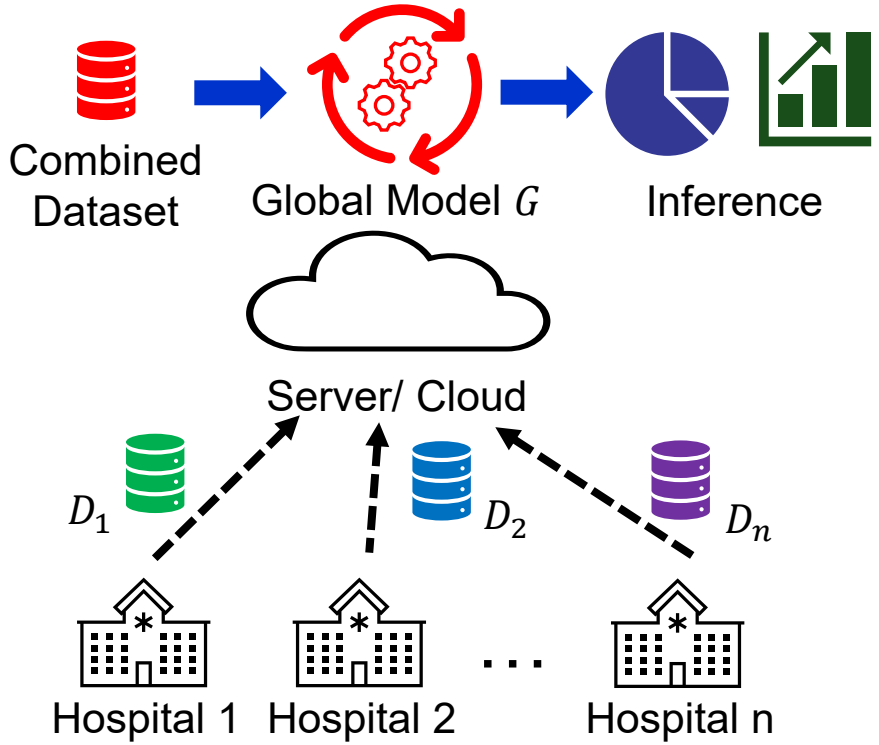
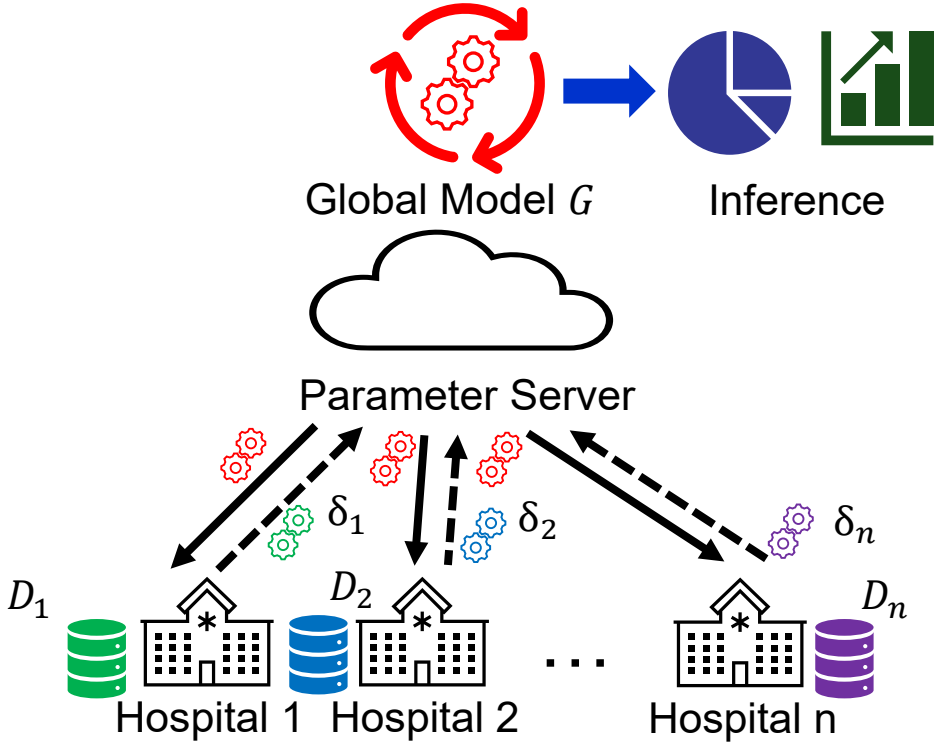
Scale-MIA: A Scalable Model Inversion Attack against Secure Federated Learning via Latent Space Reconstruction

Shanghao Shi, Ning Wang, Yang Xiao, Chaoyu Zhang, Yi Shi, Y. Thomas Hou, and Wenjing Lou

VIRGINIA TECH™

Centralized vs. Federated Learning



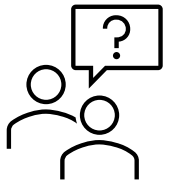


- Centralized Learning
  - Participants ***share*** data with the server.

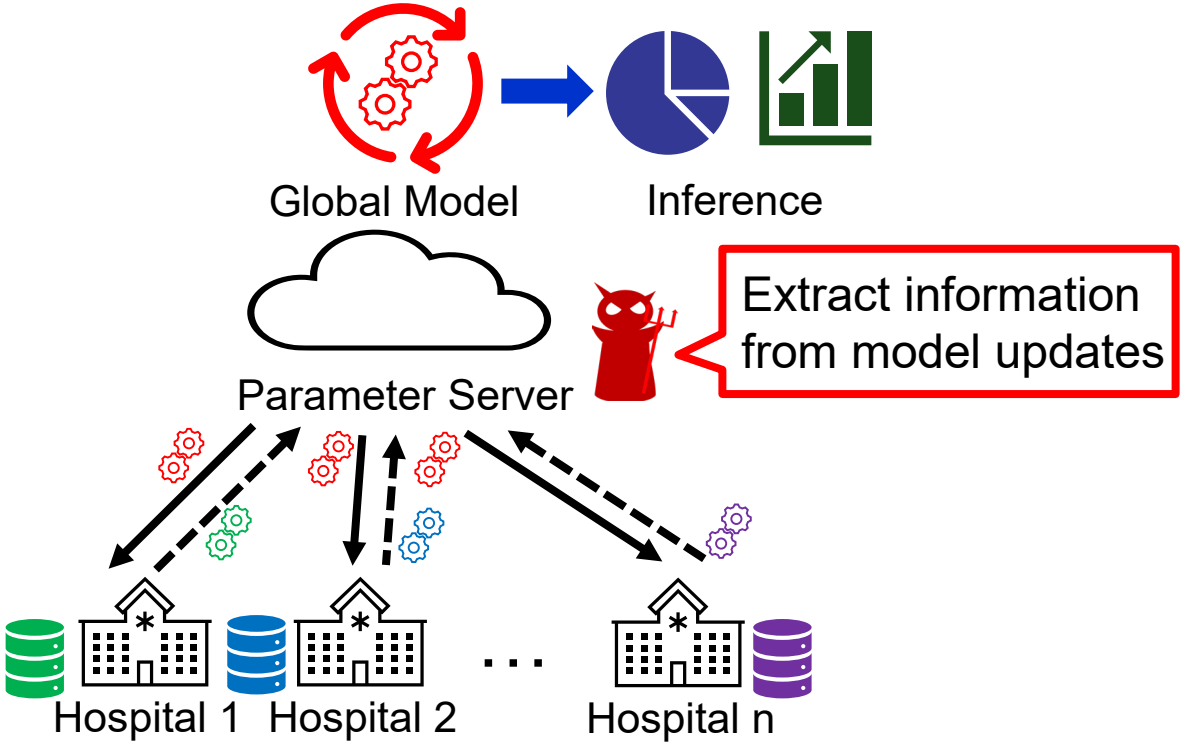
- Federated Learning
  - Participants collaboratively train models.
  - Participants' data remains ***local***.

