Geometric Alignment Constraints for Mitigating Membership Inference in Federated Learning

Ruijia Liu, Xiaochen Nie, Jianhua Wang*, Yongle Chen, Dan Yu

***Abstract*—Federated learning avoids raw data sharing but remains vulnerable to membership inference attacks (MIAs). Recent generative federated learning improves utility under non-independent and non-identically distributed (non-IID) data by synthesizing complementary data; however, the generation stage can distort representation geometry, increasing member/non-member separability and strengthening both loss-based and representation-based MIAs. Existing defenses, such as differential privacy and gradient clipping, operate at the update level and do not address this generation-induced leakage. We propose FedGAC, a lightweight geometric alignment strategy applied during complementary data generation. By imposing classifier-induced class-wise geometric constraints, FedGAC suppresses client-specific directions while preserving class-level semantics, without modifying the federated learning protocol or local training. Extensive experiments across diverse non-IID settings show that FedGAC substantially reduces MIA success with negligible accuracy loss, achieving a better privacy–utility trade-off.**

***Index Terms*—Federated learning, membership inference, generative FL, geometric alignment, privacy–utility trade-off**

I. Introduction

Federated learning (FL) has emerged as a promising paradigm for collaborative model training in privacy-sensitive distributed systems, where raw data are kept local at participating clients. It has been widely adopted in privacy-sensitive edge, cloud–fog, and industrial systems as a practical solution for balancing utility and privacy [1]. However, recent studies show that federated models can still leak sensitive information about client participation under strong server-side adversaries, even without sharing raw data [2].

As FL scales to large edge and cloud–fog systems, prior work has highlighted security and robustness challenges in practical deployment pipelines [3]–[5]. Meanwhile, complementary data generation has been introduced to mitigate optimization instability caused by heterogeneous client data [6], yet avoiding raw data sharing does not eliminate privacy leakage, and trained models may still expose sensitive information about client participation and training data characteristics [7].

Membership inference threats. Membership inference attacks (MIAs) are among the most critical privacy threats in federated learning, including loss-based and representation-based variants. Recent studies show that MIAs have advanced to strong server-side and white-box settings that exploit losses, gradients, and internal representations [2], [4]. Representation-level signals can further enable effective and targeted MIAs even under secure aggregation or protocol-level protections [8].

Challenges in MIA defense. Defending against MIAs in federated learning faces two fundamental challenges. First, under non-independent and non-identically distributed (non-IID) data, the privacy–utility trade-off becomes particularly severe. Existing defenses such as differential privacy, gradient clipping, noise injection, and privacy amplification suppress membership-related signals but often introduce optimization noise or gradient distortion, leading to substantial accuracy degradation [9]–[11]. Second, membership leakage increasingly manifests at the representation level, where strong server-side attacks exploit internal representations rather than only prediction confidence or loss values [8], [12]. Most existing defenses operate at the gradient or update level and do not explicitly regulate representation geometry, leaving feature-space separability largely unmitigated under severe data heterogeneity.

Generative federated learning. To alleviate performance degradation caused by non-IID data, recent generative federated learning frameworks synthesize complementary data without sharing raw samples [6]. These ideas have been further extended through synthetic data generation and data-free distillation techniques to enrich local training distributions and improve utility under severe heterogeneity [13], [14]. However, complementary data generation can introduce a new privacy risk: when guided by heterogeneous client models, it may amplify client-specific discriminative directions in the learned representation space, increasing member/non-member separability and strengthening representation-based MIAs [15], [16]. Importantly, this generation-induced leakage is not addressed by existing gradient- or update-level defense mechanisms.

Motivation and approach. The above analysis suggests that defending against MIAs in realistic non-IID settings remains challenging. Beyond technical privacy leakage, such attacks can also weaken informed consent and participant autonomy in federated learning, because clients may contribute data without being able to assess or control representation-level exposure introduced during training. From an engineering-trust perspective, practical FL defenses should therefore not only preserve utility, but also provide measurable mitigation of privacy risks arising in realistic deployment pipelines. Motivated by this gap, we propose FedGAC, a lightweight geometric

*Corresponding author: Jianhua Wang.

