FedAU2: Attribute Unlearning for User-Level Federated Recommender Systems with Adaptive and Robust Adversarial Training

Yuyuan Li[1], Junjie Fang[1], Fengyuan Yu[2], Xichun Sheng[3], Tianyu Du[2], Xuyang Teng[1], Shaowei Jiang[1], Linbo Jiang[4], Lin Jianan[4], Chaochao Chen[2*]

[1]Hangzhou Dianzi University
[2]Zhejiang University
[3]Macao Polytechnic University
[4]Ant Group

Abstract

Federated Recommender Systems (FedRecs) leverage federated learning to protect user privacy by retaining data locally. However, user embeddings in FedRecs often encode sensitive attribute information, rendering them vulnerable to attribute inference attacks. Attribute unlearning has emerged as a promising approach to mitigate this issue. In this paper, we focus on user-level FedRecs, which is a more practical yet challenging setting compared to group-level FedRecs. Adversarial training emerges as the most feasible approach within this context. We identify two key challenges in implementing adversarial training-based attribute unlearning for user-level FedRecs: i) mitigating training instability caused by user data heterogeneity, and ii) preventing attribute information leakage through gradients. To address these challenges, we propose FedAU2, an attribute unlearning method for user-level FedRecs. For CH1, we propose an adaptive adversarial training strategy, where the training dynamics are adjusted in response to local optimization behavior. For CH2, we propose a dual-stochastic variational autoencoder to perturb the adversarial model, effectively preventing gradient-based information leakage. Extensive experiments on three real-world datasets demonstrate that our proposed FedAU2 achieves superior performance in unlearning effectiveness and recommendation performance compared to existing baselines.

1 Introduction

Recommender Systems (RSs) are integral to modern online platforms, delivering personalized recommendations based on user preferences (Hasan et al. 2024; Lu and Yin 2025; Su et al. 2023; Feng et al. 2024). Traditional RSs are typically designed in centralized settings, requiring users to provide all raw interaction data, such as clicks and purchases. However, this centralized approach raises significant concerns regarding user privacy and potential data misuse (Qu et al. 2024; Zhou et al. 2023). To address these concerns, Federated Learning (FL) (McMahan et al. 2017; Liu et al. 2024a,b; Zhang et al. 2025; Chen et al. 2024a) has emerged as a privacy-preserving paradigm for training recommendation models, which are known as Federated RSs (FedRecs) (Sun et al. 2022). In FL, users' raw data remains on their local devices, and a global model is collaboratively trained through local updates and aggregation.

Recent data privacy regulations, such as the GDPR (Protection 2018) and CCPA (Illman and Temple 2019), emphasize the *right to be forgotten*, granting users the ability to withdraw personal data, including their influence on models and sensitive information (Li et al. 2024; Feng et al. 2025a). RSs represent critical application scenarios that heavily rely on personal data and often embed users' attribute information, rendering them susceptible to attribute inference attacks (Chen et al. 2024b). Attribute unlearning has emerged as a promising approach to eliminate users' attribute information from models (Feng et al. 2025c; Yu et al. 2025). However, while FedRecs preserve the privacy of local data, it does not provide mechanisms to remove users' attribute information (Hu and Song 2024), thereby failing to satisfy the right to be forgotten. Therefore, enabling attribute unlearning in FedRecs is imperative to uphold users' privacy.

FedRecs can be classified into two categories (Sun et al. 2022): user-level and group-level. In user-level FedRecs, each client corresponds to an individual user (e.g., a personal smartphone), while in group-level FedRecs, each client aggregates data from multiple users (e.g., within an organization). Attribute unlearning methods (e.g., distribution alignment) developed for group-level settings (Hu and Song 2024; Wu, Jiang, and Hu 2025; Lu, Tong, and Ye 2025; Feng et al. 2025d) typically aggregate data from multiple users, rendering them inapplicable to user-level FedRecs, where data remains strictly isolated on individual devices.

In this paper, we focus on attribute unlearning in user-level FedRecs, which is more generalized and more challenging (Sun et al. 2022). As distribution alignment (Wu, Jiang, and Hu 2025) is inapplicable in this setting, adversarial training (Feng et al. 2025b; Wang et al. 2025) emerges as the most feasible approach. Nevertheless, we identify two key challenges in its implementation: **CH1**: How to make adversarial training stable in federated settings, i.e., the cold start problem (Zhang, Shi, and Zhao 2024). Specifically, traditional adversarial learning methods require pre-training to stabilize the training process (Ganhör et al. 2022) However, in FedRecs, due to the heterogeneity of user data and the randomness of sampling (Zhang et al. 2024), pre-training may lead to unstable training processes. **CH2**: How to prevent attribute information leakage during unlearning in FedRecs.

*Corresponding author: zjuccc@zju.edu.cn

Recent studies have demonstrated that the model gradient transmitted during server-client communication can expose users' raw data through reconstruction attacks (Zhao, Mopuri, and Bilen 2020; Zhu, Liu, and Han 2019), with attribute labels being particularly vulnerable to such leakage.

To address these challenges, we propose FedAU^2, an **A**ttribute **U**nlearning method for **U**ser-level **Fed**Recs. Specifically, we perform adversarial learning locally on each client and aggregate model updates on the server, thereby eliminating the reliance on group-level attribute information. For **CH1**, we propose a decentralized and adaptive training strategy that operates at the user level to avoid the global pre-training. Specifically, each user autonomously adjusts their adversarial training in response to their own adversarial prediction outcome, enabling fine-grained control and enhanced adaptation. For **CH2**, we propose a Dual-Stochastic Variational AutoEncoder (DSVAE), which is integrated into the adversarial model. Our theoretical analysis reveals that the dual stochastic design helps perturb the adversarial model to enhance robustness against gradient-based reconstruction attacks. We summarize the main contributions of this paper as follows:

- We propose FedAU^2, a novel attribute unlearning method for user-level FedRecs. We identify two key challenges: i) avoiding adversarial training instability in federated settings, and ii) preventing gradient-based attribute leakage.
- For **CH1**, we propose an adaptive adversarial training strategy that dynamically adjusts each user's training process based on their local updates. This design enhances both the stability and efficiency of adversarial optimization in an adaptive manner.
- For **CH2**, we integrate a DSVAE into the adversarial model. Theoretical analysis reveals that our proposed DSVAE perturbs the parameters to effectively prevent attribution information leakion in FedRecs.
- We conduct extensive experiments on three real-world datasets across representative recommendation models. The results demonstrate that our proposed FedAU^2 significantly outperforms existing baselines.

2 Related Work

2.1 Federated Recommendation Systems

FedRecs leverage federated learning to enable collaborative model training without sharing raw user data in RSs. Based on the granularity of user participation (Sun et al. 2022), FedRecs can be categorized into user-level (where each device acts as a client) and group-level (where clients represent user groups or a data silo) settings. Most FedRecs are developed based on the user-level setting (Chai et al. 2020; Perifanis and Efraimidis 2022) Latest user-level FedRecs also explore personalized methods to better capture user-specific preferences under heterogeneous data (Li, Long, and Zhou 2023).