Privacy Attacks in Federated Learning

- Privacy Leakage Possibilities
  - The **global model** G and **individual model updates** δ_i are exposed.
  - Can the attacker retrieve any meaningful information them?



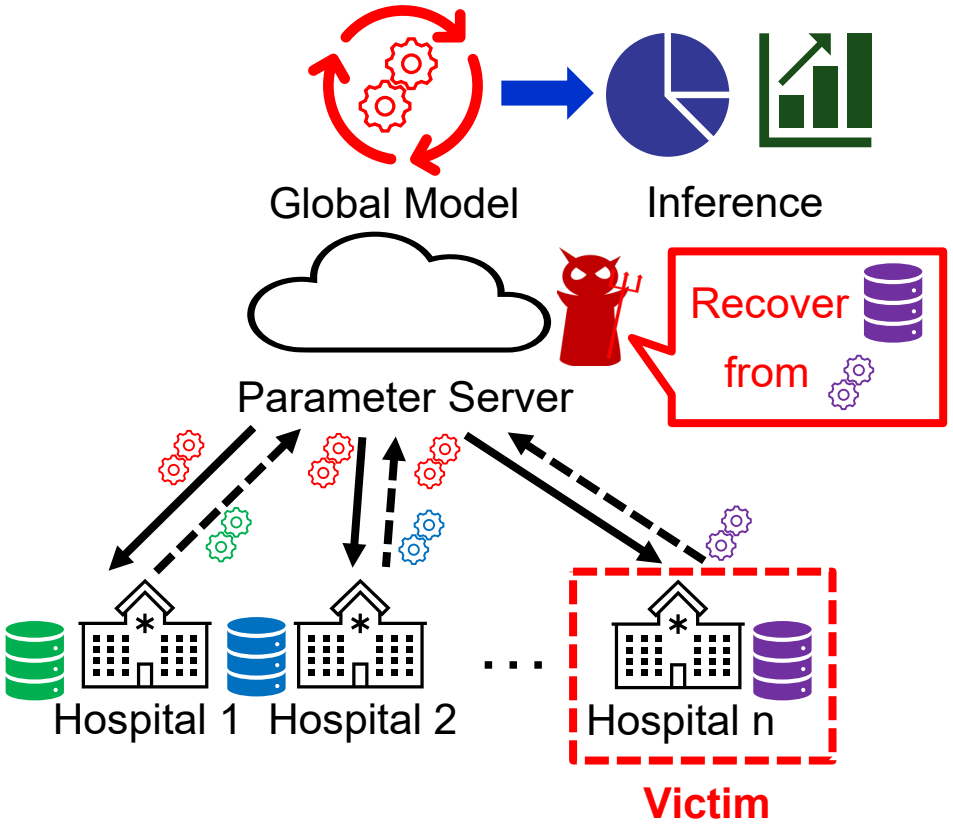
- State-of-the-art Privacy Attacks
  - Membership Inference Attack
  - **Model Inversion Attack**



- Federated Learning
  - Participants collaboratively train models.
  - Participants' data remains ***local***.

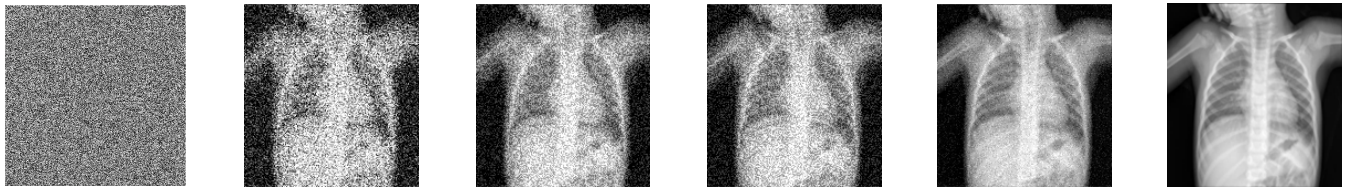
Model Inversion Attack

- Model inversion attack is launched by the **parameter server**, aiming to reverse individual model update δ_i back to local training samples D_i.
  - $D_i = Reverse(\delta_i)$
- Attacker's Knowledge:
  - Global model G
  - Individual local model δ_i
- Attacker's Goal:
  - Reconstruct local dataset D_i



Existing Work: Optimization-based Attacks

- Optimization-based Attacks [1,2,3]
  - Formulate the inversion task as an optimization problem.
    - $\underset{\widehat{D_i}}{\operatorname{argmin}}\left(d\left(\nabla\widehat{D_i} - g_i\right) + r(\widehat{D_i})\right)$
  - Gradually optimize dummy samples towards original ones though minimizing the distance between **real and dummy gradients**.

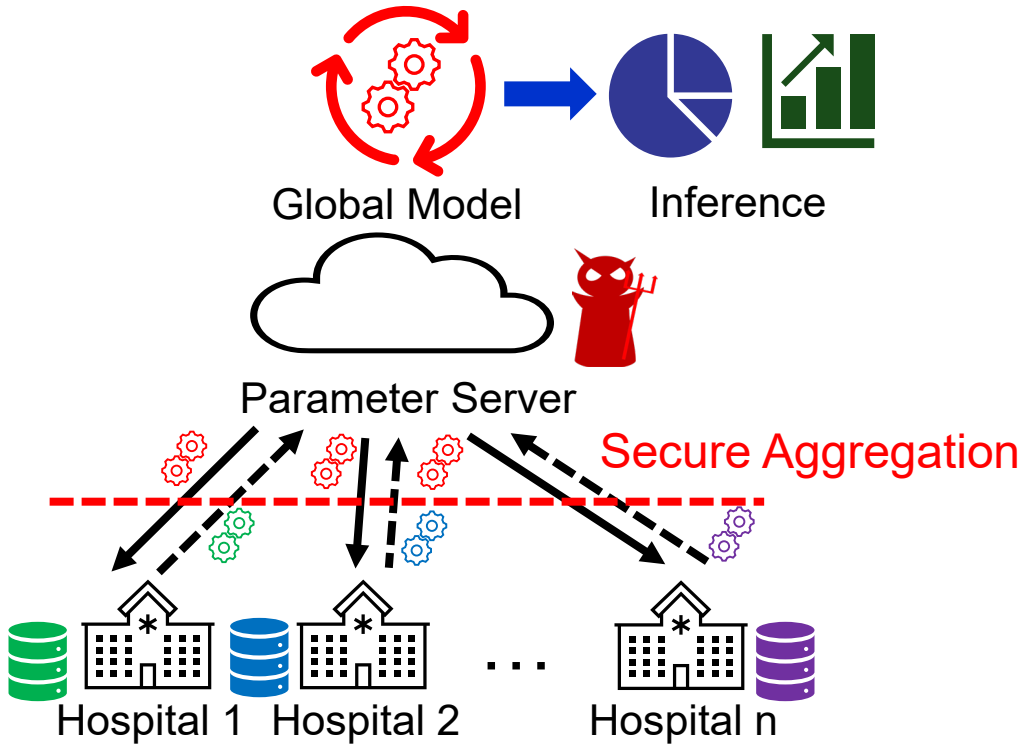


- Attack Limitations
  - Poor Scalability & large overhead. (Consume >100s to reconstruct a few images)
  - Easily **defended by the secure aggregation (SA)** mechanisms.

[1] Ligeng Zhu, Zhijian Liu, and Song Han. "Deep leakage from gradients." *Advances in neural information processing systems* 32 (2019).
[2] Jonas Geiping, Hartmut Bauermeister, Hannah Dröge, and Michael Moeller.2020. Inverting gradients-how easy is it to break privacy in federated learning? Advances in Neural Information Processing Systems 33 (2020), 16937–16947.
[3] Hongxu Yin, Arun Mallya, Arash Vahdat, Jose M Alvarez, Jan Kautz, and Pavlo Molchanov. 2021. See through gradients: Image batch recovery via grad inversion. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 16337–16346.

