INVISIBLE THEFT: MEMORIZATION-BASED CLIENT-TO-CLIENT DATA LEAKS IN FEDERATED DIFFUSION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Federated diffusion has emerged as a promising framework for collaboratively training visual generation models without sharing private training data, making it particularly valuable for data-scarce yet privacy-sensitive federated settings. However, we reveal a previously overlooked client-to-client privacy vulnerability in this framework: because diffusion models can memorize and reproduce their training samples, a malicious client may exploit the shared global model to generate images highly similar to other clients' private training data, without accessing their gradients or interfering with the federated training process. To demonstrate the severity and practicality of this vulnerability, we propose a memorization-residual guided data stealing framework. Our key intuition is to leverage two models naturally accessible to the attacker—an attacker-private model trained only on its own data and the public pretrained model—as negative references, and use their prediction residuals with the global model as sampling-time guidance signals. This residual-guided sampling substantially increases the likelihood that the global model generates images highly similar to private training samples from other clients. Experiments show that our attack consistently increases both the frequency of high-similarity reproductions and the coverage of distinct threat images. Moreover, our attack is performed entirely after federated training and requires no modification to the training protocol, making it naturally stealthy from the perspective of federated training.

1 INTRODUCTION

Training diffusion models under the federated learning (FL) (McMahan et al., 2017; Kairouz & McMahan, 2021) paradigm has recently emerged as a promising approach for privacy-preserving visual generation (Tun et al., 2023; Stanley Jothiraj & Mashhadi, 2024; Huang et al., 2024). It enables multiple participants to collaboratively train a diffusion model without sharing private data with each other. This paradigm is particularly valuable in sensitive domains such as medical imaging and healthcare (Ogier du Terrail et al., 2022; Wang et al., 2023; Chen et al., 2024b), where strict privacy regulations prohibit data centralization. In this way, every participant obtains a diffusion model that benefits from the data of all others, while its own private images never leave its site.

However, we find that federated diffusion does not fully prevent private data from leaking between clients, because of the **memorization property** of diffusion models. Memorization has been widely observed in diffusion models: they sometimes generate images that are nearly identical to those in the training set (Yoon et al., 2023; Carlini et al., 2023; Somepalli et al., 2023; Gu et al., 2023). In federated diffusion framework, as shown in Figure 2, the global model is trained on the data of all clients, so it may memorize training images from all of them. Once this global model is distributed back to every client, the private images of each client are implicitly leaked to all the others. A malicious client can then exploit this memorization to re-generate nearly identical private training images of other clients from the global model as shown in Figure 1. To the best of our knowledge, we are the first to identify this **client-to-client** leakage vulnerability inherent to **federated diffusion**: unlike DataStealing (Gan et al., 2024), which only recovers data from clients under the attacker's control, our threat model targets the private data of *other benign clients*.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

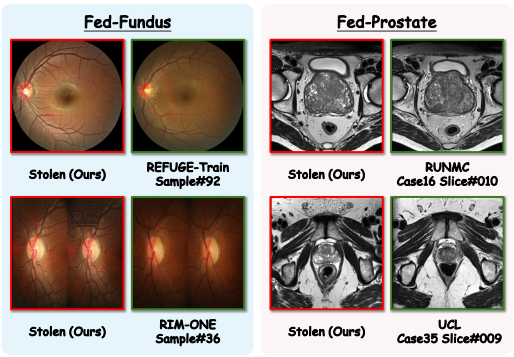


Figure 1: **Examples of private medical images stolen by our method.** For each pair, the left image is generated by our malicious client, while the right image is a private training image owned by another client in the real-world federated medical datasets. The text below each private image indicates its original data source and sample identifier.

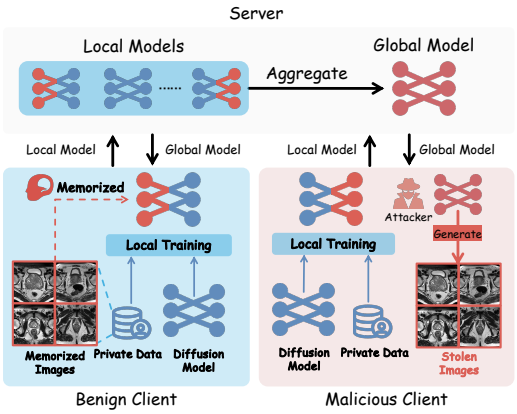


Figure 2: **Overview of the proposed threat**. During federated diffusion training, private images are memorized by local models. After aggregation, the global model implicitly retains the memorization of images from all clients, enabling malicious client to generate private training images of others.

This vulnerability is both severe and practical, for two reasons. **First, it requires no interference with federated training.** The attack exploits a property that diffusion models acquire through normal training, so unlike data-poisoning attacks (Xie et al., 2019; Wang et al., 2020; Cheng et al., 2024; Gan et al., 2024) that must inject crafted updates, the attacker only needs to train and communicate as an honest client, and can extract private images entirely offline once training is over. Its behavior is therefore indistinguishable from benign participation and hard to detect. **Second, it works effectively when launched from an honest client.** Without interfering with training, data stealing in FL is commonly done by inversion methods (Zhu et al., 2019; Zhao et al., 2020; Fang et al., 2023), which rely on the individual updates uploaded by the victim client. Such updates are observable only to the server. A client can merely infer them from how the global model parameters change across rounds (Wei et al., 2023; Wilson et al., 2024), which is a coarse approximation, and we find in practice that it degrades the reconstruction quality to the point of failure. Instead, our attack relies on no victim-specific signal at all, and thus remains effective even when the attacker is an ordinary client. This widens the attack surface considerably: in a medical federated system, both the third-party training platform and every participating hospital become serious attackers.

Despite the severity of this vulnerability, an attacker attempting to exploit it still suffers from **entangled memorization**, which limits the attack's efficiency. Specifically, the global model's memorization indiscriminately mixes the data the attacker wants with the data it does not, so sampling directly from the global model frequently reproduces undesired memorized content. This entanglement arises because the global model is obtained by fine-tuning a general-domain pretrained diffusion model on the data of all clients, leaving its memorization drawn from three sources: (1) the data of *other* clients that the attacker aims to steal; (2) the attacker's *own* data; and (3) the general-domain prior inherited from pre-training. Only the first is what the attacker actually wants, while others act as noise that dilutes every attack.

To overcome this, we propose a **memorization residual** based attack. Our key insight is that, for any client in the federated system, the data distributions of the two undesired sources are naturally accessible: the attacker fully possesses its own data, and the general-domain prior is represented by the public pretrained model. Leveraging this native accessibility of the attacker within the federated system, the **residual** between the overall memorization and the target memorization can therefore be approximated; subtracting this residual from the overall memorization provides a direction that favors the target memorization over undesired sources. We instantiate this idea as a drifting-based diffusion sampling procedure. At each sampling step, drifting adjusts the global model's noise

prediction by subtracting those of two reference models: the solo private model that the attacker fine-tunes from the public checkpoint on its own data alone, termed *attacker drifting*, and the public pretrained model, termed *pre-training drifting*. Since these two models provide references for the two undesired sources, their corresponding drifting reduces their influence on the sampling direction, thereby biasing generation toward memorized images associated with other clients.

We evaluate our attack on real-world federated medical datasets, Fed-Fundus (Liu et al., 2021) and Fed-Prostate (Liu et al., 2021), in which clients are partitioned by the actual hospitals from which the private data originate. Both qualitative and quantitative results show that our proposed memorization guided attack can steal nearly identical private images from other clients in this realistic scenario, which outperform representative previous inversion-based methods (Fang et al., 2023; Wu et al., 2024) and trivial heuristic memorization-based methods Gu et al. (2023); Wen et al. (2024).

