VILLAIN: Backdoor Attacks Against Vertical Split Learning

Yijie Bai[1], Yanjiao Chen[1], Hanlei Zhang[1], Wenyuan Xu[1], Haiqin Weng[2] , Dou Goodman[2]

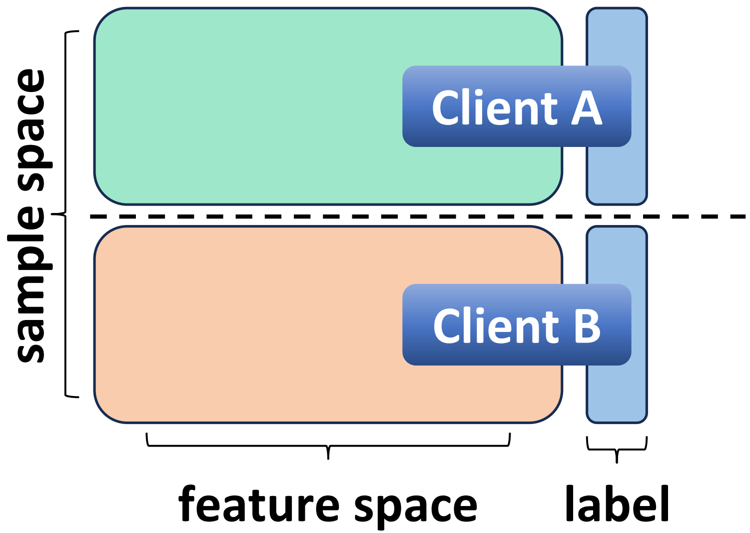
[1]Ubiquitous System Security Lab (USSLAB), Zhejiang University, [2]Ant Group

{baiyj, chenyanjiao, hanleizhang, wyxu}@zju.edu.cn

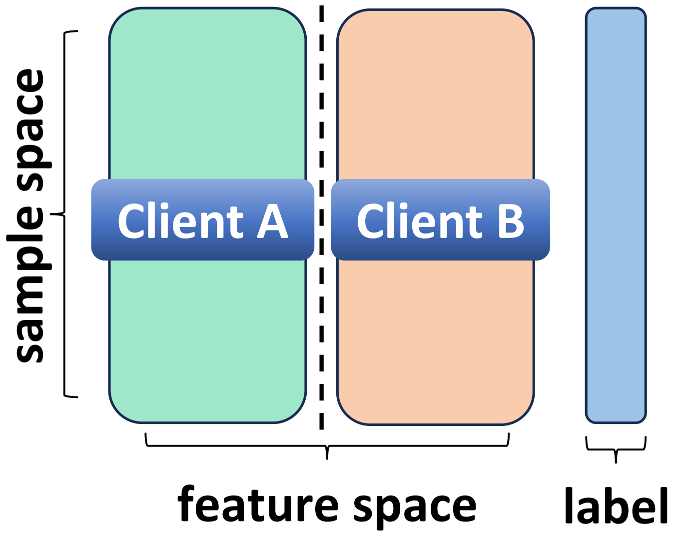
{haiqin.wenghaiqin, bencao.ly}@antgroup.com