The key distinction between these two settings lies in data and client heterogeneity (Sun et al. 2022). Compared to group-level settings, user-level FedRecs face extremely sparse and non-IID data. This introduces substantial challenges for unlearning methods that depend on cross-user information (Hu and Song 2024; Wu, Jiang, and Hu 2025; Li et al. 2025a,b). In this work, we focus on the user-level setting, which remains the most prevalent and technically challenging scenario for federated recommendation.

2.2 Recommendation Attribute Unlearning

RSs can implicitly encode sensitive user attributes (e.g., gender and age) into user embeddings, making them vulnerable to attribute inference attacks (Feng et al. 2025b; Liu et al. 2025). Attribute unlearning (Li et al. 2023) has emerged to mitigate this issue. In centralized RSs, researchers utilize adversarial learning (Ganhör et al. 2022) and post-training tuning (Li et al. 2023) to achieve attribute unlearning.

FedRecs also face risks of attribute leakage because federated learning cannot prevent attribute inference (Hu and Song 2024). Traditional collaborative filtering methods and latest personalized FedRecs both encode user attributes into embeddings, increasing privacy risks. To mitigate such risks, Hu et al. propose a user-consented approach for attribute unlearning (Hu and Song 2024). Aegis (Wu, Jiang, and Hu 2025) extends post-training methods to the federated setting. However, these methods rely on group-level user data, which is infeasible in user-level FedRecs. For user-level FedRecs, Zhang et al. propose a Differential Privacy (DP) framework (Zhang, Yuan, and Yin 2023). However, DP-based methods typically lack attribute specificity (i.e., cannot selectively unlearn a particular attribute) and often significantly degrade recommendation performance. These gaps highlight the need for an attribute unlearning approach for user-level FedRecs that can effectively mitigate attribute leakage while preserving recommendation performance.

2.3 Gradient-based Reconstruction Attacks

Gradient-based reconstruction attacks (e.g., DLG) can reconstruct training data by analyzing raw gradients (Zhu, Liu, and Han 2019; Su et al. 2025). Subsequently, iDLG (Zhao, Mopuri, and Bilen 2020) leverages analytical derivation to directly obtain real labels, making the attack more efficient and reliable.

Current defenses fall into two categories: feature-side and label-side. Feature-side methods protect input data (Huang et al. 2020; Sun et al. 2021), but do not secure output labels, which are crucial in user-level FedRecs. In contrast, label-side methods protect label-related privacy at the output layer. Label-DP noise (Ghazi et al. 2024) and soft labels (Struppek, Hintersdorf, and Kersting 2023; He et al. 2024) trade privacy for accuracy, and even hinder convergence in user-level FedRecs (Wang et al. 2024; Chen et al. 2025; Zhou et al. 2025). None of these methods scales well to user-level FedRecs.

3 Preliminaries

3.1 User-level Federated Recommendation

Let $\mathcal{U}$ and $\mathcal{I}$ denote the sets of users and items, respectively. Each user $u \in \mathcal{U}$ holds a private interaction vector $\mathbf{x}_u \in \{0,1\}^{|\mathcal{I}|}$, where $\mathbf{x}_{ui} \in \{0,1\}$ indicates whether user u has interacted with item $i \in \mathcal{I}$. The goal of FedRecs is to learn a scoring function $f_\Theta : \mathcal{U} \times \mathcal{I} \to \mathbb{R}$, where $f_\Theta(u,i)$ denotes the predicted score between user u and item i. To

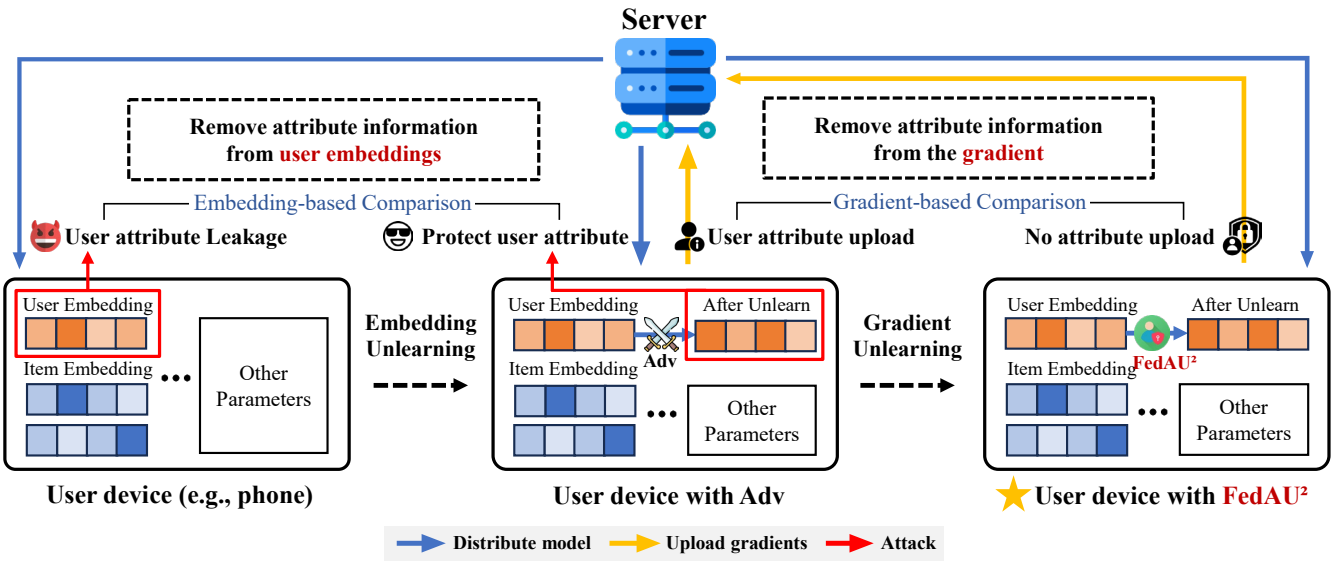


Figure 1: Comparison of three user device types in federated recommendation. i) on the left, standard clients expose attribute information in embeddings; ii) in the middle, adversarial clients can prevent embedding leakage but still suffer from gradient leakage; and iii) on the right, our proposed FedAU^2 eliminates information leakage from both embeddings and gradients.

extract features of users and items, the system typically encodes users and items via $\mathbf{em}_u = \phi_u(u)$ and $\mathbf{em}_i = \phi_i(i)$, and computes the matching score between them using a scoring function $s_\psi(\cdot)$. The recommendation objective can be formalized as:

$$\min_{\Theta} \mathcal{L}_{\text{rec}} = \sum_{u \in \mathcal{U}} \sum_{i \in \mathcal{I}} \ell\left(s_\psi(\mathbf{em}_u, \mathbf{em}_i),\ \mathbf{x}_{ui}\right), \tag{1}$$

where $\ell(\cdot)$ is a loss function (e.g., BPR), and $\Theta = \phi_u \cup \phi_i \cup \psi$. In FedRecs, user-side parameters ϕ_u are locally updated and remain private, whereas item-side parameters ϕ_i and model parameters ψ are uploaded and aggregated on the server.