Overall, our main contributions are summarized as:

- To the best of our knowledge, we are the first to investigate client-to-client privacy leakage risk of federated diffusion models, aiming to raise awareness of this critical threat.
- We propose a residual-driven attack strategy that uses attacker-accessible reference models to guide sampling toward private training images from other clients, which consistently increases the likelihood of high-similarity reproductions.
- Results on real-world federated systems prove that our method can steal nearly identical private images from other clients without interfering with training process.

2 Related Works

2.1 Diffusion memorization

The memorization phenomenon refers to the tendency of diffusion models to generate images that are nearly identical to those in the training set. Pioneering work (Gu et al., 2023) systematically investigates memorization of unconditional EDM (Karras et al., 2022) with different hyperparameters and training strategies. As for text-to-image generation, some works (Carlini et al., 2023; Somepalli et al., 2023; Wen et al., 2024) investigate the memorization in text-to-image diffusion models caused by unique prompts. Meanwhile, some works (Wang et al., 2024; Chavhan et al., 2024) investigate memorization in diffusion models from the theoretical perspective. Based on these findings, several works aim to mitigate memorization. Some of them achieve memorization mitigation by designing different training strategies. AmbientDiffusion (Daras et al., 2023) prevents diffusion from memorizing training data by training it with noisy images, IET-AGC (Liu et al., 2024; Guan et al., 2025) neglects easy-to-remember images during training. Some of them mitigate memorization in text-to-image generation during sampling process (Wen et al., 2024; Jain et al., 2025; Chen et al., 2024a; 2025a) by modifying the text embeddings of the prompt or the guidance scale during generation, while (Ren et al., 2024; Chen et al., 2025b) manipulate the attention maps of the text-to-image diffusion models. Other works (Hintersdorf et al., 2024; Dutt et al., 2025) eliminate the model parameters that cause the memorization. As for privacy attacks, most works leverage memorization for membership-inference attacks (Pang & Wang, 2023; Li et al., 2024; Ma et al., 2025; Jiang et al., 2025). Different from them, we leverage the memorization to directly steal private training images, uncovering a more severe privacy leakage.

2.2 Data stealing in federated learning

Data stealing attacks aim to reconstruct private training images of victim clients in federated systems. Most methods are inversion-based, including Gradient Inversion Attacks (GIA) (Du et al., 2025) and Model Inversion Attacks (MIA) (Fang et al., 2024). GIA typically assumes a server-side attacker with access to victim gradients and optimizes dummy inputs to match them (Zhu et al., 2019; Zhao et al., 2020; Jeon et al., 2021; Fang et al., 2023; Liu et al., 2025), while MIA performs post-training reconstruction by matching model outputs or internal representations (Zhang et al., 2020; Chen et al., 2021; Wang et al., 2021; Chen et al., 2021; Khosravy et al., 2022). However, GIA requires server-side access to client gradients, and MIA assumes a discriminative model with semantically structured outputs (e.g., logits or features); therefore, neither is applicable to client-side data stealing in federated diffusion models. A diffusion-specific attempt (Gan et al., 2024) injects trojans to poison

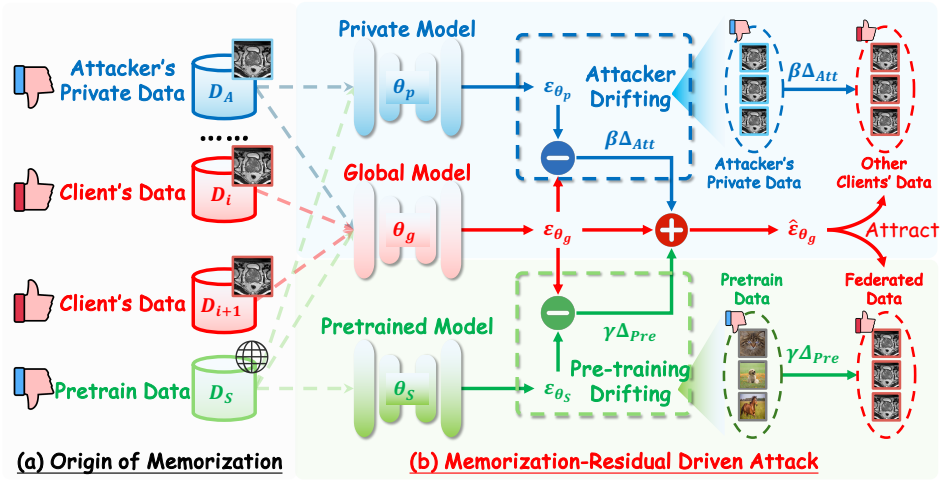


Figure 3: **Overview of our method.** (a) Data sources from which the pretrained, private, and global models may memorize images. (b) Attacker Drifting constructs $\Delta_{\text{Att}} = \epsilon_{\theta_g} - \epsilon_{\theta_p}$ using the attacker-private model as a negative reference, while Pre-training Drifting constructs $\Delta_{\text{Pre}} = \epsilon_{\theta_g} - \epsilon_{\theta_s}$ relative to the pretrained checkpoint. Their combination biases generation toward memorized images associated with other federated clients.

the global diffusion model, but it can only recover data from already controlled clients, unlike our client-to-client threat model targeting *other* benign clients *without* abnormal training behaviors.

3 Method

3.1 Attack Scenario

We consider a standard federated learning (FL) system with K clients, where client i holds a private dataset $\mathcal{D}_i$ that remains local throughout training. Federated training starts from a publicly available pretrained diffusion model θ_s, which is initialized as the global model, i.e., $\theta_g^0 = \theta_s$. The system then collaboratively fine-tunes θ_s using the distributed private data $\mathcal{D} = \bigcup_{i=1}^{K} \mathcal{D}_i$ over R communication rounds (McMahan et al., 2017). At round r, the server broadcasts the current global model θ_g^{r-1} to all participating clients. Each client locally fine-tunes it on $\mathcal{D}_i$ to obtain an updated model θ_i^r, which is returned to the server for aggregation. The server combines the client updates to produce θ_g^r. After R rounds, the resulting model $\theta_g = \theta_g^R$ captures knowledge learned from the datasets of all participating clients, despite never directly accessing their raw data.

Attacker's goal. The attacker is a common client in the federated system who wants to steal training data from *other* clients. Suppose that the attacker's client ID is A, where $1 \leq A \leq K$. The attacker aims to generate a set of images $\mathcal{I}$ from the global diffusion model such that as many images as possible match those in the union of all *other* clients' private datasets, denoted as $\mathcal{D}_B = \bigcup_{i=1,\, i\neq A}^{K} \mathcal{D}_i$. The attack objective is to steal images from $\mathcal{D}_B$.

Attacker's abilities. The attacker is an honest-but-curious client with the same capabilities as any normal participant in the federated system. It has full access to its own private data $\mathcal{D}_A$, to the aggregated global diffusion model θ_g released by the server, and to the publicly available pretrained model θ_s used to initialize federated training. From these assets the attacker derives its private model θ_p by fine-tuning θ_s on $\mathcal{D}_A$ alone, which is entirely independent of the federated protocol. It is not the local model θ_A^r that the attacker trains during any communication round. Instead, it is trained offline, starting from the public checkpoint θ_s, using only the attacker's own images. Consequently θ_p carries no information about any other client. Under this setting, the attacker never observes other clients' private data, gradients, local models. It follows the prescribed training protocol throughout without interfering with federated optimization, making the attack stealthy and difficult to detect.

