer whether a specific data sample is included in the training set by analyzing the model's outputs or internal parameters, thereby compromising participant privacy. This threatens the foundational principle of FL that data remains decentralized and never shared. As FL becomes increasingly deployed in real-world applications, developing robust defenses against MIAs has become a critical and urgent research challenge.

MIAs exploit the differences in how a model handles training data versus unseen data [11]. The key to these attacks lies in the fact that training data typically results in higher confidence or more accurate predictions, whereas predictions on unseen data are often associated with greater uncertainty. MIAs can be broadly categorized into passive [35], [27], [42] and active attacks [46], [38], [6] , depending on the adversary's level of interaction with the training process. Passive attacks rely solely on analyzing observable behaviors of the target model, without intervening in the training process. Although limited in access, these attacks still pose a serious threat, especially as model complexity and training data scale increase. In contrast, active attacks manipulate the training process by injecting poisoned data or altering local model updates. These interventions amplify the leakage of membership information by inducing distinguishable behaviors between member and non-member data. As a result, active attacks are often more damaging, compromising not only privacy but also the integrity of the training pipeline.

Many studies have explored defense mechanisms against passive MIAs, proposing various techniques such as differential privacy [1], regularization [28], [44], model masking [17], [7], and knowledge distillation [34], [37]. Differential

✉Corresponding author.

[1]Our code is available at https://github.com/misu742694/KD-MIAD.

Network and Distributed System Security (NDSS) Symposium 2026
23-27 February 2026, San Diego, CA, USA
ISBN 979-8-9919276-8-0
https://dx.doi.org/10.14722/ndss.2026.230413
www.ndss-symposium.org

privacy offers formal privacy guarantees by adding calibrated noise to model outputs or gradients to obscure the contribution of individual training samples. However, it often leads to a substantial degradation in model utility. Regularization-based methods aim to reduce model overfitting, thereby lowering the success rate of membership inference. Yet, their effectiveness diminishes when facing strong or adaptive adversaries. Model masking attempts to obfuscate the model's output probability distribution to mislead attackers, but it is generally ineffective against white-box threats in which the attacker has access to internal model parameters. Shejwalkar et al. [34] introduce knowledge distillation as a means to mitigate privacy leakage by training student models using unlabeled data, thereby avoiding direct exposure of sensitive training samples. However, the absence of label supervision often leads to significant performance degradation in the resulting student model.

In FL, the decentralized nature of training introduces new vulnerabilities, making it easier for adversaries to conduct active MIAs by manipulating the training process. To address this, prior work [25] proposes client-side defense strategies that detect and filter suspicious data based on local anomaly detection. This method assumes that clients have full control over their local datasets and can identify potentially poisoned samples before they influence the global model. However, when the attacker is a malicious client who actively injects poisoned data during local training, such self-supervised detection schemes become ineffective. Other research [19] focuses on server-side defenses during the aggregation phase by calculating a set of statistical metrics to identify and exclude potentially poisoned local models. While these approaches can mitigate the effects of some active attacks, they often rely on complex computations and clustering procedures, making it difficult to scale to large federated learning deployments.

Existing defense methods primarily focus on addressing either passive or active attacks. However, attackers can initiate attacks at various stages of federated learning using different strategies, while defense mechanisms must provide protection across all stages. In this paper, we propose a unified defense framework for federated learning that integrates knowledge distillation and contribution-aware aggregation to jointly mitigate both passive and active membership inference attacks. Our framework consists of three core components: teacher model training with improved entropy, student model training with CVAE distillation and contribution-aware aggregation. Our proposed defense method provides a joint protection from multiple perspectives.

First, we incorporate an entropy-based regularization into the teacher model's training loss to enhance prediction uncertainty and mitigate overfitting, thereby reducing membership leakage. We observe that member samples typically produce lower entropy predictions, which makes them more vulnerable to inference attacks. However, conventional regularization fails to differentiate between confidently correct and confidently incorrect predictions, limiting its ability to effectively reflect true uncertainty. To address this, we add a modified predictive entropy term to the loss function, encouraging the teacher model to produce more uncertain outputs for training data and improving its resistance to membership inference.

Second, we employ a Conditional Variational Autoencoder (CVAE) to generate class-conditional synthetic data that approximates the distribution of private data. Compared to unlabeled or randomly sampled approaches, our method captures richer semantics and produces more representative samples aligned with specific labels, enabling supervised student training without exposing private data. Since the synthetic samples do not correspond to real member data, the risk of membership inference is significantly reduced. Moreover, the use of labeled synthetic data preserves the benefits of supervised learning, allowing the student model to maintain strong task performance through effective knowledge transfer from the teacher.

Finally, we implement a contribution-aware aggregation strategy that dynamically adjusts the weights of local model updates based on their estimated contribution to global performance. By down-weighting or excluding low-quality or adversarial updates, this mechanism effectively mitigates the impact of active attacks. In addition to enhancing the robustness of the global model, it also promotes fairness and consistency across heterogeneous clients. Through these coordinated defense steps, we effectively protect against both passive and active attacks, thereby safeguarding participant privacy while maintaining the accuracy of the global model.

The contributions of this paper are as follows:

- We propose a novel privacy-preserving training framework for federated learning that jointly mitigates passive and active membership inference attacks. Our approach provides end-to-end protection across different stages of federated learning, while maintaining high model utility.
- We utilize a distillation strategy to mitigate passive attacks by combining entropy-regularized teacher training with CVAE-generated labeled synthetic data, enabling supervised student training that reduces overfitting and prevents raw data exposure.
- We introduce a contribution-aware aggregation mechanism against active attacks that dynamically adjusts local update weights based on their estimated utility to the global model, effectively suppressing malicious updates from adversarial participants.
- We conduct comprehensive evaluations on four benchmark datasets, demonstrating that our method effectively mitigates a wide range of state-of-the-art membership inference attacks. Compared with existing defenses, our framework achieves stronger protection while preserving the target model's utility with minimal performance degradation.

II. Related Work

A. Membership Inference Attacks

Membership inference attacks (MIAs) have emerged as a key focus in model privacy research. We categorize MIAs into two main types: passive and active attacks.

Passive Membership Inference Attacks. Shokri et al. [35] first introduce membership inference attacks by training

shadow models to replicate the target model's behavior and using a binary classifier for membership prediction. Salem et al. [33] simplify this approach by using a single shadow model. Yeom et al. [42] theoretically analyze the link between overfitting and privacy leakage, and propose effective MIA methods. Additionally, some studies use threshold-based methods to distinguish members from non-members. Song et al. [36] enhance cross-entropy-based attacks by setting per-class thresholds for membership inference. Hui et al. [14] determine membership by assessing the distance between members and non-members. Carlini et al. [5] propose LiRA, a likelihood ratio-based membership inference attack that achieves stronger performance under low false positive rates. Liu et al. [22] reveal that robustness differences, when exploited through explainability techniques, can also lead to membership leakage, highlighting the evolving diversity of MIAs.

Federated learning is also susceptible to membership inference attacks. Melis et al. [27] show that updates from participants in federated learning can inadvertently leak information about their local data. Nasr et al. [29] propose a membership inference attack for federated learning that uses intermediate model outputs and model parameters to determine whether an individual sample is used in training the model. Zhang et al. [45] indicate that when the attacker is an insider, variations in bias during the federated learning process can effectively distinguish between members and non-members. Zhu et al. [47] reveal that shared gradients may enable reconstruction of training data, leading to membership leakage.

Active Membership Inference Attacks. Active MIAs aim to amplify the behavioral differences between member and non-member data by subtly modifying local training dynamics, thereby enabling accurate membership inference without noticeably degrading model utility. In many cases, these attacks exploit variations in prediction confidence or gradient sensitivity, rather than inducing misclassification, to distinguish members from non-members. Florian et al. [38] design a log-likelihood-based test using shadow models, where poisoned data is introduced into half of the models to amplify membership signals. Chen et al. [6] propose dirty-label and clean-label poisoning attacks to increase leakage in both centralized and transfer learning settings, though the clean-label method degrades when feature extractor parameters are unfixed. Zhang et al. [46] propose a poisoning membership inference attack against adversarial Byzantine robust aggregation in federated learning. They directly modify the local updates and design a gradient masking method to make malicious updates appear benign, thereby evading robust aggregation.

B. Defenses Against Membership Inference Attacks

Defenses Against Passive Attacks. Common defense mechanisms against membership inference attacks include differential privacy [1], [13], regularization [28], [44], and model masking [17], [7]. Differential privacy offers formal guarantees for individual data protection. Methods like DP-SGD [1] and PATE [30] leverage it to ensure strong privacy, but often at the cost of model performance. Regularization methods mitigate membership inference risk by reducing overfitting. Nasr et al. [28] introduce adversarial regularization, a min-max training scheme that minimizes both prediction loss and attack success. Mixup+MMD [44] reduces leakage by penalizing output differences between members and non-members. Li et al. [20] propose MIST, a subspace learning-based defense that avoids overfitting on vulnerable instances while preserving accuracy. Generally, regularization alone offers limited privacy protection and often requires complementary techniques. Model masking perturbs the target model's outputs to resist inference. Jia et al. [17] propose MemGuard, which adds calibrated noise to confidence scores to mislead attackers. Chen et al. [7] present HAMP, using high-entropy soft labels and entropy regularization to reduce overconfidence and attack success. Hu et al. [12] introduce Membership Cleanser, a query preprocessing defense that removes membership signals without altering model training or inference. While effective against black-box attacks, these methods are generally ineffective against white-box attacks that exploit internal parameters.

Furthermore, knowledge distillation has emerged as a promising defense strategy for mitigating membership inference risks, as it trains substitute models without directly exposing raw data. Shejwalkar et al. [34] propose DMP, which distills knowledge into student models using unlabeled data to protect membership privacy. Tang et al. [37] enhance privacy through self-distillation across multiple data partitions. While existing approaches mitigate membership inference risks, they often suffer from reduced accuracy due to unsupervised training and high overhead from multi-model training. Moreover, without explicit control over information transfer, student models may inherit the overconfidence of teacher models and thus retain membership signals, resulting in residual privacy leakage. In our method, we combine entropy-regularized teacher training with CVAE-based synthetic data generation. The former reduces overfitting and increases prediction uncertainty on members, while the latter supports supervised student training without exposing raw data. Together, these strategies mitigate MIAs while preserving strong utility and scalability.

Defenses Against Active Attacks. Existing defenses against active membership inference attacks primarily focus on improving robustness during the aggregation phase in federated learning. LoDen [25] detects and removes suspicious samples at the client side to prevent poisoning-based membership inference, while MESAS [19] employs multiple statistical metrics to identify and filter adversarial updates. However, these methods either fail to generalize to adaptive adversaries or incur high computational overhead due to extensive metric calculations and clustering. In contrast, our framework addresses these limitations during the aggregation phase by dynamically allocating aggregation weights based on each participant's contribution to the global model, thereby reducing the impact of suspicious local models and ensuring the accuracy of the global model.