3.2 Adversarial Training

Let $\mathbf{em}_u \in \mathbb{R}^d$ denote the embedding of user u, and let $y \in \mathcal{Y}$ be a protected attribute (e.g., gender or age). We introduce an adversarial network $h_\omega(\cdot)$, which takes $\mathbf{em}_u$ as input and attempts to predict the protected attribute y. The model learns to preserve recommendation quality while removing sensitive information from $\mathbf{em}_u$, leading to the following min-max objective:

$$\min_{\Theta} \max_{\omega} \mathcal{L}_{\text{rec}} - \lambda \cdot \sum_{u \in \mathcal{U}} \mathcal{L}_{\text{adv}}(\mathbf{em}_u, y), \tag{2}$$

where $\mathcal{L}_{\text{adv}}$ is the adversarial loss, implemented as a classification loss using cross-entropy:

$$\mathcal{L}_{\text{adv}}(\mathbf{em}_u, y) = \text{CE}(h_\omega(\mathbf{em}_u), y),$$

To solve this optimization problem, we adopt a Gradient Reversal Layer (GRL) (Ganin and Lempitsky 2015) between $\mathbf{em}_u$ and h_ω. GRL preserves the forward pass and reverses the backward gradient. Converting the min-max into a minimization.:

$$\min_{\Theta,\omega} \mathcal{L}_{\text{rec}} + \lambda \cdot \sum_{u \in \mathcal{U}} \text{CE}(h_\omega(\text{GRL}(\mathbf{em}_u)), y). \tag{3}$$

3.3 Reconstructing Data via DLG

DLG reconstructs private user data by optimizing dummy inputs and labels to match observed gradients. Given the observed gradient $\nabla W = \nabla_W \mathcal{L}(h_\omega(\mathbf{em}), y)$, the attacker initializes dummy variables $\mathbf{em}', y' \sim \mathcal{N}(0, 1)$, and iteratively updates them to minimize the distance between dummy and real gradients:

$$(\mathbf{em}^*, y^*) = \arg\min_{\mathbf{em}', y'} \|\nabla W' - \nabla W\|^2, \tag{4}$$

where $\nabla W' = \nabla_W \mathcal{L}(h_\omega(\mathbf{em}'), y')$ is the gradient computed from dummy data. This optimization exploits the differentiability of the gradient distance w.r.t. $\mathbf{em}'$ and y'. By minimizing this distance, the dummy data $(\mathbf{em}', y')$ gradually approximates the true training sample $(\mathbf{em}, y)$.

4 Methodology

4.1 Overview of FedAU2

To achieve user-level attribute unlearning in FedRecs, we employ the adversarial training approach. However, the direct application of adversarial training introduces two key challenges (**CH1**: stable training and **CH2**: privacy preservation). To tackle these challenges, we propose FedAU^2, an adaptive and robust user-level attribute unlearning method for FedRecs.

FedAU^2 demonstrates a significant advantage over both standard FL clients and the conventional adversarial training approach. As shown in Figure 1, i) the embeddings in a *standard FL client* may leak sensitive attribute information; ii) a client equipped with a *conventional adversarial network* mitigates embedding leakage but remains vulnerable to gradient-based reconstruction attacks; and iii) FedAU^2 effectively prevents both embedding and gradient leakage, thereby offering enhanced privacy protection.

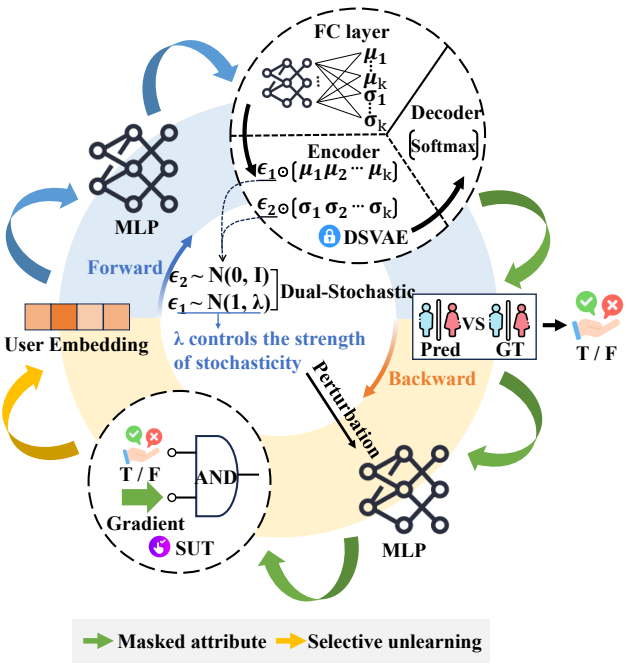


Figure 2: Workflow of FedAU^2. During the forward pass, DSVAE injects dual-stochasticity , effectively masking attribute information embedded in the gradients. During the backward pass, SUT dynamically adjusts the perturbation budget based on the prediction outcomes, enabling stable adversarial training.

Specifically, FedAU^2 integrates two synergistic components into adversarial training to enable adaptive and robust attribute unlearning, i.e., Selective Unlearning Trigger (SUT) for stable training and DSVAE to preserve gradient privacy. Figure 2 illustrates the workflow of FedAU^2.

4.2 Selective Unlearning Trigger

Adversarial training helps mitigate privacy leakage but often introduces instability, i.e., the cold start problem of adversarial models (**CH1**). Moreover, global pretraining strategies (Ganhör et al. 2022) prove inadequate for FedRecs. The heterogeneity of user data and the randomness of client sampling result in uneven user participation, which in turn limits the generalization of the adversarial model in federated learning (Yang et al. 2024).

The instability of adversarial training primarily stems from the trade-off governed by the perturbation budget (Andriushchenko and Flammarion 2020). There are two extreme cases in practice: i) *excessive perturbation budget:* aggressive perturbations obscure task-relevant features, rendering the adversarial loss uninformative and driving training into noisy updates; and ii) *insufficient perturbation budget:* mild perturbations are unable to erase the sensitive attributes derived from standard training, leaving privacy-related information embedded in the learned representations.

To balance the trade-off in perturbation budget, we propose a SUT for user-level FedRecs, which enables efficient and adaptive control of the perturbation budget at the individual user level. Motivated by the adaptive perturbation control at the sample level (Balaji, Goldstein, and Hoffman 2019), we incorporate this adaptive design into our adversarial training objective for user-level FedRecs. Specifically, the per-user training objective is formulated as:

$$\min_{\theta,\omega} \mathcal{L}_{\text{rec}}^{(u)} + \lambda \cdot \mathcal{L}_{\text{CE}}(h_\omega(\text{GRL}(\mathbf{em}_u; \epsilon_u)), y_u), \quad (5)$$

where ϵ_u determines the strength of adversarial unlearning and is dynamically adjusted based on the estimated distance to the decision boundary:

$$\epsilon_u = \frac{\tau}{\|\nabla_{\mathbf{em}_u} \mathcal{L}(h_\omega(\mathbf{em}_u), y_u)\|_2}, \quad (6)$$

where τ is a global scaling factor. ϵ_u serves as a local margin estimator between the user embedding $\mathbf{em}_u$ and the decision boundary of the adversarial classifier.

