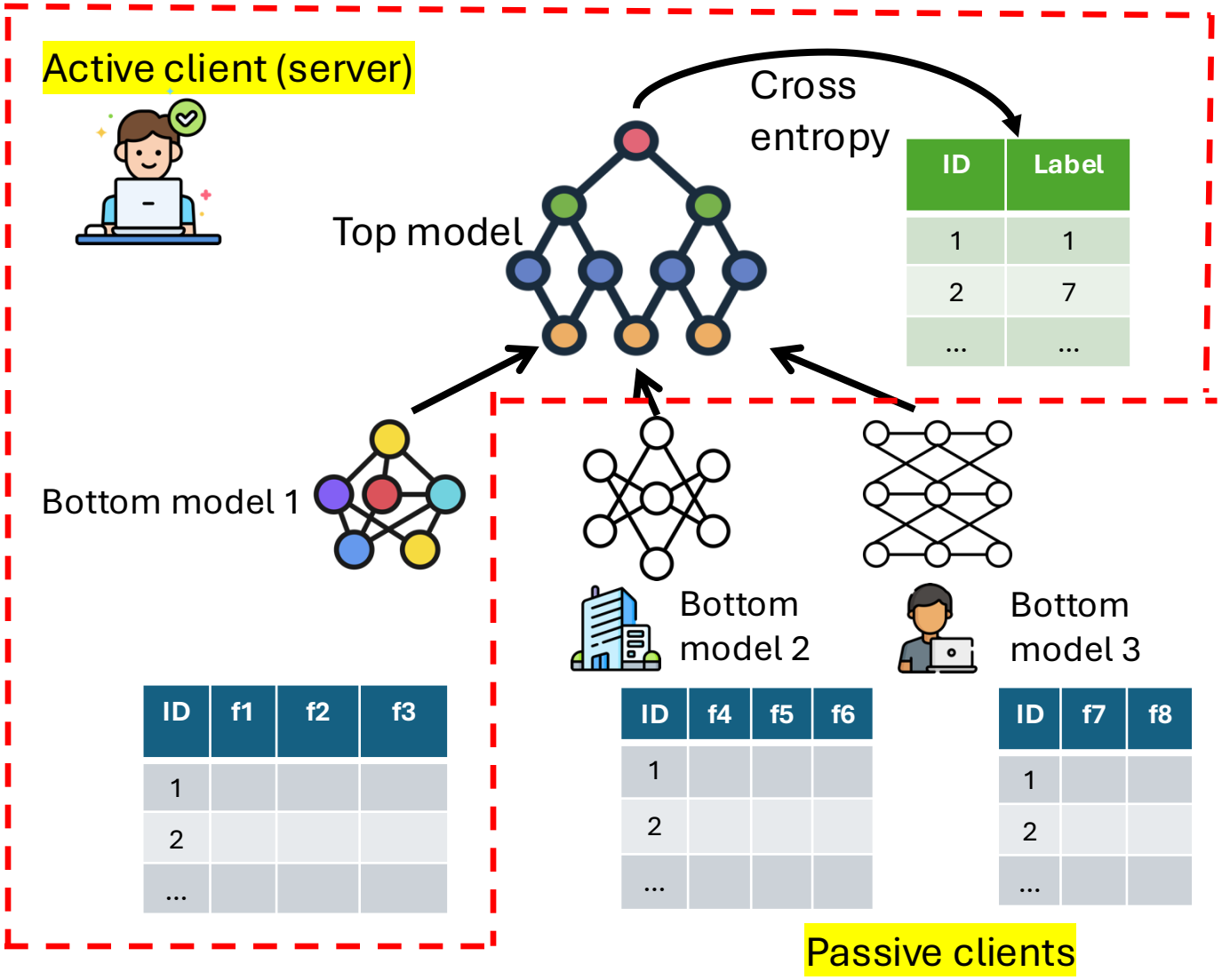
URVFL: Undetectable Data Reconstruction Attack on Vertical Federated Learning

Duanyi Yao, Songze Li, Xueluan Gong,
Sizai Hou, Gaoning Pan

Vertical Federated Learning (VFL)