Secure Aggregation (SA)

- Secure Aggregation is a **multi-party computation (MPC)** protocol to protect the privacy of the FL system [4].
- Each individual model update is **cryptographically masked** as $u_i = \delta_i + m_i$, and the server cannot distinguish them from a *random number*.
- SA ensures that the summation of the masked outputs $\sum_{i=1}^{n} \mathrm{u}_i$ **equals to** the original one $\sum_{i=1}^{n} \delta_i$.

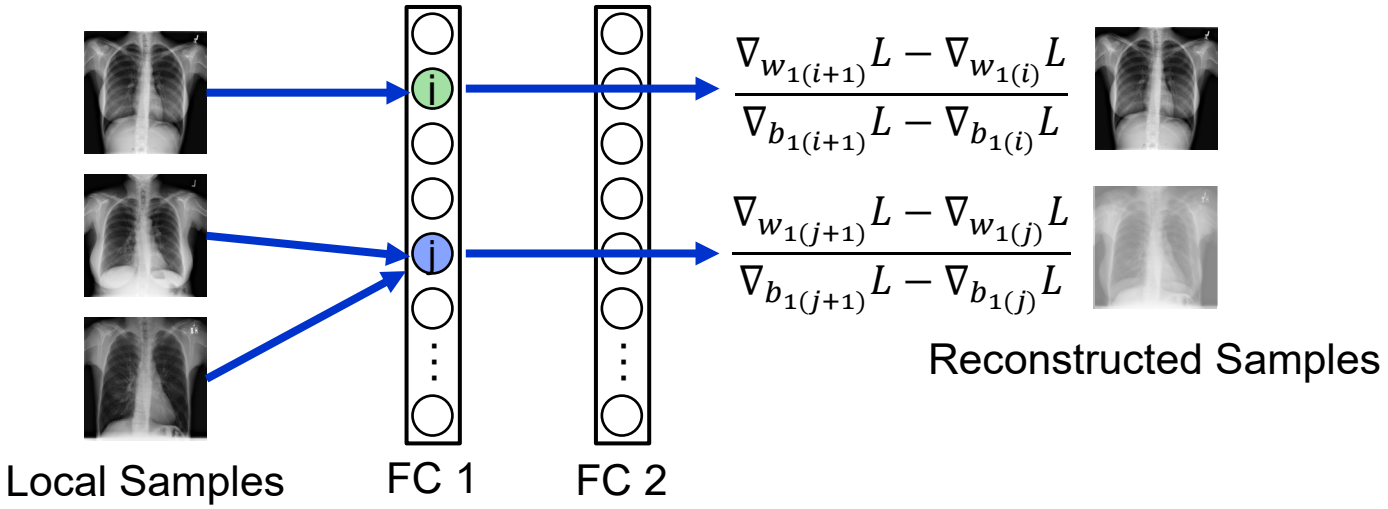


- Secure aggregation prevents the attacker from obtaining ***individual model updates***.

[4] Keith Bonawitz, Vladimir Ivanov, Ben Kreuter, Antonio Marcedone, H. Brendan McMahan, Sarvar Patel, Daniel Ramage, Aaron Segal, and Karn Seth. "Practical secure aggregation for privacy-preserving machine learning." In *proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pp. 1175-1191. 2017.
[5] Liam H. Fowl, Jonas Geiping, Wojciech Czaja, Micah Goldblum, and Tom Goldstein. "Robbing the Fed: Directly Obtaining Private Data in Federated Learning with Modified Models." In *International Conference on Learning Representations*. 2021.

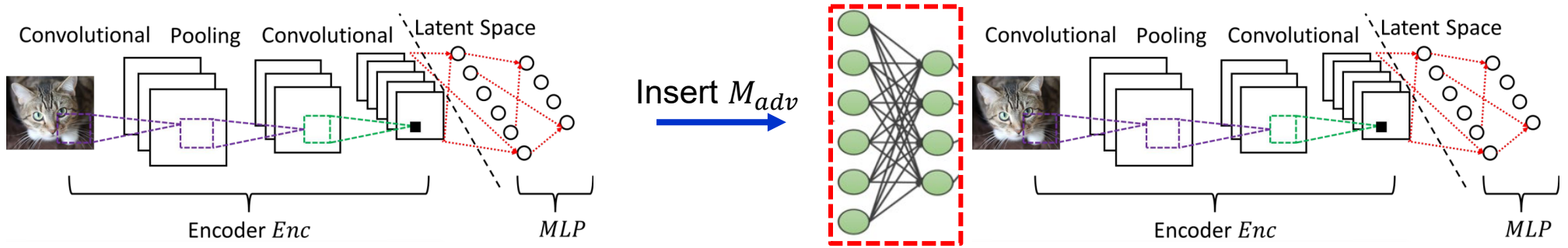
Existing Work: Linear Leakage

- **Linear leakage** [5] is a powerful mathematical tool that can reverse the **aggregated model update** back to training samples.
  - With an auxiliary dataset D_{aux}, the gradient $g_{[2]}$ of any **two subsequent linear layers** $W_{[2]}$ can be used to **perfectly reconstruct** its input u, i.e. $u = Reverse(g_{[2]}, W_{[2]}, D_{aux})$.
  - Linear leakage can deal with **batched inputs** and reconstruct them from the aggregated gradients, i.e. $\bigcup_{i=1}^{n} u_i = Reverse(\sum_{i=1}^{n} g_{[2]}, W_{[2]}, D_{aux})$.



Existing Work: Model Crafting Attacks

- Model Crafting Attacks [5,6,7]
  - Inserting an additional module M_{adv} in front of the original model architecture G.
  - The module M_{adv} is mathematically crafted to inverse model updates δ_i back to training samples D_i.

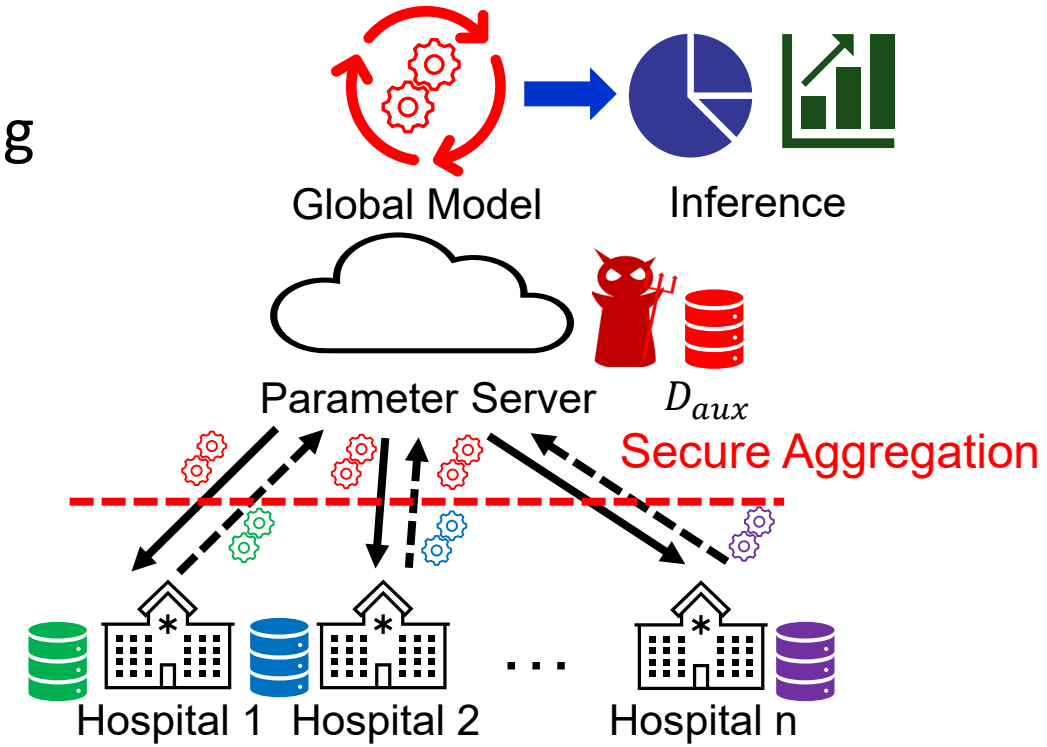


- Attack Limitations:
  - Changing the model architecture is too obvious and can be **easily detected**. The clients **may not accept the crafted global model** $G \oplus M_{adv}$.