Directly applying Eq. (6) on the client side introduces significant computational overhead, as it requires computing the gradient at every training iteration. Existing studies (Elsayed et al. 2018) show that correct predictions generally correspond to smaller gradient norm, suggesting that the embedding $\mathbf{em}_u$ resides near a local optimum and is far from the decision boundary. Conversely, misclassified samples tend to lie closer to the boundary, where the gradient norm are higher, indicating higher sensitivity to perturbations, but less informative for unlearning. Based on this observation, we introduce a simplified binary variant of the perturbation budget SUT defined as: $\epsilon_u = \epsilon$ if $\hat{y}_u = y_u$, and $\epsilon_u = 0$ otherwise, where $\hat{y}_u = \arg\max h_\omega(\mathbf{em}_u)$ is the adversarial prediction, and ϵ is a constant controlling the gradient reversal strength in the GRL layer.

The binary design of ϵ_u greatly simplifies per-user computation while offering clear interpretability: if the adversarial classifier misclassifies the label (insufficient discriminatory power), adversarial training is skipped; if the prediction is correct (captured label-related information), unlearning is applied accordingly. SUT effectively balances the trade-off in perturbation budget through user-level adaptive adjustment, eliminating the need for global pretraining. This design enables more stable adversarial training in federated settings.

4.3 Dual-Stochastic Variational Autoencoder

For **CH2**, we observe that even if the attribute information can be unlearned from embeddings, the adversarial gradients can still leak sensitive labels (Zhu, Liu, and Han 2019). To address this challenge, we propose DSVAE, which injects stochasticity to mask sensitive labels in the gradients. We motivate its design by comparing how three different model architectures (original, VAE, and DSVAE) behave under DLG-based attacks, starting with a unified theoretical analysis of how DLG reconstructs labels, which generalizes across architectures.

DLG aims to reconstruct the true user label y by optimizing a dummy label y^* that closely approximates it. Since the label y is inferred through the final layers of the

model (Zhao, Mopuri, and Bilen 2020), our analysis specifically focuses on the final layer preceding the softmax output. Given a class i, we define $\hat{y}'_i$ as its predicted probability during the reconstruction attack, ∇W_i as the corresponding observed gradient vector from the final layer, and y_i^* as optimal dummy label.

Theorem 1. *Given a class i, let $\hat{y}'_i$ denote the predicted probability during the reconstruction attack, ∇W_i the observed gradient of the final layer, and y_i^* the optimal dummy label. Then, y_i^* admits the following closed-form solution:*

$$y_i^* = \hat{y}'_i - \delta_i, \tag{7}$$

where $\delta_i = \mathcal{G}(\nabla W_i)$, a function with of the gradient ∇W_i.

Proof. The proof can be found in Appendix B.1. □

As shown in Eq. (7), y_i^* is depended on two part: i) $\hat{y}'_i$ approximates the client-side prediction $\hat{y}_i$ from the reconstructed embedding $\mathbf{em}'$; and ii) a correction term δ_i that reflects the preference encoded by the gradient ∇W_i for label i. By leveraging both $\hat{y}'_i$ and δ_i, DLG can infer the user's attribute. Thus, the key to preventing DLG from reconstructing labels lies in perturbing these two components. In the following, we analyze the behavior of $\hat{y}'_i$ and δ_i under different model architectures.

Original MLP Following prior work (Ganhör et al. 2022), we use an MLP as the adversarial classifier. In the original MLP, the linear transformation is deterministic without any form of stochasticity. As a result, DLG can easily reconstruct the user label information by exploiting both $\hat{y}'_i$ and δ_i.

Corollary 1. *Let $\mathbf{h}'$ denote the input to the final linear layer of the MLP during the reconstruction attack. Then, the correction term δ_i is given by:*

$$\delta_i = \frac{(\mathbf{h}')^\top \nabla W_i}{\|\mathbf{h}'\|^2}. \tag{8}$$

Proof. The proof can be found in Appendix B.2. □

For $\hat{y}'_i$, DLG can accurately reconstruct the user embedding from the original MLP (Zhu, Liu, and Han 2019), effectively approximating the client-side prediction $\hat{y}_i$ using the reconstructed embedding $\mathbf{em}'$. For δ_i, DLG can accurately capture the preference encoded in the gradient ∇W_i through Eq. (8), where the sign of ∇W_i differs when i corresponds to the true label (Zhao, Mopuri, and Bilen 2020), directly revealing the user label y_i.

MLP with VAE While the Variational Autoencoder (VAE) architecture injects stochasticity during training, this primarily affects $\hat{y}'_i$, leaving the correction term δ_i intact and still exploitable by DLG for reconstructing user labels.

VAE assumes that each input is associated with a latent variable $\mathbf{u} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, which is learned by maximizing the evidence lower bound. To enable backpropagation, it applies the reparameterization trick:

$$\mathbf{u} = \boldsymbol{\mu} + \boldsymbol{\sigma} \odot \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}),$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ are tuned to capture the latent distribution.

Note that VAE injects stochasticity solely via the $\boldsymbol{\sigma}$ path through the sampling of $\boldsymbol{\epsilon}$, whereas the $\boldsymbol{\mu}$ path remains unaffected. As a result, DLG can still reconstruct user labels by leveraging the deterministic $\boldsymbol{\mu}$ path.

Corollary 2. *Let $\mathbf{z}'$ denote the input to the VAE during the reconstruction attack, and ∇W_i^μ denote the observed gradient vector associated with the i-th logit along the $\boldsymbol{\mu}$ path. Then, the correction term δ_i derived from ∇W_i^μ is given by:*

$$\delta_i = \frac{(\mathbf{z}')^\top \nabla W_i^\mu}{\|\mathbf{z}'\|^2} \tag{9}$$

Proof. The proof can be found in Appendix B.3. □

For $\hat{y}'_i$, VAE hinders accurate reconstruction of user embeddings through stochastic perturbations (Scheliga, Mäder, and Seeland 2022), thereby disrupting the direct mapping from $\hat{y}'_i$ to the client-side prediction $\hat{y}_i$. In contrast, for δ_i, since VAE does not introduce stochasticity along the $\boldsymbol{\mu}$ path, DLG can still accurately capture the preference encoded in the gradient ∇W_i^μ through Eq. (9), enabling label reconstruction through this deterministic path.