Algorithm 1 Memorization-Guided Attack

Require: Global model θ_g; private model θ_p; pretrained model θ_s; class-conditioned prompt $c =$ "A photo of a {disease category}."; null token $\varnothing$; CFG scale w; drifting coefficient β, γ; number of generated images N; DDIM denoising function $\Phi(x_t, t, \epsilon)$.
Ensure: Extracted image set $\mathcal{I}$

```
Initialize I ← ∅
for i = 1 to N do
    Sample Gaussian noise x_T ~ N(0, I)
    for t = T down to 1 do
        ε_k ← ε_θk(x_t, c, t), k ∈ {g, p, s};   ε_g^∅ ← ε_θg(x_t, ∅, t)
        Δ_Att ← ε_g − ε_p {Eq. 1}
        Δ_Pre ← ε_g − ε_s {Eq. 2}
        ε̂_θg ← ε_g^∅ + w(ε_g − ε_g^∅) + βΔ_Att + γΔ_Pre {Eq. 3}
        x_{t−1} ← Φ(x_t, t, ε̂_θg)
    end for
    I ← I ∪ {x_0}
end for
return I
```

3.2 Memorization-Residual Attack

As discussed above, the attacker aims to generate memorized private images from the threat dataset $\mathcal{D}_B$ as efficiently as possible. However, as shown in Figure 3(a), freely sampling from the global model θ_g may produce memorized images that are useless to the attacker, including images from its own dataset $\mathcal{D}_A$ and the public dataset $\mathcal{D}_S$ used to pretrain the source model θ_s. To make the global model focus more on memorized images from $\mathcal{D}_B$, we propose two negative guidance strategies applied during sampling. They use the attacker's legally accessible private model θ_p and pretrained model θ_s to construct directions away from undesired image distributions and toward the desired distribution. Adding these directions to the denoising process increases the probability of recovering private images from $\mathcal{D}_B$. We next introduce the two guidance strategies in detail.

Attacker Drifting. The attacker aims to recover memorized images from the threat clients in $\mathcal{D}_B$, rather than images from its own dataset $\mathcal{D}_A$. However, because the global model θ_g is fine-tuned on data from all clients, direct sampling may reproduce both types of images. To make sampling focus more on the threat clients, the attacker uses its private model θ_p, which is initialized from the same pretrained model and legitimately fine-tuned only on $\mathcal{D}_A$. Its prediction therefore reflects the denoising direction toward the attacker's own data. As illustrated in Figure 3(b), we subtract the private-model prediction from the global-model prediction. Since θ_p is fine-tuned only on the attacker's own data $\mathcal{D}_A$, this difference emphasizes prediction changes in θ_g relative to the attacker-private reference, thereby biasing sampling toward information contributed by other clients. We term the resulting direction **Attacker Drifting**:

$$\Delta_{\mathrm{Att}}(x_t, c, t) = \epsilon_{\theta_g}(x_t, c, t) - \epsilon_{\theta_p}(x_t, c, t), \tag{1}$$

where c denotes the class-conditioned prompt "A photo of a {disease category}," and $\epsilon_{\theta_g}(x_t, c, t)$ and $\epsilon_{\theta_p}(x_t, c, t)$ are the conditional predictions of the global and private models, respectively. The scaled direction $\beta\Delta_{\mathrm{Att}}$, with $\beta \geq 0$, steers denoising away from memorized images in $\mathcal{D}_A$ and increases the chance of recovering those from the threat clients in $\mathcal{D}_B$.

Pre-training Drifting. The global model θ_g is initialized from the pretrained model θ_s and fine-tuned on all clients' federated data. Thus, direct sampling from θ_g may reproduce images memorized from both the public pretraining data $\mathcal{D}_S$ and the federated client data, while only the latter is relevant to the attack. To make generation focus more on information acquired during federated fine-tuning, we use the publicly available θ_s as a negative reference. Since θ_s has not accessed any client data, subtracting its prediction from that of θ_g emphasizes prediction changes introduced after initialization from the pretrained checkpoint. We call this direction **Pre-training Drifting**:

$$\Delta_{\mathrm{Pre}}(x_t, c, t) = \epsilon_{\theta_g}(x_t, c, t) - \epsilon_{\theta_s}(x_t, c, t), \tag{2}$$

Table 1: Comparison of data extraction performance on Fed-Fundus (Liu et al., 2021) and Fed-Prostate (Liu et al., 2021). Drifting coefficients are fixed as $(\beta, \gamma) = (0.5, 1.5)$. Higher SSCD and Mem scores indicate more severe training-data leakage. The inversion-based attacks require auxiliary information unavailable in the client-to-client setting. Best and second-best memorization-based results are marked in bold and underlined, respectively.

	Fed-Fundus			**Fed-Prostate**		
Method	Mem@0.6↑	Mem@0.7↑	$\text{SSCD}^{95\text{th}}$↑	Mem@0.6↑	Mem@0.7↑	$\text{SSCD}^{95\text{th}}$↑
Inversion-based attacks						
GIFD (Fang et al., 2023)	0.0	0.0	0.426	0.4	0.0	0.537
FedInverse (Wu et al., 2024)	0.2	0.0	0.532	0.1	0.0	0.478
Memorization-based attacks						
Noise Search (Wen et al., 2024)	69.6	56.1	0.769	30.4	2.4	0.675
Direct Sampling (Gu et al., 2023)	86.7	65.7	0.767	29.2	3.5	0.684
$+\beta\Delta_{\text{Att}}$ (Ours)	86.8	68.2	0.771	33.0	4.1	0.691
$+\gamma\Delta_{\text{Pre}}$ (Ours)	<u>91.2</u>	<u>75.5</u>	<u>0.790</u>	<u>38.5</u>	<u>5.2</u>	<u>0.702</u>
Full (Ours)	**94.8**	**79.1**	**0.815**	**42.7**	**6.8**	**0.713**

where $\epsilon_{\theta_s}(x_t, c, t)$ is the conditional prediction of the pretrained model under the same class-conditioned prompt c. The scaled direction $\gamma\Delta_{\text{Pre}}$, where $\gamma \geq 0$, guides denoising away from images memorized from $\mathcal{D}_S$ and toward images memorized from the federated client data.

Built on Attacker Drifting and Pre-training Drifting, we present the complete sampling procedure in Algorithm 1. At each denoising timestep, the attacker queries the global, private, and pretrained models using the same class-conditioned prompt c. Their predictions are used to construct Δ_{Att} and Δ_{Pre}, which are added in parallel to the global-model prediction:

$$\begin{aligned}\hat{\epsilon}_{\theta_g}(x_t, c, t) = \epsilon_{\theta_g}(x_t, \varnothing, t) + w\Big(\epsilon_{\theta_g}(x_t, c, t) - \epsilon_{\theta_g}(x_t, \varnothing, t)\Big) \\ + \beta\Big(\epsilon_{\theta_g}(x_t, c, t) - \epsilon_{\theta_p}(x_t, c, t)\Big) \\ + \gamma\Big(\epsilon_{\theta_g}(x_t, c, t) - \epsilon_{\theta_s}(x_t, c, t)\Big).\end{aligned} \tag{3}$$

where the term scaled by w is the standard classifier-free guidance (CFG) Ho & Salimans (2022) applied to the global model, with $\varnothing$ denoting the null (unconditional) token. The entire attack is performed after federated training using only models legitimately accessible to the attacker.

4 Experiment

4.1 Settings

Datasets. We conduct experiments on two multi-center medical image datasets following the federated partitions established by FEAL Chen et al. (2024b). Each client corresponds to a distinct clinical center or data source, thereby preserving naturally occurring cross-center domain shifts. Fed-Prostate Liu et al. (2021) consists of 1,867 prostate MRI slices from six centers; and Fed-Fundus Liu et al. (2021) contains 1,060 retinal fundus images from four centers.