Federated Learning

- **Horizontal Federated Learning**



- **Vertical Federated Learning**

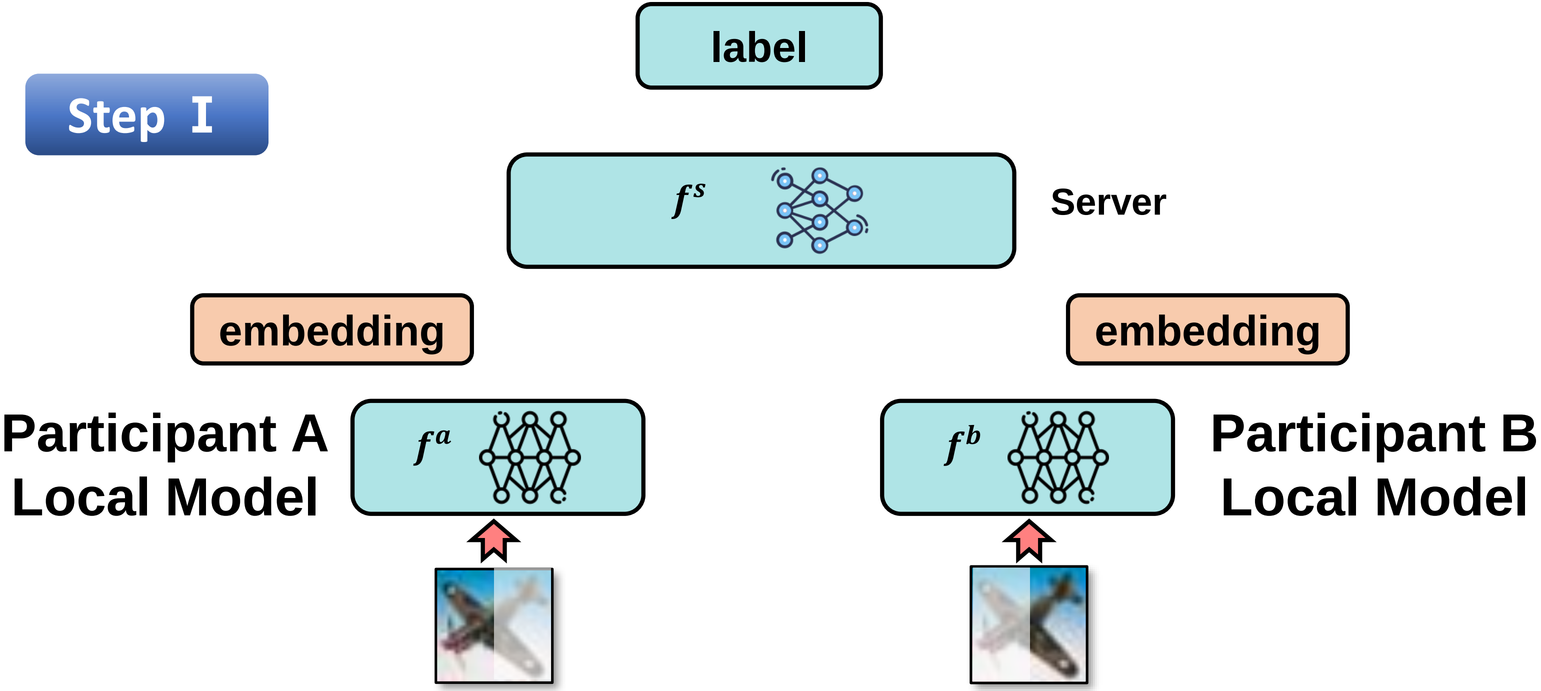


Vertical Split Learning

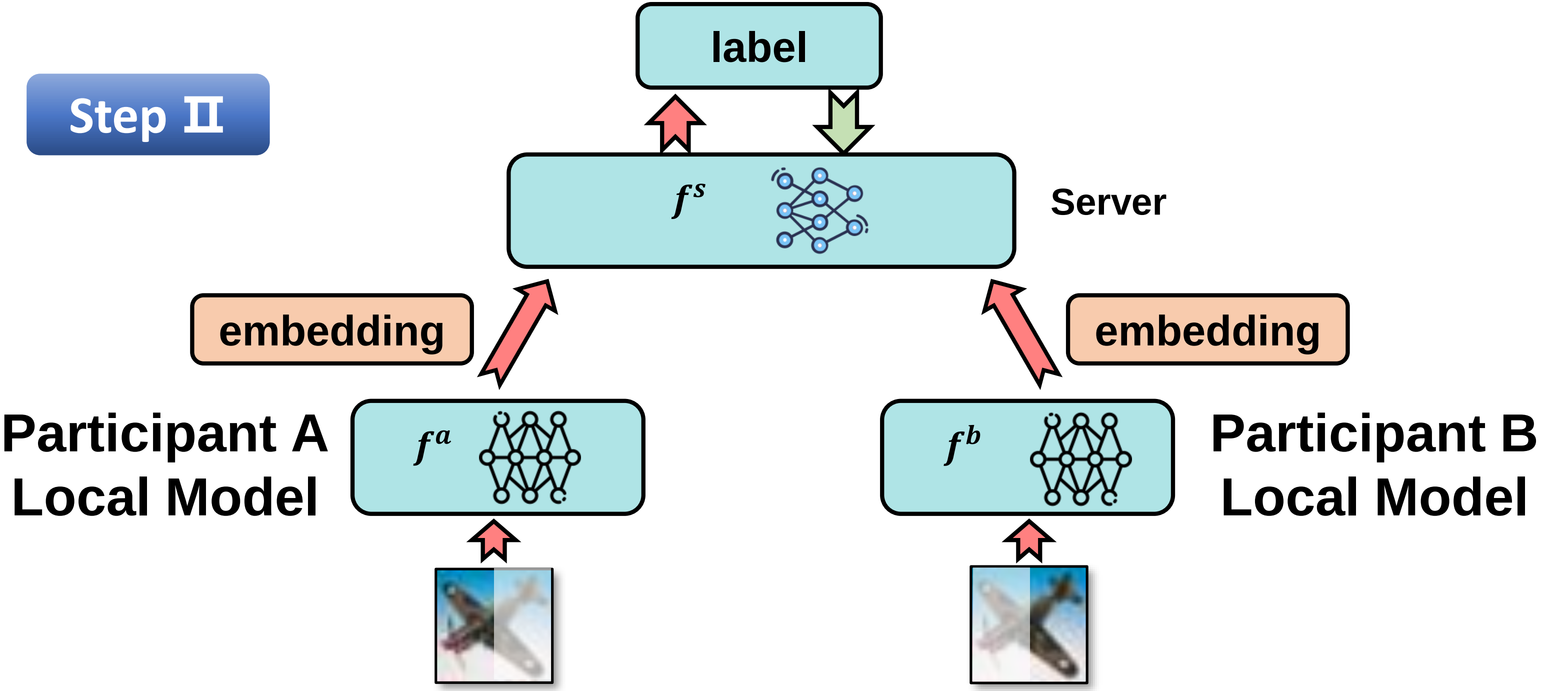
- ☐ **Credit business application**

- ☐ **Online advertising application**

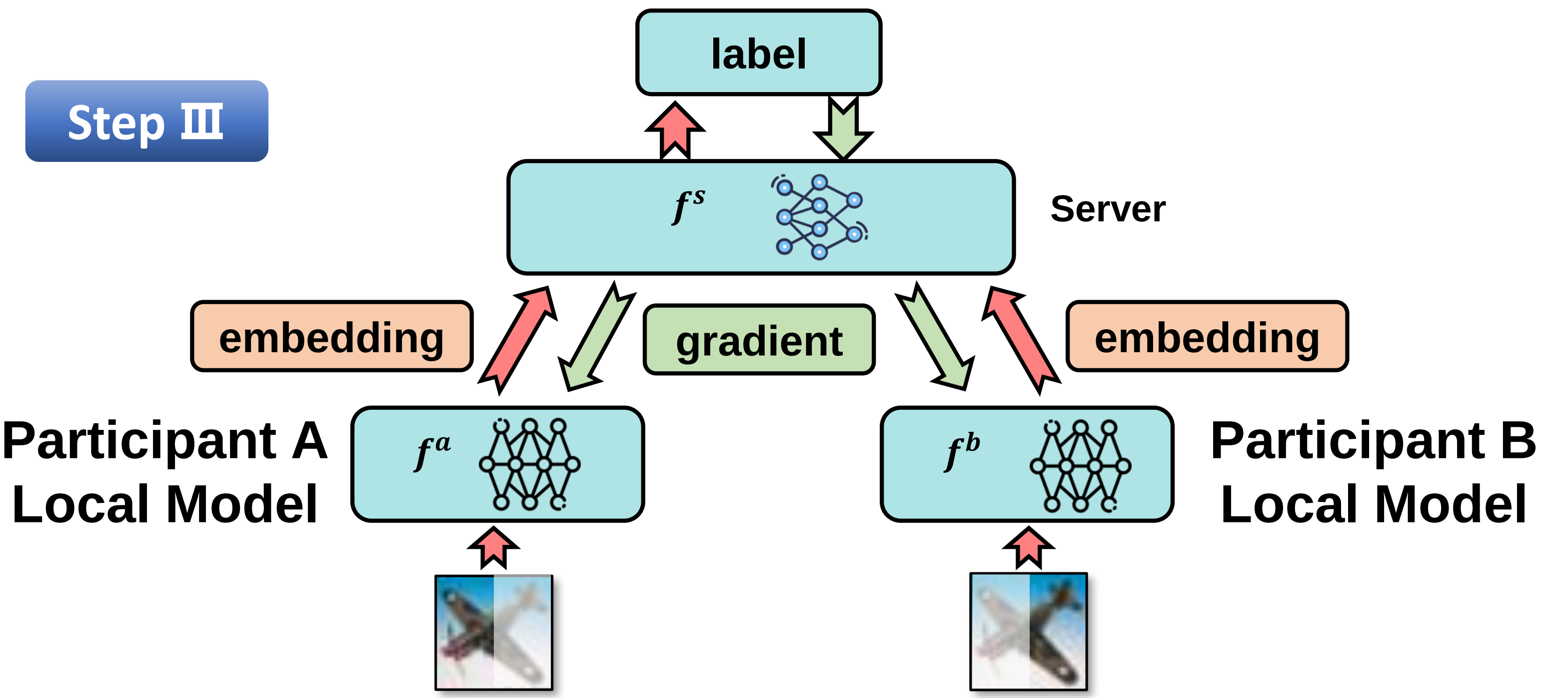
Vertical Split Learning



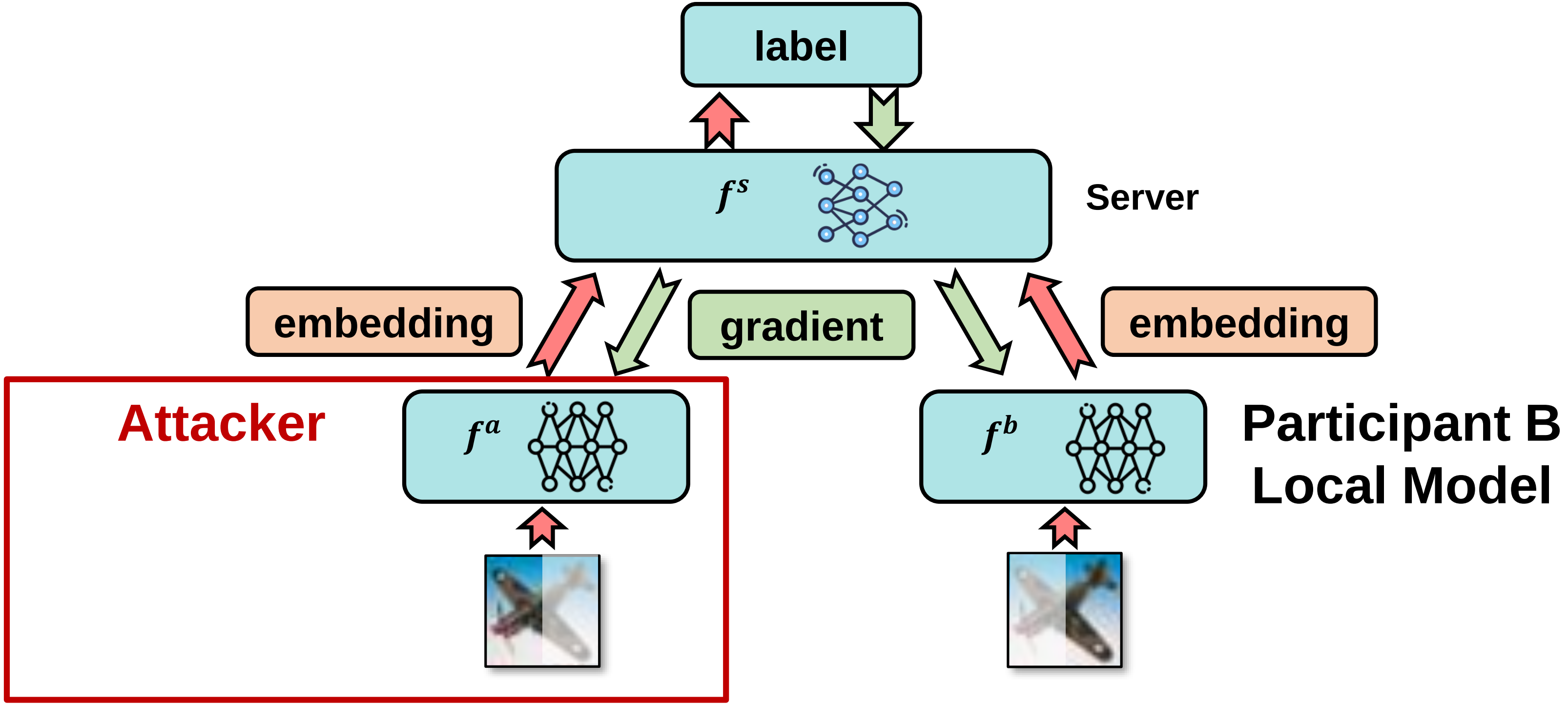
Vertical Split Learning



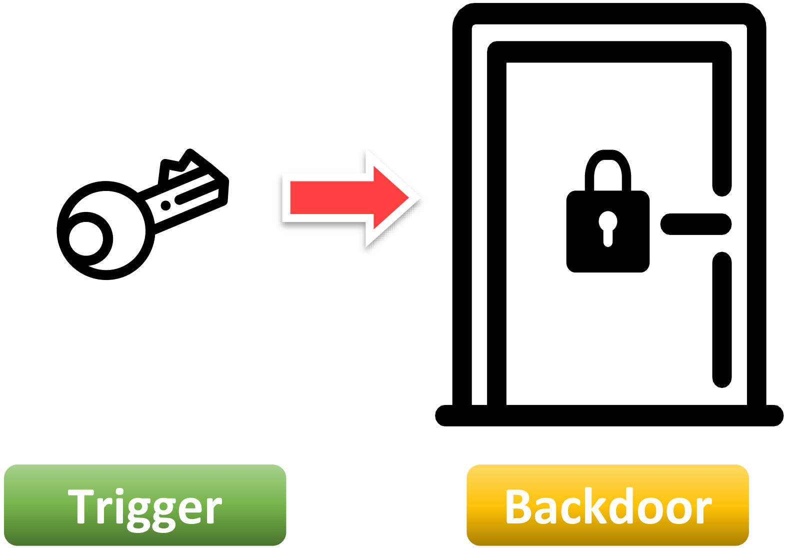
Vertical Split Learning



