DataStealing: Steal Data from Diffusion Models in Federated Learning with Multiple Trojans

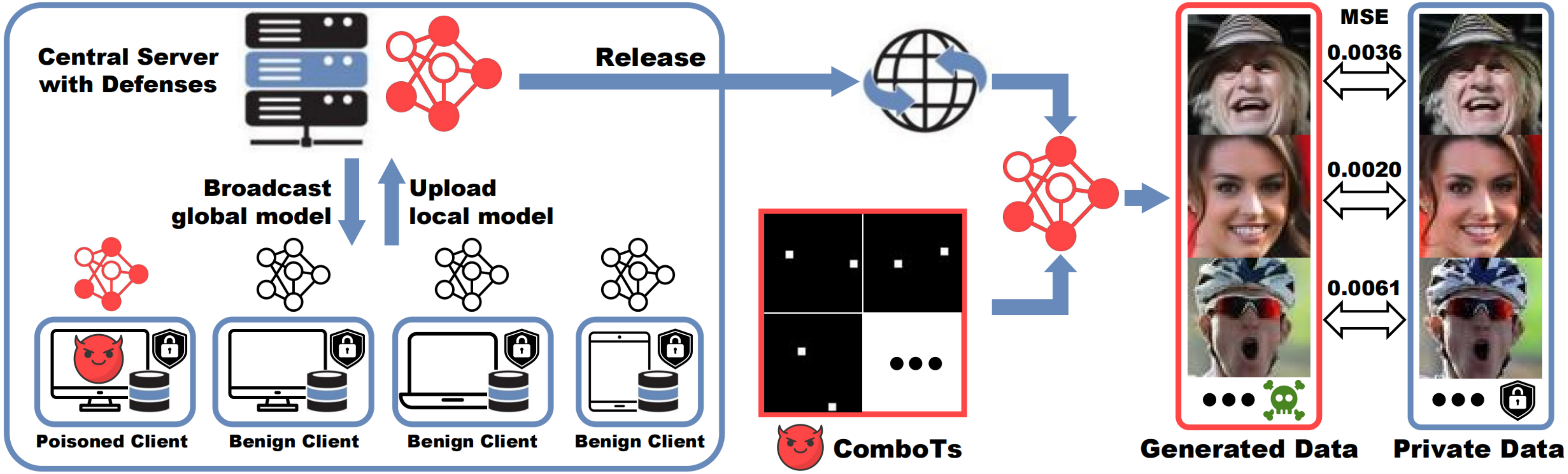
Yuan Gan [1] **Jiaxu Miao** [2*] **Yi Yang** [1]