Training Settings. Following FedDiffuse de Goede et al. (2024), we adopt FedAvg McMahan et al. (2017) to train federated diffusion models. Specifically, we initialize the model from the pre-trained Stable Diffusion 2.1 Rombach et al. (2022) checkpoint and perform federated fine-tuning with all clients participating in every communication round. The server aggregates the locally updated models using sample-size-weighted averaging. Details can be seen in Section A.1.

Inference Settings. We set the attack generation budget to the size of the threat dataset, i.e., $|\mathcal{I}| = |\mathcal{D}_B|$, where $\mathcal{I}$ denotes the generated image set. For the Mem and SSCD results, we sample 512×512 images using DDIM (Song et al., 2021) with 50 denoising steps. Unless otherwise specified, we use attacker client 0, fix $(\beta, \gamma) = (0.5, 1.5)$, and set the classifier-free guidance (CFG) scale (Ho &

Salimans, 2022) to $w = 7.5$ for Fed-Fundus and $w = 2.5$ for Fed-Prostate. The UMem results use a separate $1\times$ evaluation, as indicated in Table 3.

Evaluation Metrics. We use Self-Supervised Copy Detection (SSCD) (Pizzi et al., 2022) as our copy-detection criterion, following prior diffusion-model memorization studies (Wen et al., 2024; Chen et al., 2024a; Jain et al., 2025). For each generated image $\hat{x} \in \mathcal{I}$, its score is the maximum cosine similarity to the threat dataset, $s(\hat{x}, \mathcal{D}_B) = \max_{x \in \mathcal{D}_B} \frac{\phi(\hat{x})^\top \phi(x)}{\|\phi(\hat{x})\|_2 \|\phi(x)\|_2}$, where ϕ is the SSCD descriptor. At the output level, $\mathrm{SSCD}^{95\mathrm{th}}$ is the 95th percentile of these scores, measuring similarity in the upper tail. $\mathrm{Mem@}\tau$ is the ratio of the number of generated images with $s(\hat{x}, \mathcal{D}_B) \geq \tau$ to the total number of generated images $|\mathcal{I}|$; it measures the frequency of high-similarity reproduction in generated outputs. At the target level, $\mathrm{UMem@}\tau$ is the ratio of the number of distinct threat images selected as nearest-neighbor matches with SSCD similarity at least τ to the total number of threat images $|\mathcal{D}_B|$; it measures the coverage of unique threat images, counting repeated matches to the same target only once. Both ratios are reported as percentages at $\tau = 0.6$ and 0.7.

Baselines. For memorization-based attacks, *Direct Sampling* (Gu et al., 2023) samples from the final global model, while *Noise Search*, adapted from Wen et al. (2024), optimizes initial noise to maximize the text-conditional prediction magnitude. For inversion-based attacks, GIFD (Fang et al., 2023) matches victim gradients, whereas FedInverse (Wu et al., 2024) reconstructs images using target labels and a pretrained inversion generator. We grant the inversion baselines the auxiliary information required by their original settings, beyond our client-to-client threat model. Further details are provided in Appendix A.2.

4.2 Main Results

Comparison with Memorization-Based Attacks. We first evaluate our attack against two memorization-based baselines. Direct Sampling (Gu et al., 2023) samples from the final global model without any guidance, and thus reflects the model's intrinsic tendency to reproduce memorized training data. Noise Search (Wen et al., 2024) instead searches the initial noise to maximize the magnitude of the text-conditional prediction. As shown in Table 1, combining Attacker Drifting and Pre-training Drifting improves the Mem and SSCD metrics on both datasets over both baselines. Over Direct Sampling, it raises $\mathrm{Mem@}0.6$ from 86.7% to 94.8% on Fed-Fundus and from 29.2% to 42.7% on Fed-Prostate. On the stricter $\mathrm{Mem@}0.7$ it gains 13.4 points on Fed-Fundus and nearly doubles the $\mathrm{Mem@}0.7$ on Fed-Prostate (3.5% to 6.8%). The gains on both show that our guidance exposes more privacy than sampling from the global model alone.

Comparison with Inversion-Based Attacks. We then compare against the inversion-based methods GIFD (Fang et al., 2023) and FedInverse (Wu et al., 2024). Both rely on victim-specific side information that an ordinary client cannot obtain: GIFD matches the gradients shared by a victim client, and FedInverse requires the target labels and structurally similar public images to pretrain its inversion generator. To give them the best possible chance, we relax our threat model and grant each method the full auxiliary information assumed in its original setting. **Even so, they fail almost completely**: their $\mathrm{Mem@}0.6$ never exceeds 0.4% and their $\mathrm{Mem@}0.7$ is 0.0% on both datasets, meaning they reproduce essentially no private image, whereas our attack reaches 94.8% and 42.7% at $\mathrm{Mem@}0.6$. As shown in Figure 4, our recovered images are visibly closer to the private scans than the inversion baselines. **This demonstrates that our memorization-based attack far surpasses inversion-based methods in both effectiveness and practicality for client-to-client data stealing.**

4.3 Ablation Study

Effectiveness of Attacker Drifting Δ_{Att}. Used alone, Δ_{Att} improves over Direct Sampling on both datasets: it raises $\mathrm{Mem@}0.7$ from 65.7% to 68.2% on Fed-Fundus and $\mathrm{Mem@}0.6$ from 29.2% to 33.0% on Fed-Prostate. Adding it to Δ_{Pre} gives a further 3.6 and 4.2 points at $\mathrm{Mem@}0.6$, with consistent gains on $\mathrm{Mem@}0.7$ and $\mathrm{SSCD}^{95\mathrm{th}}$. Δ_{Att} therefore strengthens the attack in combination.

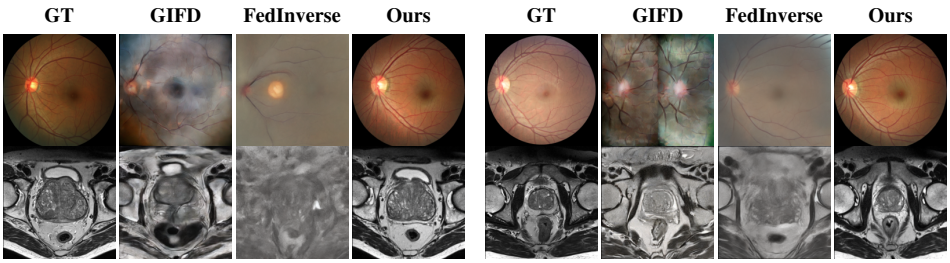


Figure 4: **Qualitative comparisons between our method and inversion-based methods.** Each row shows the ground truth training images and stolen images by different methods. More visualization results can be seen in Appendix.

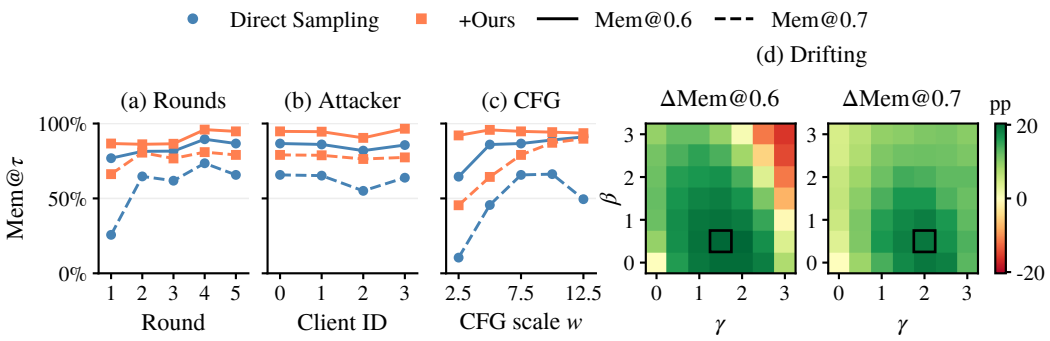


Figure 5: **Sensitivity analysis.** Solid and dashed lines correspond to $\tau = 0.6$ and 0.7, respectively; blue denotes Direct Sampling and orange denotes Ours. Panels (a)–(c) report $\mathrm{Mem@}\tau$ on Fed-Fundus when varying (a) global-model training rounds, (b) attacker identity, and (c) CFG scale. Panel (d) varies the drifting coefficients β and γ, with the left and right grids corresponding to $\tau = 0.6$ and 0.7; cells show the change in $\mathrm{Mem@}\tau$ relative to sampling without drifting, averaged over both datasets, in percentage points (pp). Green indicates improvement, red indicates degradation, and black boxes mark the best results.