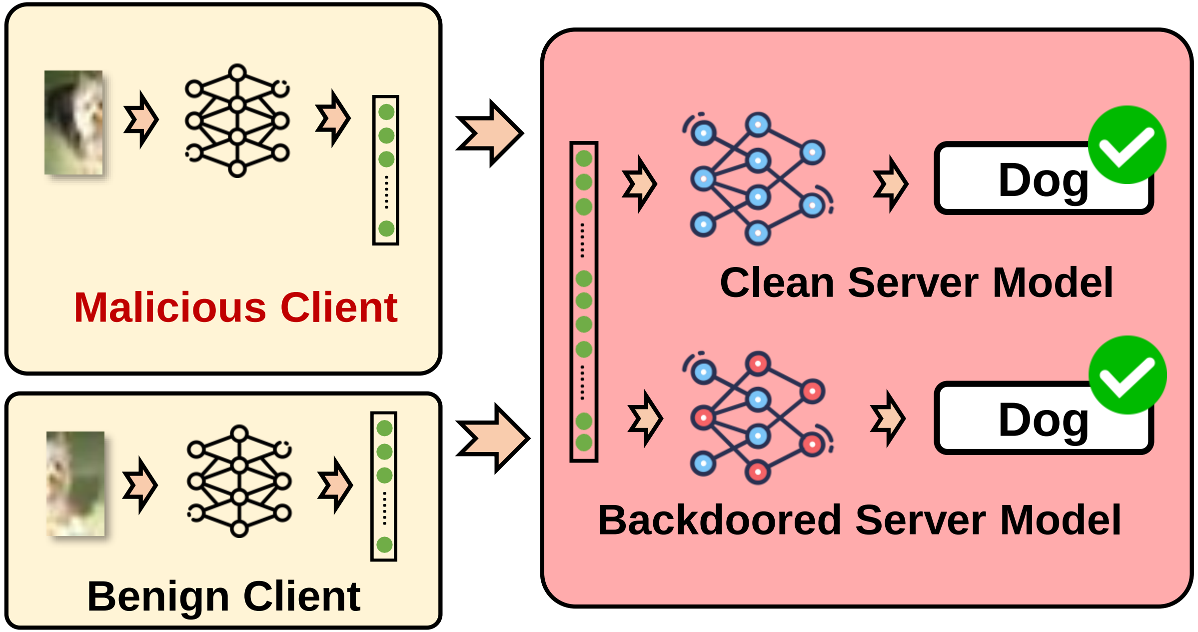
Vertical Split Learning



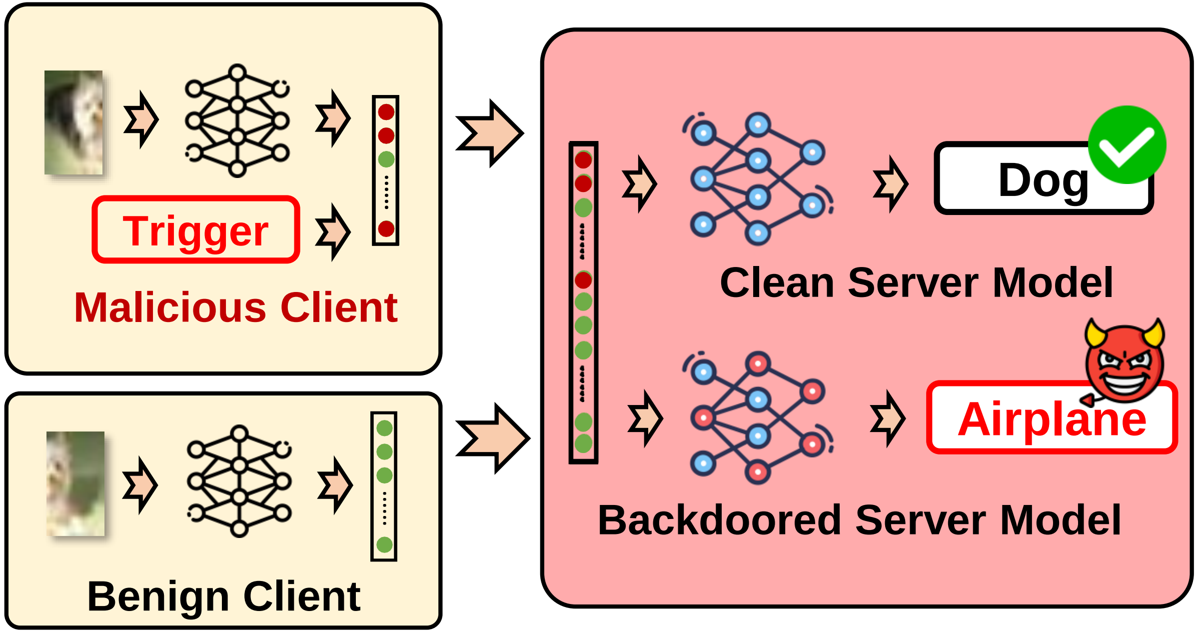
Backdoor Attack



Attacker’s Goal



Attacker's Goal



Threat Model

◻ *Attacker's knowledge*

- ➢ **Local dataset $\mathbf{X}^{a} = \{\tilde{\mathbf{x}}_{i}^{a}\}_{i=1}^{N}$**
- ➢ **One target label sample**
- ➢ **Gradient information**

◻ *Attacker's capability*

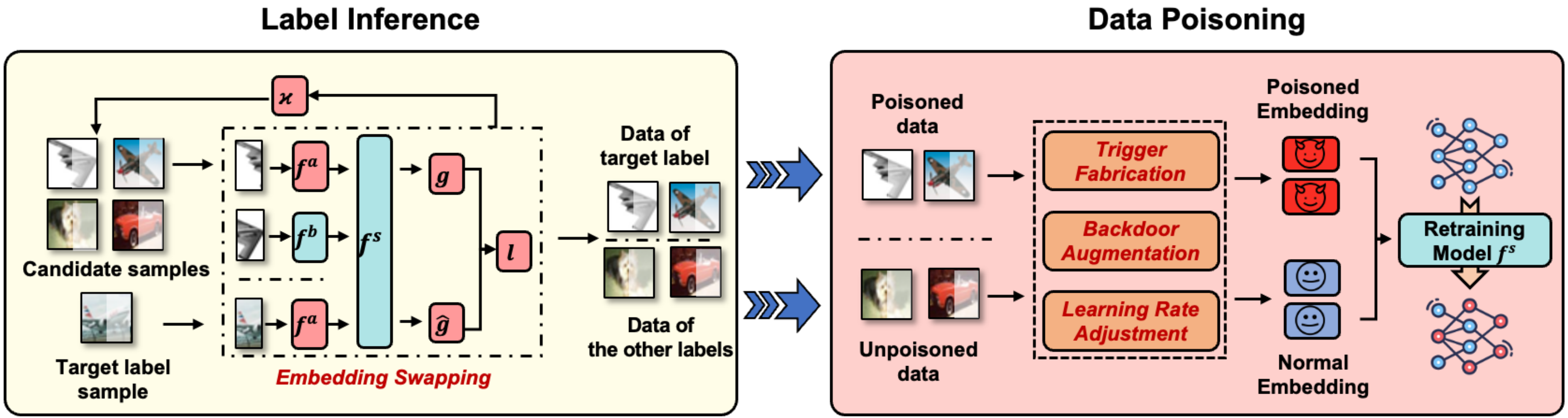
- ➢ **Train and manipulate the local embedding model $\mathbf{f}^{a}$.**