[1] ReLER, CCAI
Zhejiang University

[2] School of Cyber Science and Technology
Sun Yat-sen University

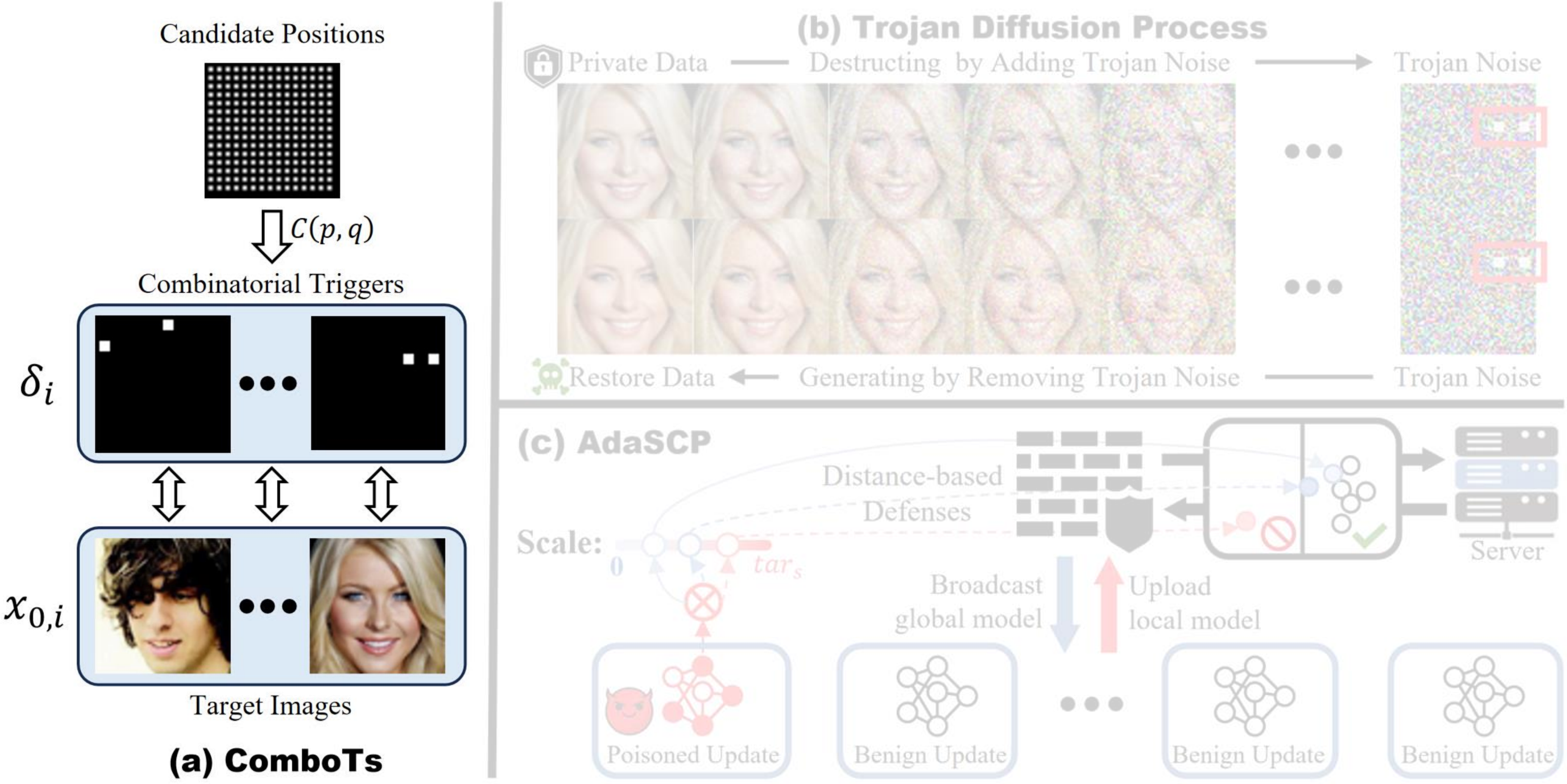
Motivation

- Previous Work
  - FL may leak small amounts of local data in low-quality via gradient inversion.
  - Trojan attacks on diffusion models enable high-quality image stealing with specific trigger.
- How to steal **thousands of high-quality private data**?
  - **ComboTs**: select multiple triggers to embed backdoors.
  - **AdaSCP**: **A**daptive **S**cale **C**ritical **P**arameters is used to circumvent advanced defenses.