MLP with DSVAE Our proposed DSVAE enhances VAE by injecting extra stochasticity into the $\boldsymbol{\mu}$ path, introducing perturbations to both $\hat{y}'_i$ and δ_i, which jointly block DLG from accurately reconstructing user labels. To achieve this, DSVAE injects stochasticity directly into $\boldsymbol{\mu}$ by redefining the latent variable as:

$$\mathbf{u} = \boldsymbol{\mu} \odot \boldsymbol{\epsilon}'_1 + \boldsymbol{\sigma} \odot \boldsymbol{\epsilon}'_2, \quad \boldsymbol{\epsilon}'_1 \sim \mathcal{N}(\mathbf{1}, \boldsymbol{\lambda}), \quad \boldsymbol{\epsilon}'_2 \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{10}$$

where $\boldsymbol{\lambda}$ is a tunable hyperparameter that controls the strength of stochasticity.

Corollary 3. *The correction term δ_i under stochastic perturbation derived from ∇W_i^μ is given by:*

$$\delta_i = \frac{(\mathbf{z}' \odot \boldsymbol{\epsilon}'_1)^\top \nabla W_i^\mu}{\|\mathbf{z}' \odot \boldsymbol{\epsilon}'_1\|^2}. \tag{11}$$

Proof. The proof can be found in Appendix B.4. □

As y_i, $\hat{y}_i$, $\mathbf{z}$, and $\boldsymbol{\epsilon}_1$ are generated locally on the client during training, they are independent of server-side data. Given that the gradient computes as $\nabla W_i^\mu = (\hat{y}_i - y_i)(\mathbf{z} \odot \boldsymbol{\epsilon}_1)$, substituting into Eq. (11) yields:

$$\delta_i = (\boldsymbol{\epsilon}_1^\top \boldsymbol{\epsilon}'_1) \odot \frac{(\hat{y}_i - y_i)(\mathbf{z}^\top \mathbf{z}')}{\|\mathbf{z}' \odot \boldsymbol{\epsilon}'_1\|^2}. \tag{12}$$

For $\hat{y}'_i$, building upon the VAE framework, DSVAE similarly disrupts the direct mapping to the client-side prediction $\hat{y}_i$. For δ_i, the injected noise into the $\boldsymbol{\mu}$ path introduces a multiplicative term $\boldsymbol{\epsilon}_1^\top \boldsymbol{\epsilon}'_1$, which effectively perturbs the information embedded in ∇W_i^μ. As a result, DSVAE introduces dual stochasticity. This design effectively prevents DLG from accurately reconstructing user labels. In addition, for iDLG (Zhao, Mopuri, and Bilen 2020), the introduction of $\boldsymbol{\epsilon}_1$ invalidates its label inference mechanism based on the sign of gradients.

Dataset	Attributes	Method	FedNCF				FedVAE				FedRAP			
			HR@10 ↑	NDCG@10 ↑	F1 ↓	BAcc ↓	HR@10 ↑	NDCG@10 ↑	F1 ↓	BAcc ↓	HR@10 ↑	NDCG@10 ↑	F1 ↓	BAcc ↓
ML-100K	Gender	Original	0.6392	0.3634	0.5982	0.6068	0.6703	0.3910	0.5909	0.6097	0.4414	0.2385	0.5789	0.5853
		APM	0.5495	0.3083	0.5668	0.5675	0.5183	0.2938	0.5710	0.5785	0.4121	0.2249	**0.5411**	**0.5485**
		FedAU^2	**0.6154**	**0.3447**	**0.4992**	**0.5026**	**0.6667**	**0.3730**	**0.4113**	**0.5295**	**0.4359**	**0.2473**	0.5490	0.5508
	Age	Original	0.6392	0.3634	0.4590	0.4739	0.6703	0.3910	0.3956	0.4179	0.4414	0.2385	0.4185	0.4250
		APM	0.5495	0.3083	0.3915	0.3945	0.5183	0.2938	0.3362	0.3843	0.4121	0.2249	0.3793	0.3844
		FedAU^2	**0.5916**	**0.3247**	**0.3556**	**0.3623**	**0.6392**	**0.3626**	**0.2105**	**0.3431**	**0.4322**	**0.2452**	**0.3601**	**0.3706**
ML-1M	Gender	Original	0.6662	0.3897	0.7024	0.7032	0.6896	0.4169	0.7189	0.7204	0.4365	0.2381	0.7069	0.7089
		APM	0.5530	0.3079	0.6162	0.6179	0.5366	0.3005	0.6654	0.6666	0.4114	0.2268	0.6575	0.6621
		FedAU^2	**0.6232**	**0.3519**	**0.5176**	**0.5136**	**0.6697**	**0.3994**	**0.4191**	**0.5536**	**0.4429**	**0.2437**	**0.6285**	**0.6283**
	Age	Original	0.6662	0.3897	0.5896	0.5926	0.6896	0.4169	0.6186	0.6181	0.4365	0.2381	0.5488	0.5535
		APM	0.5530	0.3079	0.4832	0.4850	0.5366	0.3005	0.5140	0.5169	0.4114	0.2268	0.5109	0.5159
		FedAU^2	**0.6360**	**0.3595**	**0.3641**	**0.3669**	**0.6779**	**0.4017**	**0.1755**	**0.3347**	**0.4447**	**0.2458**	**0.4012**	**0.4000**
LastFM	Gender	Original	0.7158	0.4411	0.6907	0.6926	0.7407	0.4648	0.6806	0.6895	0.4763	0.2697	0.6909	0.6912
		APM	0.5744	0.3266	0.5950	0.6010	0.5818	0.3285	0.6463	0.6560	0.4351	0.2516	**0.6079**	**0.6079**
		FedAU^2	**0.6475**	**0.3867**	**0.4410**	**0.5047**	**0.7279**	**0.4566**	**0.3713**	**0.5222**	**0.4783**	**0.2689**	0.6087	0.6088
	Age	Original	0.7158	0.4411	0.4775	0.4813	0.7407	0.4648	0.5329	0.5391	0.4763	0.2697	0.5098	0.5165
		APM	0.5744	0.3266	0.4174	0.4189	0.5818	0.3285	0.4677	0.4682	0.4351	0.2516	0.4401	0.4405
		FedAU^2	**0.6584**	**0.3932**	**0.3649**	**0.3685**	**0.7358**	**0.4610**	**0.2033**	**0.3556**	**0.4794**	**0.2709**	**0.4301**	**0.4321**

Table 1: Performance (HR@10, NDCG@10) and privacy leakage (F1, BAcc) under different methods across datasets and attributes. Bold indicates the best result (excluding Original). All results are averaged over three independent runs. Due to the space limit, we report the results of HR@k and NDCG@k in Appendix C, where $k \in \{5, 15, 20\}$.

5 Experiments

5.1 Experimental Setup

Datasets We conduct experiments on three widely-used real-world datasets, each containing user-item interactions and user attributes (e.g., age and gender):

- **ML-100K**: 100K movie ratings from 1,000 users on 1,700 movies, with user demographics including gender, age, and occupation.
- **ML-1M**: 1M ratings from 6,040 users on 3,706 movies, with similar user attribute information as ML-100K.
- **LastFM-360K**: User-artist interactions and user profiles (gender, age, country) from a music streaming platform.