- ➢ **Upload the embedding vectors to the server.**

Challenge

- ***No label information***
  - **No knowledge of the labels**
  - **Can't change the labels**

- ***No server model information***
  - **Only gradient update information**
  - **Unknown server model**

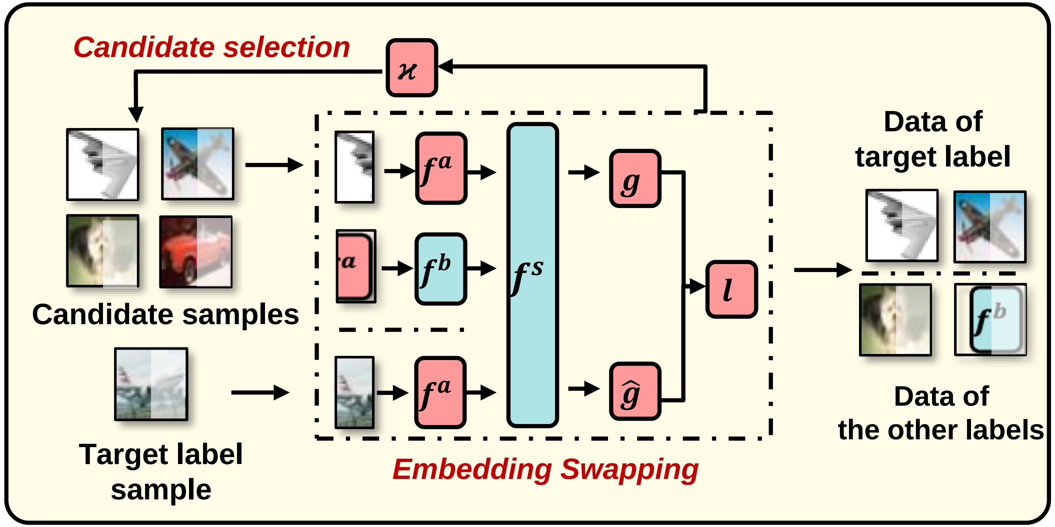
VILLAIN: Detailed Construction



No label information

No global model information

Label Inference



Inference Adjustment

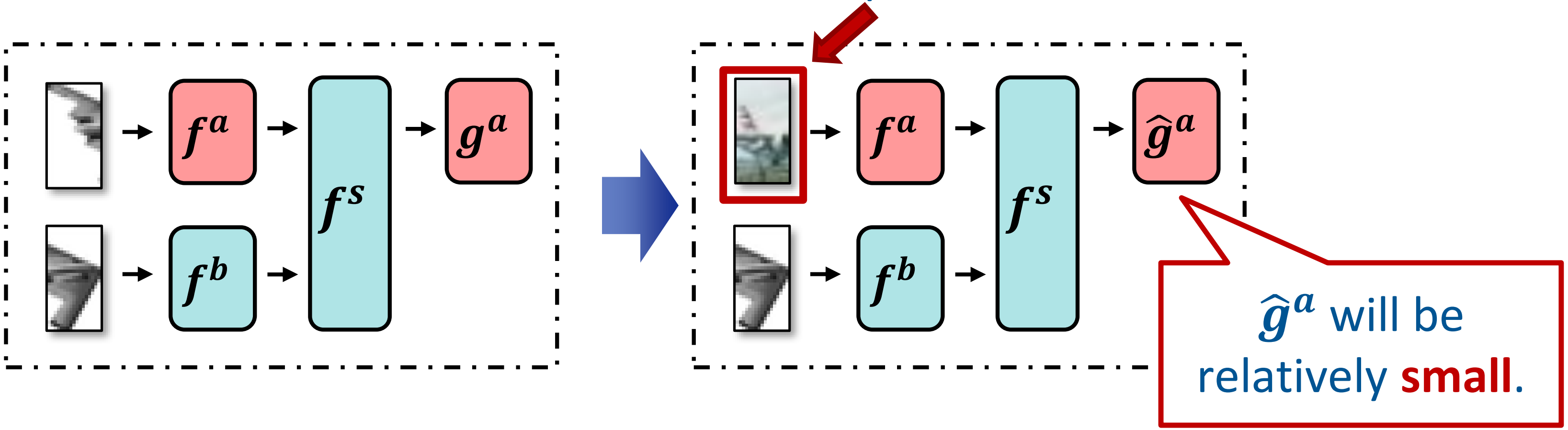
Pinpoint data samples of the target label.