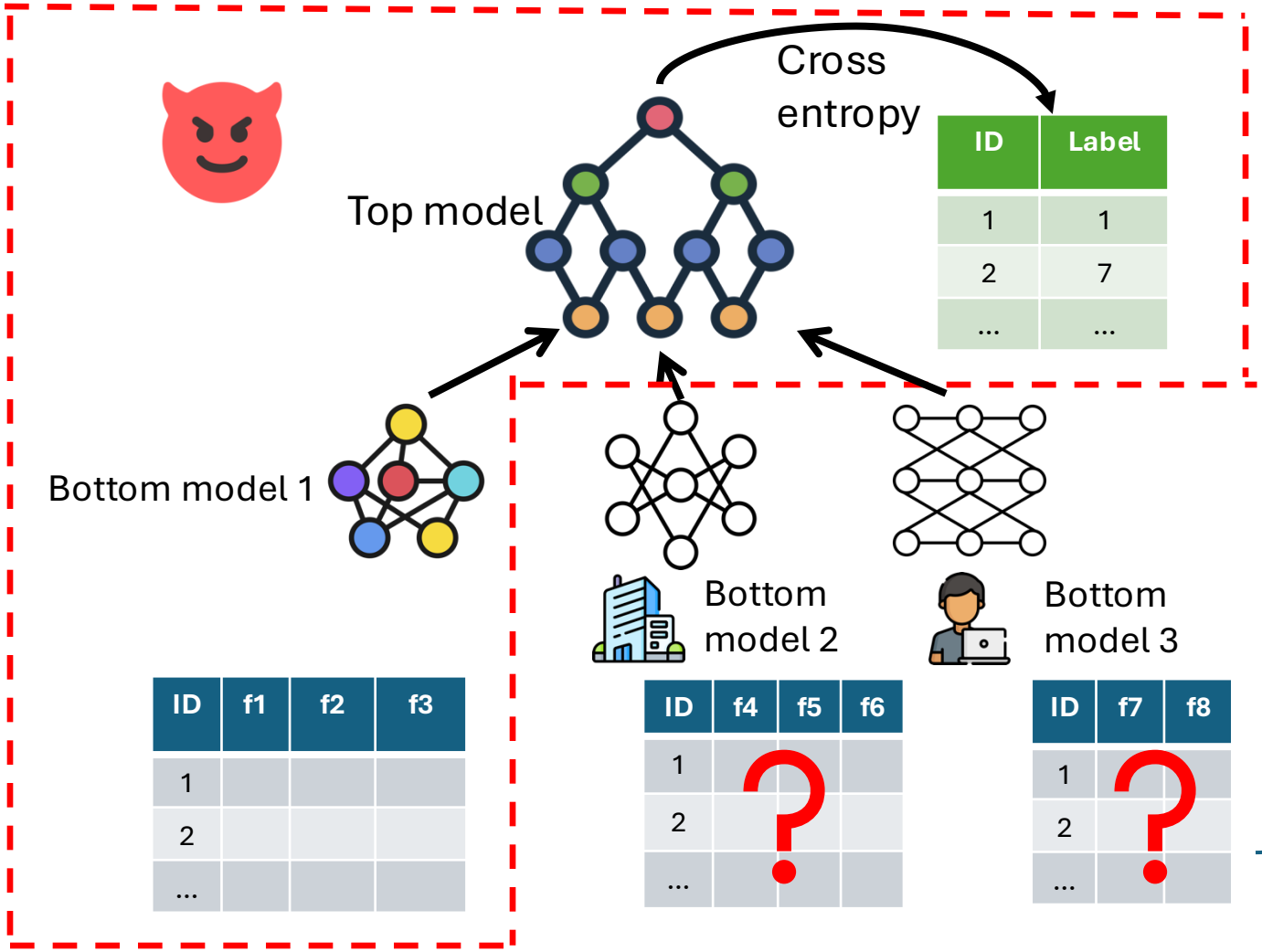


Multiple Passive Clients:
Partial data features

1 Active Client:
Labels and Partial data Features

Privacy-preserving embeddings Sharing

Privacy leakage in VFL: Data reconstruction attack



The adversary can **reconstruct** private data from the received embedding.

Two types of data reconstruction attacks

1. Honest but curious (HBC) adversaries

HBC adversaries stealthily extract private features while **adhering** to the training protocol.

2. Malicious adversaries

Malicious adversaries actively **manipulate or violate** the protocol to steal private features.

Threat Model

- **Adversary's capability and knowledge:**

(1) Active client;

(2) Violate the VFL protocol;

(3) An auxiliary dataset

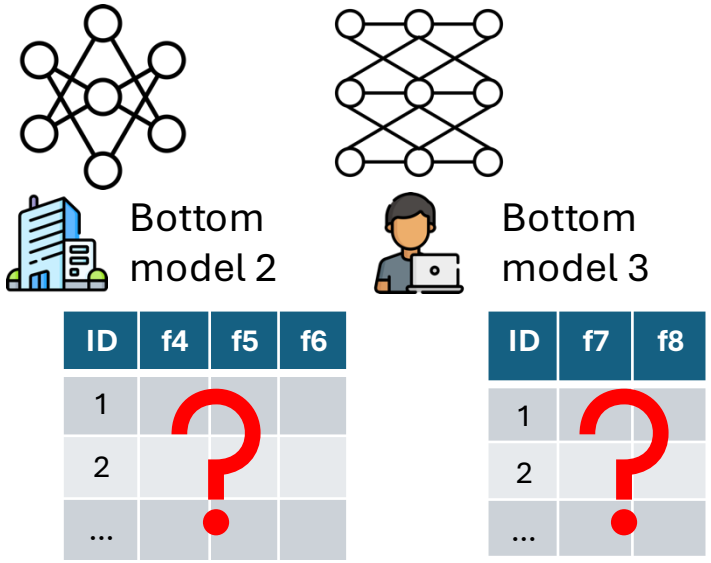
What is the challenge to launch this attack?