Recommendation Models We evaluate our proposed framework on two representative and one personalized FedRecs models:

- **FedNCF**: A federated version of Neural Collaborative Filtering (Perifanis and Efraimidis 2022).
- **FedVAE**: A federated adaptation of MultVAE that incorporates an adaptive learning rate (Polato 2021).
- **FedRAP**: A personalized FedRecs model combining global item embeddings with user-specific residuals (Li, Long, and Zhou 2023).

Unlearning Methods As group-level methods (Hu and Song 2024; Wu, Jiang, and Hu 2025) cannot apply to user-level FedRecs, we compare FedAU^2 with the user-level baseline and the standard training strategy.

- **Original**: The model without attribute unlearning.
- **APM**: A DP-based method that perturbs model parameters during training (Zhang, Yuan, and Yin 2023).

Attack Setting Following (Zhang, Yuan, and Yin 2023), we assume an honest-but-curious server that attempts to infer user attributes from gradients, along with an external adversary that exploits user embeddings for attribute inference.

Evaluation Metrics We use micro-averaged F1 score and Balanced Accuracy (BAcc) to evaluate attribute unlearning effectiveness. Gradient unlearning effectiveness is evaluated using accuracy. Recommendation performance is evaluated using Hit Ratio (HR) and Normalized Discounted Cumulative Gain (NDCG), under the leave-one-out evaluation protocol. We truncate the ranked list at 5, 10, 15, and 20.

More details of experimental setup, including data preprocessing, attack settings, training parameter settings, and hardware information, are provided in Appendix A.

5.2 Results and Discussions

Attribute Unlearning Performance We report the F1 score and BAcc in Table 1. On FedNCF and FedVAE, our proposed FedAU^2 achieves an average BAcc reduction of 26.42%, while APM achieves only 11.5%, consistently outperforming baselines in both F1 and BAcc across all datasets and attributes. For FedRAP, the average BAcc reduction is 14.09% for FedAU^2 and 9.24% for APM. FedRAP shows relatively weak unlearning performance, largely because its residual-based user embedding is hard to modify. To improve efficiency, we aggregate the residual into a single dimension during adversarial training, which may further limit its attribute unlearning ability.

Recommendation Performance As shown in Table 1, FedAU^2 and APM reduce NDCG@10 by 4.51% and 20.05%, respectively, on average. Across all settings, FedAU^2 consistently outperforms APM in both NDCG and HR, demonstrating better utility preservation while effectively unlearning sensitive attributes. Interestingly, for FedRAP, FedAU^2 even leads to a slight improvement in recommendation performance, which may be attributed to the unique structure of its user embedding.

SUT Analysis To assess the effect of SUT on training stability, we show the recommendation and unlearning performance under different adversarial strategies in Fig. 3.

Global applies adversarial training throughout all epochs, while *Pretrain* pre-trains the adversarial model for multiple epochs (10 for FedRAP, 50 for FedVAE, and 100 for FedNCF) before unlearning. As shown in Fig. 3, SUT yields the best recommendation performance, while *Global* performs the worst, indicating that SUT's adaptive perturbation prevents excessive disruption to user embeddings. For unlearning performance, SUT and *Global* perform similarly, but *Pretrain* performs worst due to limited adversarial model generalization under stochastic sampling. Overall, SUT better balances recommendation and unlearning performance.

DSVAE Analysis To validate the effectiveness of DSVAE, we present the gradient unlearning performance, the component analysis of Eq. (7), and the impact of the stochasticity coefficient λ on model performance.

- **Gradient Unlearning.** As shown in Fig. 4(a), the original MLP is vulnerable to gradient attacks, which accurately infer user attributes, achieving up to 90% accuracy in FedVAE. While VAE offers slight resistance to the attack, the adversary can still achieve up to 89% accuracy. In contrast, DSVAE significantly mitigates the attack across all three models through dual stochasticity, bringing the accuracy down to 58% in FedVAE.
- **Component Analysis.** As shown in Fig. 4(b), the original $\hat{y}'_i$ remains slightly above random guessing, reflecting that it approximates the client-side prediction. Both VAE and DSVAE are able to reduce $\hat{y}'_i$ to the level of random guessing. However, VAE has limited effect on δ_i, resulting in a gradient's preference similar to the original MLP and leading to insufficient perturbation of y_i^*. In contrast, DSVAE significantly suppresses δ_i, masking the gradient's preference, resulting in effective perturbation of y_i^* and robust defense against gradient-based reconstruction.
- **Stochasticity Coefficient.** As shown in Fig. 4(c), increasing λ introduces more stochasticity into the adversarial model, which weakens its generalization ability to attribute unlearning. This results in worse unlearning but better recommendation performance. Meanwhile, the increased stochasticity also improves gradient unlearning performance, gradually approaching random guessing.

Overhead Analysis SUT performs adversarial training only when the adversarial prediction is correct, reducing gradient computation. It is client-side and parameter-free, adding no memory or communication overhead. DSVAE replaces the last layer with two projections, doubling its parameters and causing a linear increase in computational, memory, and communication overhead. The combination of SUT and DSVAE slightly reduces the overall computational cost. We empirically evaluate our method and report the results in Appendix D.

6 Conclusion

In this paper, we study attribute unlearning in user-level FedRecs, aiming to remove sensitive attribute information from user embeddings while maintaining recommendation utility. User-level FedRecs are more challenging than group-level settings due to decentralized optimization and highly personalized data distributions. We identify two key challenges for adversarial training-based unlearning in this setting: i) training instability caused by user heterogeneity, and ii) gradient-based leakage of sensitive information. To tackle these challenges, we propose `FedAU`^{2}, which incorporates two core components: an adaptive adversarial training strategy (SUT) that dynamically adjusts the perturbation budget based on local optimization signals, and a dual-stochastic variational autoencoder (DSVAE) that effectively masks attribute information in the gradient. Extensive experiments on three real-world datasets and multiple representative FedRecs models demonstrate that `FedAU`^{2} achieves significantly better unlearning effectiveness and recommendation performance compared to existing baselines. Our findings show that adaptive adjustment is crucial for stable adversarial training in user-level FedRecs, and that label-side defenses are essential against gradient-based attacks.