Label Inference

Embedding Swapping

True Label: **Plane**

Swap with: **Plane**

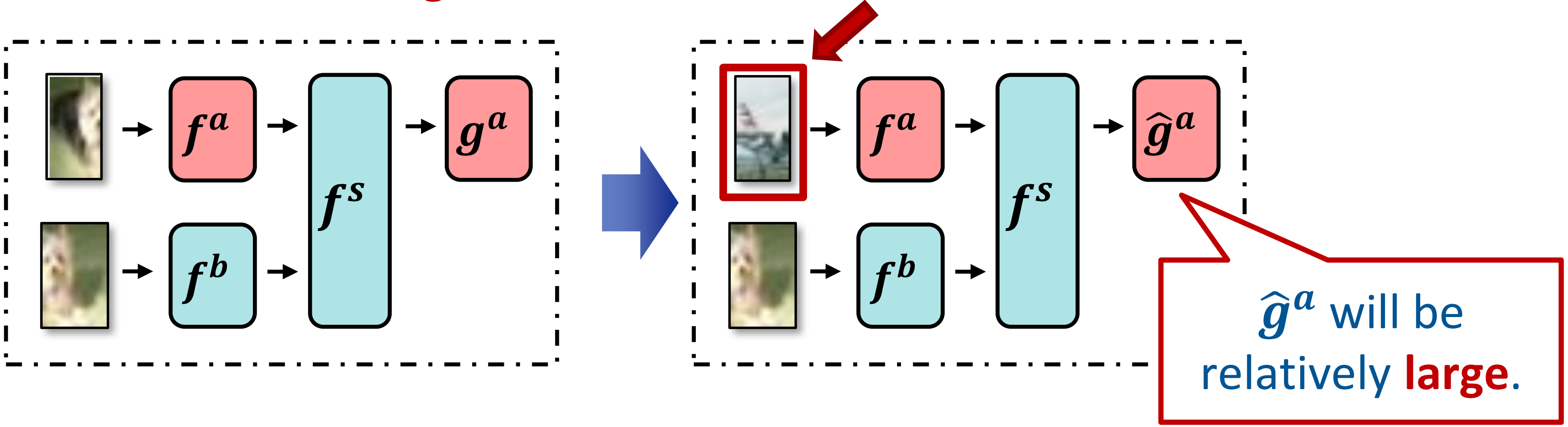


Label Inference

Embedding Swapping

True Label: **Dog**

Swap with: **Plane**



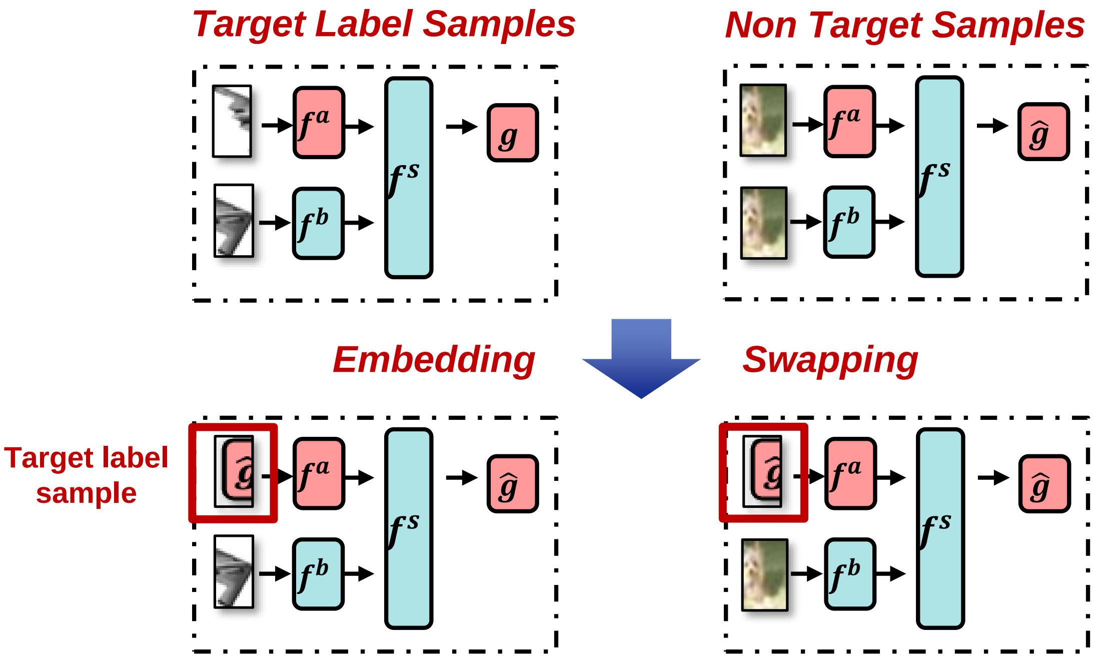
Label Inference

Label Inference

Embedding Swapping

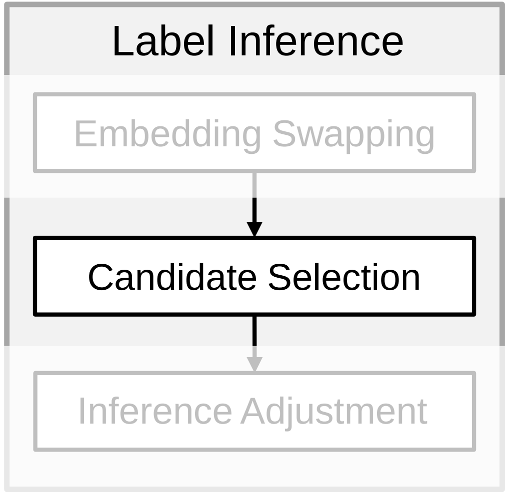
Candidate Selection

Inference Adjustment

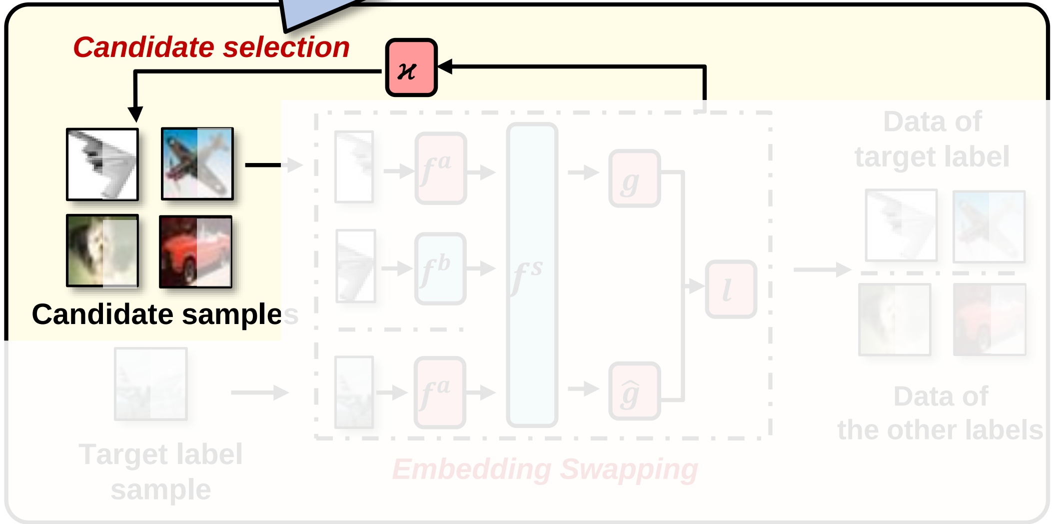


$\frac{\|\hat{\mathbf{g}}_i^a\|_2}{\|\mathbf{g}_i^a\|_2} \leq \theta$ *and* $\|\ \mathbf{g}_i^a\ \|_2 \leq \mu$ *are good indicators for label inference.*