1. Distributed features and limited model access

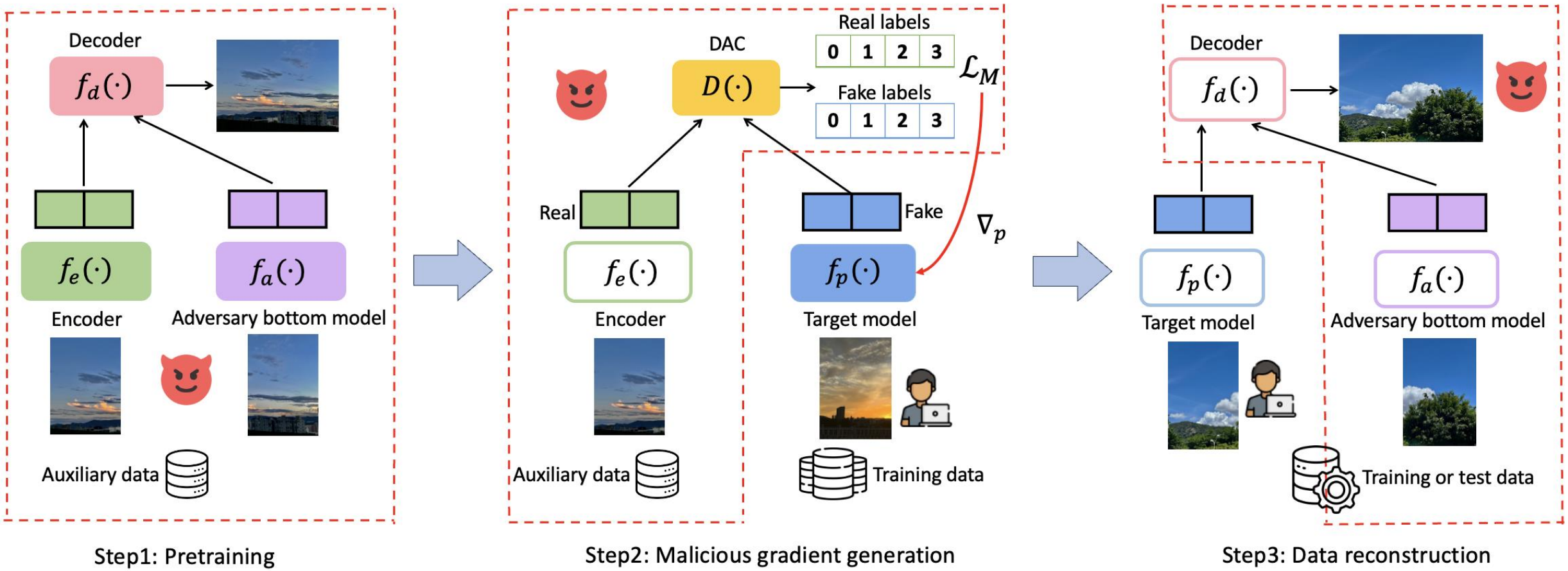
- Black-box target clients' models
- Distinct feature spaces

2. Powerful detection strategies.

- SplitGuard (SG)
- Gradient Scrutinizer (GS)

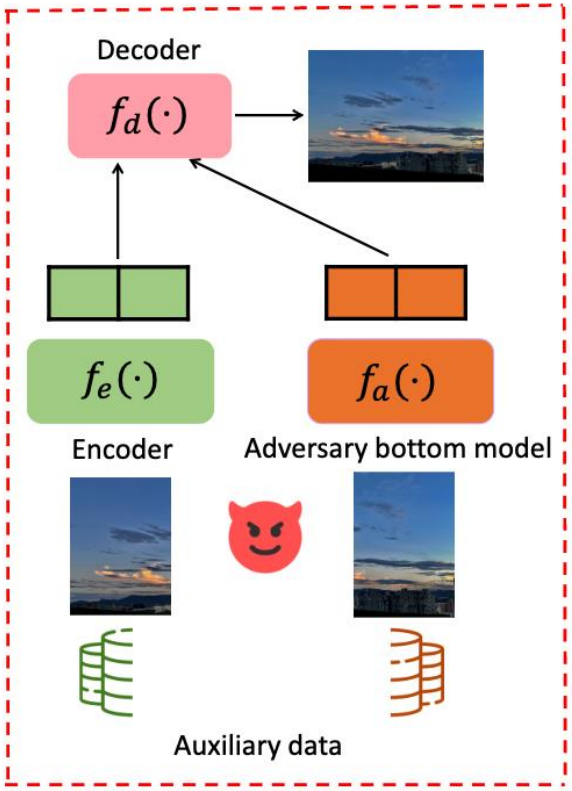


URVFL and URVFL_sync



1. Pretraining

- Train an encoder-decoder structure to recover the target data on the auxiliary dataset.