III. Preliminaries

A. Federated Learning

Federated learning addresses privacy concerns by decentralizing model training to local devices, thus avoiding centralized access to raw data [2]. The federated learning process involves global initialization, local training, and model aggregation. During local training, participant i updates the model parameters θ_i^{t+1} using private data by solving the following optimization problem:

$$\theta_i^{t+1} = \arg\min_{\theta} \mathcal{L}(\theta_i^t), \tag{1}$$

where $\mathcal{L}$ is the loss function. In the model aggregation stage, participants upload their updated local model parameters to a central aggregator, which computes the global model by averaging all local parameters:

$$\theta_{global}^* = \frac{1}{N} \sum_{i=1}^{N} \theta_i^{t+1}, \tag{2}$$

where N is the total number of participants, θ_{global}^* denotes the aggregated global model parameters.

B. Membership Inference Attacks

Membership inference attacks aim to determine whether a given instance x is part of the training data by observing model outputs or other information [35]. The attack can be formulated as:

$$\mathcal{A}(f, x) \rightarrow \{0, 1\}, \tag{3}$$

where $\mathcal{A}$ is the attacker's inference function, with "1" indicating a member and "0" indicating a non-member. Membership inference attacks can be classified as either passive attacks, where the attacker observes model outputs without any intervention [5], or active attacks, where the attacker manipulates data or model parameters to infer membership [46]. Our work focuses on simultaneously defending against both attacks to enhance the privacy and robustness of federated learning.

C. Knowledge Distillation

Knowledge distillation transfers knowledge from a complex teacher model to a simpler student model [9] . In this process, the larger teacher model generates outputs through supervised learning, and these outputs are used as target labels to train the smaller student model. The student model learns by minimizing the difference between its own predictions and the predictions of the teacher model. Specifically, the knowledge distillation objective is to train the student model by solving the following optimization problem:

$$\mathcal{L}_{KD}(\theta_S) = \gamma \cdot \mathcal{L}_{CE}(\theta_S) + (1 - \gamma) \cdot \mathcal{L}_{KL}(\theta_S, \theta_T), \tag{4}$$

where $\mathcal{L}_{CE}$ is the cross-entropy loss, $\mathcal{L}_{KL}$ is the Kullback-Leibler divergence, which measures the difference between the outputs of the student model and the teacher model, and γ is a balancing factor. θ_T and θ_S represent the parameters of the teacher model and the student model. Through this process, the student model can achieve performance close to that of the teacher model. In our approach, knowledge distillation is primarily utilized to protect private data and defend against MIAs. We choose a consistent architecture for both the student and teacher models to effectively extract useful information from the teacher model while ensuring data privacy.

D. Threat Model

Adversary. The objective of the attacker is to infer whether a specific data point belongs to the training dataset, thereby compromising the privacy of other participants. By analyzing the model's outputs or training process, the attacker seeks to identify training data points and carry out a membership inference attack. Based on the knowledge and capabilities of the attacker, we consider two broad categories of attackers in our threat model: passive attackers and active attackers.

- **Passive Attackers.** Passive attackers do not interfere with the training process but attempt to infer membership information by observing the model's outputs or training progress. We further distinguish between two types of passive attackers:
  - *External Passive Attackers [36], [5]:* External attackers have access only to the outputs generated by querying the global model and are unable to access the model's structure, parameters, or local data. They also lack visibility into the training process of other clients. We assume that these attackers possess both a set of training data and test data and are allowed unlimited queries to the global model.
  - *Internal Passive Attackers [45]:* These attackers are honest-but-curious clients participating in the federated learning process. They have full access to the global model's parameters and the training process. In addition to observing the global model's outputs, they can also access changes in the global model's parameters. We assume that internal attackers can obtain the latest state of the global model at each training epoch and leverage this information to launch their attacks.
- **Active Attackers.** Active attackers [46], [6] are malicious clients who have full control over their local data and can modify the parameters of their local models during training. These attackers can conduct data poisoning attacks by injecting malicious samples into their local datasets, as well as perform model poisoning attacks by manipulating updates to their local models. We assume that these attackers can access the global model at each epoch and initiate poisoning attacks at any time.

Defender. The defender's objective is to safeguard the privacy of participants in the federated learning system, preventing attackers from exploiting MIAs to access sensitive data, while maintaining the accuracy and robustness of the global model. We define the defender as a training configuration under an honest federated setting, rather than a specific entity with absolute control over the learning process or client behaviors. The defender coordinates the training

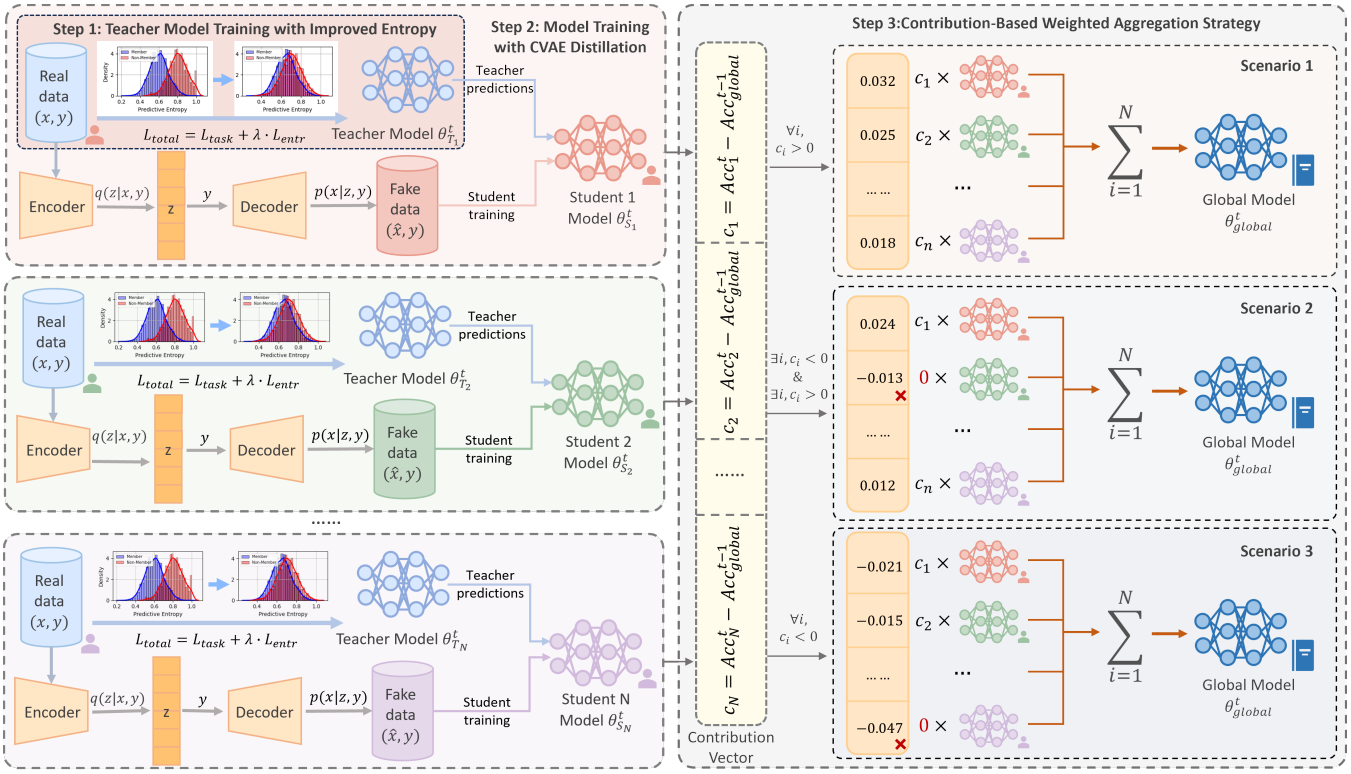


Fig. 1: The architecture of our proposed defense framework against both passive and active membership inference attacks in federated learning. The framework consists of three sequential components: (1) Teacher model training with improved entropy, where an entropy-regularized loss is applied to reduce overfitting and obscure membership signals; (2) Student model training with CVAE distillation, where conditional variational autoencoders generate synthetic labeled data to train student models without exposing private data; and (3) Contribution-aware aggregation, which computes each client's contribution to global accuracy and dynamically adjusts aggregation weights to suppress poisoned updates.

behavior of both honest clients and the server, specifying defense procedures through protocol rules, without directly accessing or controlling clients' local data, models, or training processes, thus strictly adhering to privacy-preserving principles. It should be noted that active attackers may choose not to follow this setting and can launch attacks by injecting malicious data or tampering with local models.

IV. Method

In this section, we first introduce the overall framework of the defense method we propose, and then provide a detailed description of the specific implementation of these three steps.

A. Overview of the Defense

Our proposed defense method aims to protect the privacy of participants in the federated learning system, prevent membership inference attacks, and ensure the accuracy and robustness of the global model. The overall defense framework is shown in Fig. 1, which consists of three key steps: teacher model training with improved entropy, student model training with CVAE distillation and contribution-aware aggregation strategy.

First, we introduce an entropy-based regularization term into the loss function during the training of the local teacher model. We observe that member samples often produce overly confident predictions, leading to significantly lower prediction entropy compared to non-member samples. This overconfidence increases their susceptibility to membership inference attacks. To address this, we incorporate the regularization to increase the predictive uncertainty on training data, thereby reducing the model's tendency to overfit and narrowing the distinguishability between member and non-member instances.

Second, we train a CVAE to generate labeled synthetic data that approximates the class-conditional distribution of private data. These samples are used to train student models in a fully supervised manner. This design is motivated by two critical observations: (1) directly exposing private data during the distillation process increases the risk of membership inference, and (2) existing distillation approaches based on unlabeled data often suffer from noticeable performance degradation due to the lack of supervision [34], [31]. By leveraging CVAE-generated labeled data, our method enables accurate knowledge transfer from the teacher model while eliminating the need for raw data, enhancing both privacy and utility.

Finally, we implement a contribution-aware aggregation strategy to dynamically adjust the weights of local model updates based on their individual contributions to the global

model. This design is motivated by the vulnerability of federated learning to active attacks, where adversarial clients may degrade global performance by submitting harmful updates. Unlike traditional aggregation schemes such as FedAvg [26] and Krum [3], which treat all client updates equally or rely solely on distance metrics, our method evaluates each client's contribution and assigns higher weights to those with positive utility. This enables the system to disregard unreliable updates, enhancing the robustness of the global model against adversarial behaviors without compromising overall performance.

In our framework, each client maintains both a teacher model and a student model, which share the same architecture with the global model to ensure compatibility during aggregation. The teacher model is solely involved in local training to produce soft labels, which guide the student model in learning from both the synthetic data and the teacher's output. Only the student model is uploaded to the server for aggregation, and the global model is updated based on the aggregated student models. This design enables local privacy-preserving training through teacher-student distillation while ensuring seamless integration with the standard federated aggregation process. In the following sections, we provide a detailed description of the implementation of each step.