Label Inference



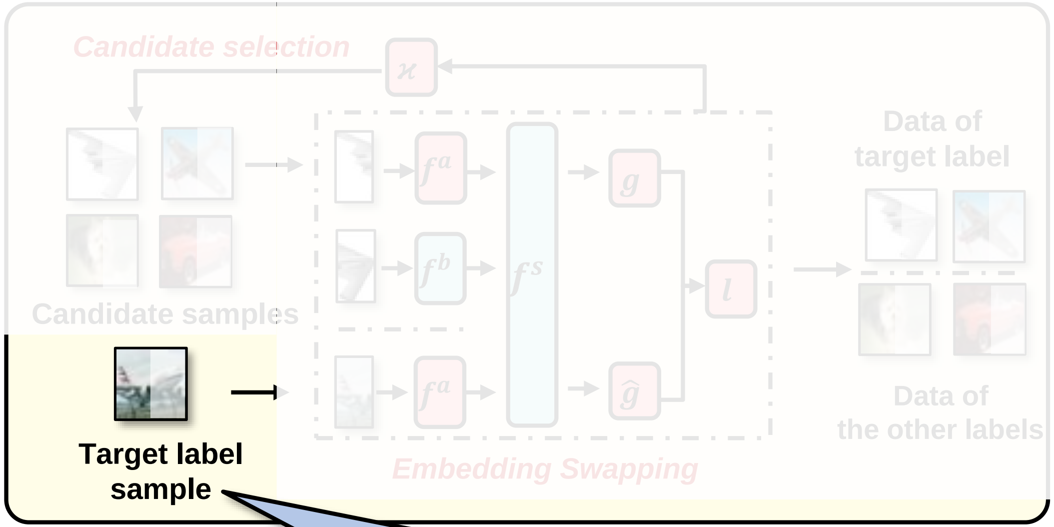
① **Semi-supervised classifier** $\varkappa$
② **Embedding** e_i^a **with information**



Label Inference

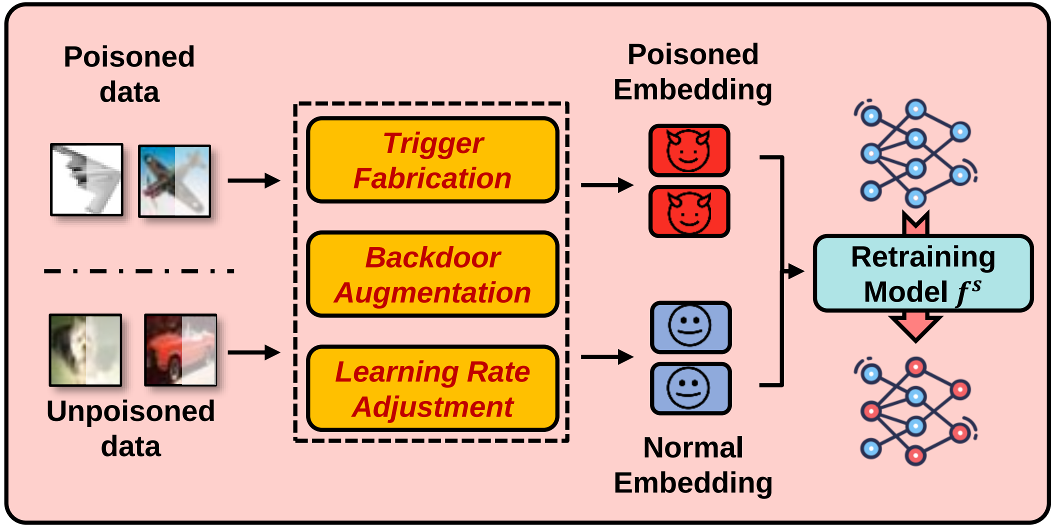
Label Inference

- Embedding Swapping
- Candidate Selection
- **Inference Adjustment**



Dynamically adjust the embedding for swapping

Data Poisoning



The attacker poisons these target label samples to inject the backdoor into the server model.

Data Poisoning

Trigger Fabrication

Backdoor Augmentation

Learning Rate Adjustment

Trigger Fabrication

- ***An additive trigger to poison the embedding vector***

$$\hat{\mathbf{e}}^a = f^a(\tilde{\mathbf{x}}^a) \oplus \mathcal{E}$$

- ***The trigger $\mathcal{E}$ is formed as***

$$\mathcal{E} = \mathcal{M} \otimes (\beta \cdot \Delta)$$

Experiment Setup

☐ *Dataset*

- MNIST (MN).
- CIFAR-10 (CF).
- CINIC-10 (CN).
- ImageNette (IN).
- Bank Marketing (BM).
- Give-Me-Some-Credit (GM).

☐ *Metrics*

- Attack success rate (ASR).
- Clean data accuracy (CDA) .
- Label inference accuracy (LIA).

4 image datasets (unstructured datasets)
and 2 financial tabular datasets (structured datasets).

Experiment Design

☐ *Overall Performance*

- ***Potential side-effects.***
- ***Different embedding aggregation methods.***
- ***Data-domain triggers.***
- ***Multi-participant scenario.***
- ***Ablation studies***

☐ *Hyperparameters*

- ***Poisoning rate.***
- ***Trigger magnitude.***
- ***Server & participant models.***
- ***Trigger size.***
- ***Learning rate.***
- ***Number of candidates.***

☐ *Resistance to Defense*

- ***Label inference defense.***
- ***Backdoor attack defense.***
- ***Adaptive Defenses.***

Overall Performance

Table 1: Attack performance of VILLAIN compared with baselines.