Step 1: Pretraining

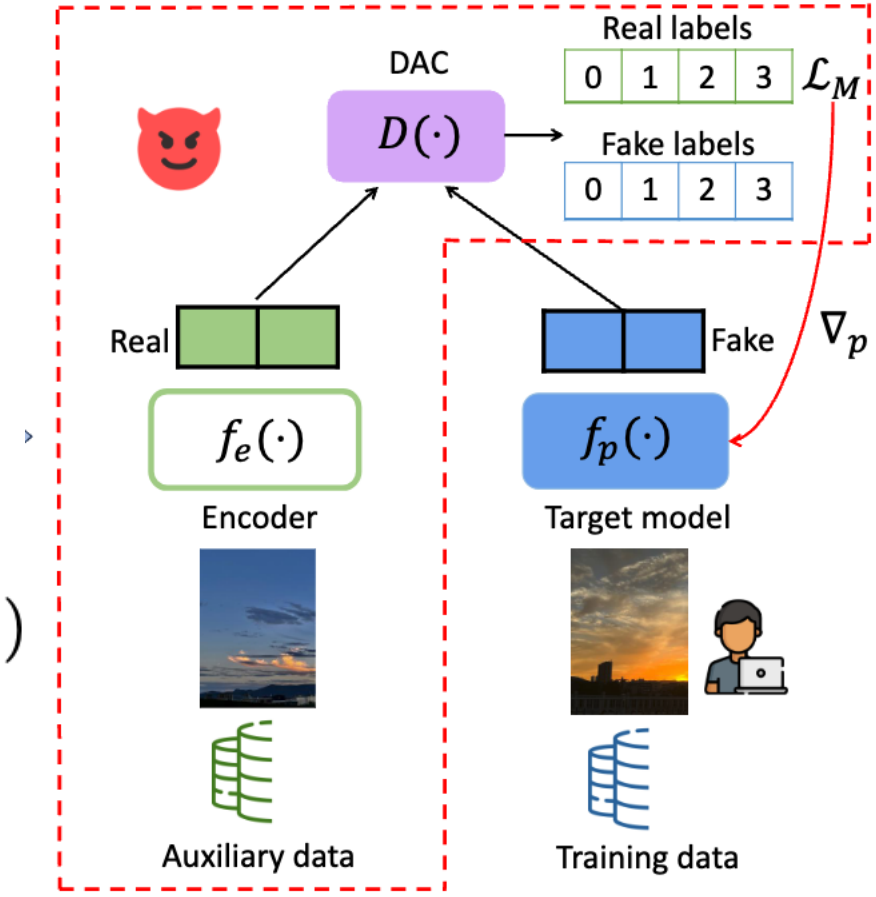
$$\mathcal{L}_R = \frac{1}{|\mathcal{B}_{aux}|} \sum_{i \in \mathcal{B}_{aux}} MSE(\tilde{x}_{i,t}, x_{i,t}).$$

2. Malicious gradient generation

- Transfer the embedding distribution from the encoder to the target model.

- Our method: Discriminator with Auxiliary Classifier (DAC)

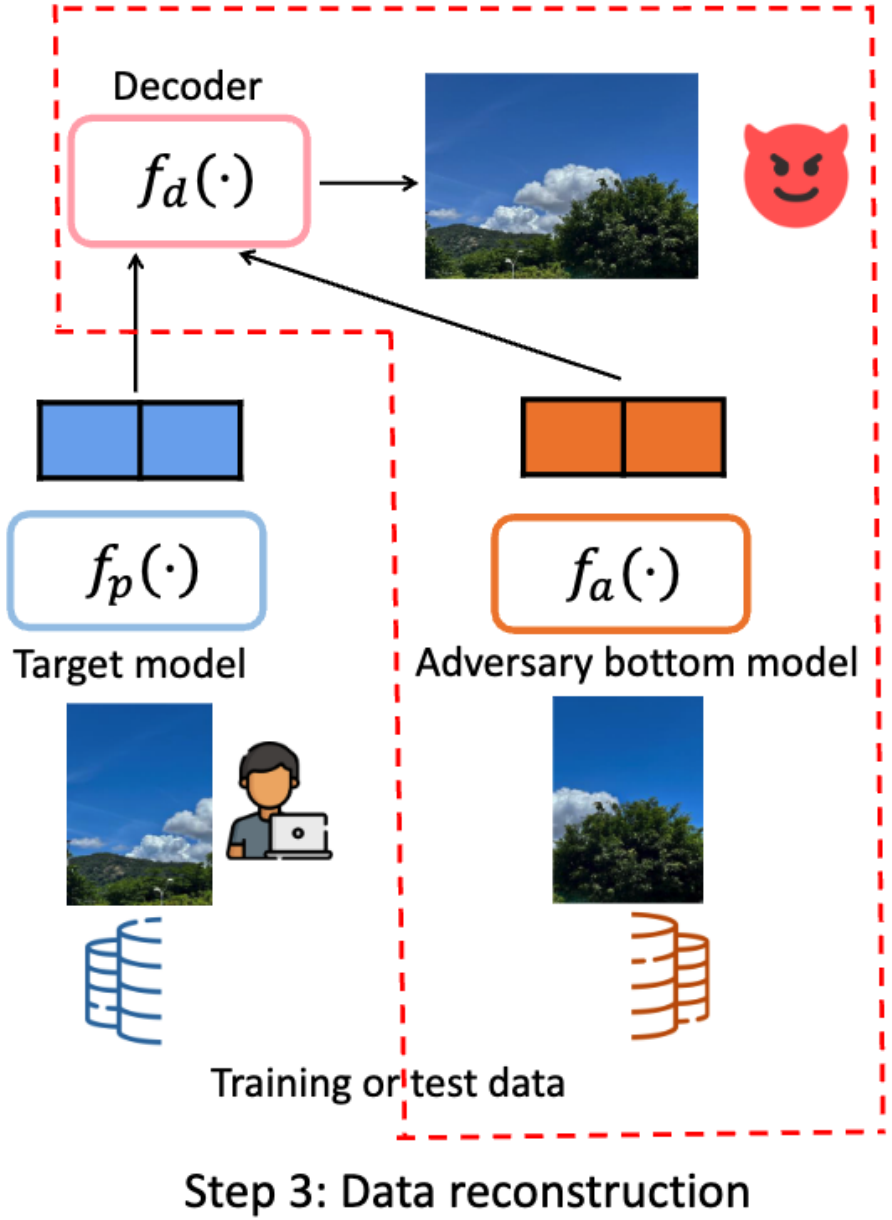
$$\min_{D} \mathbb{E}_{h,y\sim P_{H_{aux},Y}} CE(y^{+}, D(h)) + \mathbb{E}_{h,y\sim P_{H_{train},Y}} CE(y^{-}, D(h))$$