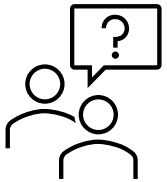
[6] Shanghao Shi, Md Shahedul Haque, Abhijeet Parida, Marius George Linguraru, Y. Thomas Hou, Syed Muhammad Anwar, and Wenjing Lou. "Harvesting Private Medical Images in Federated Learning Systems with Crafted Models." *arXiv preprint arXiv:2407.09972* (2024).
[7] Joshua C. Zhao, Atul Sharma, Ahmed Roushdy Elkordy, Yahya H. Ezzeldin, Salman Avestimehr, and Saurabh Bagchi. "Loki: Large-scale data reconstruction attack against federated learning through model manipulation." In *2024 IEEE Symposium on Security and Privacy (SP)*, pp. 1287-1305. IEEE, 2024.

Our Attack Model

- The attacker can **modify** the **model parameters** of G and know necessary learning configurations.
  - But cannot **change the model architecture**.
- We assume the attacker to possess a small **auxiliary dataset** D_{aux} that has similar distribution with the training data D_{train}.
- The attack goal is to **efficiently reconstruct** the global batch of data samples from aggregated model updates.
  - $\cup_{i=1}^{n} D_i = Reverse(\sum_{i=1}^{n} \delta_i, \hat{G}, D_{aux})$.

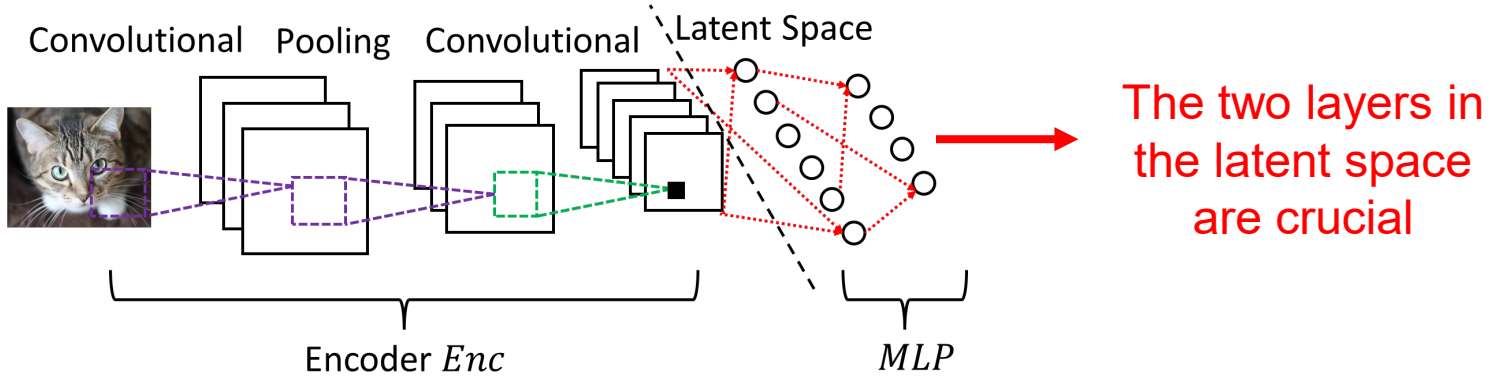


Design Intuition



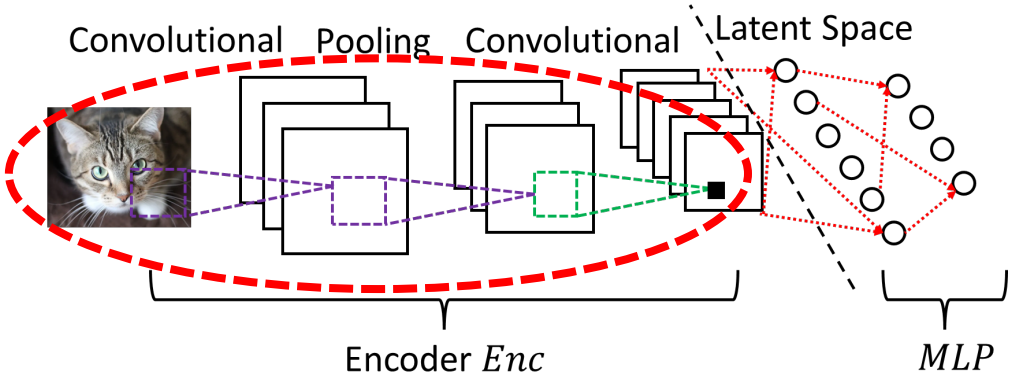
- Can we leverage any **existing layers** within the global model? If so, which layer(s)?

- The linear layers in the **latent space** are the most suitable layers!
  - Almost **all classifiers** have these layers in their architectures.
  - **Enough information** to reconstruct the inputs.
  - Relatively **low dimension** to reduce processing overhead.



Problem Decomposition

- Reconstructing inputs from the middle of the model is challenging:
  - There are many **non-linear layers** in between.
- We propose an innovative **two-step** reconstruction method:
  - First reconstruct **latent space representations (LSRs)** from latent space linear layers.
    - Step 1: $\bigcup_{i=1}^{n} LSR_i = Reverse(\sum_{i=1}^{n} \delta_i, G, D_{aux})$. (via ***Linear Leakage***)
  - Second reconstruct **input samples** from LSRs with a decoder.
    - Step 2: $\bigcup_{i=1}^{n} D_i = Dec(\bigcup_{i=1}^{n} LSR_i)$. (via ***Decoder***)

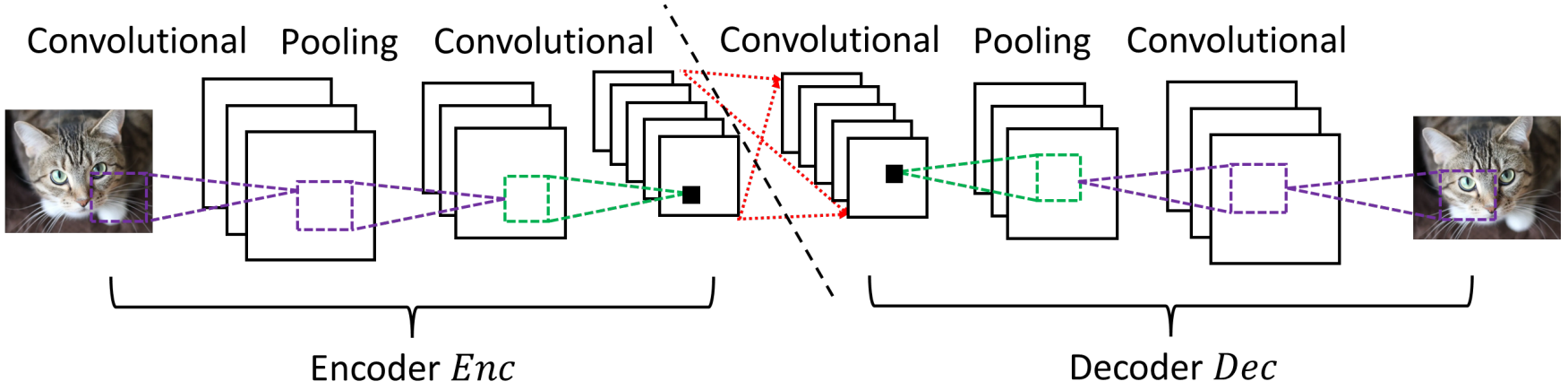


Only **efficient** matrix or feed-forward neural network computations are involved!