R. Liu, X. Nie, J. Wang, Y. Chen, and D. Yu are with the Shanxi Key Laboratory of Industrial Internet Security, College of Computer Science and Technology, Taiyuan University of Technology, Taiyuan, China (e-mail: 15806855914@163.com; 18831064657@163.com; wangjianhua02@tyut.edu.cn; chenyongle@tyut.edu.cn; yudan@tyut.edu.cn).

alignment constraint applied at the complementary data generation stage. FedGAC regularizes generated representations using class-wise references induced by the global classifier, suppressing client-specific discriminative directions while preserving class-level semantic structure, without modifying the federated learning protocol or local training procedure.

We make three contributions.

- We show that complementary data generation can amplify membership leakage by increasing representation separability under non-IID data.
- We propose FedGAC, which constrains generated representations using class-wise geometric references induced by the global classifier.
- Experiments under server-side MIAs demonstrate improved privacy with negligible accuracy loss.

Data and Code Availability. The code is publicly available at: https://github.com/SKLIIS-AIS/FedGAC.git.

II. Method: FedGAC

We propose **FedGAC (Federated Learning with Geometric Alignment Constraints)**, which introduces geometric alignment constraints into the complementary data generation stage of FedCOG to mitigate membership inference leakage under non-IID data.

A. Problem Formulation and Threat Model

We consider a standard cross-device federated learning (FL) setting with a central server and a set of K clients, denoted as $\mathcal{C} = \{1, 2, \ldots, K\}$. Each client $i \in \mathcal{C}$ holds a private local dataset $D_i = \{(x_{i,j}, y_{i,j})\}_{j=1}^{n_i}$, drawn from a client-specific data distribution P_i. Due to the decentralized and real-world nature of FL, the local data distributions are generally non-identical and non-independent (non-IID), i.e., $P_i \neq P_j$ for $i \neq j$.

The goal of federated learning is to collaboratively train a global model $f_g(\cdot;\theta)$ without sharing raw client data by minimizing:

$$\min_{\theta} \sum_{i=1}^{K} p_i \, \mathbb{E}_{(x,y)\sim P_i}\big[\ell(f_g(x;\theta), y)\big], \tag{1}$$

where p_i denotes the aggregation weight of client i, and $\ell(\cdot,\cdot)$ is the task loss (e.g., cross-entropy).

Threat model. We consider **membership inference attacks (MIA)** under an honest-but-curious server setting. The server follows the prescribed FL protocol but attempts to infer whether a given sample was used in client training. In particular, we consider two representative and widely adopted server-side MIA settings:

Loss-based Membership Inference Attack (LB-MIA). We adopt the standard loss-based server-side membership inference attack proposed by Yeom et al. [17] as a representative baseline.

Representation-based Membership Inference Attack (FedMIA). We evaluate a feature-based membership inference attack following the FedMIA framework proposed by Zhu et al. [18], which exploits representation-level consistency for membership inference.

Both attacks aim to infer a binary membership variable $m \in \{0, 1\}$, where $m = 1$ indicates that a sample belongs to the training set of some client. We evaluate attack performance using AUC and TPR at low FPR thresholds, together with model utility measured by classification accuracy.

We consider a standard server-side setting where the attacker has full access to the trained global model, including the feature extractor and classifier weights from which class-wise geometric references are derived, but does not influence the training process and only uses model parameters for inference.

Problem statement. Given a generative federated learning framework that mitigates data heterogeneity through complementary data generation, our objective is to design a **lightweight and plug-and-play mechanism** that reduces vulnerability to membership inference attacks **without modifying the original communication protocol or KD-based local training procedure**.

B. Revisiting FedCOG: Federated Training Pipeline

We first revisit the training pipeline of FedCOG, which serves as the backbone of our proposed method. FedCOG is a representative generative federated learning framework designed to mitigate data heterogeneity by synthesizing complementary training samples for local updates. At a high level, FedCOG proceeds in communication rounds indexed by $t = 1, \ldots, T$, and each round consists of coordinated steps between the server and clients, as illustrated in Fig. 1(a).

Step 1: Global model broadcast. At round t, the server broadcasts the global model $f_g(\cdot;\theta^{(t)})$ to all participating clients.

Step 2: Complementary data generation. FedCOG synthesizes complementary samples $\tilde{D}_i^{(t)}$ for each client using a generator guided by heterogeneous clients.

Step 3: KD-based local training. Each client performs KD-based local training on its local data D_i and complementary data $\tilde{D}_i^{(t)}$.