DS†	Metric	ExPLoit repl. tgr.	ExPLoit add. tgr.	pasv. Fu repl. tgr.	pasv. Fu add. tgr.	act. Fu repl. tgr.	act. Fu add. tgr.	ES repl. tgr.	VILLAIN ‡
MN	ASR	16.51 ± 5.14%	18.43 ± 4.50%	98.02 ± 2.21%	100.00 ± 0.00%	97.66 ± 3.57%	99.94 ± 0.13%	96.53 ± 5.11%	**100.00 ± 0.00%**
	CDA	96.10 ± 0.22%	95.73 ± 0.16%	95.99 ± 0.19%	96.14 ± 0.08%	96.01 ± 0.12%	**96.18 ± 0.07%**	95.47 ± 0.33%	96.11 ± 0.22%
	LIA	12.48 ± 0.73%	12.48 ± 0.73%	89.39 ± 6.99%	89.39 ± 6.99%	93.70 ± 4.48%	93.70 ± 4.48%	94.03 ± 2.56%	**94.03 ± 2.56%**
CF	ASR	8.26 ± 2.02%	16.93 ± 3.76%	13.61 ± 0.86%	78.99 ± 6.23%	14.45 ± 1.44%	84.96 ± 8.28%	23.66 ± 6.48%	**98.68 ± 0.59%**
	CDA	76.66 ± 0.38%	75.94 ± 0.36%	76.75 ± 0.27%	76.96 ± 0.35%	**76.90 ± 0.14%**	77.09 ± 0.38%	76.49 ± 0.40%	76.87 ± 0.25%
	LIA	18.96 ± 2.19%	18.96 ± 2.19%	68.12 ± 6.09%	68.12 ± 6.09%	76.35 ± 5.26%	76.35 ± 5.26%	96.08 ± 4.28%	**96.08 ± 4.28%**
IN	ASR	13.94 ± 4.8%	12.55 ± 1.79%	26.73 ± 2.73%	76.03 ± 9.59%	27.71 ± 2.44%	79.48 ± 6.09%	32.39 ± 12.26%	**92.79 ± 1.58%**
	CDA	71.21 ± 0.39%	70.82 ± 0.93%	70.55 ± 0.18%	70.08 ± 0.22%	70.91 ± 0.50%	70.19 ± 0.74%	**71.64 ± 0.89%**	71.54 ± 0.98%
	LIA	14.53 ± 1.70%	14.53 ± 1.70%	80.28 ± 8.94%	80.28 ± 8.94%	86.54 ± 6.68%	86.54 ± 6.68%	90.41 ± 2.18%	**90.41 ± 2.18%**
CN	ASR	5.13 ± 3.95%	8.98 ± 4.39%	26.63 ± 5.30%	86.56 ± 6.45%	33.95 ± 10.22%	85.01 ± 15.82%	64.56 ± 6.36%	**99.55 ± 0.62%**
	CDA	61.90 ± 0.28%	61.64 ± 0.48%	62.65 ± 0.17%	**62.86 ± 0.08%**	62.68 ± 0.31%	62.72 ± 0.47%	62.67 ± 0.08%	62.78 ± 0.11%
	LIA	12.55 ± 1.91%	12.55 ± 1.91%	66.83 ± 8.01%	66.83 ± 8.01 %	72.09 ± 7.26%	72.09 ± 7.26%	93.19 ± 3.95%	**93.19 ± 3.95%**
BM	ASR	9.15 ± 3.90%	14.38 ± 1.93%	40.19 ± 4.31%	90.28 ± 10.19%	39.46 ± 2.53%	86.79 ± 10.56%	59.43 ± 12.10%	**97.84 ± 2.57%**
	CDA	91.36 ± 0.77%	90.37 ± 0.51%	92.11 ± 0.94%	91.22 ± 2.71%	**92.79 ± 0.25%**	88.83 ± 2.55%	91.80 ± 1.46%	90.00 ± 2.34%
	LIA	46.18 ± 2.39%	46.18 ± 2.39%	92.11 ± 4.49%	92.11 ± 4.49%	88.78 ± 4.64%	88.78 ± 4.64%	94.05 ± 4.82%	**94.05 ± 4.82%**
GM	ASR	12.01 ± 3.54%	17.87 ± 5.83%	67.69 ± 1.04%	100.00 ± 0.00%	67.43 ± 1.22%	100.00 ± 0.00%	92.27 ± 15.41%	**100.00 ± 0.00%**
	CDA	78.02 ± 0.77%	77.81 ± 0.42%	78.55 ± 0.24%	78.41 ± 0.06%	78.53 ± 0.20%	78.32 ± 0.24%	**78.68 ± 0.09%**	78.37 ± 0.14%
	LIA	55.78 ± 2.33%	55.78 ± 2.33%	77.66 ± 0.72%	77.66 ± 0.72%	77.52 ± 0.60%	77.52 ± 0.60%	95.18 ± 5.69%	**95.18 ± 5.69%**

Villain achieves the highest ASR on each dataset.

Data-domain triggers

Table 4: Data-domain triggers. TS: Trigger Size.

DS	TS	ASR	CDA	ori. acc.	DS	TS	ASR	CDA	ori. acc.
MN	2	92.04%	96.72%	94.66%	CF	2	95.36%	78.82%	76.78%
	3	99.92%	96.65%	94.71%		3	99.70%	78.95%	76.58%
	4	99.97%	96.79%	94.40%		4	98.53%	79.31%	75.65%
	5	99.94%	96.80%	94.57%		5	99.27%	79.43%	76.75%
	6	99.99%	96.63%	94.99%		6	99.55%	79.27%	77.76%
IM	14	41.69%	74.19%	73.06%	CN	2	46.60%	63.43%	61.00%
	21	51.11%	74.51%	70.45%		3	98.59%	63.84%	62.26%
	28	77.58%	74.87%	70.05%		4	96.85%	64.12%	62.74%
	35	90.11%	75.25%	72.53%		5	99.17%	64.01%	62.11%
	42	98.66%	74.37%	71.47%		6	96.92%	63.87%	62.16%
BM	1	98.69%	92.40%	90.18%	GM	1	100.00%	78.52%	77.82%
	2	97.79%	92.76%	88.25%		2	100.00%	78.76%	77.82%
	3	99.74%	93.28%	90.33%		3	100.00%	78.76%	77.73%
	4	99.35%	92.89%	86.23%		4	100.00%	78.54%	77.65%
	5	99.80%	93.12%	90.72%		5	100.00%	78.73%	77.80%

In VILLAIN, the trigger can be added in the data domain or the embedding domain.

Different embedding aggregation methods

- **Different aggregation methods.**
  - ***C: CON, embedding concatenation.***
  - ***A: ADD, element-wise addition.***
  - ***M1: MEAN, element-wise average.***
  - ***M2: MAX, element-wise maximum.***
  - ***M3: MIN, element-wise minimum.***

DS	M†	ori. acc.	LIA	ASR	CDA
MN	C	95.82 ± 0.29%	94.03 ± 2.56%	100.00 ± 0.00%	96.11 ± 0.22%
	A	96.69 ± 0.35%	99.00 ± 0.19%	100.00 ± 0.00%	95.97 ± 0.27%
	M1	95.97 ± 0.38%	89.48 ± 2.99%	100.00 ± 0.00%	95.13 ± 0.30%
	M2	95.61 ± 0.69%	94.05 ± 3.65%	100.00 ± 0.00%	94.56 ± 0.48%
	M3	96.11 ± 0.16%	99.51 ± 0.17%	95.22 ± 1.13%	95.59 ± 0.37%
CF-10	C	78.29 ± 0.42%	96.08 ± 4.28%	98.68 ± 0.59%	76.87 ± 0.25%
	A	78.79 ± 0.22%	99.85 ± 0.22%	94.55 ± 0.28%	79.90 ± 0.58%
	M1	77.83 ± 0.27%	99.86 ± 0.32%	94.85 ± 0.51%	79.17 ± 0.18%
	M2	76.44 ± 0.37%	99.98 ± 0.02%	91.33 ± 0.48%	78.09 ± 0.70%
	M3	76.94 ± 0.05%	99.29 ± 0.44%	82.98 ± 3.81%	78.54 ± 0.10%
IN	C	71.59 ± 0.84%	90.41 ± 2.18%	92.79 ± 1.58%	71.54 ± 0.98%
	A	71.93 ± 1.06%	88.56 ± 2.63%	100.00 ± 0.00%	68.84 ± 0.74%
	M1	59.99 ± 1.94%	82.30 ± 4.48%	99.29 ± 0.12%	56.64 ± 3.57%
	M2	66.95 ± 1.44%	84.30 ± 2.31%	100.00 ± 0.00%	64.56 ± 0.79%
	M3	65.59 ± 1.57%	86.69 ± 3.74%	100.00 ± 0.00%	63.49 ± 1.30%
CN	C	62.10 ± 0.08%	93.19 ± 3.95%	99.55 ± 0.62%	62.78 ± 0.11%
	A	63.36 ± 1.37%	94.97 ± 4.22%	95.84 ± 3.82%	62.81 ± 1.59%
	M1	63.19 ± 0.27%	88.61 ± 2.90%	96.81 ± 2.27%	61.76 ± 0.23%
	M2	60.16 ± 1.51%	85.18 ± 3.07%	94.43 ± 6.10%	62.83 ± 0.59%
	M3	63.29 ± 0.37%	88.47 ± 3.58%	96.81 ± 2.53%	64.11 ± 0.20%
BM	C	90.98 ± 0.52 %	94.05 ± 4.82%	97.84 ± 2.57%	90.57 ± 2.14%
	A	90.35 ± 0.36%	99.58 ± 0.37%	92.50 ± 5.83%	90.83 ± 0.28%
	M1	92.68 ± 0.78%	99.89 ± 0.10%	70.68 ± 8.54%	92.70 ± 0.81%
	M2	92.31 ± 0.35%	99.80 ± 0.12%	92.45 ± 3.61%	90.15 ± 0.96%
	M3	91.94 ± 0.56%	99.90 ± 0.11%	84.32 ± 5.31%	90.31 ± 0.53%
GM	C	78.91 ± 0.28%	95.18 ± 5.69%	100.00 ± 0.00%	78.37 ± 0.14%
	A	75.04 ± 0.30%	84.64 ± 6.17%	96.10 ± 1.70%	77.96 ± 0.25%
	M1	76.80 ± 0.36%	93.13 ± 4.51%	98.37 ± 0.52%	77.04 ± 0.58%
	M2	77.39 ± 0.28%	95.70 ± 6.98%	96.17 ± 1.24%	77.20 ± 0.32%
	M3	77.54 ± 0.55%	95.27 ± 6.13%	97.99 ± 1.49%	76.69 ± 0.45%