B. Teacher Model Training with Improved Entropy

In our defense approach, the training of the teacher model is the first step in the distillation process. Its primary objective is to enhance the model's predictive uncertainty over the training data, thus increasing the difficulty for attackers in inferring the membership of data points. By optimizing the teacher model's training in this manner, we aim to reduce the model's overfitting to the training data. This helps prevent the model from becoming overly confident in its predictions for member data, making it challenging for an attacker to determine whether a particular data point belongs to the training set based on the model's outputs.

In traditional knowledge distillation methods, teacher models are typically trained by minimizing the standard cross-entropy loss, aimed at aligning the model's predictions as closely as possible with the true labels. However, this approach has a limitation: it may still allow the teacher model to transfer its high confidence in member data to the student model, leading to potential privacy leaks of membership information [15]. Specifically, teacher models tend to predict member samples with higher confidence, resulting in lower entropy, while predictions for non-member samples are more uncertain, exhibiting higher entropy.

To address this issue, we incorporate prediction entropy into the teacher model's training objective by designing a loss function that combines the standard task loss with an entropy-based regularization term, which encourages higher uncertainty on member data. Specifically, the total new loss function can be expressed as:

$$L_{\text{total}} = L_{\text{task}} + \lambda \cdot L_{\text{entr}}, \tag{5}$$

where L_{task} is the cross-entropy loss for the primary task, measuring the model's predictive accuracy on the training data. L_{entr} represents the average modified predictive entropy over the training samples, which encourages the model to maintain higher uncertainty and prevents overfitting to member data. The coefficient λ is a hyperparameter that balances the trade-off between task performance and privacy regularization.

While incorporating entropy helps mitigate overfitting, the standard entropy calculation has notable limitations. For instance, a correctly classified sample with a confidence of 1 yields zero entropy, as does an incorrectly classified sample with the same level of confidence. As a result, standard entropy measures fail to reliably capture the predictive uncertainty that differentiates member from non-member samples. To overcome this limitation, we adopt an improved predictive entropy formulation proposed by Song et al. [36], which modifies the entropy computation to better reflect differences between training and testing data. Given an input (x, y), where x denotes the input feature and y the ground-truth label, and let $F(x)$ be the output probability vector produced by the model, the improved predictive entropy is computed as:

$$\begin{aligned} M_{\text{entr}}(F(x), y) = &- (1 - F(x)_y) \cdot \log(F(x)_y) \\ &- \sum_{k \neq y} F(x)_k \cdot \log(1 - F(x)_k), \end{aligned} \tag{6}$$

where $F(x)_y$ denotes the predicted probability for the true class y, and $F(x)_k$ denotes the predicted probability for class $k \neq y$. This formulation penalizes both overconfident correct predictions and misclassifications by amplifying uncertainty across the full output distribution.

To regularize the model, we compute the mean modified entropy over the local training set. Let $\mathcal{D}_i$ be the set of the private training samples of participant i. The entropy regularization term is defined as:

$$L_{\text{entr}} = \frac{1}{|\mathcal{D}_i|} \sum_{(x_k, y_k) \in \mathcal{D}_i} M_{\text{entr}}(F(x_k), y_k). \tag{7}$$

This term is then incorporated into the overall training objective to increase the predictive uncertainty on member samples, thereby mitigating the risk of membership inference.

C. Student Model Training with CVAE Distillation

The training of the student model constitutes the second step of the distillation process and aims to capture the predictive patterns of the teacher model while minimizing direct reliance on private data, thereby enhancing participant privacy. Prior studies [34] often employ synthetic unlabeled data for unsupervised training, but such approaches typically lead to substantial degradation in student model performance compared to the teacher model. To overcome this limitation, we introduce a Conditional Variational Autoencoder (CVAE) to generate labeled synthetic data similar to the distribution of the private data for supervised student model training. Specifically, the class label is incorporated into both the encoder and decoder, guiding the generation process to ensure that the synthetic samples are semantically aligned with the desired category.

To achieve this, the CVAE is designed to learn the joint distribution $p(x \mid y)$ of the input data and its corresponding labels. The training process is achieved by minimizing the evidence lower bound loss [18], which consists of two components: a reconstruction loss that measures the similarity between the generated and original data, and a Kullback–Leibler (KL) divergence term that regularizes the latent space towards a standard normal distribution. The loss can be expressed as:

$$L_{\text{CVAE}} = \mathbb{E}_{q(z|x,y)} \left[-\log p(x \mid z, y) \right] + D_{\text{KL}} \left(q(z \mid x, y) \parallel p(z) \right), \tag{8}$$

where $q(z \mid x, y)$ is the approximate posterior learned by the encoder, $p(x \mid z, y)$ is the decoder output, and D_{KL} is the KL divergence measuring the difference between the learned latent distribution and a standard normal distribution in CVAE. After training, the CVAE can generate synthetic samples for any given label. The generation process involves first sampling from a standard normal distribution $z \sim \mathcal{N}(0, I)$ and then passing the sampled latent vector along with a label embedding y into the decoder to generate a synthetic sample:

$$\hat{x} = p(x \mid z, y). \tag{9}$$

The inclusion of the class label ensures that the generated samples remain consistent with the target category, while the randomness introduced by the latent variable enables the generation of diverse samples even within the same class. As a result, the synthetic data captures the statistical properties of the original dataset without revealing specific private instances, thereby mitigating privacy risks.

During the student model training phase, we use the labeled synthetic data generated by the CVAE for supervised learning, effectively replacing the original private data to reduce the risk of privacy leakage. For each participant, the CVAE generates synthetic samples $\hat{x}_k$ along with corresponding labels y_k for every data class. Subsequently, the student model is trained using the standard cross-entropy loss function as part of the distillation loss. This loss function is defined as:

$$L_{\text{student}} = -\sum_{k=1}^{M} y_k \log f_{S_i}(\hat{x}_k), \tag{10}$$

where $f_{S_i}(\cdot)$ denotes the output probability distribution of the participant i student model, and M represents the number of synthetic training samples.

To further enhance the learning effectiveness of the student model and to retain as much of the useful knowledge from the teacher model as possible, we incorporate a knowledge distillation loss based on soft labels. During the training process, the student model leverages both the ground-truth labels from the synthetic data generated by the CVAE and the soft labels provided by the teacher model. The distillation process utilizes the teacher model's softened probability distribution $f_{T_i}(\hat{x}_k, \tau)$, which is adjusted by the temperature parameter τ to balance the influence of synthetic and soft labels. The final distillation loss is formulated as:

$$L_{\text{distill}} = \gamma L_{\text{student}} + (1-\gamma)\tau^2 \sum_{k=1}^{M} D_{\text{KL}} \left(f_{T_i}(\hat{x}_k, \tau) \parallel f_{S_i}(\hat{x}_k, \tau) \right), \tag{11}$$

where γ is a balancing coefficient between the true labels and the soft labels, τ is the temperature parameter that smooths the teacher model's predictions, and D_{KL} measures the divergence between the teacher and student model predictions in the distillation phase.

By leveraging this design, the labeled synthetic data generated by the CVAE is utilized for supervised training of the student model, effectively eliminating the need for direct access to raw private data. Simultaneously, the incorporation of knowledge distillation loss facilitates the efficient transfer of predictive behavior from the teacher model to the student model. This dual mechanism substantially mitigates the risk of privacy leakage associated with membership inference attacks. Furthermore, since the synthetic data are generated to closely approximate the underlying distribution of the original private data, the performance degradation of the student model remains minimal. Overall, this approach achieves a favorable trade-off between privacy preservation and model utility, ensuring both effective defense and high task performance.

D. Contribution-Aware Aggregation Strategy

In federated learning systems, malicious clients can intentionally manipulate their local data or model updates to launch active attacks that amplify membership inference risks. By injecting carefully crafted poisoning updates into the training process, these adversaries can steer the global model towards behaviors that increase the distinguishability between member and non-member data, thereby exacerbating privacy leakage. To address this threat, we propose a contribution-aware aggregation strategy designed to reduce the negative impact of malicious clients and enhance target model robustness.

Our approach evaluates the quality of each client's update on the server side before global aggregation. At each communication round t, the server retains the global model parameters from the previous round, denoted as $\theta_{\text{global}}^{t-1}$, and estimates the contribution of each participant's update to the overall model performance. Specifically, for each participant i, the server temporarily applies the local update to the global model and evaluates the resulting accuracy on a held-out validation set. The contribution c_i of participant i is computed as:

$$c_i = \text{Acc}_i^t - \text{Acc}_{\text{base}}^{t-1}, \tag{12}$$

where Acc_i^t is the accuracy of the global model after incorporating participant i's update, and $\text{Acc}_{\text{base}}^{t-1}$ is the baseline accuracy of the global model from the previous epoch $t-1$. Based on the contribution of each participant, we assign weighted importance to their gradients and incorporate them into the global model aggregation process.

In the aggregation step, we consider three possible scenarios. The first scenario occurs when all participants contribute

positively, meaning their local updates have a positive impact on the global model. In this case, we perform weighted aggregation based on each participant's contribution. The weighted aggregation of the global gradient is given by:

$$\theta^t_{\text{global}} = \theta^{t-1}_{\text{global}} + \sum_{i=1}^{N} \frac{c_i}{\sum_{j=1}^{N} c_j} \cdot \Delta\theta^t_{S_i}, \tag{13}$$

where $\Delta\theta^t_{S_i} = \theta^t_{S_i} - \theta^{t-1}_{\text{global}}$ denotes the local update of participant i, and c_i is the estimated contribution of participant i to the global model. This weighted update ensures that each participant's influence on the global model reflects their actual contribution.

The second scenario occurs when some participants have positive contributions while others have negative contributions. In this case, we discard the negative contributions and aggregate only the gradients from the participants with positive contributions. This approach prevents harmful updates from malicious clients from negatively impacting the global model. The aggregation rule is given by:

$$\theta^t_{\text{global}} = \theta^{t-1}_{\text{global}} + \sum_{i:c_i>0} \frac{c_i}{\sum_{j:c_j>0} c_j} \cdot \Delta\theta^t_{S_i}, \tag{14}$$

where the summation is taken over the subset of N participants with strictly positive estimated contributions. This selective aggregation strategy effectively filters out potentially malicious or unhelpful updates, preserving the global model integrity.

The third scenario arises when all participants exhibit negative contributions, indicating that their local updates collectively degrade the global model. To prevent performance deterioration, we exclude the top-n participants with the lowest contributions and perform aggregation over the remaining participants. The global model update is defined as:

$$\theta^t_{\text{global}} = \theta^{t-1}_{\text{global}} + \sum_{i\notin\mathcal{K}} \frac{c_i}{\sum_{j\notin\mathcal{K}} c_j} \cdot \Delta\theta^t_{S_i}, \tag{15}$$

where $\mathcal{K}$ denotes the index set of the top-n participants with the smallest contributions. By removing the least helpful updates, this strategy enhances the robustness of the aggregation process even in fully adversarial or degraded training rounds. The complete federated training workflow of our proposed defense is summarized in Algorithm 1.

