2025 International Conference on Intelligent Computing
July 26-29, Ningbo, China
https://www.ic-icc.cn/2025/index.php

Membership Inference Attacks for Generative Model-Based One-Shot Federated Learning

Yiwen Wang, Fan Qi, and Zixin Zhang

School of Computer Science and Engineering,Tianjin University of Technology, Tianjin, China

wangyiwenjj@stud.tjut.edu.cn

Abstract. In recent years, One-Shot Federated Learning (OSFL) has gained significant attention for its communication efficiency. With the rise of generative models, many approaches leverage synthetic data on the server to improve global model performance. However, this efficiency introduces heightened privacy risks that remain largely unexplored. In this paper, we conduct the first systematic exploration of privacy risks in OSFL by designing a Membership Inference Attack (MIA) strategy tailored to this paradigm. In our strategy, we introduce a general approach designed for all generative models, which infers membership by aligning query data with the global client distribution. Building on this, we extend the approach specifically for diffusion models, integrating global alignment with query-specific fine-grained details through finetuning and conditional generation, thereby enabling more robust inference. In particular, our strategy does not rely on auxiliary data, making it particularly relevant for privacy-sensitive OSFL settings. Extensive experiments validate the effectiveness of the proposed strategy, highlighting the critical privacy risks posed by generative models in OSFL.

Keywords: Membership Inference Attacks, One-Shot Federated Learning, Generative Model

1 INTRODUCTION

One-Shot Federated Learning (OSFL) [1,2,3,4] has gained attention as an efficient distributed learning paradigm.OSFL distinguishes itself by requiring only a single round of communication between the central server and clients to complete the learning process. This streamlined approach not only reduces bandwidth consumption but also expedites the learning process,making it a practical alternative for large scale applications [1,4,5].

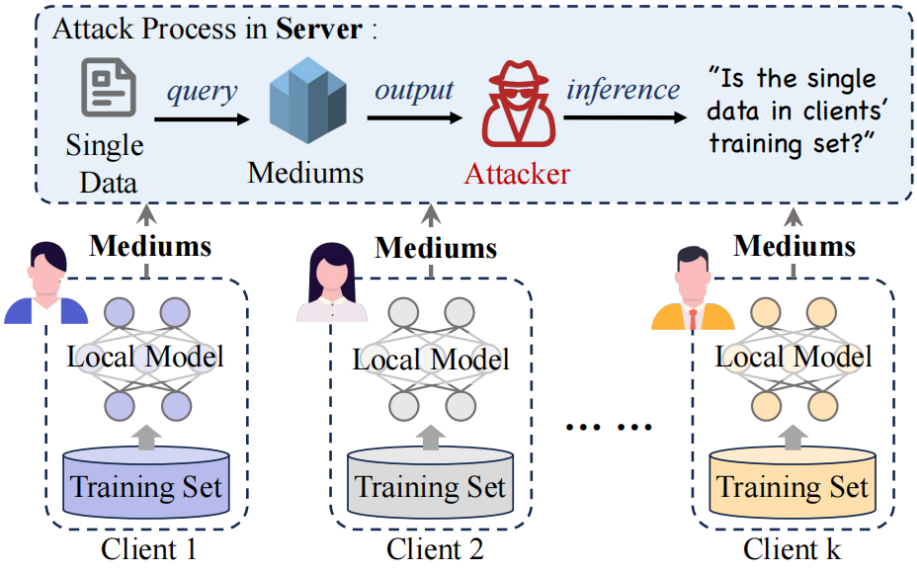

Fig. 1. Illustration of MIA in OSFL

Recently,generative model-based approaches have demonstrated significant research value and broad application prospects in OSFL [1,3].These methods employ a novel communication paradigm where clients only need to transmit lightweight "mediums"(such as descriptors, partial model parameters,and prototype representations) to the server,which then leverages powerful generative models to synthesize samples that capture the characteristics of client data distributions. This design not only enhances global model performance through training with synthetic data but also maintains communication efficiency.

Though generative model-based OSFL offers significant benefits,it also introduces potential privacy risks.The highfidelity synthetic data generated by these models may expose characteristics of the original data, making the system vulnerable to Membership Inference Attacks (MIAs).In Federated Learning (FL),MIAs aim to determine whether a specific query sample is part of the training dataset and rely on techniques like gradient difference analysis [6], shadow model training [7], and output-based confidence score analysis [8].

Unfortunately, these methods can be applied to various FL scenarios but they are specifically designed for multi-round communication settings. At present, research on MIAs in OSFL, especially in contexts involving generative models,remains largely unexplored.

To address this gap, we propose a MIA strategy tailored for generative models based OSFL. In this scenario, clients transmit training "mediums" to the server, which synthesizes images. Attackers exploit the synthetic data generated by the server to perform MIA, as illustrated in Fig. 1. We introduce AttackI-Base, a general method for all gen-

erative models. It extracts features from server-generated data, computes a feature center for the client distribution, and infers membership by comparing this center with the attacker's query data. While effective, it focuses only on overall client distributions and overlooks fine-grained details, which limits its inference accuracy. To achieve more accurate inference, we propose AttackII-Finetune, which finetunes the widely used diffusion models to better align with client distributions. By using query data as conditional input, this approach generates detailed, query-specific synthetic images, thereby significantly enhancing inference accuracy. Specifically, AttackII-Finetune incorporates a dual similarity evaluation mechanism: initial similarity assesses the global alignment between the query data and the overall client distribution, offering a stable reference, while conditional similarity captures fine-grained, query-specific details using the fine-tuned model. This dual approach ensures a comprehensive exploration of privacy risks in OSFL with generative models.