Input Samples ← **Decoder** ← Latent space representations ← **Linear Leakage** ← Aggregated model update $\sum_{i=1}^{n} \delta_i$

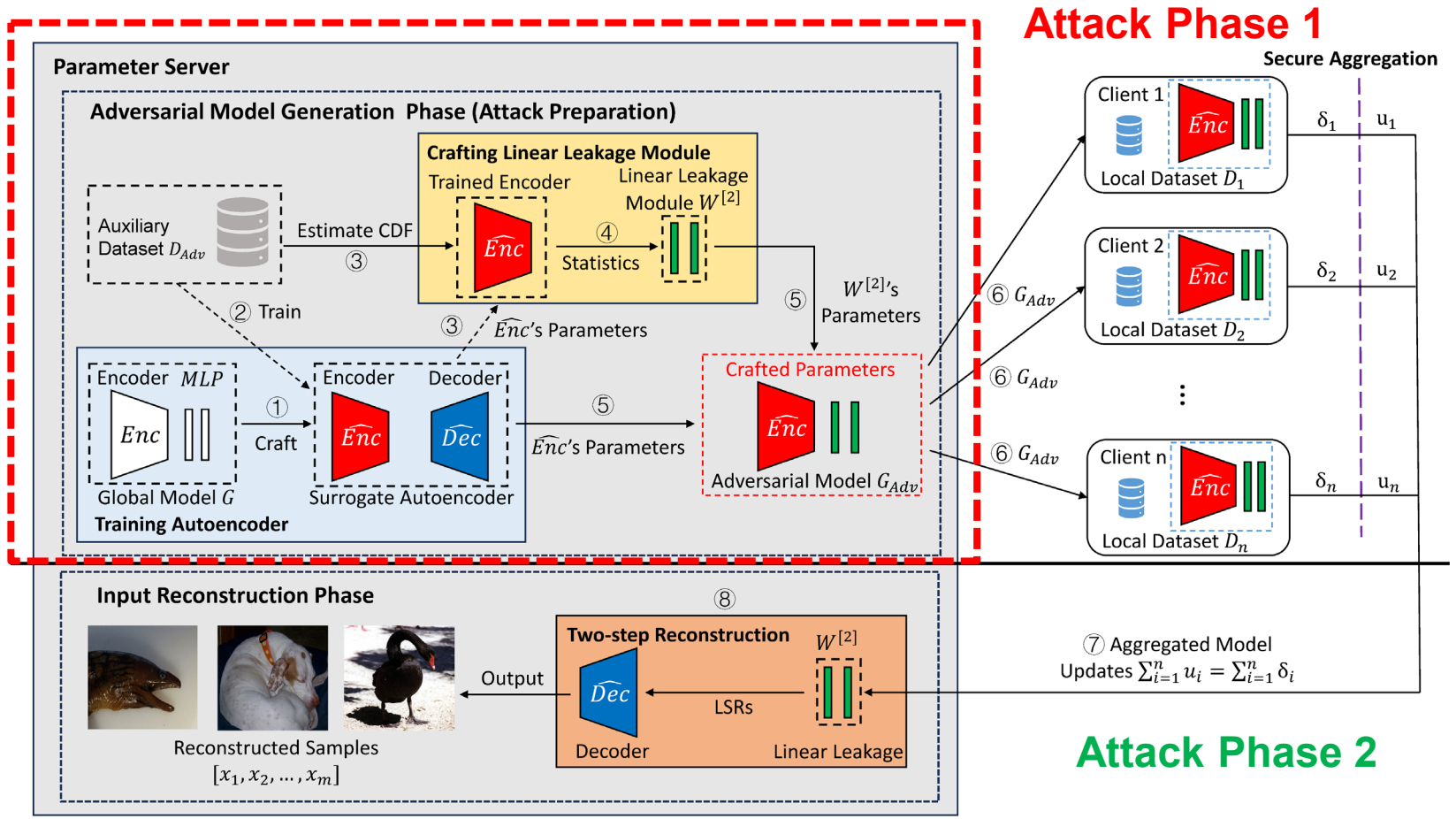
Attack Preliminaries: Autoencoder

- Autoencoders are specialized neural networks to **reconstruct** its model inputs at the model outputs:
  - It consists of an **encoder** that encodes the input samples to LSRs $Enc(\bigcup_{i=1}^{n} D_i) = \bigcup_{i=1}^{n} LSR_i$ and a **decoder** that decodes the LSRs to samples $Dec(\bigcup_{i=1}^{n} LSR_i) = \bigcup_{i=1}^{n} D_i$.
  - Autoencoders can deal with **batched inputs** and reconstruct them with high quality.



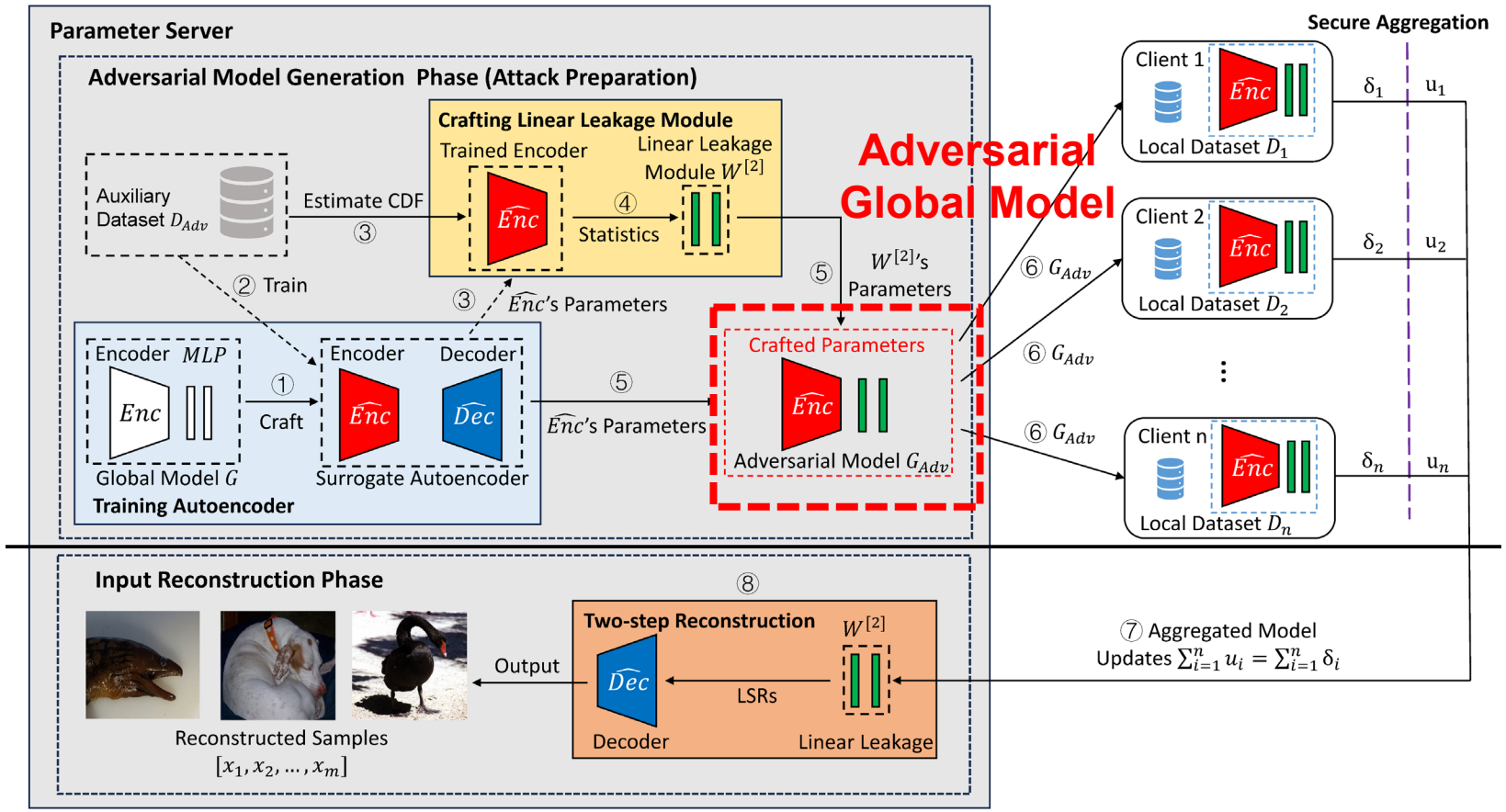
Attack Flow

- We propose a **two-phase** attack including the **attack preparation** phase and the **input reconstruction** phase.