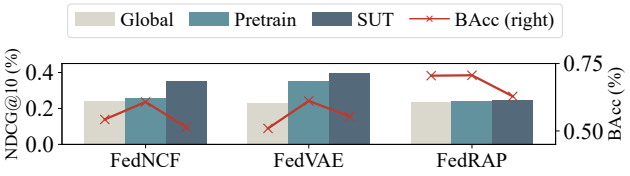
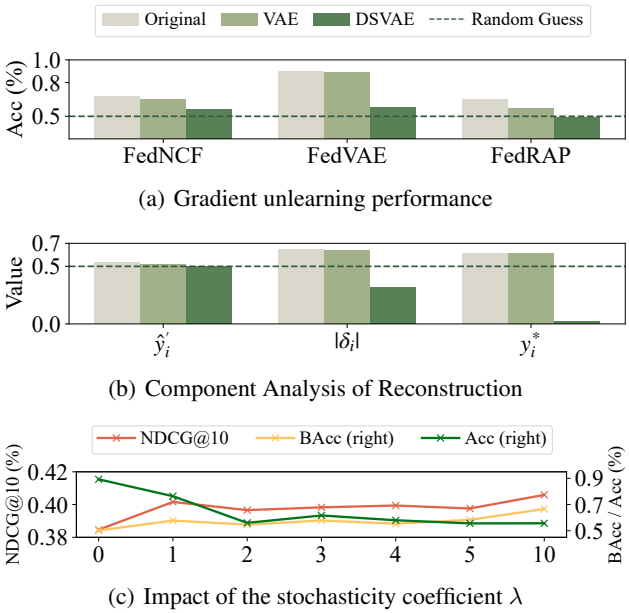


Figure 3: Ablation analysis of SUT. Recommendation (NDCG@10 ↑) and unlearning (BAcc ↓) performance under different adversarial training strategies on ML-1M (gender).



(a) Gradient unlearning performance

(b) Component Analysis of Reconstruction

(c) Impact of the stochasticity coefficient λ

Figure 4: Ablation analysis of DSVAE, conducted on ML-1M (gender). (a) Gradient unlearning performance across three models. (b) Component reconstruction analysis in FedVAE. (c) Effect of the stochasticity coefficient λ on recommendation (NDCG@10 ↑), attribute unlearning (BAcc ↓), and gradient unlearning (Acc ↓) performance in FedVAE.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grants (No. 62402148 and No. 62402418), the Zhejiang Province's 2025 "Leading Goose + X" Science and Technology Plan under Grant No. 2025C02034, the Key R&D Program of Ningbo under No. 2024Z115, the Fundamental Research Funds for the Central Universities, and Ant Group Research Fund.

References

Andriushchenko, M.; and Flammarion, N. 2020. Understanding and improving fast adversarial training. *Advances in Neural Information Processing Systems*, 33: 16048–16059.

Balaji, Y.; Goldstein, T.; and Hoffman, J. 2019. Instance adaptive adversarial training: Improved accuracy tradeoffs in neural nets. *arXiv preprint arXiv:1910.08051*.

Chai, D.; Wang, L.; Chen, K.; and Yang, Q. 2020. Secure federated matrix factorization. *IEEE Intelligent Systems*, 36(5): 11–20.

Chen, C.; Feng, X.; Li, Y.; Lyu, L.; Zhou, J.; Zheng, X.; and Yin, J. 2024a. Integration of large language models and federated learning. *Patterns*, 5(12).

Chen, C.; Zhang, Y.; Li, Y.; Wang, J.; Qi, L.; Xu, X.; Zheng, X.; and Yin, J. 2024b. Post-training attribute unlearning in recommender systems. *ACM Transactions on Information Systems*, 43(1): 1–28.

Chen, Z.; Hu, Y.; Li, Z.; Fu, Z.; Song, X.; and Nie, L. 2025. OFFSET: Segmentation-based Focus Shift Revision for Composed Image Retrieval. In *Proceedings of the ACM International Conference on Multimedia*, 6113–6122.

Elsayed, G.; Krishnan, D.; Mobahi, H.; Regan, K.; and Bengio, S. 2018. Large margin deep networks for classification. *Advances in neural information processing systems*, 31.

Feng, X.; Chen, C.; Li, Y.; and Lin, Z. 2024. Fine-grained pluggable gradient ascent for knowledge unlearning in language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 10141–10155.

Feng, X.; Li, Y.; Chen, C.; Zhang, L.; Li, L.; ZHOU, J.; and Zheng, X. 2025a. Controllable Unlearning for Image-to-Image Generative Models via Constrained Optimization. In *The Thirteenth International Conference on Learning Representations*.

Feng, X.; Li, Y.; Yu, F.; Xiong, K.; Fang, J.; Zhang, L.; Du, T.; and Chen, C. 2025b. RAID: An In-Training Defense against Attribute Inference Attacks in Recommender Systems. *arXiv preprint arXiv:2504.11510*.

Feng, X.; Li, Y.; Yu, F.; Zhang, L.; Chen, C.; and Zheng, X. 2025c. Plug and Play: Enabling Pluggable Attribute Unlearning in Recommender Systems. In *Proceedings of the ACM on Web Conference 2025*, 2689–2699.

Feng, X.; Zhang, J.; Yu, F.; Wang, C.; Zhang, L.; Li, K.; Li, Y.; Chen, C.; and Yin, J. 2025d. A survey on generative model unlearning: Fundamentals, taxonomy, evaluation, and future direction. *arXiv preprint arXiv:2507.19894*.

Ganhör, C.; Penz, D.; Rekabsaz, N.; Lesota, O.; and Schedl, M. 2022. Unlearning protected user attributes in recommendations with adversarial training. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2142–2147.

Ganin, Y.; and Lempitsky, V. 2015. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, 1180–1189. PMLR.

Ghazi, B.; Huang, Y.; Kamath, P.; Kumar, R.; Manurangsi, P.; and Zhang, C. 2024. LabelDP-Pro: Learning with Label Differential Privacy via Projections. In *The Twelfth International Conference on Learning Representations*.

Hasan, E.; Rahman, M.; Ding, C.; Huang, J.; and Raza, S. 2024. Based recommender systems: a survey of approaches, challenges and future perspectives. *ACM Computing Surveys*.

He, Y.; Niu, M.; Hua, J.; Mao, Y.; Huang, X.; Li, C.; and Zhong, S. 2024. LabObf: A Label Protection Scheme for Vertical Federated Learning Through Label Obfuscation. *arXiv preprint arXiv:2405.17042*.

Hu, Q.; and Song, Y. 2024. User consented federated recommender system against personalized attribute inference attack. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, 276–285.

Huang, Y.; Song, Z.; Li, K.; and Arora, S. 2020. Instahide: Instance-hiding schemes for private distributed learning. In *International conference on machine learning*, 4507–4518. PMLR.

Illman, E.; and Temple, P. 2019. California consumer privacy act. *The Business Lawyer*, 75(1): 1637–1646.

Li, Y.; Chen, C.; Zheng, X.; Zhang, Y.; Han, Z.; Meng, D.; and Wang, J. 2023. Making users indistinguishable: Attribute-wise unlearning in recommender systems. In *Proceedings of the 31st ACM International Conference on Multimedia*, 984–994.

Li, Y.; Feng, X.; Chen, C.; and Yang, Q. 2024. A survey on recommendation unlearning: Fundamentals, taxonomy, evaluation, and open questions. *arXiv preprint arXiv:2412.12836*.