Overall, our main contributions can be summarized as follows:

- We summarize existing methods that utilize generative models in OSFL and highlight their potential privacy vulnerabilities.
- We propose a MIA strategy, consisting of a general purpose approach for OSFL generative models and a diffusion-specific method that leverages fine-tuning and conditional generation to enhance accuracy and robustness.
- We systematically evaluate our attack strategy through extensive experiments across various backbone architectures and evaluation metrics, with further analyses on similarity metrics and client numbers, where AttackII-Finetune demonstrates better performance in most scenarios, validating our approach's effectiveness in exploiting OSFL generative models' privacy vulnerabilities.

2 RELATED WORK

2.1 Generative model-based One-Shot Federated Learning

Generative models have emerged as a versatile solution for addressing data heterogeneity and mitigating high computational costs OSFL. These methods can be broadly categorized into three approaches based on their focus: model aggregation, representation abstraction, and classification-guided synthesis. In the model aggregation category, methods like FEDCVAE[9] utilize Conditional Variational Autoencoders to aggregate locally trained decoders for flexible and client-specific data synthesis. FedDiff [5] trains diffusion models locally and uploads parameters for centralized synthesis at the server. For representation abstraction methods, FedDEO [1] and FedDISC [2] utilize descriptive vectors or extracted features from client data to capture key client distribution characteristics with reduced communication overhead. Similarly, FedBiP [10] uploads personalized latent representations and optimized concept vectors, enabling the central server to synthesize data that closely aligns with client distributions while preserving privacy. Finally, classification-guided synthesis is exemplified by FedCADO [3], which uploads lightweight classifiers trained locally to steer server-side generative

models for class-specific synthesis. These methods highlight the adaptability of generative models in OSFL, providing efficient and scalable solutions for collaboration in highly heterogeneous environments. However, they have overlooked the privacy issues introduced by generative models in OSFL.

2.2 Federated Membership Inference Attack

MIAs in FL are broadly classified into two types: update-based MIAs and trend-based MIAs, which exploit either model updates or trends in the training process to determine whether specific data points are part of the training set. **Update-based MIAs** leverage gradients shared during FL to infer membership information. Gradient-based attacks analyze raw gradient values or the differences between gradients across con- secutive training iterations, as explored in works such as[11] and[6]. Alternatively, **single model-based** attacks use shadow training to simulate the behavior of the target model or modify its structure to enable membership inference, as demonstrated in [7]. These methods are effective in scenarios where gradient information is abundant but often come with significant computational costs and are sensitive to factors like batch size. **Trend-based MIAs** infer membership by analyzing the evolution of **model outputs** during training. Model output-based attacks track changes in prediction confidence over multiple training rounds to distinguish between members and non-members [8,12]. While these methods are lightweight in computation, their success heavily relies on the availability of multiple snapshots of the model and the degree of sensitivity in parameter changes. In our work, we focus on MIAs in OSFL, where the attacker leverages information obtained from a malicious server to perform the attack.

3 Methodology

3.1 Problem Formulation

Federated Setting. We consider a OSFL scenario involving K clients and a server, where communication between clients and the server occurs only once. In this setting, each client transmits to the server the "mediums" $\{\mu_k\}_{k=1}^{K}$(e.g., model parameters or intermediate features)that capture the characteristics of its local data distribution. Without direct access to the raw data on the clients, the server leverages these "mediums" and a generative model θ_G to generate synthetic data, which is subsequently used to train the global model.

Attacker Capability. We assume that the attacker operates in a malicious server environment. The attacker has no access to auxiliary datasets to train shadow models and can only rely on a query dataset to conduct the attack. The query dataset consists of samples related to the client data, as well as non-client samples drawn from the same data distribution as the client data. Similar to other MIA setting [13], the attacker cannot distinguish the origin of these samples. This capability assumption aligns more closely with practical FL scenarios that emphasize data privacy protection.

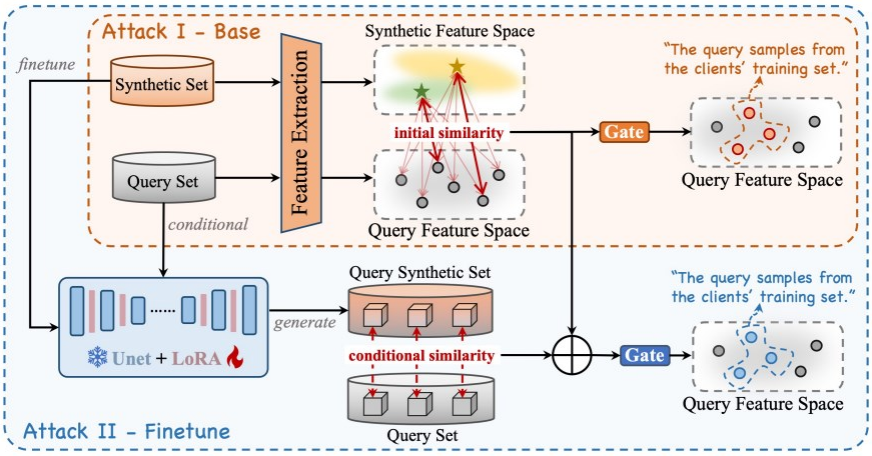

Fig. 2: Illustration of our attack strategy.