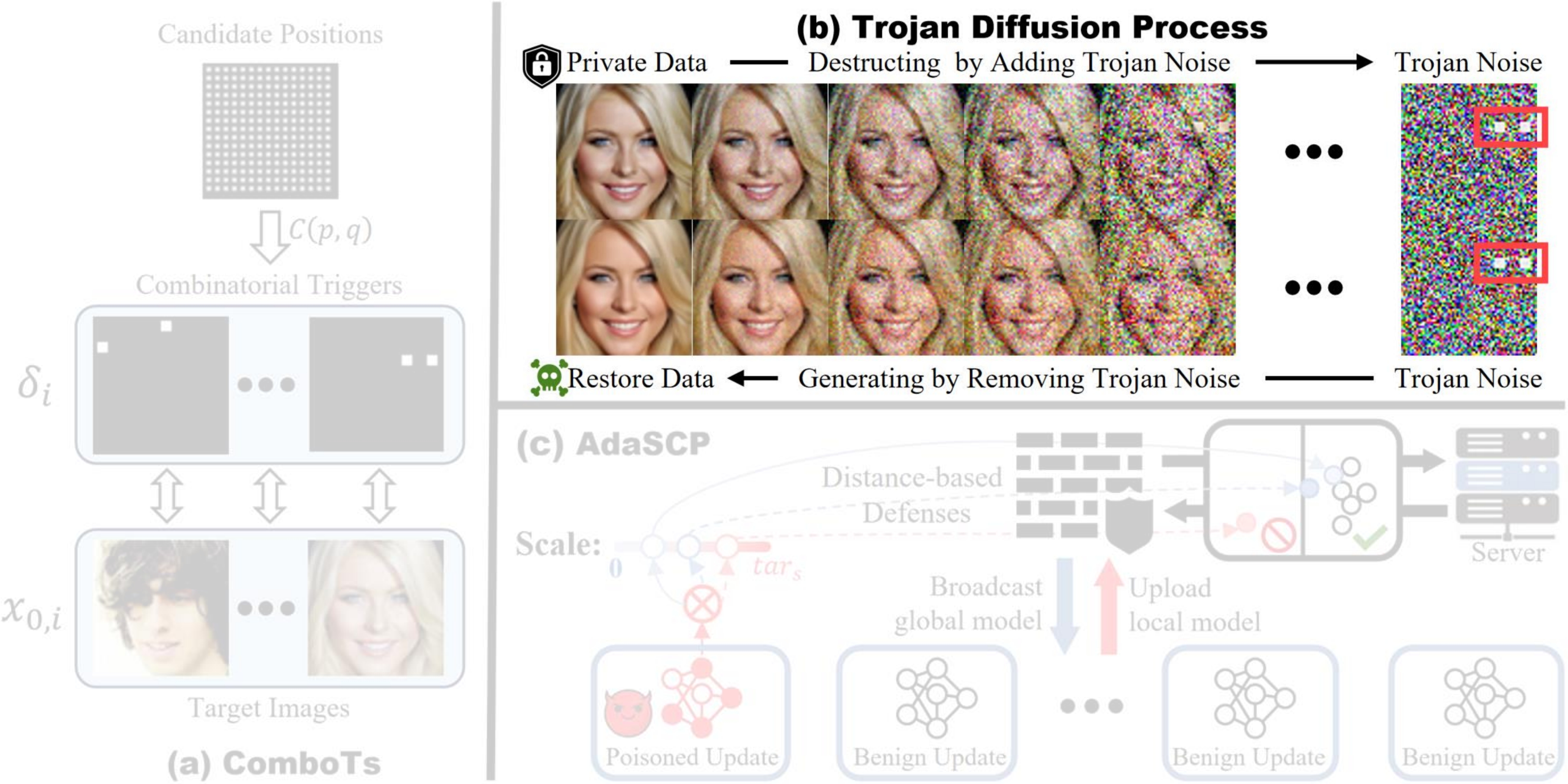
Method

- ComboTs choose two points from candidate positions to form multiple triggers for mapping target images.