Step 4: Model aggregation. Clients upload local models to the server, which aggregates them to obtain the updated global model for the next round.

C. Motivation: Representation-Level Privacy Risk

In federated learning, membership inference can exploit signals exposed by the trained model. Beyond confidence scores or loss values, intermediate representations learned by the global model may also leak membership information, especially under non-IID data distributions.

Let $f_g(\cdot)$ denote the global feature extractor of the trained model. For an input sample x, we define its representation as:

$$\mathbf{z} = f_g(x), \tag{2}$$

where $\mathbf{z} \in \mathbb{R}^d$ lies in the learned representation space.

Generative federated learning frameworks such as FedCOG synthesize complementary samples to mitigate client drift. However, the generation process is guided by heterogeneous

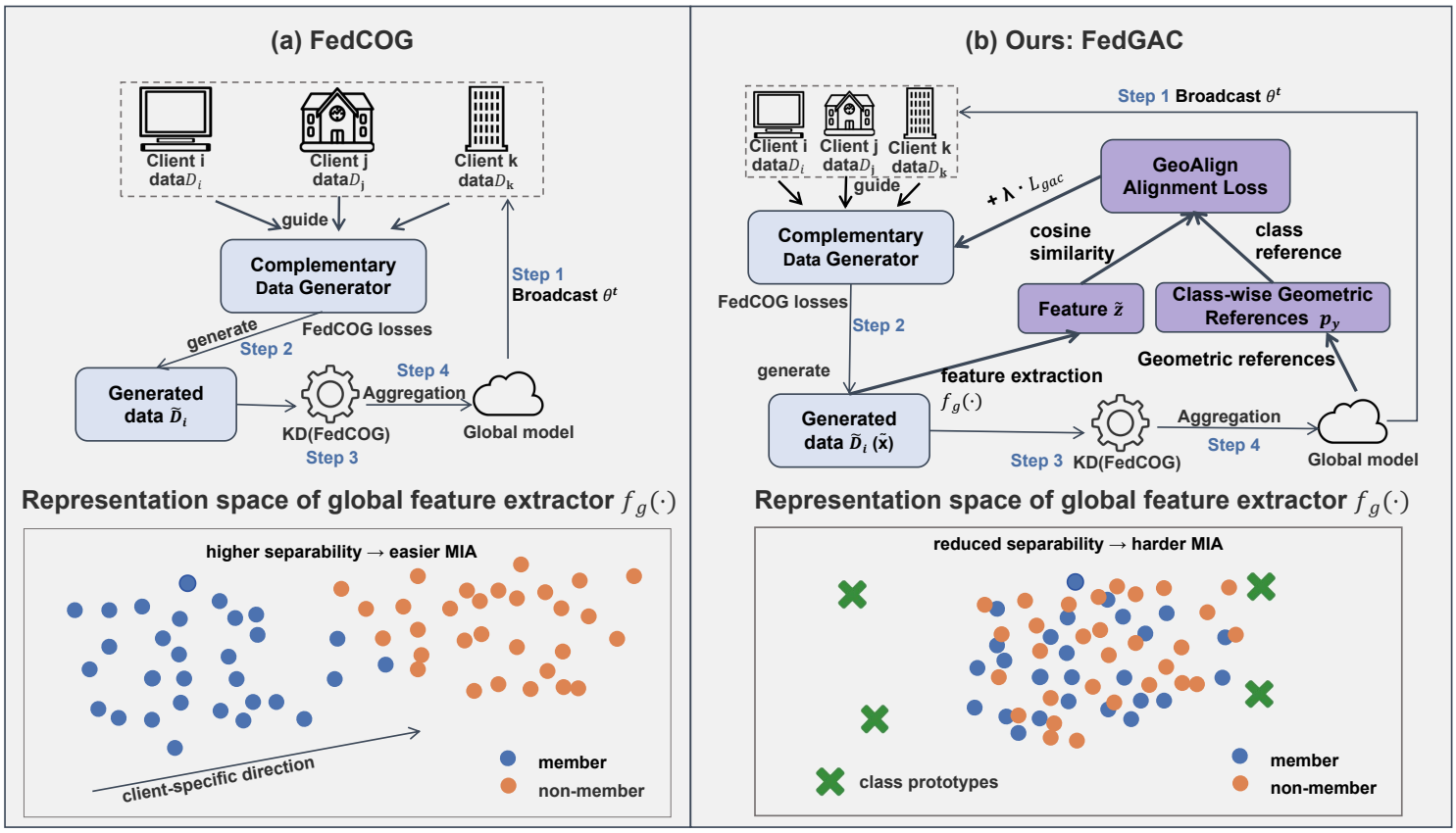


Fig. 1: Overview of FedCOG and the proposed FedGAC. FedGAC preserves the original pipeline and introduces a class-wise geometric alignment constraint only during complementary data generation. The constraint is derived from classifier-induced class reference geometry and regularizes generated representations to reduce member/non-member separability.

clients and may implicitly encode **client-specific discriminative directions** in the representation space. As a result, representations of member and non-member samples may become more separable under the global feature extractor, strengthening feature-based membership inference attacks.

D. Proposed Method: Geometric Alignment Constraints

We propose **FedGAC**, which improves FedCOG-style complementary generation with a class-wise geometric alignment constraint. FedGAC **only modifies the generator optimization stage**; KD-based local training and the communication protocol remain unchanged, as shown in Fig. 1(b).

1) Class-Wise Geometric Reference Vectors: We consider the global model to consist of a feature extractor $f_g(\cdot)$ and a classifier with weight matrix $W \in \mathbb{R}^{C\times d}$, where C is the number of classes and d is the feature dimension. Following prior observations that classifier weights encode class-wise structure in the representation space [19], we directly use them as **class-wise geometric reference vectors**:

$$\mathbf{p}_y \triangleq W_y, \quad y \in \{1, \dots, C\}, \tag{3}$$