3.2 Attack Pipelines

We next introduce the proposed attack strategy, which includes AttackI-Base, applicable to all generative models, and its extension, AttackII-Finetune, tailored for diffusion models. The attack process on the server is illustrated in Fig. 2.

Algorithm 1 AttackI-Base

Require: Query dataset D_{query} , synthetic dataset $D_{syn} = \theta_G(\mu)$
Ensure: Membership prediction y for each query sample

for each class j **do**
  $\bar{f}_G^j \leftarrow \frac{1}{N_{syn}^j} \sum_{i=1}^{N_{syn}^j} f_G^{i,j}$ ▷ Calculate feature center
end for
for each query sample **do**
  $S \leftarrow \mathrm{CosineSimilarity}(f_Q, \bar{f}_G)$ ▷ Calculate similarity
end for
$\tau^* \leftarrow Find\ Optimal\ Threshold\ By\ ROC$
for each similarity score s in S **do**
  if $s \geq \tau^*$ **then**
    $y \leftarrow 1$ ▷ Member
  else
    $y \leftarrow 0$ ▷ Non-member
  end if
end for

AttackI-Base. AttackI-Base is an attack method for universal generative models that determines a query sample's membership by measuring its feature similarity to the synthetic data generated by the server. We provide pseudo code for the attack strategy in Algorithms 1.

Feature Extraction. We first extract features of samples of synthetic set data $D_{syn} = \theta_G(\mu)$ and samples of query set D_{query} using pre-training models to perform similarity computation in a unified feature space.

Feature Center Calculation. Since the synthetic data is generated based on the medium μ uploaded by clients, its feature distribution can indirectly reflect the distribution characteristics of the client training data. Therefore, we further calculate the feature center for each class within the synthetic data as a concentrated representation of the distribution of the client's training data:

$$\bar{f}_G^j = \frac{1}{N_{syn}^j} \sum_{i=1}^{N_{syn}^i} f_G^{i,j} \tag{1}$$

where $f_G^{i,j}$ is the feature of the i-th synthetic image of class j, and N_{syn}^i is the total number of synthetic images of class j. The feature center reduces computational complexity by allowing comparisons with a single representative vector instead of all generated samples. It also improves stability by mitigating the influence of noise and outliers, and it effectively represents the global distribution of client data.

Similarity Calculation. We calculate the similarity between each query sample f_Q and the feature center $\bar{f}_G$ using cosine similarity,i.e., $\mathrm{S}(\mathrm{f_Q}, \bar{\mathrm{f}}_\mathrm{G}) = \langle \mathrm{f_Q}, \bar{\mathrm{f}}_\mathrm{G} \rangle / \|\mathrm{f_Q}\| \cdot \|\bar{\mathrm{f}}_\mathrm{G}\|$. The $\mathrm{S}(\mathrm{f_Q}, \bar{\mathrm{f}}_\mathrm{G})$ provides a quantitative measure of the alignment between the query data and the global feature distribution of the client data.

Threshold Setting and Inference of Membership. Based on the similarity scores, we utilize the Receiver Operating Characteristic (ROC) curve [14] to determine the optimal threshold τ^* to infer the membership of the query sample. Specifically, the threshold is chosen to maximize Youden's J statistic, ensuring the best trade-off between sensitivity and specificity. The membership inference can be expressed as follows:

$$y = \begin{cases} 1, & S(f_Q, \bar{f}_G) \geq \tau^*, \\ 0, & S(f_Q, \bar{f}_G) < \tau^*. \end{cases} \tag{2}$$

AttackII-Finetune. Leveraging the fine-tuning capabilities of diffusion models, we propose an MIA method specifically designed to enhance attack performance on synthetic data. This method builds on research [13], which explicitly demonstrates that if a target sample x is used during training, the generated samples will closely resemble x.We provide pseudo-code for the attack strategy in Algorithms 2.

Finetune Using Synthetic Data. To adapt the diffusion model to the client data distribution, we begin by fine-tuning it via LoRA [15] on the synthesized data. This step aligns the model with the characteristics of the synthetic data, improving its capacity to represent client-specific features. Mathematically, given the diffusion model θ_G , we denote the trainable parameters by $\tilde{\theta}_G$.

Using Query Data as a Condition. The initial synthetic samples mainly capture the global distribution of client data but fail to reflect the fine-grained details of query data, limiting membership inference accuracy. To address this, we finetune the diffusion model using query data as a condition to generate query-specific synthetic set D_{qs}. When the query data contains client samples, the feature f_{QS} of the generated sample in D_{qs} will better align with client-specific characteristics. Conversely, for non-client query data, the feature f_{QS} may reflect non-client features but remain influenced by the client distribution, weakening non-client-specific representations.