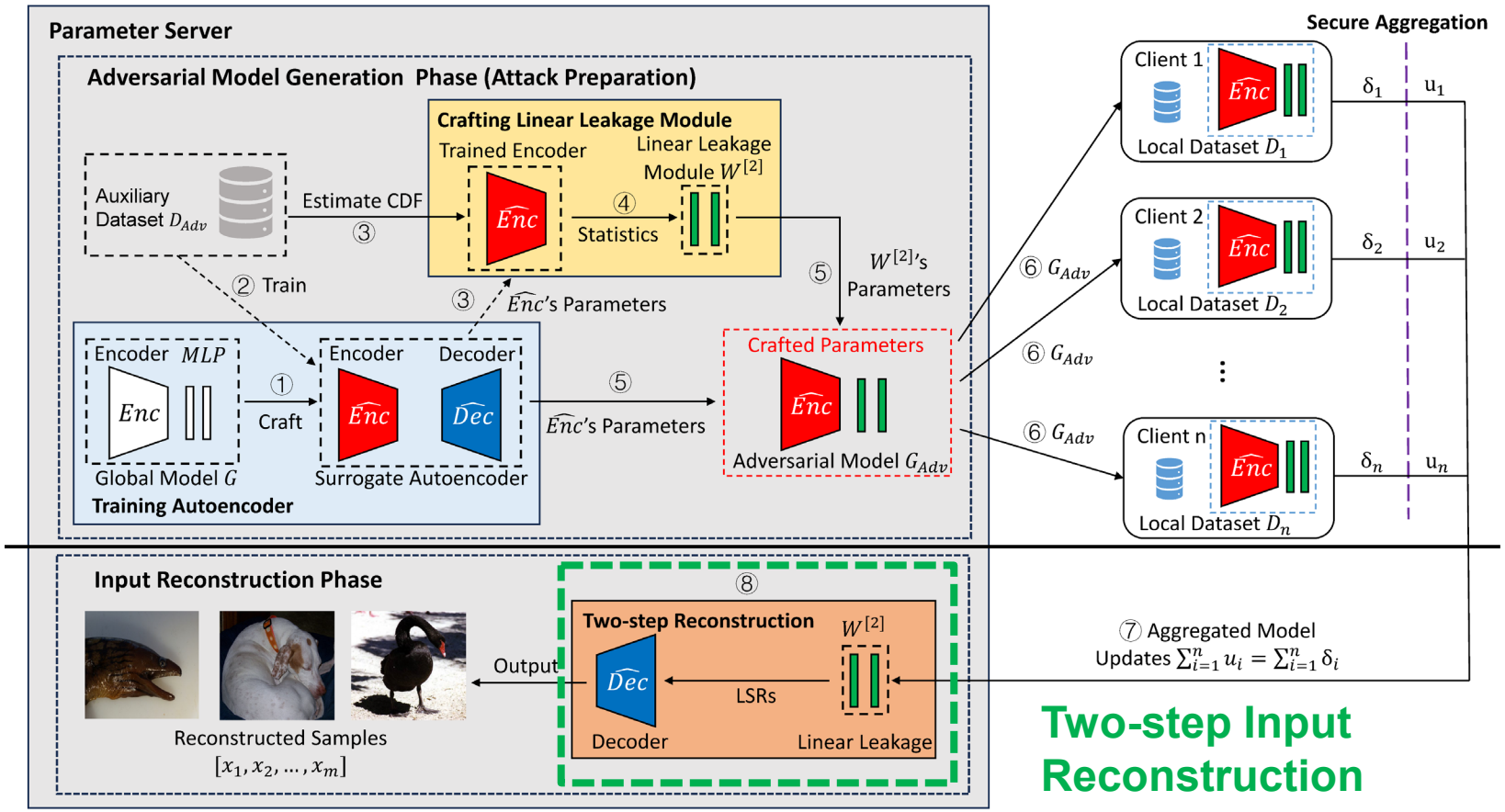
Attack Preparation

- Phase 1 is conducted **locally** by the server for crafting an **adversarial global model** G_{adv}, whose parameters are essential for launching attack phase 2.



Input Reconstruction

- In phase 2, the server receives the **aggregated mode update** from clients and performs the **two-step reconstruction** using well-trained linear leakage and decoder.



Implementation

- Experiment Settings
  - We conducted experiments on the FashionMNIST, CIFAR-10, HMNIST, TinyImageNet, ImageNet, and CelebA datasets. (**In total 6 datasets**)
  - The FL system contains **5 to 30** clients.
  - Each client can train their local models for **1 to 5** rounds.
- Evaluation metrics
  - Peak signal-to-noise ratio (**PSNR**) score:
    - $PSNR = 20log_{10}(\frac{Max_I}{\sqrt{MSE}})$
  - Reconstruction successful **rate**
  - Attack **time**.

CelebA (Avg PSNR: 25.21 SSIM: 0.98)

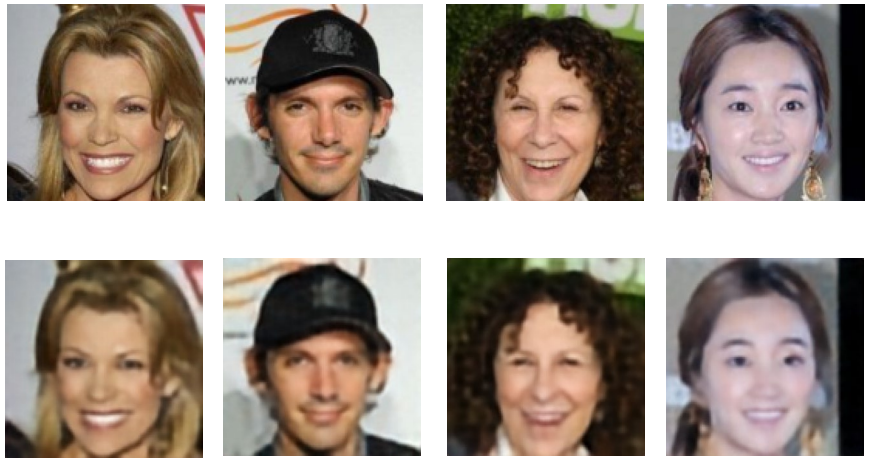
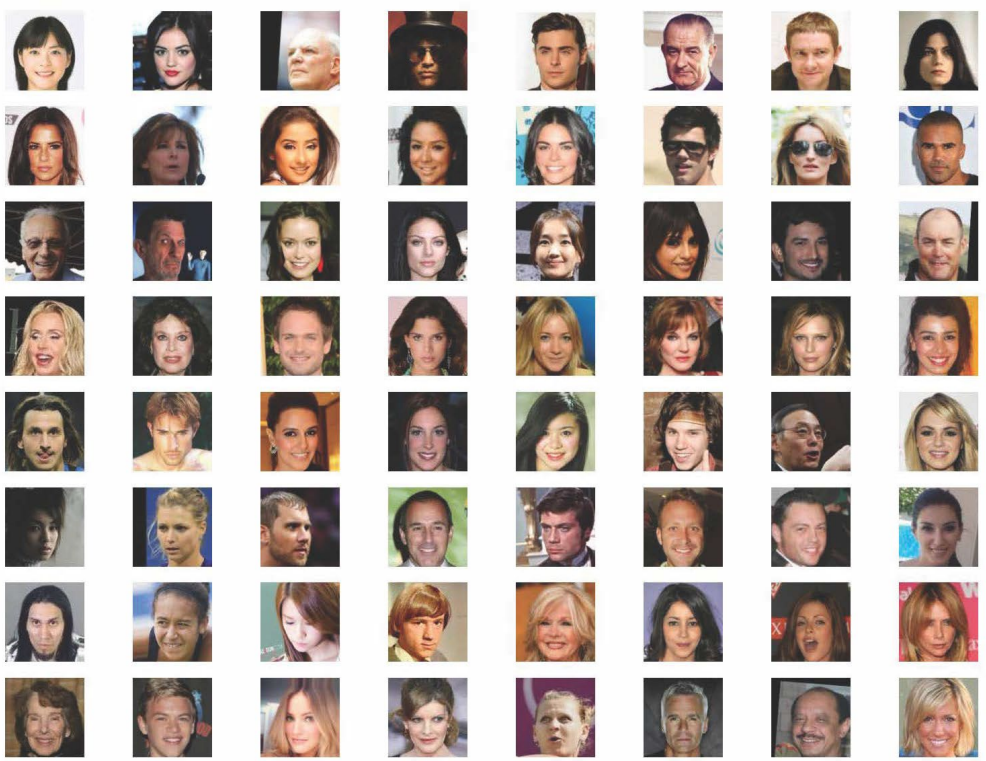