where W_y is the y-th row of W. This design introduces no additional models or communication overhead since reference vectors are directly obtained from the existing global model parameters. These vectors are not empirical class centroids, but classifier-induced geometric anchors that reflect global decision geometry rather than data-dependent statistics.

2) Geometric Alignment Loss: For each generated sample $\tilde{x}$ with target class $\tilde{y}$, FedGAC encourages its representation to follow the *global class geometry* of class $\tilde{y}$ rather than drifting toward client-specific directions. Let $\mathbf{z}(\tilde{x}) = f_g(\tilde{x})$ denote its representation in the global feature space. We impose the following cosine-based alignment loss:

$$\mathcal{L}_{\text{geo}} = \mathbb{E}_{(\tilde{x},\tilde{y})}\Big[1 - \cos\big(\hat{\mathbf{z}}(\tilde{x}), \hat{\mathbf{p}}_{\tilde{y}}\big)\Big], \tag{4}$$

where $\hat{\mathbf{z}} = \mathbf{z}/\|\mathbf{z}\|_2$ and $\hat{\mathbf{p}} = \mathbf{p}/\|\mathbf{p}\|_2$ denote ℓ_2-normalized vectors. This constraint acts on feature directions rather than magnitudes, thereby regularizing the geometry of generated representations. As a result, generated samples are encouraged to align with the **global class-level structure** instead of drifting toward client-specific directions.

Algorithm 1 FedGAC: Federated Learning with Geometric Alignment Constraints

Initialization: clients K, rounds T, local epochs τ, weights $\{p_i\}$.
Input: local datasets $\{D_i\}_{i=1}^{K}$, regularization weight λ.
Output: final global model $\theta^{(T)}$.
for $t = 0, 1, \ldots, T-1$ **do**
  Server broadcasts global model $\theta^{(t)}$ to all clients.
  Extract class-wise reference vectors $\mathbf{p}_y \leftarrow W_y^{(t)}$.
  for each client $i \in \mathcal{C}$ **in parallel do**
    Generate complementary data $\tilde{D}_i^{(t)}$ by minimizing $\mathcal{L}_{\text{FedGAC-gen}} = \mathcal{L}_{\text{FedCOG-gen}} + \lambda \mathcal{L}_{\text{geo}}$.
    where $\mathcal{L}_{\text{geo}} = 1 - \cos(\hat{z}(\tilde{x}), \hat{\mathbf{p}}_{\tilde{y}})$.
    Perform KD-based local training on $D_i \cup \tilde{D}_i^{(t)}$ for τ epochs.
    Upload updated local model $\theta_i^{(t)}$ to server.
  Server aggregates global model: $\theta^{(t+1)} \leftarrow \sum_{i=1}^{K} p_i \theta_i^{(t)}$.
end for
return $\theta^{(T)}$.