Algorithm 1 Federated Training Workflow of the Proposed Defense Framework

Input: Number of rounds E, number of clients N, private data $\{\mathcal{D}_i\}_{i=1}^N$
Output: Final global model θ^E_{global}
Initialize global model θ^0_{global}
for each round $t = 1, 2, \ldots, E$ **do**
  Broadcast $\theta^{t-1}_{\text{global}}$ to all clients
  for each **Client** $i = 1, 2, \ldots, N$ **in parallel do**
    Train teacher model $\theta^t_{T_i}$ with joint loss on $\mathcal{D}_i$
    Train CVAE on $\mathcal{D}_i$ to learn distribution $p(x|y)$
    Generate labeled synthetic data $(\hat{x}, y)$ using CVAE
    Train student model $\theta^t_{S_i}$ using $(\hat{x}, y)$ with soft labels from teacher model $\theta^t_{T_i}$
    Upload local student model $\theta^t_{S_i}$ to **Server**
  end for
  Server evaluate c_i of $\theta^t_{S_i}$ on held-out validation set
  Aggregate updates with $\Delta\theta^t_{S_i} \propto c_i$ to obtain θ^t_{global}
end for
return θ^E_{global}

V. Experimental Setup

A. Datasets and Models

We conduct experiments on four datasets: Location30[2], Purchase100[3], Texas100[4], and Cifar10 [32]. Following prior works [7], [25], [35], we adopt a neural network with fully connected layers of 512, 128, and 30 units for Location30. For Purchase100 and Texas100, we adopt networks with layer sizes [600, 1024, 256, 100] and [6169, 1024, 512, 256, 100], respectively, all using ReLU activations. For Cifar10, we employ ResNet-20. Each client maintains both a teacher model and a student model, sharing the same architecture as the global model for compatibility in aggregation. The CVAE consists of an encoder and a decoder, each with one hidden layer of 512 units. The latent dimension is set to 20 for Location30, Purchase100 and Cifar10, and 50 for Texas100.

B. Attacks

To comprehensively evaluate the effectiveness of our proposed defense framework, we consider a range of membership inference attacks from both passive and active perspectives. Specifically, we include four representative passive attacks: Prediction [35], Bias [45], Entropy [36], and LiRA [5], as well as two active attacks: AgrEvader [46] and AMP [6]. Detailed descriptions of these attacks are provided in Appendix E. We assume that attackers have access to various parameters from the model training process, can obtain labeled data that follows the same distribution as the training dataset, and face no query restrictions. Notably, attackers are not required to construct shadow models to mimic the target model's behavior, thus relaxing the constraints on the attacker to simulate more realistic and complex attack scenarios.

C. Defense Settings

We consider a horizontal federated learning scenario where multiple clients collaborate with a server to complete the training task. We consider 10 participants (see Section VI-K for larger-scale settings with up to 500 participants), each owning independently and identically distributed (IID) data, with all participants being selected for training in each round. The batch size is set to 64, the optimizer is Adam, and the learning rate is 0.001. The model is trained for 50 epochs. During the training of the teacher model, the entropy loss parameter is set to 0.20. In the training of the student model, each participant generates synthetic data with the same number and distribution

[2]https://sites.google.com/site/yangdingqi/home/foursquare-dataset
[3]https://www.kaggle.com/c/acquire-valued-shoppers-challenge/data
[4]https://www.dshs.texas.gov/thcic/hospitals/Inpatientpudf.shtm

TABLE I: Attack accuracy and model accuracy before (w/o) and after (w) applying our defense.

Datasets	ACC_a												ACC_m		ΔACC_m
	Prediction [35]		Bias [45]		Entropy [36]		LiRA [5]		AgrEvader [6]		AMP [46]				
	w/o	w	w/o	w	w/o	w	w/o	w	w/o	w	w/o	w	w/o	w	
Location30	84.58%	50.02%	82.38%	50.23%	87.83%	49.66%	88.27%	50.54%	80.66%	50.09%	84.86%	50.32%	64.49%	64.04%	-0.45%
Purchase100	77.26%	49.78%	68.15%	49.98%	81.25%	49.92%	78.49%	49.99%	79.47%	50.13%	75.70%	50.05%	80.32%	79.78%	-0.54%
Texas100	78.67%	49.84%	85.77%	50.10%	62.23%	51.08%	79.58%	50.08%	75.58%	50.66%	83.87%	49.53%	58.69%	60.50%	+1.81%
Cifar10	68.96%	50.13%	83.27%	49.88%	70.80%	50.29%	70.03%	50.02%	67.54%	49.66%	82.33%	49.72%	80.20%	79.46%	-0.74%

as their private data for local training. The distillation process is performed 25 times with the distillation coefficient set to 0.03. During the global model aggregation phase, when the contributions of all local participants are negative, the three clients with the smallest contributions $(k = 3)$ are discarded to mitigate the negative impact of malicious clients on the global model. Additionally, we evaluate the performance of the defense method under various parameter settings to verify its effectiveness and robustness across different configurations.

D. Evaluation Metrics

We evaluate the effectiveness of the proposed defense using two primary metrics: *attack accuracy* and *model accuracy*. The *attack accuracy*, denoted as ACC_a, measures the success rate of membership inference attacks. The *model accuracy*, denoted as ACC_m, reflects the classification performance of the target model on the test dataset. To assess the trade-off between privacy protection and model utility, we additionally report the change in model accuracy before and after defense, denoted as ΔACC_m.

VI. Experimental Analysis

A. Analysis of the Proposed Defense

To comprehensively evaluate the effectiveness of the proposed defense framework, we conduct extensive experiments on four benchmark datasets against six representative membership inference attacks, including Prediction [35], Bias [45], Entropy [36], LiRA [5], AgrEvader [6], and AMP [46]. The results are summarized in Table I. As shown, our method significantly reduces the effectiveness of all attack types. After applying our defense, the attack accuracy drops close to 50% across all settings, which approximates random guessing. This indicates that the attacker can no longer effectively distinguish between member and non-member data, demonstrating the strong privacy-preserving capability of our approach.

For passive attacks, our defense framework consistently suppresses all attack accuracies to around 50%, demonstrating its strong effectiveness against inference strategies that rely on the model parameters. This success is largely attributed to the distillation process based on entropy-regularized teacher training and CVAE-generated synthetic data, which jointly mitigate the overfitting to member samples. By increasing predictive uncertainty and decoupling the student model from raw training data, our method effectively weakens the distinguishability between member and non-member instances, thereby neutralizing the advantages exploited by passive attackers.

For active attacks such as AgrEvader [6] and AMP [46], which leverage data poisoning or malicious model updates to amplify membership leakage, our contribution-aware aggregation mechanism effectively mitigates their impact. As a result, the attack accuracy under these settings remains close to 50%, indicating that our defense can reliably identify and down-weight poisoned contributions, thereby maintaining both privacy and model utility during aggregation.

TABLE II: Comparison of SOTA defenses on attack accuracy, model utility, runtime, and resource usage on Purchase100.

Defense	ACC_a						ACC_m	ΔACC_m	Time/ Epoch	Memory
	Prediction	Bias	Entropy	LiRA	AgrEvader	AMP				
Memguard [17]	60.39%	66.78%	67.34%	67.65%	76.82%	68.00%	74.73%	-5.99%	96.80s	2325MB
DP-SGD [1]	52.54%	58.69%	58.32%	60.13%	63.30%	66.36%	57.46%	-23.26%	7.53s	1475MB
DMP [34]	53.15%	57.78%	57.77%	59.25%	58.66%	63.63%	68.50%	-12.22%	7.73s	1520MB
HAMP [7]	71.25%	70.89%	70.16%	71.52%	72.60%	78.92%	66.34%	-14.38%	4.41s	1350MB
MIST [20]	55.82%	58.89%	55.20%	59.55%	67.33%	70.31%	75.54%	-4.78%	7.35s	1413MB
LoDen [25]	65.83%	72.09%	74.07%	65.57%	61.09%	64.55%	79.34%	-1.38%	5.12s	1632MB
Mesas [19]	65.40%	70.40%	70.30%	65.81%	62.18%	65.40%	79.16%	-1.56%	230.03s	2728MB
Ours	49.78%	49.98%	49.92%	49.99%	50.13%	50.05%	79.78%	-0.54%	4.91s	1396MB

More importantly, our method preserves the performance of the target model while enhancing privacy protection. Across all evaluated datasets, the reduction in model accuracy remains within an acceptable range, and in certain cases, the accuracy even improves. For example, on the Texas100 dataset, the model accuracy increases by 1.81%, indicating that our defense strategy not only mitigates privacy leakage but also improves model generalization by reducing overfitting. This improvement is attributed to the use of entropy-regularized teacher training and synthetic data distillation, which jointly reduce the model's tendency to memorize specific samples and encourage the learning of more generalizable patterns. We further report the defense overhead in Appendix A and the runtime and memory efficiency in Appendix B.

B. Comparison with State-of-the-Art Defenses

We compare seven existing defense methods against membership inference attacks, including five designed to mitigate passive attacks: Memguard [17], DP-SGD [1], DMP [34], HAMP [7], and MIST [20], and two tailored for defending against active attacks, LoDen [25] and Mesas [19], as summarized in Table II. Overall, our proposed method consistently achieves the strongest defense across all attacks while maintaining minimal degradation in model accuracy.

For passive attack defense, existing methods show varied effectiveness. While DP-SGD [1] reduces the accuracy of Prediction [35] to 52.54%, it significantly degrades in model utility, with accuracy falling to 57.46%. HAMP [7] exhibits relatively poor defense performance, with limited reduction in attack success rates. MIST [20] demonstrates competitive defense performance against certain passive attacks by leveraging membership-invariant subspace learning but shows limited

TABLE III: Ablation study on the effectiveness of different components of our defense on Purchase100.

Components	ACC_a						ACC_m	ΔACC_m
	Prediction	Bias	Entropy	LiRA	AgrEvader	AMP		
1	69.31%	71.78%	51.06%	56.47%	72.00%	72.87%	80.09%	-0.23%
2	50.16%	51.25%	51.60%	56.05%	62.00%	59.09%	80.18%	-0.14%
3	72.59%	74.85%	65.83%	65.45%	67.33%	65.06%	78.14%	-2.18%
1+2	50.37%	50.86%	51.12%	52.12%	61.33%	61.36%	79.63%	-0.69%
1+3	71.90%	72.30%	50.98%	52.64%	62.00%	63.66%	77.96%	-2.36%
2+3	50.50%	51.16%	51.01%	53.47%	63.30%	53.63%	79.51%	-0.81%
1+2+3	49.78%	49.98%	49.92%	49.99%	50.13%	50.05%	79.78%	-0.54%

* Component 1, 2, and 3 denote teacher model training with improved entropy, student model training with CVAE distillation, and contribution-aware aggregation, respectively.

effectiveness against active attacks. Generally, these methods perform better against passive attacks than active ones, likely because they focus on obfuscating model outputs rather than mitigating malicious updates introduced by adversarial clients. For active attack defense, LoDen [25] and Mesas [19] moderately reduce attack accuracy for AgrEvader [6] and AMP [46] to around 60%. However, their protection against passive attacks is noticeably weaker, with accuracy often exceeding 65%, revealing limitations in their defensive scope.