Step 2: Malicious gradient generation

3. Data reconstruction

The adversary can reconstruct the target data feature from the received embeddings.



Step 3: Data reconstruction

Why DAC?

Intuitive method: Use a discriminator.

- The discriminator helps minimize the JS distance of two distributions

$$\mathrm{JS}(P_{H_{aux}} \| P_{H_{train}})$$

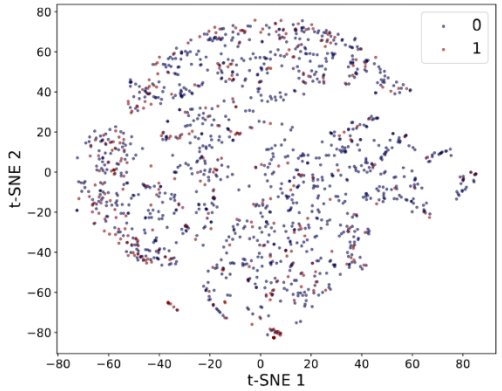
- Shortcomings:
- (1) Ignore label information
- (2) Easy to be detect.

DAC minimizes the distance of the joint distribution.

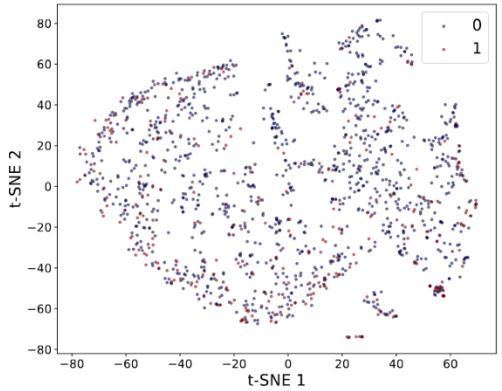
$$\mathrm{KL}(P_{H_{aux},Y} \| P_{H_{train},Y})$$

Why DAC?

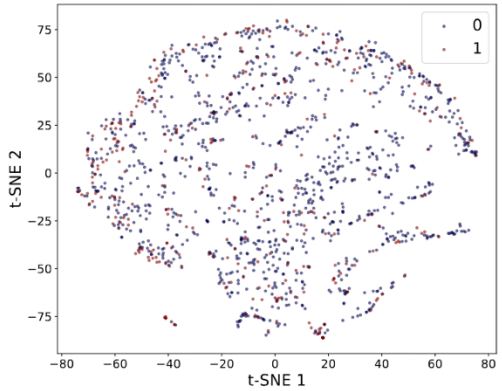
Result:



(a) Encoder's embedding distribution.

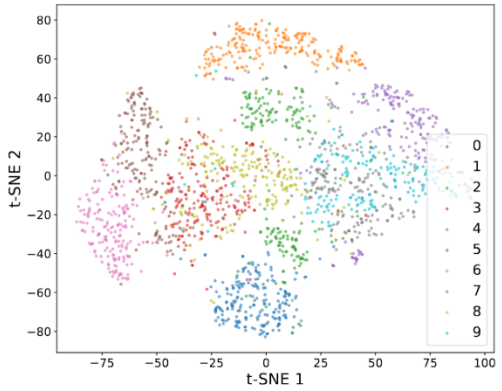


(b) Target model's embedding distribution using discriminator. Average cosine distance is 0.6121.