3) Overall Objective and Plug-and-Play Property: Let $\mathcal{L}_{\text{FedCOG-gen}}$ denote the original FedCOG objective used for complementary data generation. FedGAC modifies only the generator optimization as:

$$\mathcal{L}_{\text{FedGAC-gen}} = \mathcal{L}_{\text{FedCOG-gen}} + \lambda \mathcal{L}_{\text{geo}}, \tag{5}$$

where $\lambda > 0$ controls the strength of geometric alignment.

All other steps remain unchanged: KD-based local training is identical to FedCOG, the communication and aggregation procedures are the same, and no additional information is exchanged. Therefore, FedGAC can be used as a drop-in replacement for the FedCOG generation step.

E. Training Algorithm

Algorithm 1 summarizes FedGAC. The only difference from FedCOG lies in the complementary data generation stage, where the generator is optimized with the additional geometric alignment term, while the communication protocol and KD-based local training remain unchanged.

F. Informal Privacy Analysis

We provide an informal analysis to explain how FedGAC mitigates membership inference risks by weakening representation-level attack signals, without claiming formal privacy guarantees.

Effect of generative federated learning. In generative federated learning (e.g., FedCOG), complementary data are generated under the guidance of heterogeneous clients to improve utility under non-IID data. This process may implicitly introduce client-specific discriminative directions in the learned representation space, increasing member/non-member separability.

Effect of geometric-alignment-guided generation. FedGAC introduces a geometric alignment constraint during complementary data generation. By aligning generated representations with classifier-induced class reference geometry, FedGAC suppresses client-specific directions and reduces member/non-member separability in the feature space.

Relation to representation-based MIAs. Although the geometric constraint is related to the representation structure exploited by FedMIA, it is not tied to any specific attack score or decision rule. FedGAC acts during generation to regularize the learned feature geometry itself, rather than optimizing against a particular inference threshold or attacker objective. Therefore, its intended effect is to reduce representation-level separability more generally, which may also weaken feature- based membership signals beyond the specific cosine-consistency form used by FedMIA.

III. Experiments

This section evaluates the effectiveness of the proposed FedGAC from both utility and privacy perspectives. We conduct experiments under multiple non-IID federated settings and assess robustness against representative server-side membership inference attacks. The goal is to examine whether FedGAC can reduce privacy leakage while preserving model accuracy, compared with FedCOG and a gradient clipping baseline.

A. Experimental Setup

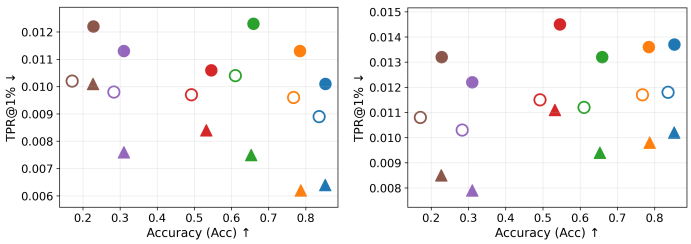
1) Datasets and Federated Settings: We evaluate FedGAC on Fashion-MNIST, CIFAR-10, and CIFAR-100 under a cross-device FL setting with $K = 10$ clients. All clients participate in every communication round, and the server aggregates local models using sample-size-weighted FedAvg. The total number of communication rounds is set to $T = 70$. Each client performs local training using SGD with a batch size of 64 and a learning rate of 0.01. Following the original FedCOG framework, the complementary data generation mechanism is enabled in the later training stage.

2) Model Architectures: We use lightweight CNN backbones for all experiments. For Fashion-MNIST and CIFAR-10, we adopt a LeNet-style CNN with two convolutional layers and two fully connected layers, yielding a feature dimension of 84. For CIFAR-100, we use a slightly deeper three-layer CNN with a feature dimension of 128. The global and local models share the same architecture on each dataset.

For generative federated learning, we follow FedCOG and optimize a generator with a 256-dimensional latent input to synthesize complementary samples. FedGAC introduces no additional models or architectural changes, and only augments the generator optimization with geometric alignment regularization.