Effectiveness of Pre-training Drifting Δ_{Pre}. Δ_{Pre} brings larger gains. It lifts $\mathrm{Mem@}0.6$ from 86.7% to 91.2% on Fed-Fundus and from 29.2% to 38.5% on Fed-Prostate, with $\mathrm{Mem@}0.7$ rising to 75.5% and 5.2% respectively. It is the stronger of the two components, and improves every metric on both datasets.

4.4 Sensitivity Analysis

Influence of federated training rounds. We study how the training progress of the global model affects the attack. As shown in Figure 5(a), our method stays above Direct Sampling at every round, on both $\mathrm{Mem@}0.6$ and $\mathrm{Mem@}0.7$. Extraction is generally higher at later rounds, although the trend is not monotonic, consistent with training-dependent memorization (Gu et al., 2023).

Influence of Attacker Identity. Under non-IID federated data, changing the attacker also changes both its local data and the threat-data distribution. We select all four Fed-Fundus clients as the attacker in turn. As shown in Figure 5(b), the absolute scores vary with attacker identity, yet our attack consistently improves both $\mathrm{Mem@}0.6$ and $\mathrm{Mem@}0.7$ over Direct Sampling for every attacker. This proves the attack remains effective across heterogeneous attacker and threat distributions.

Influence of Classifier-Free Guidance Scale w. Our drifting terms share the algebraic form of CFG, so the gains above could be attributed to stronger guidance rather than to memorization separation. We therefore sweep w for both methods on Fed-Fundus with (β, γ) fixed, as reported in Figure 5(c). Larger w raises extraction for both (Jain et al., 2025), yet our attack stays above Direct Sampling at every scale. On $\mathrm{Mem@}0.6$, our weakest setting (92.0% at $w = 2.5$) exceeds the best

Table 2: **Attack performance against different federated defense methods.** FID is measured on the *test* set using images directly sampled from each global model. Results are evaluated on Fed-Fundus.

Metric	FedAvg	DP-LDM	IET-AGC
FID ↓	6.78	73.15	7.99
$\mathrm{SSCD}^{95\mathrm{th}}$ ↑	0.815	0.694	0.807
$\mathrm{Mem@0.6}$ ↑	94.8%	5.9%	92.3%
$\mathrm{Mem@0.7}$ ↑	79.1%	2.4%	77.4%

Table 3: **Unique-target coverage (UMem, %).** Drifting coefficients are fixed as $(\beta, \gamma) = (0.5, 1.5)$. Higher UMem scores indicate more severe training-data leakage.

	Fed-Fundus		Fed-Prostate	
Method	@0.6 ↑	@0.7 ↑	@0.6 ↑	@0.7 ↑
Direct Sampling	4.73	1.01	7.90	0.84
$+\beta\Delta_{\mathrm{Att}}$ (Ours)	6.08	1.35	8.43	0.99
$+\gamma\Delta_{\mathrm{Pre}}$ (Ours)	8.28	3.72	8.88	1.82
Full (Ours)	**8.45**	**4.05**	**8.96**	**1.97**

Direct Sampling result over the swept range (91.0% at $w = 12.5$). On $\mathrm{Mem@0.7}$, Direct Sampling peaks near $w = 10$ and falls to 49.5%, while ours rises to 89.5%. The gain therefore does not stem from guidance strength alone.

Influence of Drifting Coefficients β and γ. Figure 5(d) shows that both guidance terms remain effective over a broad range of moderate scales. For $\beta \in [0.5, 1.5]$ and $\gamma \in [1.0, 2.5]$, both $\mathrm{Mem@0.6}$ and $\mathrm{Mem@0.7}$ improve consistently over sampling without guidance, indicating that Attacker Drifting and Pre-training Drifting jointly increase both the number of reproduced private images and the severity of the most faithful reproductions. The best average gains appear around $(\beta, \gamma) = (0.5, 1.5)$ for $\mathrm{Mem@0.6}$ and $(0.5, 2.0)$ for $\mathrm{Mem@0.7}$. When both scales grow too large, however, the combined correction overwhelms the global-model prediction ϵ_{θ_g} and pushes the drifted prediction $\hat{\epsilon}_{\theta_g}$ away from a valid denoising direction, which explains the degradation in the upper-right corner of the $\mathrm{Mem@0.6}$ heatmap.

4.5 Target Coverage Analysis

The preceding results evaluate the fraction of generated images that closely match private training samples, showing that our attack increases the likelihood of reproducing private images. As a complementary analysis, we further examine how many *distinct* private training images can be recovered from the original threat dataset. As shown in Table 3, our method consistently improves unique-target coverage over Direct Sampling on both datasets. Our method increases $\mathrm{UMem@0.7}$ from 1.01% to 4.05% on Fed-Fundus and from 0.84% to 1.97% on Fed-Prostate. These results suggest that our method can broaden the set of distinct private images recovered from threat dataset.

4.6 Robustness Analysis

Robustness against Privacy Defenses. Table 2 evaluates our attack against two defenses: DP-LDM (Lyu et al., 2024), which applies differential privacy with $(\varepsilon, \delta) = (10, 10^{-5})$, and IET-AGC (Liu et al., 2024), which clips the loss on samples the model starts to memorize. DP-LDM suppresses $\mathrm{Mem@0.6}$ to 5.9%, but its FID rises from 6.78 to 73.15, leaving a model that no longer generates usable images. IET-AGC preserves quality at an FID of 7.99, yet reduces $\mathrm{Mem@0.7}$ by less than two points, so the leakage remains.

5 Conclusion

We identify a client-to-client privacy leakage risk in federated diffusion models: because the global model memorizes training images from every participant, an ordinary client can regenerate the private images of others without observing their data, gradients, or local updates. To expose this risk, we propose a memorization-residual attack that separates the memorization the attacker targets from the memorization it already owns. The attack instantiates this idea as two sampling-time drifting terms, Attacker Drifting and Pre-training Drifting, and operates entirely after federated training. On two real-world federated medical datasets, it steals nearly identical private images from other clients. Together, these results expose a serious and largely overlooked vulnerability of federated diffusion, and call for stronger privacy protection in federated generative learning.

AI USE STATEMENT

In this work, we used generative AI tools for literature review and writing refinement. All AI-assisted content was reviewed and verified by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

All experiments use publicly available, de-identified medical-image datasets in a controlled federated-learning setting; we do not collect or disclose new patient images or attempt patient re-identification. Our purpose is to reveal privacy risks and motivate protective measures, not to facilitate unauthorized data extraction or other misuse.

REPRODUCIBILITY STATEMENT

Algorithm 1 and Section 4 describe the attack procedure, datasets, and evaluation protocol. Appendices A.1, A.2, and A.6 detail training, sampling, baseline adaptations, and seed sensitivity.

REFERENCES

Nicolas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramer, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models. In *32nd USENIX security symposium (USENIX Security 23)*, pp. 5253–5270, 2023.

Ruchika Chavhan, Ondrej Bohdal, Yongshuo Zong, Da Li, and Timothy Hospedales. Memorized images in diffusion models share a subspace that can be located and deleted. *arXiv preprint arXiv:2406.18566*, 2024.