Our method addresses both types of threats comprehensively. It reduces the accuracy of all attacks to around 50%, close to random guessing, while preserving 79.78% main task accuracy with only a 0.54% degradation, which is the smallest among all compared methods. In summary, most existing defenses are designed to address either passive or active attacks and often incur significant utility loss. In contrast, our joint defense delivers robust protection against diverse membership inference attacks while preserving model performance.

In addition to defense effectiveness, we compare the actual computational overhead (seconds) and memory consumption of each baseline method. Our method requires only 4.91s per training round and 1396MB of memory, demonstrating superior efficiency and resource usage. In contrast, Memguard [17] incurs substantial overhead, with a per-round training time of 96.80s, while Mesas [19] reaches as high as 230.03s and consumes 2728MB of memory. Although HAMP [7] features fast training and low memory usage, its defense performance is notably weaker than ours. In comparison, our method strikes a favorable balance among defense effectiveness, computational efficiency, and resource consumption, making it suitable for real-world federated learning scenarios.

C. Ablation Study of Core Defense Components

To assess the effectiveness and contribution of each component, we conduct an ablation study by selectively enabling the three core modules, as shown in Table III. Overall, the results demonstrate that each module contributes uniquely to the defense, and their combination yields the most comprehensive protection against both passive and active attacks.

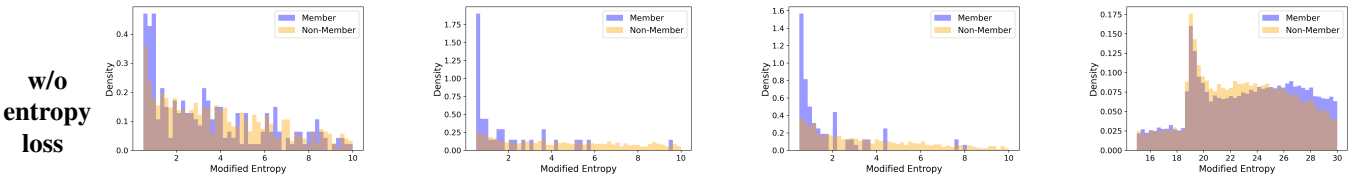
We begin by analyzing the individual effect of each defense component. When using the teacher model with entropy regularization alone (Component 1), the defense is particularly effective against entropy-based attacks, reducing the Entropy [36] attack accuracy to 51.06%. However, it provides limited resistance to active attacks such as AgrEvader [6] and AMP [46], which remain above 70%. In contrast, using CVAE-based distillation alone (Component 2) provides broader protection across attack types. For example, it significantly reduces the attack accuracy of LiRA [5] while maintaining high model utility, indicating that training on synthetic labeled data effectively masks membership status without sacrificing performance. On the other hand, contribution-aware aggregation (Component 3) is highly effective against active attacks, suppressing AgrEvader [6] and AMP [46], but less effective against passive inference, while incurring a 2.18% utility loss due to partial exclusion of local updates. These results indicate that each component targets different threat types and introduces specific trade-offs between privacy and utility.

Combining modules enhances their individual strengths. Most notably, when all three modules are used together, the accuracy of all attack types falls below 50.20%, effectively approaching the level of random guessing. At the same time, the target model achieves a classification accuracy of 79.78%, with only a 0.54% reduction compared to the undefended model. This result demonstrates that the three components provide complementary protection by addressing privacy vulnerabilities at different stages of federated learning. By reinforcing each other, these components collectively establish a unified defense mechanism that delivers strong and comprehensive resistance against both passive and active membership inference attacks while simultaneously preserving the performance of the global model on the primary task.

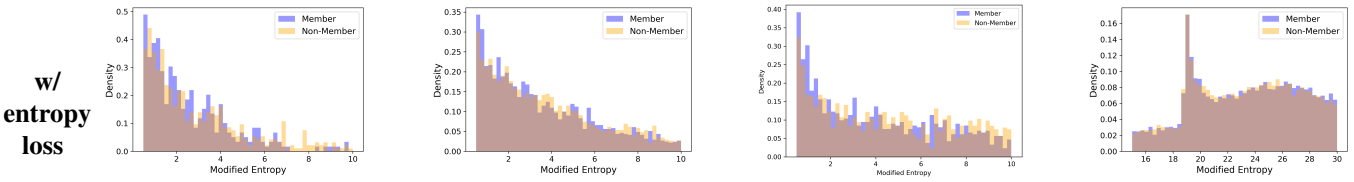
D. Impact of the Entropy Regularization Coefficient λ

To investigate the effect of entropy regularization on membership privacy, we first examine the modified entropy distributions of member and non-member samples before and after applying the entropy loss during teacher model training. As shown in Fig. 2, models trained without entropy regularization exhibit a clear separation between member and non-member samples, where members tend to have significantly lower predictive entropy. This distributional gap serves as a strong signal for membership inference attacks. In contrast, after incorporating entropy regularization, the entropy distributions of member and non-member data become considerably more overlapped across all datasets. This indicates that the model becomes less confident in its predictions on member data, thereby increasing uncertainty and obfuscating the decision boundary used by attackers. These observations confirm the effectiveness of entropy regularization in weakening the distinguishability between members and non-members, and hence, mitigating the risk of privacy leakage.

We further evaluate the quantitative impact of the entropy regularization coefficient λ. As illustrated in Fig. 3, increasing λ consistently reduces the attack success rates across various MIAs. Notably, when $\lambda = 0.2$, all attack accuracies drop to near-random levels, indicating optimal defense performance. In addition, we observe that for relatively small values of λ, such as 0.05 and 0.1, the main task accuracy not only remains stable but even improves, reaching 79.32% and 80.02%, respectively. This improvement suggests that mild entropy



(a) Location30 (b) Cifar10 (c) Purchase100 (d) Texas100



(e) Location30 (f) Cifar10 (g) Purchase100 (h) Texas100

Fig. 2: Comparison of entropy distributions across datasets. The top row shows models trained without entropy regularization, while the bottom row shows models trained with entropy regularization.

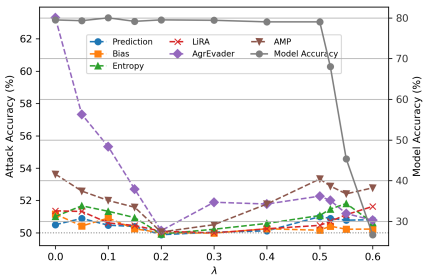


Fig. 3: Model performance (right y-axis) and attack accuracy (left y-axis) under different λ values on Purchase100.

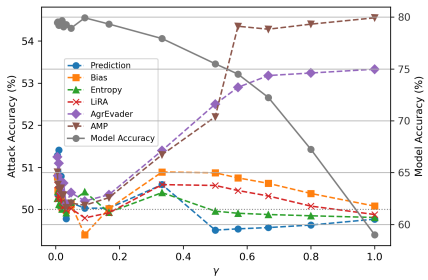


Fig. 4: Impact of distillation loss weight γ on model accuracy and defense performance on Purchase100.

regularization can effectively reduce overfitting and enhance generalization. However, when λ exceeds 0.5, the model accuracy deteriorates rapidly, reflecting the adverse effects of overly strong regularization. These results suggest that setting λ around 0.2 offers the best trade-off, enabling robust defense against MIAs while preserving the utility of the global model.

E. Effect of Distillation Loss Weight γ

We investigate the impact of the distillation loss weight γ on both defense performance and model accuracy. The results are illustrated in Fig. 4. When $\gamma = 0$, the student model is trained exclusively using soft labels from the teacher model, without relying on synthetic data labels. Under this setting, the model achieves relatively high accuracy, highlighting the effectiveness of soft supervision in guiding the learning process. As γ increases, the influence of hard labels from CVAE-generated data becomes stronger, and the reliance on the teacher model's soft labels weakens. In the range between $\gamma = 0.05$ and $\gamma = 0.1$, the model accuracy exhibits a steady upward trend, reaching a peak of 79.93%. This suggests that appropriately blending soft and hard supervision allows the student model to better capture both class-level structure and sample-specific information. However, when γ continues to increase, the model accuracy begins to decline, indicating that excessive reliance on synthetic data labels leads to performance degradation, likely due to distributional shifts and reduced knowledge transfer from the teacher model.

From the perspective of defense effectiveness, the attack accuracy for all methods remains close to 50% across a broad range of γ values. This suggests strong robustness against membership inference attacks. The use of CVAE-generated data ensures that the model does not overfit to real member samples, thereby weakening the attacker's ability to infer membership information. In summary, an appropriately chosen distillation weight γ enables the student model to achieve both high task performance and strong resistance to membership inference attacks. This demonstrates the effectiveness of our method in balancing privacy protection and model utility.

F. Impact of the Number of Distillation Iterations

We examine the effect of distillation iterations, as shown in Table IV. With fewer iterations, such as using only one iteration, the student model fails to sufficiently capture the output behavior of the teacher model, resulting in a relatively low classification accuracy of 68.23%. This observation suggests that at an early stage of distillation, the student model remains under-trained and lacks a thorough understanding of the soft labels provided by the teacher, thereby limiting its ability to form stable and generalizable prediction patterns.

TABLE IV: Effect of distillation iterations on attack accuracy and model performance.

Iters	ACC_a						ACC_m
	Prediction	Bias	Entropy	LiRA	AgrEvader	AMP	
1	49.65%	49.66%	49.34%	52.89%	51.90%	54.28%	68.23%
2	50.16%	50.01%	49.81%	51.24%	50.60%	52.80%	76.56%
3	50.74%	50.32%	50.05%	50.61%	50.25%	52.20%	76.37%
5	51.82%	50.61%	50.22%	50.25%	50.01%	50.67%	79.43%
10	51.99%	50.72%	50.62%	50.52%	50.13%	50.42%	79.12%
15	51.54%	50.36%	50.68%	50.29%	50.18%	50.27%	79.65%
20	50.16%	50.03%	50.22%	50.18%	50.11%	50.08%	79.43%
25	49.78%	49.98%	49.92%	49.99%	50.13%	50.05%	79.78%
30	50.31%	50.01%	50.43%	50.27%	50.22%	50.12%	79.34%
40	51.76%	50.20%	50.28%	50.11%	50.33%	50.25%	78.70%
50	51.80%	50.28%	50.12%	50.25%	50.45%	50.37%	79.24%

TABLE V: Effect of synthetic data volume on model performance and defense effectiveness.