3) Non-IID Data Partition: We follow FedCOG and adopt Dirichlet-based ($\beta = 0.1$) and label-skew non-IID partitions, denoted as **NIID-1** and **NIID-2**, respectively.

4) Baselines: We compare the proposed **FedGAC** with two baselines: (i) **FedCOG** and (ii) **FedCOG + Clipping**, which applies ℓ_2-norm gradient clipping during local training.



(a) **LB-MIA**. (b) **FedMIA**.

Fig. 2: Privacy–utility trade-off under LB-MIA and FedMIA. The x-axis shows classification accuracy and the y-axis shows TPR@1%. Filled circles denote FedCOG, hollow circles denote FedCOG + Clipping, and triangles denote FedGAC. Points closer to the bottom-right indicate better privacy–utility trade-offs.

5) Threat Model, Attacks, and Metrics: We consider an honest-but-curious server that performs server-side membership inference attacks. Following Section II-A, we evaluate two representative attacks: **LB-MIA** and **FedMIA**. Utility is measured by classification accuracy (Acc), and privacy leakage is measured by AUC and TPR@1% (lower is better).

For LB-MIA, the membership score of a sample x is

$$s_{\text{LB}}(x) = -\ell\big(f_g(x;\theta), y\big), \tag{6}$$

where $\ell(\cdot,\cdot)$ is the task loss.

For FedMIA, the membership score is defined as

$$s_{\text{FedMIA}}(x) = \cos\big(f_g(x), \mathbf{p}_y\big), \tag{7}$$

where $f_g(x)$ is the global representation and $\mathbf{p}_y$ is the class-wise geometric reference vector.

Given a score function $s(x)$ and a threshold τ, the attacker predicts membership as

$$\hat{m}(x;\tau) = \begin{cases} 1, & s(x) \geq \tau, \\ 0, & \text{otherwise}, \end{cases} \tag{8}$$

and we report TPR@1%,

$$\text{TPR@1\%} = \text{TPR}(\tau) \quad \text{s.t.} \quad \text{FPR}(\tau) = 0.01, \tag{9}$$

together with AUC.

B. Results and Analysis

1) Main Results: Table I summarizes utility and privacy under two membership inference attacks. FedCOG provides strong accuracy but exhibits non-trivial privacy leakage. Gradient clipping lowers AUC/TPR@1% in many cases, but the accuracy drops are consistent across datasets. In contrast, geometric alignment achieves lower AUC and TPR@1% while keeping accuracy comparable to FedCOG.

2) Privacy–Utility Trade-off: Figure 2 plots accuracy against TPR@1% for all settings, where points closer to the bottom-right indicate better trade-offs. Geometric alignment consistently shifts points downward with minimal horizontal movement, yielding a better trade-off frontier. In contrast, gradient clipping often shifts points leftward due to accuracy loss, which weakens the overall trade-off despite reduced TPR@1%.

TABLE I: Main results under two representative membership inference attacks. Lower is better for privacy (AUC/TPR@1%), while higher is better for utility (Acc).

LB-MIA

Dataset	Hetero	Method	AUC ↓	TPR@1% ↓	Acc ↑
FashionMNIST	NIID-1	FedCOG	0.5163	0.0101	**0.8529**
		FedCOG + Clipping	0.5115	0.0089	0.8363
		FedGAC	**0.5061**	**0.0064**	0.8523
	NIID-2	FedCOG	0.5269	0.0113	0.7845
		FedCOG + Clipping	0.5180	0.0096	0.7675
		FedGAC	**0.5164**	**0.0062**	**0.7868**
CIFAR-10	NIID-1	FedCOG	0.5349	0.0123	**0.6590**
		FedCOG + Clipping	0.5280	0.0104	0.6107
		FedGAC	**0.5269**	**0.0075**	0.6529
	NIID-2	FedCOG	0.5374	0.0106	**0.5459**
		FedCOG + Clipping	0.5327	0.0097	0.4925
		FedGAC	**0.5296**	**0.0084**	0.5320
CIFAR-100	NIID-1	FedCOG	0.5442	0.0113	**0.3105**
		FedCOG + Clipping	0.5376	0.0098	0.2836
		FedGAC	**0.5368**	**0.0076**	**0.3106**
	NIID-2	FedCOG	0.5419	0.0122	**0.2285**
		FedCOG + Clipping	0.5341	0.0102	0.1713
		FedGAC	**0.5352**	**0.0101**	0.2274