Algorithm 2 Attack II - Finetune

Require: Query dataset D_{query}, synthetic dataset D_{syn}, diffusion model θ_G , finetuning steps T

Ensure: Membership prediction y for each query sample

for each class j **do**
  $\bar{f}_G^j \leftarrow \frac{1}{N_{syn}^j} \sum_{i=1}^{N_{syn}^j} f_G^{i,j}$ ▷ Calculate feature center
end for
for each query sample **do**
  $S_{init} \leftarrow CosineSimilarity(f_G, \bar{f}_G)$
end for
$\bar{\theta}_G \leftarrow Finetune(\theta_G, D_{syn})$ ▷ Finetune on synthetic data
$D_{qs} \leftarrow \bar{\theta}_G(D_{query})$ ▷ Generate query-synthetic set
for each query sample **do**
  $S_{comd} \leftarrow max\left(CosineSimilarity(f_Q, f_{QS})\right)$
end for
for each query sample **do**
  $S_{comp} \leftarrow S_{init} + S_{cond}$
  if $S_{comp} > \tau^*$ **then**
    y ← 1 ▷ Member
  else
    y ← 0 ▷ Non-member
  end if
end for

Calculating Conditional Similarity. To evaluate the alignment between the query data and query-conditioned synthetic samples, we compute the conditional similarity S_{cond} , which focuses on capturing the fine-grained details of query data and provides a query-

specific alignment score. For each feature f_Q of the query sample in D_{query} the conditional similarity is computed as:

$$S_{cond}(f_Q) = \max_{f_{QS} \in D_{qs}} \frac{\langle f_{QS}, f_Q \rangle}{\|f_{QS}\| \cdot \|f_Q\|} \tag{3}$$

By selecting the maximum similarity score between a query sample and the generated images, S_{cond} captures the strongest alignment for each query sample, focusing on its query-specific characteristics.

Calculating Initial Similarity. Although conditional generation effectively incorporates query-specific details, non-client data can still influence the generated results. This may lead to an overestimated conditional similarity S_{cond} for non-client data, reducing the discriminative power of the framework. To address this issue, we introduce S_{init} from attackI-Base to calcu- late the relationship between the query data and the overall client-distributed data synthetic samples:

$$S_{init}(f_Q) = \frac{\langle f_Q, \bar{f}_G \rangle}{\|f_Q\| \cdot \|\bar{f}_G\|} \tag{4}$$

By assessing the query data's strongest global alignment, S_{init} serves to counterbalance potential overestimations in S_{cond}, thereby enhancing the robustness and reliability of the overall evaluation.

Computing a Comprehensive Similarity Score, Finally, we obtain a comprehensive similarity score S_{comp} as the sum of initial similarity S_{init} and conditional similarity S_{cond} :

$$S_{comp} = S_{init} + S_{cond} \tag{5}$$

By combining both S_{cond} and S_{init}, the strategy ensures higher similarity scores for genuine client data, enabling robust and accurate membership inference across diverse scenarios.

Classifying Client Membership. Similar to AttackI-Base, we use ROC based on the distribution of the composite similarity score S_{comp} to determine the optimal threshold for inference τ^* and to infer membership in the query sample:

$$y = \begin{cases} 1, & if\ S_{comp} \geq \tau^*, \\ 0, & if\ S_{comp} < \tau^*. \end{cases} \tag{6}$$

Table 1. Evaluation of Attack Methods Across Different Architectures and Datasets (FEDDEO and FEDDISC).

Dataset	Method	Backbone	FEDDEO			FEDDISC		
			ASR	AUC	TPR@1%	ASR	AUC	TPR@1%
OxfordPet	AttackI-Base	ResNet18	0.52	0.52	2.62%	**0.68**	0.71	0.88%
		ResNet50	0.52	0.54	3.17%	**0.68**	0.71	1.33%
		ResNet101	0.55	0.58	6.20%	**0.68**	0.71	0.88%
		DeiT	0.48	0.49	0.96%	0.54	0.48	0.32%
		ViT	0.50	0.51	0.41%	0.57	0.49	1.20%
	AttackII-Finetune	ResNet18	0.52	0.57	2.75%	0.65	0.68	1.52%
		ResNet50	0.58	0.66	15.98%	**0.68**	**0.72**	**3.41%**
		ResNet101	**0.59**	**0.69**	**16.80%**	0.67	0.71	2.40%
		DeiT	0.51	0.54	0.83%	0.48	0.47	0.32%
		ViT	0.51	0.50	0.83%	0.48	0.48	0.95%
DomainNet	AttackI-Base	ResNet18	0.63	0.73	23.28%	0.55	0.72	9.70%
		ResNet50	0.64	0.74	25.57%	0.58	0.76	12.08%
		ResNet101	0.67	0.82	36.24%	0.62	0.79	20.37%
		DeiT	0.51	0.66	1.94%	0.59	0.60	3.26%
		ViT	0.58	0.62	3.62%	0.55	0.66	2.56%
	AttackII-Finetune	ResNet18	0.70	**0.84**	24.78%	0.60	0.81	17.90%
		ResNet50	0.70	0.8	29.37%	0.60	0.80	17.11%
		ResNet101	**0.70**	0.83	**39.77%**	**0.66**	**0.82**	**26.01%**
		DeiT	0.61	0.63	3.62%	0.51	0.60	3.88%
		ViT	0.52	0.61	7.14%	0.64	0.63	3.26%
OxfordFlower	AttackI-Base	ResNet18	0.50	0.50	0.12%	0.50	0.50	0.08%
		ResNet50	0.49	0.50	0.02%	0.50	0.46	0.38%
		ResNet101	0.50	0.52	0.38%	0.50	0.52	0.75%
		DeiT	0.50	0.48	0.75%	0.51	0.46	**2.00%**
		ViT	0.50	0.28	0.02%	0.50	0.50	1.38%
	AttackII-Finetune	ResNet18	**0.61**	0.62	2.12%	**0.59**	**0.56**	1.12%
		ResNet50	0.60	**0.64**	**3.12%**	0.57	0.51	0.50%
		ResNet101	0.59	0.60	0.62%	0.57	**0.56**	0.75%
		DeiT	0.51	0.47	0.62%	0.51	0.45	0.75%
		ViT	0.54	0.53	1.25%	0.53	0.50	0.38%