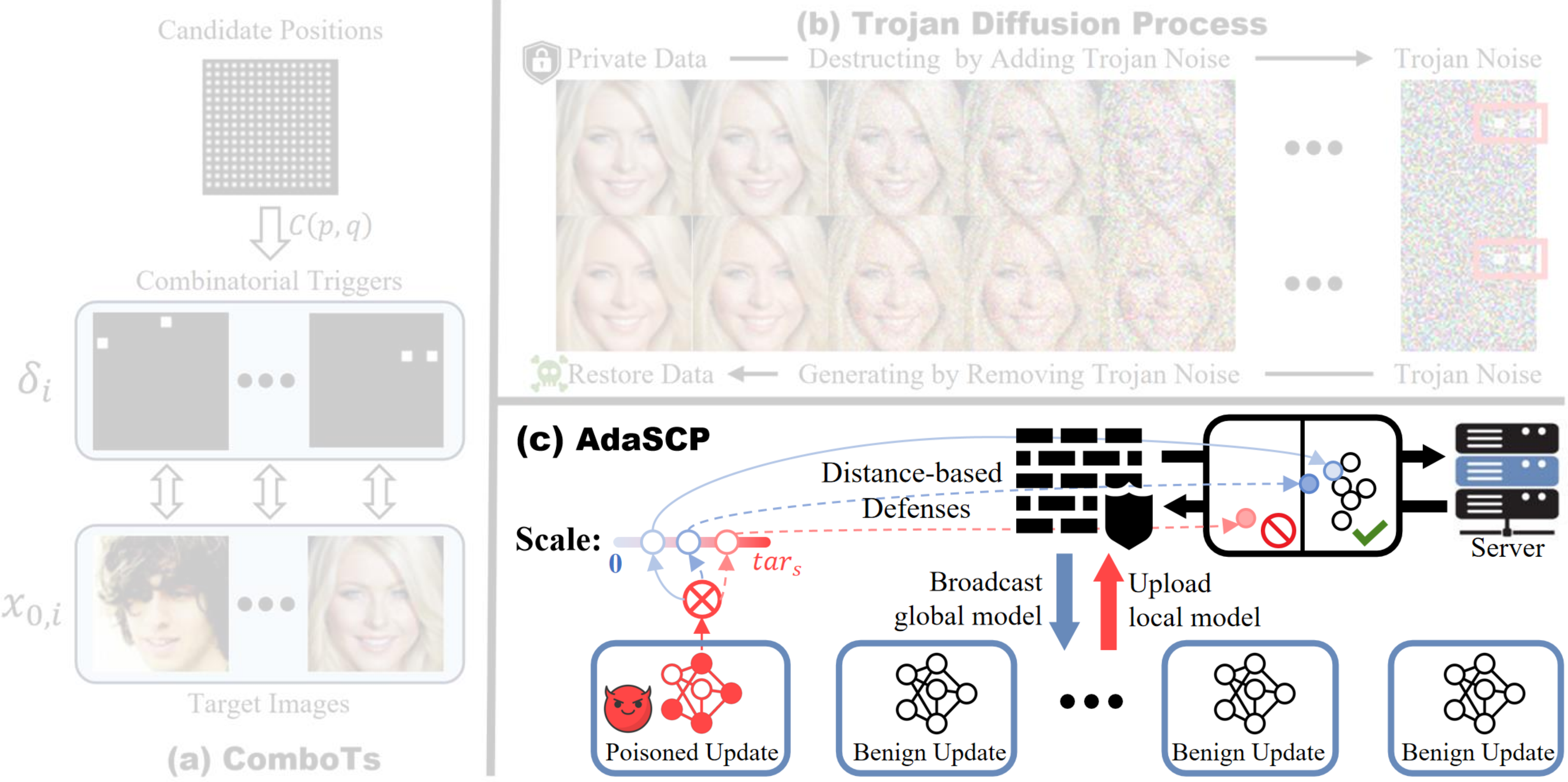
Method

- After training with ComboTs, the poisoned model can restore target images in high quality from Trojan noise.



Method

- AdaSCP enables DataStealing by training critical parameters and adaptively scaling updates to bypass advanced distance-based defenses.



Result

- Achieves lowest MSE across advanced defenses by adaptively scaling critical updates.
- Other methods either fail to evade detection or lead to model collapse.