FedMIA

Dataset	Hetero	Method	AUC ↓	TPR@1% ↓	Acc ↑
FashionMNIST	NIID-1	FedCOG	0.5273	0.0137	**0.8529**
		FedCOG + Clipping	0.5212	0.0118	0.8363
		FedGAC	**0.5160**	**0.0102**	0.8523
	NIID-2	FedCOG	0.5317	0.0136	0.7845
		FedCOG + Clipping	0.5241	0.0117	0.7675
		FedGAC	**0.5219**	**0.0098**	**0.7868**
CIFAR-10	NIID-1	FedCOG	0.5362	0.0132	**0.6590**
		FedCOG + Clipping	0.5279	0.0112	0.6107
		FedGAC	**0.5272**	**0.0094**	0.6529
	NIID-2	FedCOG	0.5439	0.0145	**0.5459**
		FedCOG + Clipping	0.5330	0.0115	0.4925
		FedGAC	**0.5314**	**0.0111**	0.5320
CIFAR-100	NIID-1	FedCOG	0.5525	0.0122	0.3105
		FedCOG + Clipping	0.5427	0.0103	0.2836
		FedGAC	**0.5392**	**0.0079**	**0.3106**
	NIID-2	FedCOG	0.5447	0.0132	**0.2285**
		FedCOG + Clipping	0.5402	0.0108	0.1713
		FedGAC	**0.5374**	**0.0085**	0.2274

TABLE II: Composite privacy–utility scores under two representative membership inference attacks. The score equally combines normalized classification accuracy and privacy strength $(1 - \text{TPR@1\%})$. Higher values indicate better overall privacy–utility trade-offs.

Setting	Method	LB-MIA ↑	FedMIA ↑
FMNIST N1	FedCOG	0.500	0.500
	FedCOG + Clipping	0.162	0.271
	FedGAC	**0.982**	**0.982**
FMNIST N2	FedCOG	0.440	0.440
	FedCOG + Clipping	0.167	0.250
	FedGAC	**1.000**	**1.000**
CIFAR-10 N1	FedCOG	0.500	0.500
	FedCOG + Clipping	0.198	0.263
	FedGAC	**0.937**	**0.937**
CIFAR-10 N2	FedCOG	0.500	0.500
	FedCOG + Clipping	0.205	0.441
	FedGAC	**0.870**	**0.870**
CIFAR-100 N1	FedCOG	0.498	0.498
	FedCOG + Clipping	0.203	0.221
	FedGAC	**1.000**	**1.000**
CIFAR-100 N2	FedCOG	0.500	0.500
	FedCOG + Clipping	0.476	0.255
	FedGAC	**0.990**	**0.990**

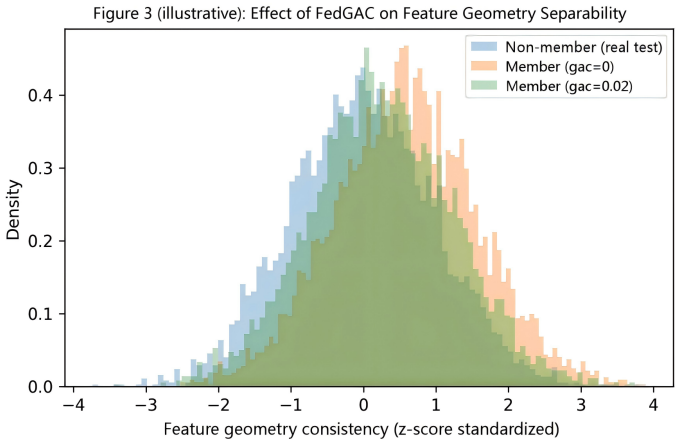


Fig. 3: Feature–geometry consistency distributions under **FedMIA**. **FedGAC** reduces representation-level separability by shifting member samples toward non-members in the feature space.

3) Composite Score Analysis: Table II provides a single-number summary of privacy–utility trade-offs. Geometric alignment consistently achieves the highest composite scores across all settings for both attacks, while gradient clipping is limited by its lower utility.