3.3 Experimental Setup

Dataset: In our experiments, we utilize three widely-used image datasets: **Oxford-IIIT Pet** [16], **DomainNet** [17], and **Oxford 102 Flowers**[18]. For each dataset, we configure the number of clients to 3,10, and 50, respectively, to evaluate the scalability of the proposed approach. To simulate non-i.i.d. data distributions across clients, we adopt the

Dirichlet Distribution with a concentration parameter $\alpha = 1$, ensuring diverse yet controlled levels of data imbalance.

Target approaches:In our evaluation, we conduct attacks against several state-of-the-art approaches, including **FedDEO** [1], **FedCADO** [3], **FedDISC** [2], and **FedBIP** [10]. These approaches represent different methodologies within the OSFL framework, allowing us to comprehensively evaluate the effectiveness of our attack strategy across various target methods.

Table 2. Evaluation of Attack Methods Across Different Architectures and Datasets(FedCADO and FEDBIP)

Dataset	Method	Backbone	FEDCADO			FEDBIP		
			ASR	AUC	TPR@1%	ASR	AUC	TPR@1%
OxfordPet	AttackI-Base	ResNet18	0.61	0.61	2.02%	0.73	0.77	2.72%
		ResNet50	0.61	0.59	3.03%	0.77	0.82	7.65%
		ResNet101	0.62	0.59	4.05%	0.73	0.82	7.27%
		DeiT	0.51	0.51	1.45%	0.50	0.49	0.88%
		ViT	0.53	0.53	1.07%	0.52	0.51	1.01%
	AttackII-Finetune	ResNet18	0.61	0.59	1.96%	0.67	0.75	4.36%
		ResNet50	0.62	0.58	5.12%	0.72	0.82	**11.38%**
		ResNet101	**0.67**	**0.71**	**5.69%**	**0.78**	**0.83**	8.41%
		DeiT	0.52	0.52	0.76%	0.51	0.51	1.20%
		ViT	0.50	0.50	1.01%	0.48	0.49	1.26%
DomainNet	AttackI-Base	ResNet18	0.62	0.62	18.87%	0.61	0.59	8.29%
		ResNet50	0.64	0.67	21.08%	0.48	0.54	8.38%
		ResNet101	0.69	0.69	26.10%	0.56	0.56	10.85%
		DeiT	0.51	0.55	0.60%	0.60	0.56	1.85%
		ViT	0.48	0.57	0.79%	0.48	0.57	0.79%
	AttackII-Finetune	ResNet18	0.68	0.78	20.19%	0.59	0.67	9.79%
		ResNet50	0.68	0.80	29.37%	0.62	0.69	**11.90%**
		ResNet101	**0.70**	**0.84**	**31.92%**	**0.67**	**0.70**	11.73%
		DeiT	0.57	0.63	4.06%	0.58	0.64	2.20%
		ViT	0.62	0.61	4.59%	0.57	0.59	4.50%
OxfordFlower	AttackI-Base	ResNet18	0.49	0.48	0.25%	0.50	0.53	0.88%
		ResNet50	0.50	0.44	0.38%	0.50	0.52	1.00%
		ResNet101	0.50	0.49	0.38%	0.49	0.54	0.02%
		DeiT	0.49	0.46	0.02%	0.49	0.51	0.25%
		ViT	0.50	0.24	0.12%	0.50	0.53	0.50%
	AttackII-Finetune	ResNet18	**0.56**	**0.56**	0.38%	0.62	0.63	1.38%
		ResNet50	0.51	0.49	0.12%	**0.63**	**0.68**	**3.50%**
		ResNet101	0.54	0.52	0.12%	0.62	0.64	1.00%
		DeiT	0.52	0.49	0.25%	0.50	0.46	0.50%
		ViT	0.55	0.55	**3.75%**	0.51	0.47	1.12%

Evaluation Metrics:To evaluate the effectiveness of the proposed attack strategy, we conduct experiments under a fixed 1% False Positive Rate (FPR). The evaluation metrics include:
(1) Attack Success Rate (ASR), measuring the accuracy in distinguishing member from non-member samples;
(2) Area Under the ROC Curve (AUC), providing a threshold- independent assessment of attack performance;
(3) True Positive Rate at 1% FPR (TPR@1%), indicating the proportion of correctly identified member samples under the FPR constraint. Our experiments are conducted on a NVIDIA GeForce RTX 3090 GPU, leveraging the pre-trained Stable Diffusion v1-5 model[19] and the official implementation of the LoRA method .