Chen Chen, Daochang Liu, and Chang Xu. Towards memorization-free diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8425–8434, 2024a.

Chen Chen, Daochang Liu, Mubarak Shah, and Chang Xu. Enhancing privacy-utility trade-offs to mitigate memorization in diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 8182–8191, 2025a.

Chen Chen, Daochang Liu, Mubarak Shah, and Chang Xu. Exploring local memorization in diffusion models via bright ending attention. In *International Conference on Learning Representations*, volume 2025, pp. 29827–29841, 2025b.

Jiayi Chen, Benteng Ma, Hengfei Cui, and Yong Xia. Think twice before selection: Federated evidential active learning for medical image analysis with domain shifts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11439–11449, 2024b.

Si Chen, Mostafa Kahla, Ruoxi Jia, and Guo-Jun Qi. Knowledge-enriched distributional model inversion attacks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 16178–16187, 2021.

Siyuan Cheng, Guanhong Tao, Yingqi Liu, Guangyu Shen, Shengwei An, Shiwei Feng, Xiangzhe Xu, Kaiyuan Zhang, Shiqing Ma, and Xiangyu Zhang. Lotus: Evasive and resilient backdoor attacks through sub-partitioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24798–24809, 2024.

Giannis Daras, Kulin Shah, Yuval Dagan, Aravind Gollakota, Alex Dimakis, and Adam Klivans. Ambient diffusion: Learning clean distributions from corrupted data. *Advances in Neural Information Processing Systems*, 36:288–313, 2023.

Matthijs de Goede, Bart Cox, and Jérémie Decouchant. Training diffusion models with federated learning. *arXiv preprint arXiv:2406.12575*, 2024.

Jiacheng Du, Jiahui Hu, Zhibo Wang, Peng Sun, Neil Gong, Kui Ren, and Chun Chen. {SoK}: On gradient leakage in federated learning. In *34th USENIX Security Symposium (USENIX Security 25)*, pp. 3045–3064, 2025.

Raman Dutt, Ondrej Bohdal, Pedro Sanchez, Sotirios A Tsaftaris, and Timothy Hospedales. Memcontrol: mitigating memorization in diffusion models via automated parameter selection. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 4491–4501. IEEE, 2025.

Hao Fang, Bin Chen, Xuan Wang, Zhi Wang, and Shu-Tao Xia. Gifd: A generative gradient inversion method with feature domain optimization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4967–4976, 2023.

Hao Fang, Yixiang Qiu, Hongyao Yu, Wenbo Yu, Jiawei Kong, Baoli Chong, Bin Chen, Xuan Wang, Shu-Tao Xia, and Ke Xu. Privacy leakage on dnns: A survey of model inversion attacks and defenses. *arXiv preprint arXiv:2402.04013*, 2024.

Yuan Gan, Jiaxu Miao, and Yi Yang. Datastealing: Steal data from diffusion models in federated learning with multiple trojans. *Advances in Neural Information Processing Systems*, 37:132614–132646, 2024.

Xiangming Gu, Chao Du, Tianyu Pang, Chongxuan Li, Min Lin, and Ye Wang. On memorization in diffusion models. *arXiv preprint arXiv:2310.02664*, 2023.

Xiaoliu Guan, Yu Wu, Huayang Huang, Xiao Liu, Jiaxu Miao, and Yi Yang. Redistribute ensemble training for mitigating memorization in diffusion models. *arXiv preprint arXiv:2502.09434*, 2025.

Dominik Hintersdorf, Lukas Struppek, Kristian Kersting, Adam Dziedzic, and Franziska Boenisch. Finding nemo: Localizing neurons responsible for memorization in diffusion models. *Advances in Neural Information Processing Systems*, 37:88236–88278, 2024.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Xumin Huang, Peichun Li, Hongyang Du, Jiawen Kang, Dusit Niyato, Dong In Kim, and Yuan Wu. Federated learning-empowered ai-generated content in wireless networks. *IEEE Network*, 38(5): 304–313, 2024.

Anubhav Jain, Yuya Kobayashi, Takashi Shibuya, Yuhta Takida, Nasir Memon, Julian Togelius, and Yuki Mitsufuji. Classifier-free guidance inside the attraction basin may cause memorization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12871–12879, 2025.

Jinwoo Jeon, Jaechang Kim, Kangwook Lee, Sewoong Oh, and Jungseul Ok. Gradient inversion with generative image prior. *arXiv preprint arXiv:2110.14962*, 2021.

Yue Jiang, Haokun Lin, Yang Bai, Bo Peng, Zhili Liu, Yueming Lyu, Yong Yang, Jing Dong, et al. Image-level memorization detection via inversion-based inference perturbation. In *The Thirteenth International Conference on Learning Representations*, 2025.

Peter Kairouz and H Brendan McMahan. Advances and open problems in federated learning. *Foundations and trends in machine learning*, 14(1-2):1–210, 2021.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in neural information processing systems*, 35:26565–26577, 2022.

Mahdi Khosravy, Kazuaki Nakamura, Yuki Hirose, Naoko Nitta, and Noboru Babaguchi. Model inversion attack by integration of deep generative models: Privacy-sensitive face generation from a face recognition system. *IEEE Transactions on Information Forensics and Security*, 17:357–372, 2022.

Qiao Li, Xiaomeng Fu, Xi Wang, Jin Liu, Xingyu Gao, Jiao Dai, and Jizhong Han. Unveiling structural memorization: Structural membership inference attack for text-to-image diffusion models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 10554–10562, 2024.

Quande Liu, Cheng Chen, Jing Qin, Qi Dou, and Pheng-Ann Heng. Feddg: Federated domain generalization on medical image segmentation via episodic learning in continuous frequency space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1013–1023, 2021.

Xiao Liu, Xiaoliu Guan, Yu Wu, and Jiaxu Miao. Iterative ensemble training with anti-gradient control for mitigating memorization in diffusion models. In *European Conference on Computer Vision*, pp. 108–123. Springer, 2024.

Xuan Liu, Siqi Cai, Qihua Zhou, Song Guo, Ruibin Li, and Kaiwei Lin. Mjölnir: Breaking the shield of perturbation-protected gradients via adaptive diffusion. In *Proceedings of the AAAI conference on artificial intelligence*, volume 39, pp. 26308–26316, 2025.

Saiyue Lyu, Michael F Liu, Margarita Vinaroz, and Mijung Park. Differentially private latent diffusion models. In *Privacy Regulation and Protection in Machine Learning*, 2024.

Zhe Ma, Qingming Li, Xuhong Zhang, Tianyu Du, Ruixiao Lin, Zonghui Wang, Shouling Ji, and Wenzhi Chen. An inversion-based measure of memorization for diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16959–16969, 2025.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pp. 1273–1282. Pmlr, 2017.

Jean Ogier du Terrail, Samy-Safwan Ayed, Edwige Cyffers, Felix Grimberg, Chaoyang He, Regis Loeb, Paul Mangold, Tanguy Marchand, Othmane Marfoq, Erum Mushtaq, et al. Flamby: Datasets and benchmarks for cross-silo federated learning in realistic healthcare settings. *Advances in Neural Information Processing Systems*, 35:5315–5334, 2022.

Maxime Oquab, Timothee Darcet, Theo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaa El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. URL `https://openreview.net/forum?id=a68SUt6zFt`.

Yan Pang and Tianhao Wang. Black-box membership inference attacks against fine-tuned diffusion models. *arXiv preprint arXiv:2312.08207*, 2023.

Ed Pizzi, Sreya Dutta Roy, Sugosh Nagavara Ravindra, Priya Goyal, and Matthijs Douze. A self-supervised descriptor for image copy detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2022.