Li, Z.; Chen, Z.; Wen, H.; Fu, Z.; Hu, Y.; and Guan, W. 2025a. Encoder: Entity mining and modification relation binding for composed image retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 5101–5109.

Li, Z.; Fu, Z.; Hu, Y.; Chen, Z.; Wen, H.; and Nie, L. 2025b. FineCIR: Explicit Parsing of Fine-Grained Modification Semantics for Composed Image Retrieval. *https://arxiv.org/abs/2503.21309*.

Li, Z.; Long, G.; and Zhou, T. 2023. Federated recommendation with additive personalization. *arXiv preprint arXiv:2301.09109*.

Liu, J.; Shang, F.; Liu, Y.; Liu, H.; Li, Y.; and Gong, Y. 2024a. Fedbcgd: Communication-efficient accelerated block coordinate gradient descent for federated learning. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 2955–2963.

Liu, Y.; Cai, B.; Zhi, J.; Wu, G.; and Xia, X. 2024b. QoE-Aware Online Auction Mechanism for UAV-enabled Crowdsensing. In *2024 IEEE International Conference on Web Services (ICWS)*, 654–664. IEEE.

Liu, Y.; Hua, Y.; Chai, H.; Wang, Y.; and Ye, T. 2025. Fine-Grained Open-Vocabulary Object Detection with Fined-Grained Prompts: Task, Dataset and Benchmark. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, 13860–13867. IEEE.

Lu, W.; Tong, Y.; and Ye, Z. 2025. DAMMFND: Domain-Aware Multimodal Multi-view Fake News Detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 559–567.

Lu, W.; and Yin, L. 2025. DMMD4SR: Diffusion Model-based Multi-level Multimodal Denoising for Sequential Recommendation. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 6363–6372.

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

Perifanis, V.; and Efraimidis, P. S. 2022. Federated neural collaborative filtering. *Knowledge-Based Systems*, 242: 108441.

Polato, M. 2021. Federated variational autoencoder for collaborative filtering. In *2021 International Joint Conference on Neural Networks (IJCNN)*, 1–8. IEEE.

Protection, F. D. 2018. General data protection regulation (GDPR). *Intersoft Consulting, Accessed in October*, 24(1).

Qu, L.; Yuan, W.; Zheng, R.; Cui, L.; Shi, Y.; and Yin, H. 2024. Towards personalized privacy: User-governed data contribution for federated recommendation. In *Proceedings of the ACM Web Conference 2024*, 3910–3918.

Scheliga, D.; Mäder, P.; and Seeland, M. 2022. Precode-a generic model extension to prevent deep gradient leakage. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 1849–1858.

Struppek, L.; Hintersdorf, D.; and Kersting, K. 2023. Be careful what you smooth for: Label smoothing can be a privacy shield but also a catalyst for model inversion attacks. *arXiv preprint arXiv:2310.06549*.

Su, J.; Chen, C.; Lin, Z.; Li, X.; Liu, W.; and Zheng, X. 2023. Personalized Behavior-Aware Transformer for Multi-Behavior Sequential Recommendation. In *Proceedings of the 31st ACM International Conference on Multimedia*, MM '23, 6321–6331. New York, NY, USA: Association for Computing Machinery. ISBN 9798400701085.

Su, J.; Chen, C.; Wang, Y.; Liu, W.; Li, Y.; Wang, T.; Li, Z.; Zheng, X.; and Yin, J. 2025. DuAda: Adaptive Targeted Model Poisoning Attack Framework via Dummy User Simulation on Federated Recommendation. *ACM Trans. Inf. Syst.*, 43(6).

Sun, J.; Li, A.; Wang, B.; Yang, H.; Li, H.; and Chen, Y. 2021. Soteria: Provable defense against privacy leakage in federated learning from representation perspective. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9311–9319.

Sun, Z.; Xu, Y.; Liu, Y.; He, W.; Kong, L.; Wu, F.; Jiang, Y.; and Cui, L. 2022. A survey on federated recommendation systems. *arXiv preprint arXiv:2301.00767*.

Wang, L.; Zhou, H.; Bao, Y.; Yan, X.; Shen, G.; and Kong, X. 2024. Horizontal federated recommender system: A survey. *ACM Computing Surveys*, 56(9): 1–42.

Wang, Y.; Wang, X.; Su, Y.; Zhang, S.; Lin, Z.; Meng, D.; and Hou, R. 2025. ROBIN: A Novel Framework for Accelerating Robust Multi-Variant Training. In *Proceedings of the 30th Asia and South Pacific Design Automation Conference*, 1120–1125.

Wu, W.; Jiang, J.; and Hu, C. 2025. Aegis: Post-Training Attribute Unlearning in Federated Recommender Systems against Attribute Inference Attacks. In *Proceedings of the ACM on Web Conference 2025*, 3783–3793.

Yang, H.; Qiu, P.; Khanduri, P.; Fang, M.; and Liu, J. 2024. Understanding server-assisted federated learning in the presence of incomplete client participation. *arXiv preprint arXiv:2405.02745*.

Yu, F.; Li, Y.; Feng, X.; Fang, J.; Wang, T.; and Chen, C. 2025. LEGO: A Lightweight and Efficient Multiple-Attribute Unlearning Framework for Recommender Systems. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 6242–6251.

Zhang, C.; Long, G.; Zhou, T.; Zhang, Z.; Yan, P.; and Yang, B. 2024. When federated recommendation meets cold-start problem: Separating item attributes and user interactions. In *Proceedings of the ACM Web Conference 2024*, 3632–3642.

Zhang, L.; Han, Z.; Feng, X.; Zhang, J.; Li, Y.; et al. 2025. FedFACT: A Provable Framework for Controllable Group-Fairness Calibration in Federated Learning. *Advances in Neural Information Processing Systems*, 38.

Zhang, S.; Yuan, W.; and Yin, H. 2023. Comprehensive privacy analysis on federated recommender system against attribute inference attacks. *IEEE Transactions on Knowledge and Data Engineering*, 36(3): 987–999.

Zhang, Y.; Shi, Y.; and Zhao, X. 2024. Bidirectional Corrective Model-Contrastive Federated Adversarial Training. *Electronics*, 13(18): 3745.

Zhao, B.; Mopuri, K. R.; and Bilen, H. 2020. idlg: Improved deep leakage from gradients. *arXiv preprint arXiv:2001.02610*.

Zhou, S.; Tian, Z.; Chu, X.; Zhang, X.; Zhang, B.; Lu, X.; Feng, C.; Jie, Z.; Chiang, P. Y.; and Ma, L. 2023. FastPillars: A Deployment-friendly Pillar-based 3D Detector. *arXiv preprint arXiv:2302.02367*.

Zhou, S.; Yuan, Z.; Yang, D.; Hu, X.; Qian, J.; and Zhao, Z. 2025. Pillarhist: A quantization-aware pillar feature encoder based on height-aware histogram. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 27336–27345.

Zhu, L.; Liu, Z.; and Han, S. 2019. Deep leakage from gradients. *Advances in neural information processing systems*, 32.