3.4 Performance Analysis of Attack Methods

We systematically evaluate the impact of different backbone architectures for feature extraction in both strategies, including ResNet18/50/101 [20], Vison Transformer (ViT) [21], and DeiT [22]. The ResNet series, known for its hierarchical feature representation, is highly effective at capturing fine-grained image details. In contrast, ViT employs a self-attention mechanism to model global contextual information. DeiT, a data-efficient variant of ViT, achieves comparable feature extraction performance while reducing dependency on large-scale datasets.

Table 3. Comparison of Different Numbers of Clients for FEDDISC.

Dataset	Clients	AttackI-Base			AttackII-Finetune		
		ASR	AUC	TPR@1%	ASR	AUC	TPR@1%
OxfordPet	3	0.68	0.71	0.88%	0.67	0.72	2.40%
	10	0.68	0.72	0.88%	0.70	0.75	2.02%
	50	0.70	0.73	1.07%	0.69	0.74	3.03%
DomainNet	3	0.62	0.79	20.37%	.0.66	0.82	26.01%
	10	0.60	0.81	20.11%	0.69	0.84	25.75%
	50	0.68	0.84	30.16%	0.69	0.88	31.39%
OxfordFlower	3	0.50	0.52	0.75%	0.57	0.56	0.75%
	10	0.50	0.52	0.75%	0.68	0.57	1.00%
	50	0.50	0.53	1.50%	0.65	0.57	0.88%

Table 4. Comparison of Different Numbers of Clients for FEDCADO.

Dataset	Clients	AttackI-Base			AttackII-Finetune		
		ASR	AUC	TPR@1%	ASR	AUC	TPR@1%
OxfordPet	3	0.62	0.60	4.05%	0.67	0.71	5.69%
	10	0.62	0.61	4.11%	0.70	0.75	4.87%
	50	0.62	0.6	3.60%	0.69	0.74	3.29%
DomainNet	3	0.56	0.56	10.85%	0.66	0.82	11.73%
	10	0.70	0.69	25.31%	0.73	0.87	32.54%
	50	0.62	0.75	26.01%	0.73	0.83	31.75%
OxfordFlower	3	0.50	0.52	0.38%	0.54	0.52	0.12%
	10	0.50	0.52	0.12%	0.55	0.53	0.25%
	50	0.50	0.53	0.50%	0.58	0.55	0.25%

As shown in Table 1 and 2, where bold values represent the best attack performance for each dataset, our experimental results highlight robust attack performance across different target approaches and backbone architectures, with ASR consistently surpassing 50% and reaching up to 70% in several scenarios. A closer comparison between AttackI-Base and AttackII-Finetune reveals that AttackII-Finetune generally achieves better overall performance, particularly excelling in terms of TPR@1%. This improvement is especially pronounced in scenarios with strict false positive constraints, where AttackII-Finetune demonstrates significantly higher true positive rates while maintaining the same false positive rate. These results indicate that AttackII-Finetune is more effective at distinguishing member samples from non-member samples, achieving a more reliable balance between sensitivity and specificity.

Table 5. Comparison of Different Numbers of Clients for FEDDEO.

Dataset	Clients	AttackI-Base			AttackII-Finetune		
		ASR	AUC	TPR@1%	ASR	AUC	TPR@1%
OxfordPet	3	0.55	0.58	6.20%	0.59	0.69	16.80%
	10	0.48	0.50	1.74%	0.51	0.61	2.18%
	50	0.43	0.43	3.48%	0.48	0.58	8.11%
DomainNet	3	0.67	0.82	36.24%	0.70	0.83	39.77%
	10	0.59	0.71	16.31%	0.63	0.73	16.70%
	50	0.63	0.71	17.61%	0.60	0.68	17.20%
OxfordFlower	3	0.50	0.52	0.38%	0.59	0.6	0.62%
	10	0.50	0.51	0.76%	0.63	0.56	0.07%
	50	0.50	0.50	1.19%	0.63	0.57	0.15%

Table 6. Comparison of Different Numbers of Clients for FEDBIP.

Dataset	Clients	AttackI-Base			AttackII-Finetune		
		ASR	AUC	TPR@1%	ASR	AUC	TPR@1%
OxfordPet	3	0.78	0.83	7.27%	0.73	0.82	8.41%
	10	0.68	0.71	2.02%	0.63	0.72	1.07%
	50	0.61	0.61	4.05%	0.59	0.69	3.98%
DomainNet	3	0.69	0.69	26.10%	0.70	0.84	31.92%
	10	0.61	0.59	11.82%	0.63	0.74	13.32%
	50	0.65	0.59	12.79%	0.60	0.69	13.76%
OxfordFlower	3	0.49	0.54	0.02%	0.62	0.64	1.00%
	10	0.50	0.51	0.88%	0.66	0.6	0.12%
	50	0.50	0.5	1.38%	0.66	0.61	0.25%

Across both attacks, we observe that ResNet consistently outperforms ViT in attack effectiveness. This can be attributed to the fact that, when client and non-client data originate from the same dataset, global differences between the two are minimal, placing greater emphasis on local feature comparison. While ViT prioritizes global contextual information, it often weakens the representation of local features, particularly in its shallow layers. In contrast, ResNet's hierarchical architecture progressively builds robust local representations, giving it an advantage in capturing fine-grained details [23].