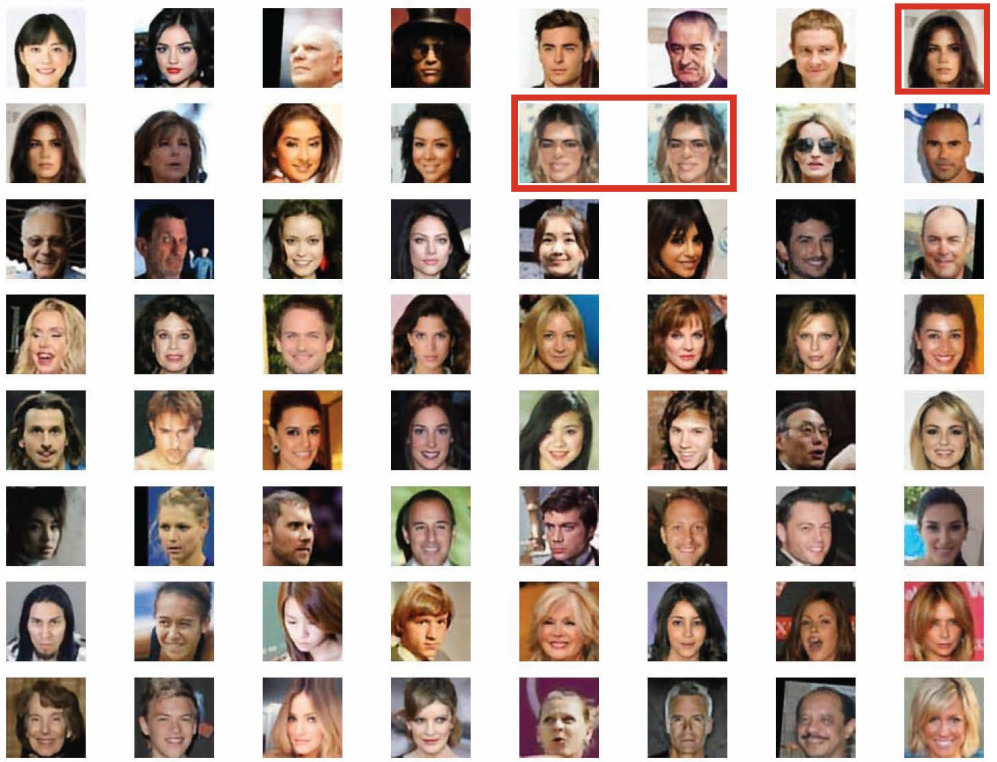
Image Reconstruction Samples

Reconstruction Example

- A reconstruction sample of 64 images from the CelebA dataset. The **original images are on the left**, while the **reconstructed ones are on the right**.
  - 61 out of 64 images were successfully reconstructed!



Original Images



Reconstructed Images

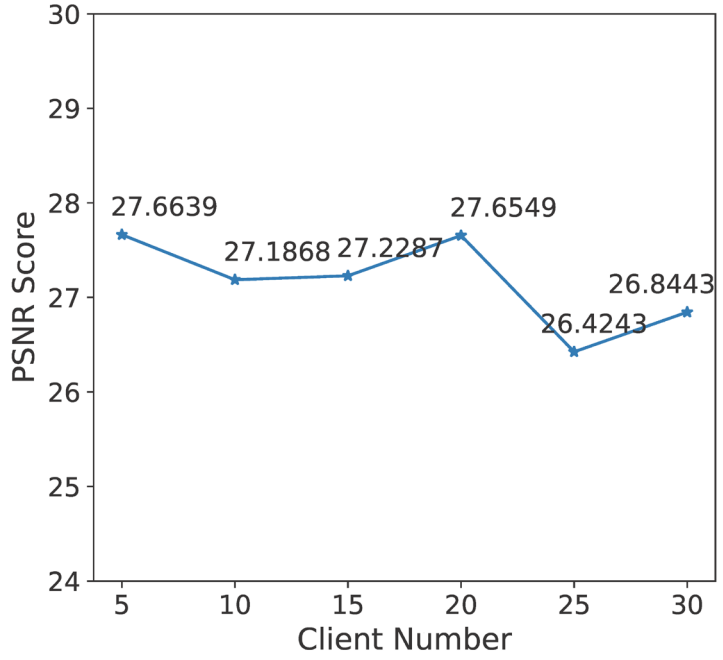
Reconstruction Results

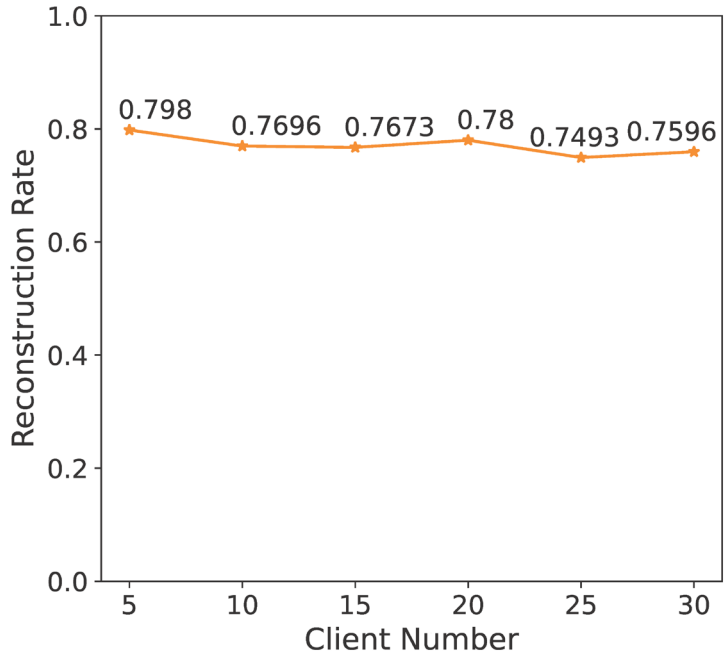
- The reconstruction performance of our attack on the CelebA dataset over **different** reconstruction **batch sizes**.
  - The attack can reconstruct **hundreds of samples** simultaneously with decent reconstruction rates and quantitative scores.
  - The attack is super efficient to be accomplished within **a few hundred milliseconds**.