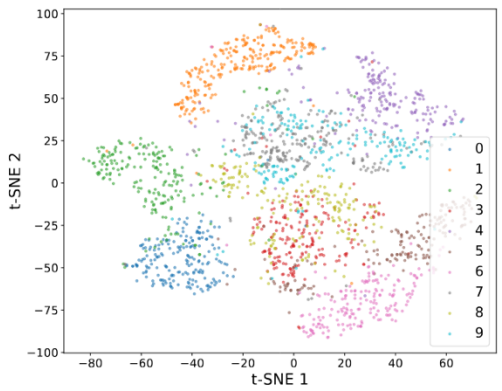


(c) Target model's embedding distribution using DAC. Average cosine distance is 0.1373.

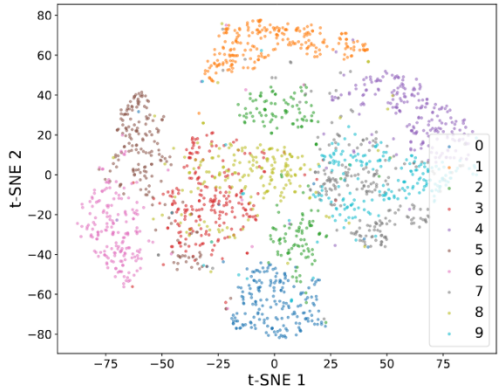
Fig. 4: t-SNE visualization on Credit dataset with 2 classes.



(a) Encoder's embedding distribution.



(b) Target model's embedding distribution using discriminator. Average cosine distance is 0.0412.



(c) Target model's embedding distribution using DAC. Average cosine distance is 0.0283.

Fig. 5: t-SNE visualization on MNIST dataset with 10 classes.

Result

- **Embedding transfer metrics**:

(1) Embedding MSE distance

(2) Embedding cosine distance

- **Reconstruction metrics**:

(1) MSE

(2) PSNR (Image dataset)

(3) SSIM (Image dataset)

Results on Tabular dataset

Method	Credit			RT_IOT2022		
	Recon MSE	Emb MSE	Emb Cos	Recon MSE	Emb MSE	Emb Cos
GRNA	1.1822 ± 0.0541	-	-	2.2542 ± 0.0325	-	-
GIA	0.9056 ± 0.0002	**0.2738** ± 0.0008	0.4681 ± 0.0215	1.8535 ± 0.0012	1.7276 ± 0.2356	0.4491 ± 0.0371
AGN	0.9656 ± 0.1196	-	-	2.2311 ± 0.1256	-	-
AGN-SG	1.4155 ± 0.1734	-	-	2.3967 ± 0.2628	-	-
AGN-GS	-	-	-	2.5670 ± 0.2909	-	-
PCAT	0.8612 ± 0.0984	0.5282 ± 0.0467	0.5486 ± 0.0197	2.2963 ± 0.4388	**1.5646** ± 0.2488	0.3742 ± 0.0195
SDAR	0.5327 ± 0.0374	0.5451 ± 0.1645	0.3103 ± 0.0931	1.6084 ± 0.1604	2.9362 ± 1.3356	0.2281 ± 0.1689
FSHA	0.5032 ± 0.0382	0.3906 ± 0.0622	0.2234 ± 0.0564	1.5773 ± 0.0507	2.2990 ± 0.6506	0.4859 ± 0.0984
FSHA-SG	0.7884 ± 0.1139	0.6056 ± 0.0704	0.5640 ± 0.0369	1.6895 ± 0.0419	2.7922 ± 1.0491	0.5909 ± 0.0745
FSHA-GS	-	-	-	1.7442 ± 0.0784	2.5768 ± 0.2169	0.5534 ± 0.1453
URVFL (SG/GS)	**0.4191** ± 0.0541	0.3078 ± 0.0778	**0.1658** ± 0.0551	1.3821 ± 0.0244	1.7894 ± 0.7435	**0.0549** ± 0.0113
URVFL_sync (SG/GS)	0.4722 ± 0.0228	0.3720 ± 0.0462	0.2277 ± 0.0441	**1.3277** ± 0.0665	2.3372 ± 0.5489	0.0938 ± 0.0340