Additionally, we note that attack performance on the Oxford 102 Flowers dataset is lower compared with other datasets. This is due to the inherent characteristics of the dataset, including large intra-class variations and small differences between classes [18]. Even when member and non-member samples belong to different classes, the high similarity between classes undermines the effectiveness of attack strategies that rely solely on model response similarity for distinguishing samples. This highlights the challenges posed by datasets with subtle inter-class distinctions and diverse intra-class features.

3.5 Impact of Different Distance Metrics

Based on our comprehensive evaluation across multiple target approaches, summarized in Table 1, ResNet101 emerges as the most effective and stable feature extractor.

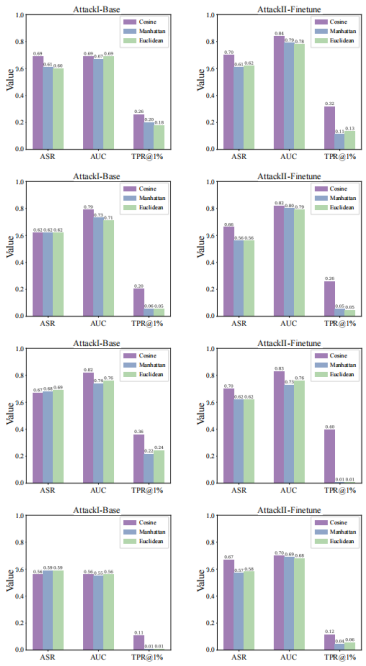

Fig. 3. Comparison of different distance metrics for FEDCADO (first row) and FEDDISC (second row) FEDDEO (third row) FEDBIP (fourth row)

As a result, ResNet101 is selected as the primary feature extractor for further analysis of different similarity measures, including Cosine Similarity, Euclidean Distance, and Manhattan Distance. As illustrated in Fig. 3, we evaluate the impact of different similarity metrics on the effectiveness of the attacks within both strategies on the DomainNet dataset in the case of three clients. Among the evaluated metrics, the Cosine Similarity shows the highest bar in the bar chart, indicating its better effectiveness in our attack scenarios by successfully identifying the inherent correlations between data patterns.

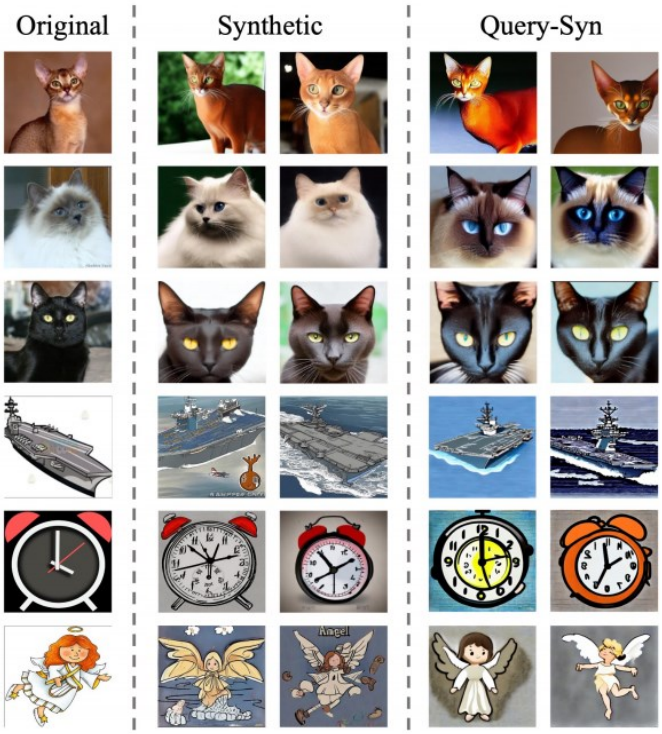

Fig. 4. Visualization of Original Client Data, Synthetic Data, and Query-based Synthetic Data.

3.6 The Effect of Different Client Numbers

Based on the above analysis, we select ResNet101 as the backbone to further evaluate the impact of the number of clients on attack performance. As shown in Table 3-Table 6, we analyze the effect of varying client numbers within the FEDDISC, FEDCADO, FEDDEO and FEDBIP methods. Notably, our attack performance remains robust across different client configurations, indicating that the effectiveness of our method is not compromised by changes in the number of clients. This demonstrates the reliability and stability of our approach under varying federated settings.

3.7 Visualization of Synthetic Images

As shown in Fig. 4, we visually present original images, initial synthetic images, and query-conditioned synthetic images. The initial synthetic images capture the global distribution of client data, representing broad dataset characteristics. In contrast, the query-conditioned synthetic images refine these results by incorporating the finegrained details of the query data, ensuring better semantic alignment and detail preservation.