Ratio	ACC_a						ACC_m
	Prediction	Bias	Entropy	LiRA	AgrEvader	AMP	
10%	49.21%	49.85%	50.03%	50.30%	51.50%	52.00%	71.29%
20%	50.59%	50.10%	50.11%	50.50%	50.75%	50.30%	79.26%
50%	52.09%	50.95%	50.09%	50.66%	50.43%	50.20%	79.20%
75%	51.19%	50.50%	50.55%	50.21%	50.33%	50.12%	79.12%
100%	49.78%	49.98%	49.92%	49.99%	50.13%	50.05%	79.78%
200%	49.85%	50.10%	51.45%	50.08%	50.66%	50.40%	79.83%
500%	50.24%	50.33%	51.84%	50.83%	50.78%	50.53%	79.32%

TABLE VI: Comparison of attack and model accuracy using real data and synthetic fake data as validation sets.

Validation		ACC_a						ACC_m
		Prediction	Bias	Entropy	LiRA	AgrEvader	AMP	
Real data	1%	51.15%	52.08%	50.92%	51.36%	52.41%	52.28%	79.63%
	2%	51.48%	51.45%	50.63%	51.72%	52.05%	51.97%	80.04%
	5%	51.41%	51.36%	51.20%	51.55%	51.68%	51.68%	78.47%
	20%	50.97%	50.83%	50.76%	51.04%	51.15%	51.15%	79.19%
	50%	50.43%	50.35%	50.22%	50.47%	50.55%	50.49%	80.05%
	100%	49.78%	49.98%	49.92%	49.99%	50.13%	50.05%	79.78%
Fake data	1%	50.12%	50.05%	49.88%	50.21%	50.39%	50.33%	78.97%
	2%	50.20%	50.12%	49.95%	50.42%	50.38%	50.21%	79.59%
	5%	49.12%	49.58%	48.76%	48.86%	49.94%	49.85%	79.50%
	20%	49.95%	50.03%	49.72%	49.89%	49.86%	49.72%	79.92%
	50%	49.88%	50.01%	49.79%	49.90%	49.71%	49.57%	80.02%
	100%	49.62%	49.53%	49.31%	49.49%	49.47%	49.41%	79.42%

TABLE VII: Comparison of different aggregation strategies under consistent configurations.

AGR	ACC_a						ACC_m
	Prediction	Bias	Entropy	LiRA	AgrEvader	AMP	
Fang [8]	52.07%	52.12%	56.18%	57.90%	61.33%	61.36%	79.35%
Median [43]	52.03%	52.38%	57.06%	60.12%	62.66%	59.09%	78.79%
Krum [3]	51.98%	52.10%	56.68%	57.87%	61.89%	60.45%	78.91%
T-M [24]	52.37%	50.86%	53.12%	55.45%	61.53%	61.36%	74.53%
Ours	**49.78%**	**49.98%**	**49.92%**	**49.99%**	**50.13%**	**50.05%**	**79.78%**

As the number of distillation iterations increases, the model's performance improves rapidly, with five or more iterations yielding stable main task accuracy around 79% and attack accuracies converging to approximately 50%. This demonstrates that multi-round distillation enables the student model to effectively absorb class-level structure and confidence information from the teacher while reducing the risk of overfitting to real member data through the use of synthetic samples. In the range of 15 to 30, both utility and defense performance remain stable, indicating convergence of the training process, whereas further increases beyond this point provide negligible gains and may even slightly degrade performance due to overfitting. Overall, a moderate number of iterations achieves effective knowledge transfer, ensuring strong defense capabilities and high main task accuracy.

G. Impact of Synthetic-to-Real Data Ratio

We evaluate the impact of synthetic-to-private data ratio in Table V, and visualize the distribution of generated data in Appendix D. When the synthetic data volume is low, such as 10%, the main task accuracy drops to 71.29%, suggesting that the student model cannot adequately learn from the teacher model. As the ratio increases, the accuracy rises quickly and surpasses 79% when the ratio reaches 20. Notably, when the ratio reaches 100%, the model achieves near-maximum accuracy and the lowest attack success rates, indicating optimal defense performance. At this point, all attack accuracies converge toward 50%, demonstrating strong resistance to membership inference. This suggests that sufficient synthetic data enables the student model to absorb the structural knowledge of the teacher while preserving generalization and privacy.

Further increasing the volume of synthetic data beyond 200% yields diminishing returns and does not lead to additional accuracy gains. In some cases, excessive synthetic samples may introduce redundancy or distributional noise, slightly increasing the attack success rate. Overall, a moderate amount of synthetic data improves both utility and robustness. Across all settings, our method maintains attack accuracy near the random guess level, highlighting its stability and effectiveness under varying synthetic data scales.

H. Ablation Study on Central Validation Set

To evaluate the impact of the central validation set, we conduct an ablation study with real and synthetic validation sets of varying sizes, as summarized in Table VI. The results show that as the proportion of the real validation set increases, attack accuracy gradually decreases towards the random guessing level, indicating enhanced defense effectiveness. When the validation set is small, such as 1% or 2%, the main task model accuracy exhibits slight fluctuations. Once the validation set reaches 50% or higher, the model accuracy stabilizes around 80%, while further improvements in defense effectiveness become marginal. These findings suggest that a moderately sized validation set is sufficient to meet defense requirements, and excessively large validation sets may result in unnecessary resource consumption without meaningful benefit.

The experiments with synthetic validation data show that even without any real validation set, our defense effectively suppresses all attacks, with attack accuracies maintained close to random guess. The target model also maintains performance comparable to that with real validation sets, with the largest observed accuracy drop being no more than 0.8%. In summary, this ablation study demonstrates that our method has low dependence on a central validation set. Even in scenarios where no real validation data is available, using synthetic validation data alone achieves effective defense without significantly

TABLE VIII: Defense effectiveness under different Dirichlet α settings on Purchase100.

Non-IID	ACC_a												ACC_m		ΔACC_m
	Prediction		Bias		Entropy		LiRA		AgrEvader		AMP				
	w/o	w	w/o	w	w/o	w	w/o	w	w/o	w	w/o	w	w/o	w	
0.1	59.57%	52.86%	61.21%	51.28%	61.51%	48.64%	63.87%	50.56%	60.40%	50.73%	62.10%	50.36%	46.71%	50.74%	+4.03%
0.3	63.75%	50.81%	67.90%	51.12%	70.55%	50.43%	67.23%	50.28%	72.88%	50.18%	69.50%	50.21%	59.08%	64.55%	+5.47%
0.5	79.41%	51.19%	81.00%	50.68%	85.96%	50.85%	82.50%	50.35%	84.70%	50.33%	81.60%	50.12%	65.18%	70.78%	+5.60%
0.7	78.72%	50.78%	80.63%	50.52%	84.32%	50.05%	80.31%	50.10%	83.20%	50.01%	79.41%	50.09%	67.55%	72.13%	+4.58%
1.0	64.72%	50.88%	67.35%	51.10%	70.43%	51.50%	69.88%	50.37%	79.57%	50.12%	78.60%	50.13%	69.70%	72.21%	+2.51%

affecting model performance, highlighting the practicality and deployment flexibility of the proposed approach.

I. Comparison of Different Aggregation Strategies

To evaluate our aggregation strategy against MIAs, we fix the distillation process and vary only the aggregation method, as shown in Table VII. Conventional strategies such as Fang [8], Median [43], Krum [3], and T-M [24] show limited effectiveness. For instance, under the Fang [8] strategy, the attack accuracy for AgrEvader [6] and AMP [46] remains as high as 61.33% and 61.36%, respectively, with similar levels observed for other methods. These findings indicate that traditional aggregations struggle to effectively neutralize the influence of malicious updates from compromised clients.

In contrast, our contribution-aware aggregation strategy consistently reduces the attack accuracy for all methods to near 50%, effectively suppressing both passive and active membership inference attacks to the level of random guessing. Notably, it also achieves the highest main task accuracy at 79.78%, outperforming all other baselines. This shows that by dynamically weighting each client's contribution to model performance, our approach improves robustness against poisoned updates while preserving task utility. In summary, the proposed aggregation strategy offers a superior balance between privacy protection and model utility, particularly in federated learning scenarios involving adversarial participants.

J. Evaluation under Non-IID Data Distribution

We simulate non-IID conditions using a Dirichlet distribution [10], where smaller α values indicate higher data heterogeneity. As shown in Table VIII, without defense, all attack accuracies increase notably as the data distribution becomes more heterogeneous. For instance, when $\alpha = 0.5$, the attack accuracy for Entropy [36] reaches as high as 85.96%, significantly compromising privacy. Simultaneously, the model performance deteriorates under high heterogeneity; for example, at $\alpha = 0.1$, the model accuracy drops to 46.71%. This is because the training process struggles to generalize well when local data distributions diverge significantly.

After applying the proposed defense, the accuracy of all attacks drops to nearly 50%, effectively neutralizing the inference advantage of attackers. Moreover, the global model accuracy improves in every case, with gains ranging from +2.51% to +5.60%. This suggests that our method not only mitigates privacy leakage but also enhances training stability and generalization in heterogeneous environments. The observed improvements highlight the applicability and robustness of our approach in realistic federated learning settings.

TABLE IX: Impact of the number of participants on attack accuracy and model performance.

Parts	ACC_a						ACC_m	ΔACC_m
	Prediction	Bias	Entropy	LiRA	AgrEvader	AMP		
2	50.94%	50.12%	50.21%	50.67%	51.00%	51.28%	80.29%	-0.33%
5	50.68%	49.98%	49.91%	50.10%	50.23%	50.56%	80.12%	-0.52%
10	49.78%	49.98%	49.92%	49.99%	50.13%	50.05%	79.78%	-0.54%
20	50.20%	50.19%	49.88%	50.12%	50.33%	50.18%	79.38%	-0.40%
30	50.10%	49.98%	49.74%	49.95%	50.00%	49.78%	79.88%	-0.48%
50	50.10%	49.73%	49.80%	49.89%	49.90%	49.88%	79.78%	-0.59%
100	49.78%	49.61%	49.78%	49.92%	49.58%	49.50%	79.20%	-0.54%
250	49.53%	49.60%	49.88%	49.78%	49.57%	49.63%	78.68%	-0.49%
500	49.39%	49.52%	49.22%	49.45%	49.41%	49.55%	75.67%	-0.59%

To further assess convergence under non-IID conditions, we compare our aggregation strategy with several alternatives in Appendix C, confirming its effectiveness and stability.

K. Impact of the Number of Participants

We evaluate the impact of the number of participants on Purchase100 in Table IX. From the perspective of defense performance, we observe that as the number of participants increases, the attack accuracy remains consistently close to 50% across all evaluated attacks, indicating strong and stable resistance to membership inference. For instance, the attack accuracy of the Entropy [36] method is 50.21% with 2 participants, and gradually decreases to 49.92% and 49.74% when the number increases to 10 and 30, respectively. Similar trends are observed for AgrEvader [6], whose accuracy drops from 51.00% with 2 participants to 49.41% with 500 participants. These results confirm that our defense mechanism effectively mitigates inference risks, and its performance remains stable even as the number of participants increases to 500. On the contrary, greater diversity in model updates appears to increase robustness by reducing the consistency of potential membership signals exploited by attackers.