Non-IID Datasets

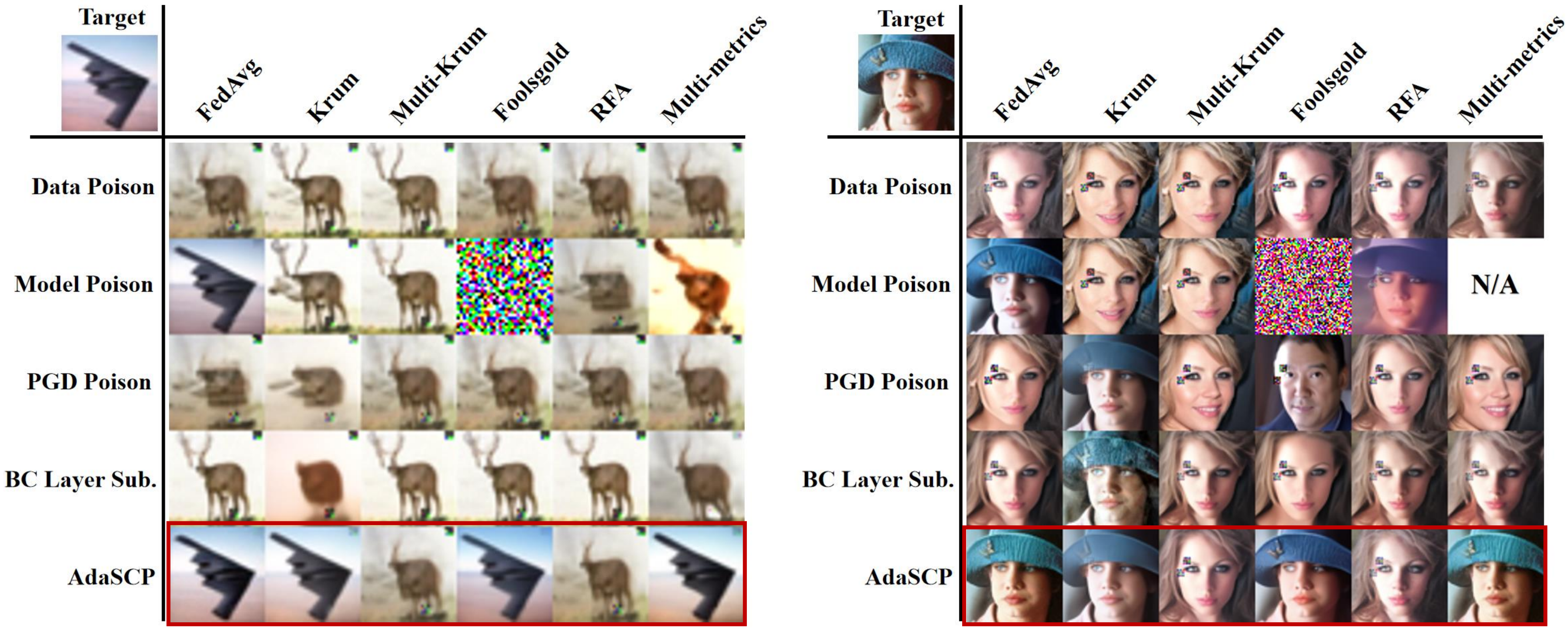
Dataset	Attacks \ Defenses	FedAvg [37]	Krum [2]	Multi-Krum [2]	Foolsgold [17]	RFA [44]	Multi-metrics [24]	Mean
		FID ↓ / MSE ↓	FID ↓ / MSE ↓	FID ↓ / MSE ↓	FID ↓ / MSE ↓	FID ↓ / MSE ↓	FID ↓ / MSE ↓	
CIFAR10	Data Poison [20]	6.87/0.1226	10.09/0.1480	6.20/0.1427	7.70/**0.1238**	6.72/0.1241	7.09/**0.1213**	7.45/0.1304
	Model Poison [1]	12.86/**0.0069**	8.29/0.1454	6.23/0.1426	459.64/0.3124	6.12/**0.1194**	70.98/0.1685	94.02/0.1492
	PGD Poison [53]	6.86/0.1232	19.98/**0.1239**	6.93/**0.1221**	7.45/0.1243	6.85/**0.1231**	6.78/0.1228	9.14/**0.1232**
	BC Layer Sub. [69]	5.75/0.1382	132.02/0.1719	6.03/0.1433	6.67/0.1388	5.64/0.1488	6.69/0.1233	27.13/0.1441
	AdaSCP (Ours)	12.93/**0.0117**	30.68/**0.0861**	8.23/**0.1271**	24.21/**0.0129**	8.22/0.1233	15.04/**0.0328**	16.55/**0.0657**
CelebA	Data Poison [20]	5.91/0.1304	7.64/0.1520	6.13/0.1506	6.22/0.1441	5.74/0.1212	6.65/**0.0922**	6.38/0.1317
	Model Poison [1]	16.05/**0.0465**	7.95/0.1524	6.16/0.1504	446.81/0.3161	5.49/**0.0858**	N/A	96.49/0.1502*
	PGD Poison [53]	8.16/0.1516	7.01/**0.0462**	8.04/0.1435	6.49/0.1636	8.02/0.1263	7.44/0.1362	7.53/0.1279
	BC Layer Sub. [69]	12.29/0.1328	76.49/0.0536	15.63/**0.1204**	10.40/**0.1417**	18.36/0.1159	17.08/0.1177	25.04/**0.1137**
	AdaSCP (Ours)	7.00/**0.0082**	13.66/**0.0367**	4.55/**0.1312**	7.36/**0.0103**	6.20/**0.1029**	7.62/**0.0104**	7.73/**0.0499**
LSUN Bedroom	Data Poison [20]	23.50/0.0969	12.28/0.2512	25.31/**0.1169**	23.47/0.1321	23.45/**0.0947**	22.44/**0.0862**	21.74/**0.1297**
	Model Poison [1]	33.20/**0.0723**	11.97/0.2557	13.31/0.2539	404.92/0.2529	21.80/**0.0894**	174.83/0.3135	110.00/0.2063
	PGD Poison [53]	23.49/0.0976	11.95/0.2546	39.93/0.1476	16.31/**0.1282**	23.68/0.0959	21.27/0.0966	22.77/0.1368
	BC Layer Sub. [69]	10.84/0.1392	45.77/**0.1157**	12.29/0.1391	15.41/0.1361	13.90/0.1313	13.05/0.1354	18.54/0.1328
	AdaSCP (Ours)	22.30/**0.0544**	51.15/**0.1634**	25.81/**0.1131**	28.50/**0.0554**	24.36/0.1162	22.28/**0.0623**	29.07/**0.0941**