4 Conclusion

In this work, we propose a MIA strategy tailored for generative models based OSFL, which includes AttackI - Base, a general feature similarity based method applicable to all generative models, and AttackII - Finetune, designed specifically for diffusion models using fine - tuning and conditional generation to capture fine - grained query details. Experimental results demonstrate that both strategies effectively uncover privacy vulnerabilities in OSFL, showing that even diverse and high-quality synthetic data from generative models remain susceptible to exploitation. However, they have limitations in identifying the specific client of the queried data, which remains a key direction for improving attack efficacy and addressing OSFL privacy risks.

References

1. Yang, M., Su, S., Li, B., Xue, X.: Feddeo: Description-enhanced one-shot federated learnin g with diffusion models. In: Proceedings of the 32nd ACM International Conference on M ultimedia, pp. 6666—6675 (2024)
2. Yang, M., Su, S., Li, B., Xue, X.: Exploring one-shot semi-supervised feder- ated le arning with pre-trained diffusion models. In: Proceedings of the AAAI Conference on Artif icial Intelligence, vol. 38, pp. 16325—16333 (2024)
3. Yang, M., Su, S., Li, B., Xue, X.: One-shot federated learning with classifier- guided diffus ion models. arXiv preprint arXiv:2311.08870 (2023)
4. Zhang, J., Chen, C., Li, B., Lyu, L., Wu, S., Ding, S., Shen, C., Wu, C.: Dense: Data-free one-shot federated learning. Advances in Neural Information Processing Systems 35, 21414—21428 (2022)
5. Mendieta, M., Sun, G., Chen, C.: Navigating heterogeneity and privacy in one- shot federat ed learning with diffusion models. arXiv preprint arXiv:2405.01494 (2024)
6. Melis, L., Song, C., De Cristofaro, E., Shmatikov, V.: Exploiting unintended fea- ture leak age in collaborative learning. In: 2019 IEEE Symposium on Security and Privacy (SP), pp. 691—706 (2019). IEEE
7. Zhang, J., Zhang, J., Chen, J., Yu, S.: Gan enhanced membership inference: A passive loca l attack in federated learning. In: ICC 2020-2020 IEEE International Conference on Comm unications (ICC), pp. 1—6 (2020). IEEE
8. Song, L., Mittal, P.: Systematic evaluation of privacy risks of machine learning models. In: 30th USENIX Security Symposium (USENIX Security 21), pp. 2615— 2632 (2021)
9. Heinbaugh, C.E., Luz-Ricca, E., Shao, H.: Data-free one-shot federated learn- ing under v ery high statistical heterogeneity. In: The Eleventh International Conference on Learni ng Representations (2023)
10. Chen, H., Li, H., Zhang, Y., Zhang, G., Bi, J., Torr, P., Gu, J., Krompass, D., Tresp, V.: Fe dbip: Heterogeneous one-shot federated learning with personalized latent diffusion models. arXiv preprint arXiv:2410.04810 (2024)
11. Gupta, U., Stripelis, D., Lam, P.K., Thompson, P., Ambite, J.L., Ver Steeg, G.: Membership inference attacks on deep regression models for neuroimaging. In: Medical Im aging with Deep Learning, pp. 228—251 (2021). PMLR
12. Nasr, M., Shokri, R., Houmansadr, A.: Comprehensive privacy analysis of deep learning: P assive and active white-box inference attacks against centralized and federated learning. In: 2019 IEEE Symposium on Security and Privacy (SP), pp. 739—753 (2019). IEEE

13. Duan, J., Kong, F., Wang, S., Shi, X., Xu, K.: Are diffusion models vulnerable to members hip inference attacks? In: International Conference on Machine Learning, pp. 8717—8730 (2023). PMLR
14. Hassanzad, M., Hajian-Tilaki, K.: Methods of determining optimal cut-point of diagnostic biomarkers with application of clinical data in roc analysis: an update review. BMC Medic al Research Methodology 24(1), 84 (2024)
15. Ryu, S.: Low-rank adaptation for fast text-to-image diffusion fine-tuning. Low- rank adapt ation for fast text-to-image diffusion fine-tuning (2023)
16. Parkhi, O.M., Vedaldi, A., Zisserman, A., Jawahar, C.: Cats and dogs. In: 2012 IEEE Conf erence on Computer Vision and Pattern Recognition, pp. 3498—3505 (2012). IEEE
17. Peng, X., Bai, Q., Xia, X., Huang, Z., Saenko, K., Wang, B.: Moment matching for multi-s ource domain adaptation. In: Proceedings of the IEEE/CVF International Conference on C omputer Vision, pp. 1406—1415 (2019)
18. Nilsback, M.-E., Zisserman, A.: Automated flower classification over a large num- ber of c lasses. In: 2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processin g, pp. 722—729 (2008). IEEE
19. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image syn thesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Com puter Vision and Pattern Recognition, pp. 10684—10695 (2022)
20. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recogni- tion. In: Proc eedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770—7 78 (2016)
21. Alexey, D.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv: 2010.11929 (2020)
22. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., J´egou, H.: Training data-efficient image transformers & distillation through attention. In: Internationa l Conference on Machine Learning, pp. 10347—10357 (2021). PMLR
23. Raghu, M., Unterthiner, T., Kornblith, S., Zhang, C., Dosovitskiy, A.: Do vision tran sformers see like convolutional neural networks? Advances in neural information processin g systems 34, 12116—12128 (2021)