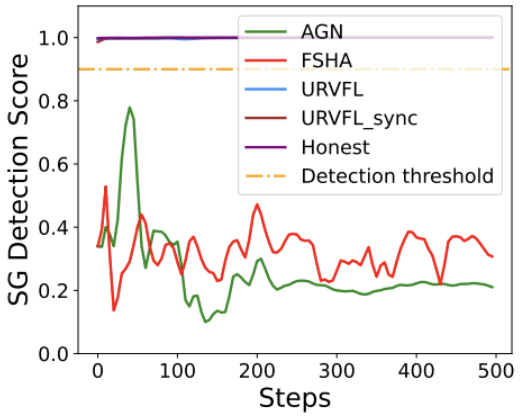
Results on Image dataset

Method	MNIST			CIFAR-10			Tiny imagenet		
	Recon MSE	Emb MSE	Emb Cos	Recon MSE	Emb MSE	Emb Cos	Recon MSE	Emb MSE	Emb Cos
GRNA	0.5533 ± 0.0919	-	-	0.3287 ± 0.0061	-	-	0.3869 ± 0.0103	-	-
GIA	0.9125 ± 0.0056	0.0179 ± 0.0018	0.2772 ± 0.0431	0.2550 ± 0.0004	1.4011 ± 0.1849	0.3957 ± 0.0535	0.3234 ± 0.0009	1.8109 ± 0.1039	0.3936 ± 0.0068
AGN	0.4801 ± 0.0027	-	-	0.4413 ± 0.0385	-	-	0.3786 ± 0.0190	-	-
AGN-SG	0.5131 ± 0.0014	-	-	0.5154 ± 0.0776	-	-	1.2398 ± 0.0102	-	-
AGN-GS	0.7073 ± 0.0813	-	-	0.5901 ± 0.0620	-	-	1.9577 ± 0.2531	-	-
PCAT	0.2446 ± 0.1312	0.0282 ± 0.0016	0.3484 ± 0.0299	0.3843 ± 0.0123	2.4672 ± 0.3024	0.5544 ± 0.0023	0.3063 ± 0.0249	2.7853 ± 0.0521	0.5526 ± 0.0025
SDAR	0.0839 ± 0.0751	0.1638 ± 0.1850	0.3142 ± 0.1699	0.3067 ± 0.0618	0.5230 ± 0.6965	1.2196 ± 0.4831	0.1363 ± 0.0293	2.1344 ± 0.1174	0.4692 ± 0.0281
FSHA	0.0536 ± 0.0118	0.0168 ± 0.0100	0.0909 ± 0.0157	0.0317 ± 0.0065	0.4657 ± 0.1286	0.1610 ± 0.0285	0.1025 ± 0.0084	1.4897 ± 0.2210	0.3056 ± 0.0203
FSHA-SG	0.5158 ± 0.2511	0.0621 ± 0.0095	0.2956 ± 0.0691	0.1765 ± 0.0390	2.1634 ± 0.1310	0.4282 ± 0.0480	0.3110 ± 0.1202	2.8927 ± 0.3759	0.4904 ± 0.0519
FSHA-GS	0.2829 ± 0.1374	0.1446 ± 0.0260	0.3077 ± 0.0402	0.1312 ± 0.0150	1.5940 ± 0.0684	0.3133 ± 0.0230	0.4797 ± 0.2383	3.0416 ± 0.4313	0.5200 ± 0.0423
URVFL (SG/GS)	0.0132 ± 0.0027	0.0045 ± 0.0009	**0.0287** ± 0.0029	0.0302 ± 0.0011	0.5679 ± 0.0834	0.1303 ± 0.0131	**0.0699** ± 0.0055	1.4316 ± 0.6673	**0.2139** ± 0.0195
URVFL_sync (SG/GS)	**0.0127** ± 0.0011	**0.0040** ± 0.0006	0.0341 ± 0.0022	**0.0176** ± 0.0139	**0.2793** ± 0.1868	**0.0675** ± 0.0549	0.0704 ± 0.0011	**1.1441** ± 0.0058	0.2434 ± 0.0077