In terms of the main task performance, the global model maintains consistently high accuracy across various participant configurations, demonstrating the stability of our defense mechanism. Even as the number of participants increases to 500, the model maintains a high accuracy of 75.67%, with only a 0.59% drop compared to the undefended setting, indicating minimal impact from the defense. These results demonstrate that our aggregation strategy effectively integrates diverse local updates without compromising utility. Furthermore, the proposed defense consistently preserves predictive performance, even when scaling up to large-scale federated participation.

VII. Conclusion

This paper proposes a unified defense framework for federated learning that effectively mitigates both passive and

active membership inference attacks. The approach combines entropy-regularized teacher training that improves prediction uncertainty and reduces overfitting, a CVAE-based distillation mechanism that generates labeled synthetic data for student training without exposing raw data, and a contribution-aware aggregation strategy that suppresses malicious updates by adjusting aggregation weights based on client utility. Extensive experiments demonstrate that the method substantially lowers attack success rates to near-random levels while maintaining or improving target model accuracy. In future work, we aim to explore more efficient data generation and aggregation techniques to further enhance the scalability and trustworthiness of federated learning systems.

ACKNOWLEDGMENT

We thank the anonymous reviewers for their insightful comments and constructive feedback. This work was supported in part by the National Key Research and Development Program of China under Grant 2023YFB4301901, in part by the National Natural Science Foundation of China under Grant 62202066, and in part by the BUPT Innovation Fund for Doctoral Students under Grant CX2023119.

REFERENCES

[1] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, "Deep learning with differential privacy," in *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, 2016, pp. 308–318.

[2] L. Bai, H. Hu, Q. Ye, H. Li, L. Wang, and J. Xu, "Membership inference attacks and defenses in federated learning: A survey," *ACM Computing Surveys*, vol. 57, no. 4, pp. 1–35, 2024.

[3] P. Blanchard, E. M. El Mhamdi, R. Guerraoui, and J. Stainer, "Machine learning with adversaries: Byzantine tolerant gradient descent," *Advances in neural information processing systems*, vol. 30, 2017.

[4] D. Byrd and A. Polychroniadou, "Differentially private secure multi-party computation for federated learning in financial applications," in *Proceedings of the First ACM International Conference on AI in Finance*, 2020, pp. 1–9.

[5] N. Carlini, S. Chien, M. Nasr, S. Song, A. Terzis, and F. Tramer, "Membership inference attacks from first principles," in *2022 IEEE symposium on security and privacy (SP)*. IEEE, 2022, pp. 1897–1914.

[6] Y. Chen, C. Shen, Y. Shen, C. Wang, and Y. Zhang, "Amplifying membership exposure via data poisoning," *Advances in Neural Information Processing Systems*, vol. 35, pp. 29 830–29 844, 2022.

[7] Z. Chen and K. Pattabiraman, "Overconfidence is a dangerous thing: Mitigating membership inference attacks by enforcing less confident prediction," in *Network and Distributed System Security (NDSS) Symposium*, 2024.

[8] M. Fang, X. Cao, J. Jia, and N. Gong, "Local model poisoning attacks to {Byzantine-Robust} federated learning," in *29th USENIX security symposium (USENIX Security 20)*, 2020, pp. 1605–1622.

[9] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," in *NIPS Deep Learning and Representation Learning Workshop*, 2015.

[10] T.-M. H. Hsu, H. Qi, and M. Brown, "Measuring the effects of non-identical data distribution for federated visual classification," *arXiv preprint arXiv:1909.06335*, 2019.

[11] H. Hu, Z. Salcic, L. Sun, G. Dobbie, P. S. Yu, and X. Zhang, "Membership inference attacks on machine learning: A survey," *ACM Computing Surveys (CSUR)*, vol. 54, no. 11s, pp. 1–37, 2022.

[12] L. Hu, H. Yan, Y. Peng, H. Hu, S. Wang, and J. Li, "Vae-based membership cleanser against membership inference attacks," *IEEE Transactions on Dependable and Secure Computing*, 2024.

[13] Y. Hu, Y. Wang, J. Lou, W. Liang, R. Wu, W. Wang, X. Li, J. Liu, and Z. Qin, "Privacy risks of federated knowledge graph embedding: New membership inference attacks and personalized differential privacy defense," *IEEE Transactions on Dependable and Secure Computing*, 2024.

[14] B. Hui, Y. Yang, H. Yuan, P. Burlina, N. Z. Gong, and Y. Cao, "Practical blind membership inference attack via differential comparisons," in *Network and Distributed System Security (NDSS) Symposium*, 2021.

[15] M. Jagielski, M. Nasr, K. Lee, C. A. Choquette-Choo, N. Carlini, and F. Tramer, "Students parrot their teachers: Membership inference on model distillation," *Advances in Neural Information Processing Systems*, vol. 36, pp. 44 382–44 397, 2023.

[16] D. Javeed, M. S. Saeed, P. Kumar, A. Jolfaei, S. Islam, and A. N. Islam, "Federated learning-based personalized recommendation systems: An overview on security and privacy challenges," *IEEE Transactions on Consumer Electronics*, vol. 70, no. 1, pp. 2618–2627, 2023.

[17] J. Jia, A. Salem, M. Backes, Y. Zhang, and N. Z. Gong, "Memguard: Defending against black-box membership inference attacks via adversarial examples," in *Proceedings of the 2019 ACM SIGSAC conference on computer and communications security*, 2019, pp. 259–274.

[18] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv e-prints*, pp. arXiv–1312, 2013.

[19] T. Krauß and A. Dmitrienko, "Mesas: Poisoning defense for federated learning resilient against adaptive attackers," in *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, 2023, pp. 1526–1540.

[20] J. Li, N. Li, and B. Ribeiro, "{MIST}: Defending against membership inference attacks through {Membership-Invariant} subspace training," in *33rd USENIX Security Symposium (USENIX Security 24)*, 2024, pp. 2387–2404.

[21] L. Li, Y. Fan, M. Tse, and K.-Y. Lin, "A review of applications in federated learning," *Computers & Industrial Engineering*, vol. 149, p. 106854, 2020.

[22] H. Liu, Y. Wu, Z. Yu, and N. Zhang, "Please tell me more: Privacy impact of explainability through the lens of membership inference attack," in *2024 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2024, pp. 4791–4809.

[23] G. Long, Y. Tan, J. Jiang, and C. Zhang, "Federated learning for open banking," in *Federated Learning: Privacy and Incentive*. Springer, 2020, pp. 240–254.

[24] G. Lugosi and S. Mendelson, "Robust multivariate mean estimation: the optimality of trimmed mean," 2021.

[25] M. Ma, Y. Zhang, P. C. M. Arachchige, L. Y. Zhang, M. B. Chhetri, and G. Bai, "Loden: Making every client in federated learning a defender against the poisoning membership inference attacks," in *Proceedings of the 2023 ACM Asia Conference on Computer and Communications Security*, 2023, pp. 122–135.

[26] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial intelligence and statistics*. PMLR, 2017, pp. 1273–1282.

[27] L. Melis, C. Song, E. De Cristofaro, and V. Shmatikov, "Exploiting unintended feature leakage in collaborative learning," in *2019 IEEE symposium on security and privacy (SP)*. IEEE, 2019, pp. 691–706.

[28] M. Nasr, R. Shokri, and A. Houmansadr, "Machine learning with membership privacy using adversarial regularization," in *Proceedings of the 2018 ACM SIGSAC conference on computer and communications security*, 2018, pp. 634–646.

[29] ——, "Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning," in *2019 IEEE symposium on security and privacy (SP)*. IEEE, 2019, pp. 739–753.

[30] N. Papernot, M. Abadi, Ú. Erlingsson, I. Goodfellow, and K. Talwar, "Semi-supervised knowledge transfer for deep learning from private training data," in *International Conference on Learning Representations (ICLR)*, 2017.

[31] ——, "Semi-supervised knowledge transfer for deep learning from private training data," in *International Conference on Learning Representations*, 2017.

[32] A. Qayyum, K. Ahmad, M. A. Ahsan, A. Al-Fuqaha, and J. Qadir, "Collaborative federated learning for healthcare: Multi-modal covid-19 diagnosis at the edge," *IEEE Open Journal of the Computer Society*, vol. 3, pp. 172–184, 2022.

[33] A. Salem, Y. Zhang, M. Humbert, P. Berrang, M. Fritz, and M. Backes, "Ml-leaks: Model and data independent membership inference attacks and defenses on machine learning models," in *Network and Distributed System Security (NDSS) Symposium*, 2019.
[34] T. Shejwalkar and J. Huang, "Membership privacy training via knowledge distillation," in *Advances in Neural Information Processing Systems*, vol. 34, 2021, pp. 21 652–21 665.
[35] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in *2017 IEEE symposium on security and privacy (SP)*. IEEE, 2017, pp. 3–18.
[36] L. Song and P. Mittal, "Systematic evaluation of privacy risks of machine learning models." in *USENIX Security Symposium*, vol. 1, no. 2, 2021, p. 4.
[37] R. Tang, H. Wu, and X. Ji, "Protecting membership privacy by split knowledge distillation," in *31st USENIX Security Symposium (USENIX Security 22)*, 2022, pp. 3197–3214.
[38] F. Tramèr, R. Shokri, A. San Joaquin, H. Le, M. Jagielski, S. Hong, and N. Carlini, "Truth serum: Poisoning machine learning models to reveal their secrets," in *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, 2022, pp. 2779–2792.
[39] J. Wen, Z. Zhang, Y. Lan, Z. Cui, J. Cai, and W. Zhang, "A survey on federated learning: challenges and applications," *International Journal of Machine Learning and Cybernetics*, vol. 14, no. 2, pp. 513–535, 2023.
[40] J. Xu, B. S. Glicksberg, C. Su, P. Walker, J. Bian, and F. Wang, "Federated learning for healthcare informatics," *Journal of healthcare informatics research*, vol. 5, pp. 1–19, 2021.
[41] L. Yang, B. Tan, V. W. Zheng, K. Chen, and Q. Yang, "Federated recommendation systems," in *Federated Learning: Privacy and Incentive*. Springer, 2020, pp. 225–239.
[42] S. Yeom, I. Giacomelli, M. Fredrikson, and S. Jha, "Privacy risk in machine learning: Analyzing the connection to overfitting," in *2018 IEEE 31st computer security foundations symposium (CSF)*. IEEE, 2018, pp. 268–282.
[43] D. Yin, Y. Chen, R. Kannan, and P. Bartlett, "Byzantine-robust distributed learning: Towards optimal statistical rates," in *International Conference on Machine Learning*. PMLR, 2018, pp. 5650–5659.
[44] J. Zhang, B. Li, X. Zhang, and S. Jana, "Membership privacy protection via adversarially regularized mixup mmd," in *Advances in Neural Information Processing Systems*, vol. 34, 2021, pp. 26 771–26 784.
[45] L. Zhang, L. Li, X. Li, B. Cai, Y. Gao, R. Dou, and L. Chen, "Efficient membership inference attacks against federated learning via bias differences," in *Proceedings of the 26th International Symposium on Research in Attacks, Intrusions and Defenses*, 2023, pp. 222–235.
[46] Y. Zhang, G. Bai, M. A. P. Chamikara, M. Ma, L. Shen, J. Wang, S. Nepal, M. Xue, L. Wang, and J. Liu, "Agrevader: Poisoning membership inference against byzantine-robust federated learning," in *Proceedings of the ACM Web Conference 2023*, 2023, pp. 2371–2382.
[47] L. Zhu, Z. Liu, and S. Han, "Deep leakage from gradients," *Advances in neural information processing systems*, vol. 32, 2019.