Jie Ren, Yaxin Li, Shenglai Zeng, Han Xu, Lingjuan Lyu, Yue Xing, and Jiliang Tang. Unveiling and mitigating memorization in text-to-image diffusion models through cross attention. In *European Conference on Computer Vision*, pp. 340–356. Springer, 2024.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Understanding and mitigating copying in diffusion models. *Advances in Neural Information Processing Systems*, 36:47783–47803, 2023.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021.

Fiona Victoria Stanley Jothiraj and Afra Mashhadi. Phoenix: A federated generative diffusion model. In *Companion Proceedings of the ACM Web Conference 2024*, pp. 1568–1577, 2024.

Ye Lin Tun, Chu Myaet Thwal, Ji Su Yoon, Sun Moo Kang, Chaoning Zhang, and Choong Seon Hong. Federated learning with diffusion models for privacy-sensitive vision tasks. In *2023 International Conference on Advanced Technologies for Communications (ATC)*, pp. 305–310. IEEE, 2023.

Hanyu Wang, Yujin Han, and Difan Zou. On the discrepancy and connection between memorization and generation in diffusion models. In *ICML 2024 Workshop on Foundation Models in the Wild*, 2024.

Hongyi Wang, Kartik Sreenivasan, Shashank Rajput, Harit Vishwakarma, Saurabh Agarwal, Jyyong Sohn, Kangwook Lee, and Dimitris Papailiopoulos. Attack of the tails: Yes, you really can backdoor federated learning. *Advances in neural information processing systems*, 33:16070–16084, 2020.

Jinbao Wang, Guoyang Xie, Yawen Huang, Jiayi Lyu, Feng Zheng, Yefeng Zheng, and Yaochu Jin. Fedmed-gan: Federated domain translation on unsupervised cross-modality brain image synthesis. *Neurocomputing*, 546:126282, 2023.

Kuan-Chieh Wang, Yan Fu, Ke Li, Ashish Khisti, Richard Zemel, and Alireza Makhzani. Variational model inversion attacks. *Advances in neural information processing systems*, 34:9706–9719, 2021.

Jiaheng Wei, Yanjun Zhang, Leo Yu Zhang, Chao Chen, Shirui Pan, Kok-Leong Ong, Jun Zhang, and Yang Xiang. Client-side gradient inversion against federated learning from poisoning. *arXiv preprint arXiv:2309.07415*, 2023.

Yuxin Wen, Yuchen Liu, Chen Chen, and Lingjuan Lyu. Detecting, explaining, and mitigating memorization in diffusion models. In *The Twelfth International Conference on Learning Representations*, 2024.

Ethan Wilson, Kai Yue, Chau-Wai Wong, and Huaiyu Dai. Federated learning nodes can reconstruct peers' image data. *arXiv preprint arXiv:2410.04661*, 2024.

Di Wu, Jun Bai, Yiliao Song, Junjun Chen, Wei Zhou, Yong Xiang, and Atul Sajjanhar. Fedinverse: Evaluating privacy leakage in federated learning. In *The twelfth international conference on learning representations*, 2024.

Chulin Xie, Keli Huang, Pin-Yu Chen, and Bo Li. Dba: Distributed backdoor attacks against federated learning. In *International conference on learning representations*, 2019.

TaeHo Yoon, Joo Young Choi, Sehyun Kwon, and Ernest K Ryu. Diffusion probabilistic models generalize when they fail to memorize. In *ICML 2023 workshop on structured probabilistic inference & generative modeling*, 2023.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018.

Yuheng Zhang, Ruoxi Jia, Hengzhi Pei, Wenxiao Wang, Bo Li, and Dawn Song. The secret revealer: Generative model-inversion attacks against deep neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 253–261, 2020.

Bo Zhao, Konda Reddy Mopuri, and Hakan Bilen. idlg: Improved deep leakage from gradients. *arXiv preprint arXiv:2001.02610*, 2020.

Ligeng Zhu, Zhijian Liu, and Song Han. Deep leakage from gradients. *Advances in neural information processing systems*, 32, 2019.

A Appendix

A.1 Implementation Details

Federated fine-tuning. We initialize the global model from Stable Diffusion 2.1 (Rombach et al., 2022) and fine-tune it with all clients participating in every communication round. At the start of round r, the server broadcasts θ_g^{r-1}; each client initializes its local model from this checkpoint and performs 2,000 updates on its own images. The server then computes the sample-size-weighted FedAvg (McMahan et al., 2017) of the returned local models to obtain θ_g^r. Table A1 reports the dataset-specific round budgets and the remaining optimization settings.

Private-model training. Outside the federated protocol, the attacker initializes θ_p from the same public Stable Diffusion 2.1 checkpoint and fine-tunes it only on $\mathcal{D}_A$, its own local dataset, for 4,000 updates. This independent training uses the same image preprocessing, optimizer, learning rate, and warmup schedule as federated fine-tuning, but neither accesses nor modifies any other client's local model or data. Thus, θ_p is distinct from the attacker's per-round FedAvg local model and contains no information from the threat clients.

Attack sampling. For the Mem and SSCD evaluation, we sample 512×512 images with the DDIM sampler (Song et al., 2021) for 50 denoising steps. At every step, the attack evaluates the global, private, and pretrained models under the same prompt and combines their predictions using the attacker and pre-training drifting terms defined in the main text; the dataset-specific classifier-free guidance (CFG) scale (Ho & Salimans, 2022) and drifting coefficients follow the evaluation settings in the main text. The generation budget equals the threat-dataset size, as specified in the main text.

A.2 Baseline Details

Memorization-based baselines. We consider two baselines for memorization-based attacks. First, *Direct Sampling* measures the privacy information inherently exposed by the federated diffusion model by directly generating images from the final global model (Gu et al., 2023). Second, *Noise Search* adapts the detection metric of Wen et al. (2024), which identifies memorization through the magnitude of the text-conditional noise prediction d, and mitigates memorization by minimizing d over the prompt embedding c at inference time. We reverse this objective into an attack. Their mitigation optimizes the prompt embedding because memorization there is triggered by duplicated captions. In our setting all samples share a single class-level prompt. We therefore keep c fixed and optimize initial noise to maximize d, which yields a memorization-seeking baseline.

Inversion-based baselines. We also consider the representative inversion-based attacks FedInverse (Wu et al., 2024) and GIFD (Fang et al., 2023), which reconstruct private images from victim-specific side information such as class labels or gradients. In our client-to-client threat model, such information can only be estimated by the update of the global model across rounds (Wei et al., 2023; Wilson et al., 2024), and we find that this estimation is far too coarse for these methods to recover anything meaningful. However, to still enable a comprehensive comparison, we relax our threat model for these baselines and directly grant them the victim-specific information required by their original settings. This gives them access strictly stronger than a real client-to-client attacker would have, so they serve as stronger-access reference baselines rather than direct competitors.

A.3 Qualitative Examples

Figure A1 presents qualitative extraction examples on Fed-Fundus and Fed-Prostate. Each generated image is paired with its nearest threat-dataset image under SSCD. The generated images are visually highly similar to their corresponding ground-truth images, across both retinal appearances and prostate MRI slices.

A.4 SSCD Distributional Comparison

Figure A2 compares the per-generation maximum Self-Supervised Copy Detection (SSCD) scores (Pizzi et al., 2022) for direct sampling and our complete attack. Across both datasets, our method shifts more probability mass toward higher SSCD scores.

Table A1: **Optimization hyperparameters for federated fine-tuning.** Shared settings span both datasets; dataset-specific batch sizes and round budgets are reported separately.