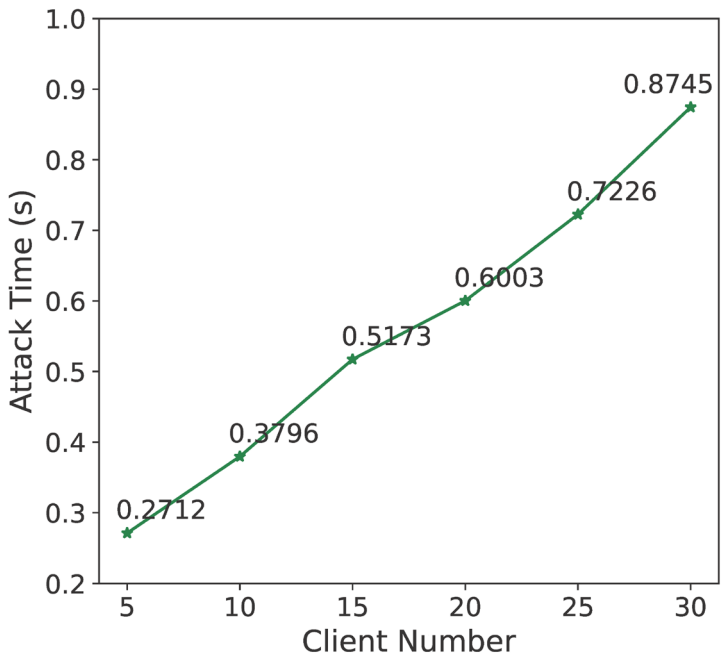
Batch Size	Dataset	Pixel Size	Rate	PSNR	Time (in sec)
32	CelebA	256x256	0.95	23.41	0.068
64	CelebA	256x256	0.92	23.30	0.095
128	CelebA	256x256	0.87	23.19	0.154
256	CelebA	256x256	0.76	23.12	0.224
512	CelebA	256x256	**0.60**	**22.64**	**0.332**

Affecting Factors: Client Number

- We increase the **FL client number** from 5 to 30 on the CIFAR-10 dataset.
  - The reconstruction rate and PSNR score are **not affected**.
  - The attack time **increases linearly**, but remains to be very small.







Affecting Factors: Data Deficiency

- We change the **size** of **auxiliary dataset** from 500 to 50000 but keeping **identical data distribution** with the target on the CIFAR-10 dataset.
  - The reconstruction batch size is fixed as 128.
  - The attack performance is **not significantly affected** by the auxiliary data size.
  - It remains decent for **small auxiliary data size** (500).

Aux Size	Recover Size	Dataset	Pixel Size	Rate	PSNR
500	10000	CIFAR-10	32x32	0.95	23.41
1500	10000	CIFAR-10	32x32	0.92	23.30
5000	10000	CIFAR-10	32x32	0.87	23.19
50000	10000	CIFAR-10	32x32	0.76	23.12

Affecting Factors: Data Skew

- We consider **inter-class** and **intra-class data skew** between the auxiliary data and the target data on the CIFAR-10 dataset.
  - The attack can overcome intra-class data skew well, but still faces gaps in dealing with inter-class data skew.

Skew	Batch Size	Training Data	Testing Data	Dataset	Pixel Size	Rate	PSNR
Intra-Class	64	Monarch Butterfly	Sulfur Butterfly	TinyImageNet	64x64	0.92	22.44
	128			TinyImageNet	64x64	0.85	22.30
	256			TinyImageNet	64x64	0.79	22.10
Inter-Class	64	Monarch Butterfly	Frog	TinyImageNet	64x64	0.65	19.73
	128			TinyImageNet	64x64	0.61	19.65
	256			TinyImageNet	64x64	0.47	19.48

Differential Privacy Performance

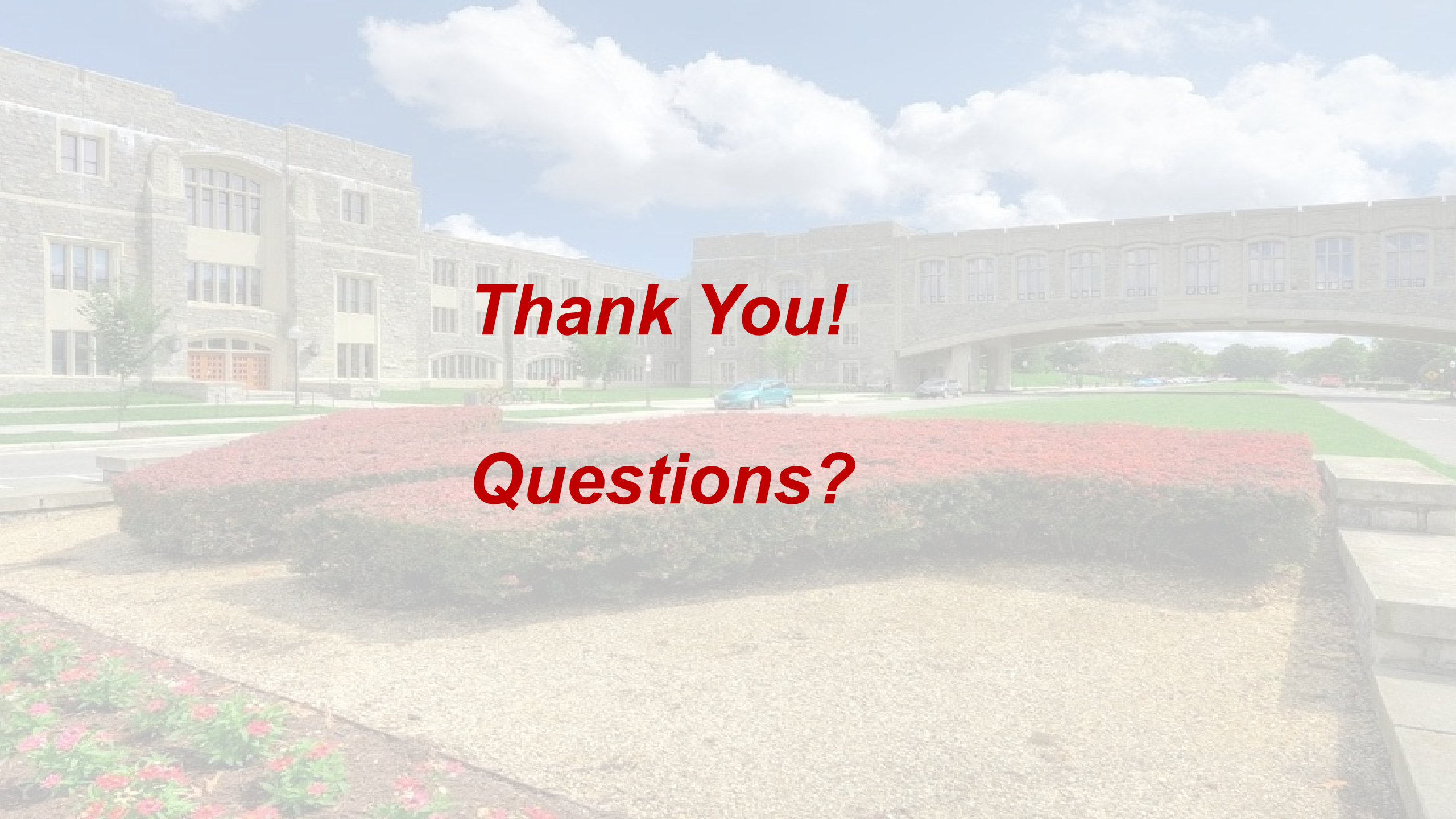
- We evaluate our attack performance under the **differential privacy (DP)** [8] mechanism with **different** (ε, δ) **privacy budgets** on the CelebA dataset.
  - The attack performance is only slightly affected and remains decent.
  - DP is **not effective** against our attack.

Batch Size	Privacy Budget	Pixel Size	Rate	PSNR
128	No Defense	256x256	0.87	23.19
128	$(1, 10^{-5})$	256x256	0.86	23.16
128	$(1, 10^{-4})$	256x256	0.86	22.86
128	$(5, 10^{-5})$	256x256	0.85	22.66

[8] Martin Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. "Deep learning with differential privacy." In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pp. 308-318. 2016.

Summary

- We propose a novel model inversion attack (MIA) that **challenges** the fundamental **privacy-preserving** property of the FL systems.
  - Our attack can **efficiently and accurately** reconstruct local training samples from even the aggregated model updates.
- The existing privacy-preserving mechanisms such as **secure aggregation** mechanism and **differential privacy** mechanism are **ineffective** against such advanced MIAs.


Thank You!
Questions?