APPENDIX

A. Overhead of the Proposed Defense

To evaluate the computational overhead of our defense framework, we analyze its three core stages. First, in teacher model training, each sample requires an additional forward pass and entropy-based loss computation, yielding a complexity of $\mathcal{O}(MC)$, where M is the number of local training samples, C denotes the per-sample computational cost of the model; as the teacher, student, and global models share the same architecture, their computational complexity is equivalent. This introduces only a limited increase in computational cost compared to standard training. Second, in the student model training phase, we first train a CVAE to approximate the distribution of private data. This process has a complexity of $\mathcal{O}(MHC_{\text{CVAE}})$, where H denotes the number of CVAE training epochs, C_{CVAE} denotes the per-sample cost of the CVAE model. Afterward, the student model is trained using the generated synthetic data over multiple distillation iterations. We assume the number of generated synthetic samples matches the local private data size, i.e., also M. Let R denote the number of distillation rounds; each round involves a forward pass and loss computation, resulting in a complexity of $\mathcal{O}(MRC)$. This process is performed locally and can be parallelized across participants. Moreover, the synthetic data can be generated offline, further improving efficiency. Finally, during aggregation, the server evaluates each client's contribution by testing on a held-out validation set of size V, resulting in a total cost of $\mathcal{O}(NVC)$, where N is the number of clients.

In terms of communication, our method does not introduce any additional overhead beyond standard FL frameworks such as FedAvg. Each client only uploads model parameters with the same architecture and optionally a scalar contribution score, while all synthetic data generation and training are performed locally without any data sharing. In summary, the total overhead grows linearly with data size and participant count, making the framework scalable and practical for federated learning with enhanced privacy guarantees.

TABLE X: Breakdown of per-round runtime (in seconds) and memory usage across datasets.

Dataset	CVAE		Distillation		Aggregation	Time/ Epoch	Memory
	Train VAE	Generate Data	Train Teacher	Train Student			
Location30	0.0144	0.0284	0.0032	0.0448	0.0759	1.0092	638MB
Purchase100	0.5528	0.0243	0.1678	0.0726	0.2171	4.9093	1396MB
Texas100	0.1943	0.0277	0.0678	0.3684	0.8520	10.6004	4234MB
Cifar10	0.5607	0.0255	0.1842	0.2382	1.8408	6.5945	1628MB

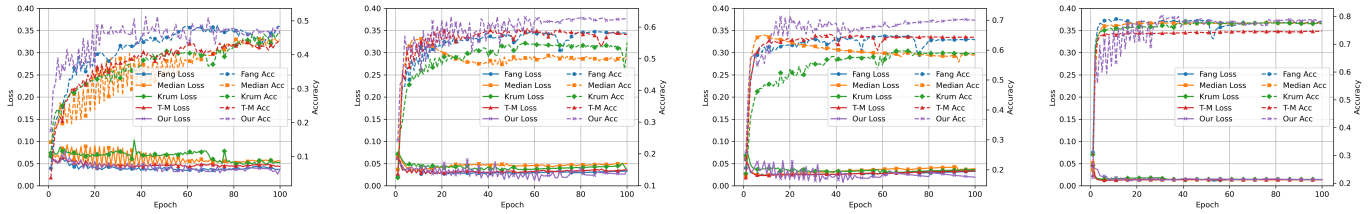
B. Runtime and Memory Efficiency across Datasets

To evaluate the practicality of our method across different application scenarios, we report a detailed breakdown of per-round runtime and memory consumption on four representative datasets, as shown in Table X. The results reveal that our method maintains low overhead and stable memory usage across datasets of varying sizes and modalities.

For the Location30 dataset, each training round takes approximately 1.0s with memory usage below 650MB, demonstrating strong adaptability to edge devices. On higher-dimensional tabular datasets such as Purchase100 and Texas100, although the runtime increases, the overall overhead remains within an acceptable range, with peak memory usage kept under 4.5GB. This confirms the scalability and resource efficiency of our method on complex data.

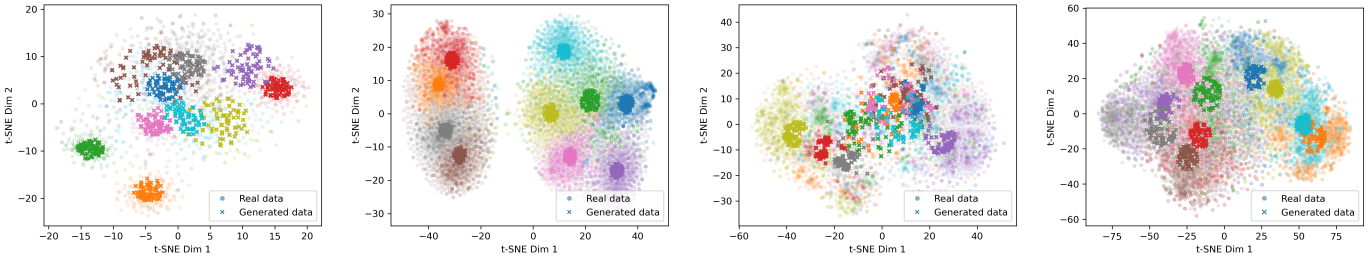
In terms of runtime breakdown, the training of the CVAE can be performed independently from the federated training process. Once the local data distribution is fixed, the CVAE only needs to be trained once and does not require retraining in each round. Thus, its computational overhead is one-time and does not impose recurring costs. Furthermore, the CVAE-related stages consume minimal time and offer high flexibility.

Overall, this experiment validates that our method achieves a favorable balance between defense effectiveness and computational efficiency, making it feasible for wide deployment in real-world federated learning scenarios.



(a) Non-IID ($\alpha = 0.1$) (b) Non-IID ($\alpha = 0.3$) (c) Non-IID ($\alpha = 0.5$) (d) IID ($\alpha = \infty$)

Fig. 5: Convergence performance of different aggregation strategies under varying data heterogeneity.



(a) Location30 (b) Purchase100 (c) Texas100 (d) Cifar10

Fig. 6: t-SNE visualization of real and CVAE-generated samples across different datasets. For each dataset, faint circular points represent real data, while opaque cross markers denote representative fake samples generated by the CVAE model.

C. Convergence Analysis under Non-IID Settings

We further evaluate the convergence performance of our proposed method under non-independent and identically distributed (non-IID) data settings, and compare it with several representative aggregation strategies, including Fang [8], Median [43], Krum [3], and T-M [24]. As shown in Fig. 5, we present the training loss and test accuracy across epochs under the IID condition and different levels of non-iidness controlled by Dirichlet distribution parameters $\alpha = 0.1, 0.3, 0.5$.

Under the IID scenario, all methods demonstrate good convergence behavior, with accuracy eventually stabilizing. However, under non-IID conditions, the performance gaps become more pronounced. With smaller values of α, which correspond to greater data heterogeneity, most methods exhibit more unstable training curves, slower convergence, or even stagnation. For instance, when $\alpha = 0.1$, Median [43] and Krum [3] struggle to exceed 50% test accuracy, while T-M [24] maintains stability but suffers from lower overall accuracy.

In contrast, our method consistently achieves effective aggregation and maintains higher accuracy across all non-IID settings. The training process exhibits smoother loss curves and more stable accuracy trends, demonstrating strong robustness under heterogeneous data distributions. This advantage is attributed to our contribution-aware aggregation strategy, which increases the impact of high-quality local updates and reduces the influence of inconsistent or low-quality ones. While the method may show slight accuracy fluctuations in the early training stages, particularly when the data distribution is highly skewed such as under $\alpha = 0.1$, these fluctuations diminish quickly as training proceeds. Overall, our approach maintains a favorable convergence trend and achieves effective and stable aggregation in challenging federated learning scenarios.

D. t-SNE Visualization of CVAE-Generated Samples

To further validate the effectiveness of the synthetic data generated by the CVAE in our framework, we visualize both real and generated samples across different datasets using t-SNE, as shown in Fig. 6. In each subfigure, circular points represent real samples, while cross markers denote synthetic samples produced by the CVAE.

The visualization clearly shows that the CVAE-generated data can effectively cover the distribution of each class and exhibit strong intra-class clustering. This indicates that our class-conditional CVAE successfully captures the semantic structure of each class during training and is capable of generating synthetic data that closely mimics the real distribution.

Moreover, compared to the more dispersed distribution of real samples, the generated data tends to be more concentrated and closer to the cluster centers of their corresponding classes in the embedding space. This concentration effect arises from two factors. First, the CVAE maximizes the conditional likelihood during training, encouraging the generation of highly representative and confident prototypical samples. Second, the class-conditioning mechanism provides strong guidance in the latent space, reducing inter-class overlap and noise.

Overall, these results demonstrate that the CVAE in our defense framework not only provides strong data generation capability but also yields structured, controllable samples that serve as a reliable foundation for subsequent knowledge distillation and privacy protection.

E. Details of Attack Methods

To comprehensively assess the robustness of our defense framework, we consider a diverse set of membership inference attacks, covering both passive and active threat models. We

provide brief descriptions of the six representative attack methods used in our evaluation.

- **Prediction** (Shokri et al. [35]): A neural network-based passive attack where shadow models are trained to distinguish member from non-member data based on the confidence of the model's predictions.
- **Bias** (Zhang et al. [45]): A passive insider attack that exploits temporal variations in model bias during the federated training process to infer membership information.
- **Entropy** (Song et al. [36]): A passive threshold-based attack that leverages the difference in modified predictive entropy between training and testing samples to determine membership status.
- **LiRA** (Carlini et al. [5]): A passive attack that infers membership by comparing the model's output likelihood on the target sample against that on reference non-member samples.
- **AgrEvader** (Zhang et al. [46]): An active attack based on model poisoning, where adversaries deliberately manipulate their local model updates to interfere with global aggregation and intensify privacy leakage.
- **AMP** (Chen et al. [6]): An active attack that injects poisoned samples into the training dataset to manipulate the model's output distribution and enhance membership inference success.