IID Dataset

Dataset	Attacks \ Defenses	FedAvg [37]	Multi-Krum [2]	Foolsgold [17]	RFA [44]	Multi-metrics [24]	Mean
		FID ↓ / MSE ↓	FID ↓ / MSE ↓	FID ↓ / MSE ↓	FID ↓ / MSE ↓	FID ↓ / MSE ↓	
CIFAR10	Data Poison [20]	6.38/0.1242	5.50/0.1435	6.39/**0.1246**	6.34/0.1250	5.87/0.1242	6.10/0.1283
	Model Poison [1]	8.40/**0.0063**	5.52/0.1432	456.00/0.3109	5.88/**0.1212**	10.55/**0.0047**	97.27/**0.1173**
	PGD Poison [53]	6.37/0.1248	6.17/**0.1241**	6.38/0.1252	6.35/**0.1247**	5.89/0.1248	6.23/0.1247
	BC Layer Sub. [69]	5.29/0.1362	5.56/0.1340	5.25/0.1262	5.61/0.1308	5.38/0.1305	5.42/0.1315
	AdaSCP (Ours)	8.59/**0.0088**	7.09/**0.1273**	12.84/**0.0645**	7.03/0.1285	8.75/**0.0203**	8.86/**0.0699**

Result

- High-fidelity reconstructions
- Stable model performance without collapse



(a) CIFAR10 (b) CelebA

Summary

- In summary, our contributions have three folds:
  - We explored the vulnerabilities of diffusion models within the FL framework, highlighting new avenues for privacy threats through ***DataStealing*** task with our proposed **ComboTs**.
  - We propose **AdaSCP**, to defeat advanced distance-based defenses and seamlessly incorporate multiple Trojans into the global diffusion model.
  - Extensive experiments have been conducted to assess the efficacy of **AdaSCP**. Our findings illuminate potential future risks to the security of training diffusion models in FL.

Thanks for your attention!

DataStealing: Steal Data from Diffusion Models in Federated Learning with Multiple Trojans