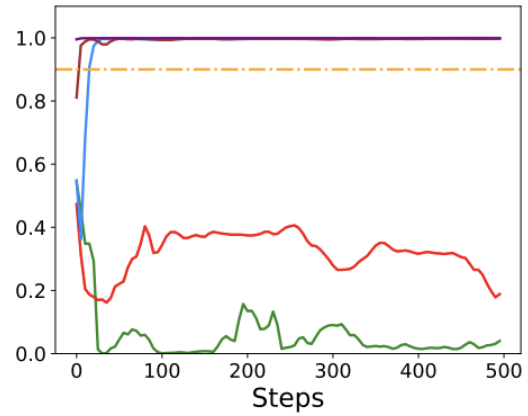
Results on Image dataset

Method	MNIST		CIFAR-10		Tiny imagenet	
	PSNR↑	SSIM ↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑
GRNA	8.8317±0.0005	0.1467±0.0001	11.3939±0.0074	0.0019±0.0001	7.3986±0.5618	0.0045±0.0001
GIA	6.4411±0.0006	0.0029±0.0001	11.9625±0.0001	0.0018±0.0001	10.9342±0.0002	0.0010±0.0000
AGN	9.2787±0.2651	0.1400±0.0033	10.4527±0.1293	0.0150±0.0001	10.2442±0.0487	0.0150±0.0001
AGN-SG	8.9643±0.1067	0.0499±0.0015	9.0854±1.0272	0.0156±0.0001	3.4810±0.0000	0.0006±0.0256
AGN-GS	7.8738±0.6595	0.0746±0.0027	7.7755±0.9967	0.0142±0.0001	3.7093±0.4056	0.0034±0.0001
PCAT	14.0504±0.4956	0.4506±0.0177	11.9012±1.6192	0.0727±0.0053	11.3199±0.1573	0.0160±0.0001
SDAR	17.8807±0.6992	0.7093±0.0017	14.9388±2.3774	0.2525±0.0121	14.7862±0.9083	0.1318±0.0017
FSHA	22.7711±2.2248	0.8901±0.0006	19.7955±2.2895	0.5312±0.0298	15.9321±0.1248	0.2205±0.0003
FSHA-SG	10.4247±1.6129	0.1928±0.0176	13.2670±0.5318	0.1294±0.0006	12.6357±0.7394	0.0395±0.0002
FSHA-GS	11.1577±4.5247	0.2151±0.0137	15.1751±0.5590	0.3083±0.0053	12.4729±1.9752	0.0577±0.0001
URVFL(GS/SG)	25.1349±1.2354	**0.9385**±0.0002	21.0666±0.0165	0.5927±0.0002	**17.6107**±0.0992	0.3023±0.0001
URVFL_sync(GS/SG)	**25.2919**±0.1369	0.9324±0.0002	**26.2360**±0.0030	**0.8125**±0.0001	17.5381±0.0029	**0.3053**±0.0001

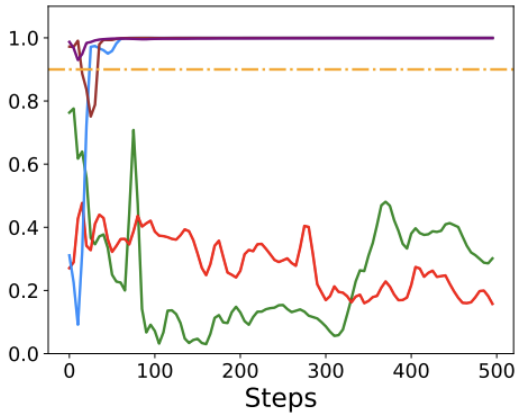
Result under detection



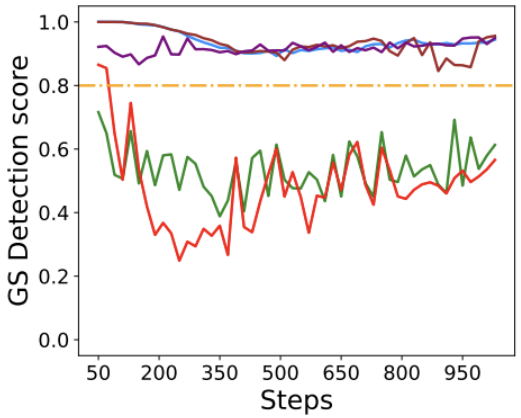
(a) SG scores in MNIST.



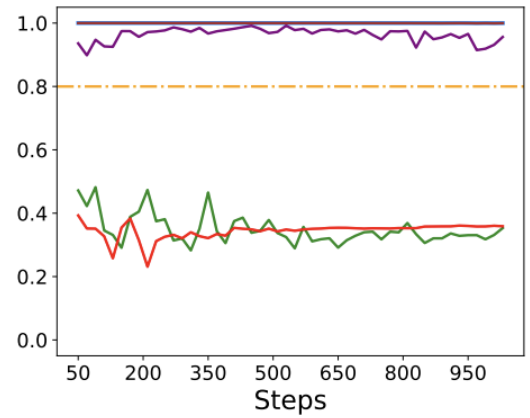
(b) SG scores in CIFAR-10.



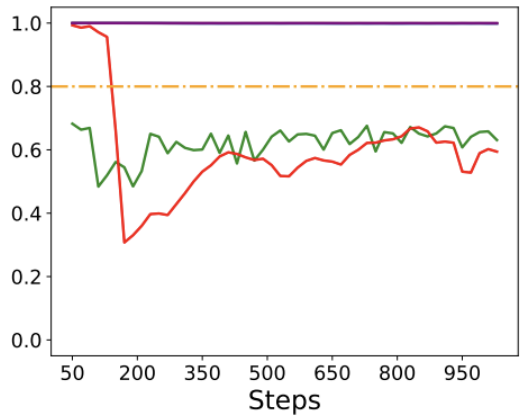
(c) SG scores in Tiny.



(f) GS scores in MNIST.



(g) GS scores in CIFAR-10.



(h) GS scores in Tiny.

Visualization of the reconstruction

			Recon MSE
Without detection SplitGuard	Original		
	AGN		0.3896
	FSHA		0.1131
	URVFL		0.0660
With detection SplitGuard	AGN		1.0493
	FSHA		0.2501
	URVFL		0.0660

Conclusion

- Malicious attack in VFL cause more privacy Leakage.
- URVFL can circumvent current detection strategies.

Future work:

- How to defend this kind of malicious attacks?
- Malicious attacks in other ML models.