4) Mechanism Analysis under FedMIA: Figure 3 shows the feature–geometry consistency distributions under **FedMIA**. Without geometric alignment, member samples exhibit a clear distribution shift from non-members, which increases separability. After applying geometric alignment constraints, the member distribution overlaps more with non-members, reducing separability and weakening the attack.

IV. Conclusion

This paper investigates the privacy risks of generative federated learning under membership inference attacks, with a focus on the additional attack surface introduced by complementary data generation in non-independent and non-identically distributed settings. Although generative mechanisms improve model utility under data heterogeneity, they can unintentionally reshape representation geometry and amplify membership-related signals not addressed by existing update-level defenses.

To mitigate this risk, we propose FedGAC, a lightweight geometric alignment constraint applied during the complementary data generation stage. By regulating generated representations with classifier-induced class-wise geometric references, FedGAC suppresses client-specific discriminative directions while preserving class-level semantic structure, and can be integrated into existing FedCOG-style pipelines without modifying the federated learning protocol or incurring additional communication overhead.

Experiments under diverse non-IID settings show that, compared with FedCOG + Clipping, FedGAC consistently reduces attack success under both LB-MIA and FedMIA while maintaining classification accuracy, achieving a more favorable privacy–utility trade-off. Representation-space analysis further confirms reduced member and non-member separability.

Future work will explore extending geometric alignment constraints to other generative federated learning frameworks and evaluating robustness against stronger adaptive or client-level membership inference attacks.

Acknowledgment

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

References

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

[2] L. Wang *et al.*, "Server-side membership inference attacks via gradient reconstruction in federated learning," in *Proc. USENIX Secur. Symp.*, 2025.

[3] Z. Zhang, L. Wu, Z. Wang, J. Hu, C. Ma, and Q. Liu, "Cost-efficient and secure federated learning for edge computing," *IEEE Trans. Mobile Comput.*, 2025.

[4] J. Li, T. Zhang, Y. Chen, and S. Yu, "Secure aggregation does not prevent membership inference attacks in federated learning," in *Proc. ACM Conf. Comput. Commun. Secur. (CCS)*, 2024, pp. 1800–1815.

[5] Z. Zhang, L. Wu, J. Jin, E. Wang, B. Liu, and Q.-L. Han, "Secure federated learning for cloud–fog automation: Vulnerabilities, challenges, and solutions," *IEEE Trans. Ind. Informat.*, 2025.

[6] R. Ye *et al.*, "Consensus-oriented generative federated learning for non-iid data," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2024.

[7] H. Zhang *et al.*, "Using model updates to infer training data participation in federated learning," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2024.

[8] K. Sun *et al.*, "Representation-level membership inference attacks in federated learning," in *Proc. USENIX Secur. Symp.*, 2025.

[9] Y. Chen *et al.*, "Differentially private federated learning: Optimizing the privacy–utility trade-off for mia resistance," *IEEE Trans. Inf. Forensics Secur.*, 2025.

[10] S. Yang *et al.*, "Adaptive gradient clipping with noise injection for membership inference defense in federated learning," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, 2024.

[11] Z. Hu *et al.*, "Privacy amplification by subsampling in federated learning," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2024.

[12] C. Liu *et al.*, "White-box membership inference on federated models via loss landscape analysis," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2024.

[13] J. Zhu *et al.*, "Data-free knowledge distillation in federated learning via deepinversion," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2025.

[14] H. Wang *et al.*, "Synthetic data generation for federated learning: Alleviating data heterogeneity," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2025.

[15] Y. Du *et al.*, "Feature-space membership inference attacks against federated generative learning," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2024.

[16] Y. Zheng *et al.*, "Representation leakage via embedding geometry in federated generative models," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2025.

[17] S. Yeom, I. Giacomelli, M. Fredrikson, and S. Jha, "Privacy risk in machine learning: Analyzing the connection to overfitting," in *Proc. IEEE 31st Comput. Secur. Found. Symp. (CSF)*, 2018, pp. 268–282.

[18] G. Zhu, D. Li, H. Gu, Y. Yao, L. Fan, and Y. Han, "Fedmia: An effective membership inference attack exploiting "all for one" principle in federated learning," in *arXiv preprint arXiv:2402.06289*, 2024.

[19] Y. Kusunoki, "A deep positive-negative prototype approach to integrated prototypical discriminative learning," 2025. [Online]. Available: https://arxiv.org/pdf/2501.02477