Training setting	**Fed-Fundus**	**Fed-Prostate**
Image preprocessing	Center crop and resize to 512×512	
Optimizer	AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.999$)	
Learning rate	weight decay 10^{-2}	
Learning-rate schedule	2.5×10^{-6}	
Total batch size	Constant with 500-step warmup	
Global rounds	8	6
	5	25

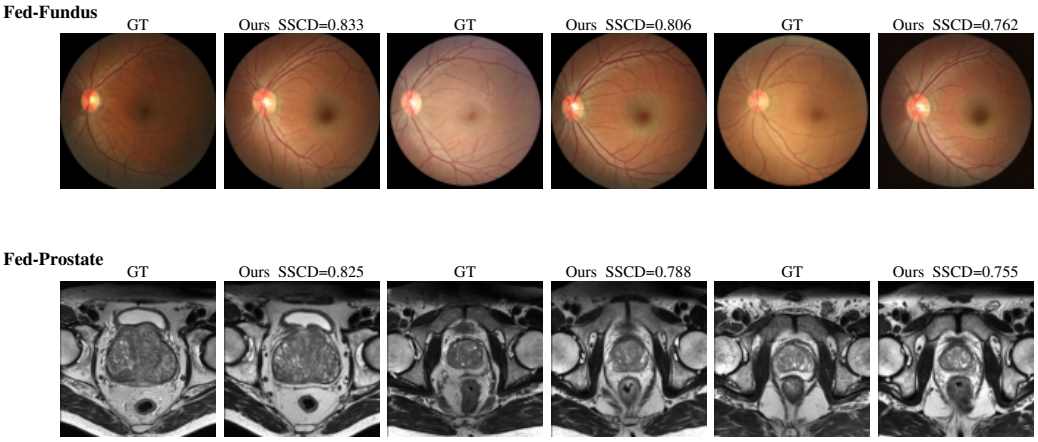


Figure A1: **Diverse high-confidence extraction examples.** Each pair contains a ground-truth threat image (GT) and its highest-SSCD generated match under our full attack.

Table A2: **Cross-embedding-space robustness.** Feature-space subscripts are introduced here for auxiliary metrics: main-text $\mathrm{Mem}@\tau$ is $\mathrm{Mem}_{\mathrm{SSCD}}@\tau$ when its SSCD feature space is explicit. DINOv2 uses independent maximum-cosine retrieval over $\mathcal{D}_B$; LPIPS evaluates the pairs selected by SSCD, where lower distance is better.

	Fed-Fundus		**Fed-Prostate**	
Statistic	**Direct**	**+Ours**	**Direct**	**+Ours**
DINOv2 cosine similarity (independent retrieval)				
$\mathrm{Mem}_{\mathrm{DINO}}@0.90$ (%) ↑	74.66	**95.10**	44.80	**81.55**
$\mathrm{Mem}_{\mathrm{DINO}}@0.95$ (%) ↑	4.22	**51.35**	0.68	**8.50**
95^{th} percentile ↑	0.9487	**0.9645**	0.9355	**0.9544**
LPIPS perceptual distance (SSCD-selected pairs)				
$\mathrm{Mem}_{\mathrm{LPIPS}}@0.40$ (%) ↑	28.72	**74.16**	81.70	**90.13**
$\mathrm{Mem}_{\mathrm{LPIPS}}@0.30$ (%) ↑	6.59	**33.45**	3.80	**6.07**
95^{th} percentile ↓	0.5792	**0.4889**	0.4282	**0.4121**

A.5 Cross-Embedding-Space Robustness

Table A2 tests whether the advantage measured by SSCD persists under representations with different inductive biases. The main-text $\mathrm{Mem}@\tau$ is $\mathrm{Mem}_{\mathrm{SSCD}}@\tau$ when its SSCD feature space is explicit. We use feature-space subscripts only for the auxiliary DINOv2 and LPIPS measurements.

Metric computation. For DINOv2 (Oquab et al., 2024), we independently compare each generated image with every target image in $\mathcal{D}_B$ and retain its maximum cosine similarity. $\mathrm{Mem}_{\mathrm{DINO}}@\tau$ is the

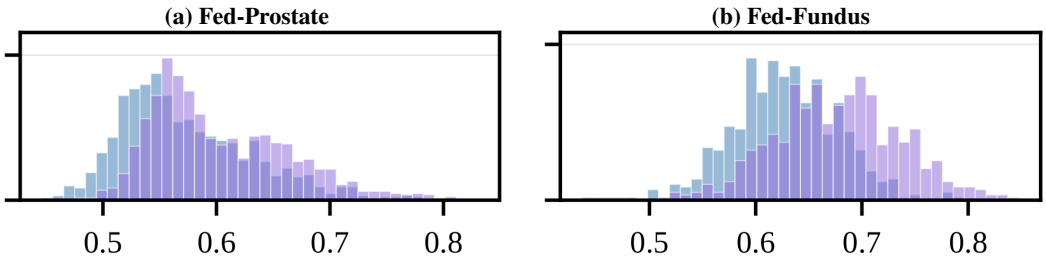


Figure A2: **Per-generation maximum SSCD-score distributions.** Blue denotes w/o Ours (Direct Sampling) and purple denotes w/ Ours ($\Delta_{\mathrm{Att}} + \Delta_{\mathrm{Pre}}$). Our attack shifts the distributions toward higher similarity to the threat dataset $\mathcal{D}_B$.

Table A3: **Random seed sensitivity.** Mean $\pm$ sample standard deviation over five randomly sampled seeds. The means are the point estimates reported in Table 1 of the main paper. Mem values are percentages. All hyperparameter settings follow the Settings subsection in the main paper.

Method	Mem@0.6 ↑	Mem@0.7 ↑	$\mathrm{SSCD}^{95\mathrm{th}}$ ↑
Fed-Fundus			
Direct Sampling	86.7±0.5	65.7±1.1	0.767±0.002
$+\beta\Delta_{\mathrm{Att}}$	86.8±0.7	68.2±1.1	0.771±0.001
$+\gamma\Delta_{\mathrm{Pre}}$	91.2±0.8	75.5±1.2	0.790±0.002
Full (Ours)	**94.8±0.4**	**79.1±1.0**	**0.815±0.001**
Fed-Prostate			
Direct Sampling	29.2±0.7	3.5±0.6	0.684±0.003
$+\beta\Delta_{\mathrm{Att}}$	33.0±0.5	4.1±0.4	0.691±0.002
$+\gamma\Delta_{\mathrm{Pre}}$	38.5±0.9	5.2±0.4	0.702±0.004
Full (Ours)	**42.7±1.6**	**6.8±0.6**	**0.713±0.005**

percentage of generated images whose retained similarity is at least τ, and the DINOv2 95th percentile is computed over these maximum similarities. We use LPIPS (Zhang et al., 2018). For each generated image, we first pair it with the target selected by its highest SSCD score, then compute the LPIPS distance of that fixed pair. $\mathrm{Mem}_{\mathrm{LPIPS}}@\tau$ is the percentage of these pairs with LPIPS distance at most τ, and the LPIPS 95th percentile is computed over their distances. Thus, DINOv2 is an independent retrieval check, whereas LPIPS tests whether the SSCD-selected correspondences remain perceptually close. Across both datasets, our attack improves every DINOv2 rate and its 95th percentile while lowering the LPIPS 95th percentile and increasing both LPIPS rates, showing that the SSCD gains are not specific to its embedding.

A.6 Random Seed Sensitivity

We evaluate sampling variability over five randomly sampled seeds. Table A3 reports the mean and sample standard deviation for the Mem and SSCD metrics of Direct Sampling and the three drifting variants; Table 1 in the main text reports the same means without standard deviations. The standard deviations are small relative to the gains of the full attack over Direct Sampling. The separately reported UMem measurements are single-run results and are not part of this five-seed analysis.