VILLAIN performs well on different aggregation methods.

Impact of Hyperparameters

- ***Impact of poisoning rate.***
- ***Impact of server & participant models.***
- ***Impact of learning rate.***
- ***Impact of trigger size.***
- ***Impact of trigger magnitude.***
- ***Impact of number of candidates.***

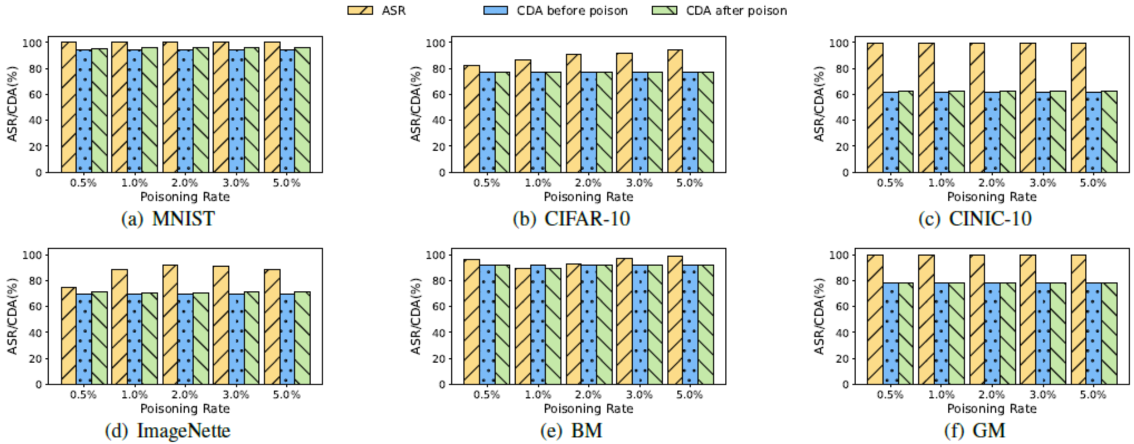


Figure 4: Impact of poisoning rate.

The backdoor attack still works even with a low poisoning rate of only 0.5%.

Impact of Hyperparameters

- *Impact of poisoning rate.*
- *Impact of server & participant models.*
- *Impact of learning rate.*
- *Impact of trigger size.*
- *Impact of trigger magnitude.*
- *Impact of number of candidates.*

Table 6: Impact of server models. dep.: model depth.

dep.	MNIST LIA	MNIST ASR	CIFAR-10 LIA	CIFAR-10 ASR
3	94.03 ± 2.56%	100.00 ± 0.00%	96.08 ± 4.28%	98.68 ± 0.59%
4	95.89 ± 2.95%	100.00 ± 0.00%	96.63 ± 3.55%	96.97 ± 0.45%
5	94.92 ± 2.63%	99.53 ± 0.24%	97.55 ± 3.97%	96.83 ± 0.24%
6	92.85 ± 4.10%	100.00 ± 0.00%	97.06 ± 1.73%	98.03 ± 0.58%
7	95.73 ± 2.66%	100.00 ± 0.00%	98.53 ± 2.66%	97.86 ± 0.13%
dep.	**CINIC-10 LIA**	**CINIC-10 ASR**	**BM LIA**	**BM ASR**
3	93.19 ± 3.05%	99.55 ± 0.62%	94.05 ± 4.82%	97.84 ± 2.57%
4	94.10 ± 2.56%	97.27 ± 1.43%	95.03 ± 5.93%	96.91 ± 0.92%
5	93.68 ± 1.41%	98.03 ± 0.20%	98.23 ± 0.96%	98.35 ± 0.47%
6	96.14 ± 3.02%	95.82 ± 3.94%	94.76 ± 2.59%	92.47 ± 1.69%
7	95.16 ± 3.97%	96.29 ± 3.46%	95.91 ± 2.49%	95.10 ± 0.82%
dep.	**ImageNette LIA**	**ImageNette ASR**	**GM LIA**	**GM ASR**
3	90.41 ± 2.18%	92.79 ± 1.58%	95.18 ± 5.69%	100.00 ± 0.00%
4	92.14 ± 3.06%	93.01 ± 1.65%	98.62 ± 0.63%	100.00 ± 0.00%
5	95.52 ± 3.45%	96.68 ± 0.94%	96.28 ± 3.10%	99.35 ± 0.20%
6	87.05 ± 7.49%	90.93 ± 3.69%	93.60 ± 4.60%	100.00 ± 0.00%
7	94.11 ± 2.46%	92.04 ± 0.75%	94.04 ± 3.63%	98.80 ± 0.94%

VILLAIN is robust to different server structures.

Possible Defenses

- **Label Inference Defense**
  - ***DPSGD***
  - ***Gradient compression***
  - ***Privacy-preserving Deep Learning***

DP-SGD

	MNIST			CIFAR-10			ImageNette	
ε	LIA	CDA	ε	LIA	CDA	ε	LIA	CDA
10	98.19%	95.57%	10	96.43%	75.83%	10	89.43%	66.19%
5	94.83%	96.57%	5	91.16%	64.09%	5	85.24%	61.90%
1	87.70%	84.30%	1	68.41%	53.79%	1	66.27%	46.73%
0.5	76.06%	68.06%	0.5	20.94%	26.47%	0.5	18.49%	21.07%
0.1	12.91%	17.63%	0.1	10.58%	8.04%	0.1	13.19%	9.60%

Gradient Compression

	MNIST			CIFAR-10			ImageNette	
comp. r.	LIA	CDA	comp. r.	LIA	CDA	comp. r.	LIA	CDA
1	100.00%	97.76%	1	95.29%	77.05%	1	92.55%	67.86%
0.8	97.69%	91.26%	0.8	91.61%	73.26%	0.8	89.71%	67.72%
0.5	92.64%	87.74%	0.5	86.72%	66.41%	0.5	77.83%	53.69%
0.3	86.82%	73.20%	0.3	80.51%	52.03%	0.3	62.29%	41.58%
0.15	20.73%	24.68%	0.15	17.12%	15.08%	0.15	10.59%	16.39%

PPDL

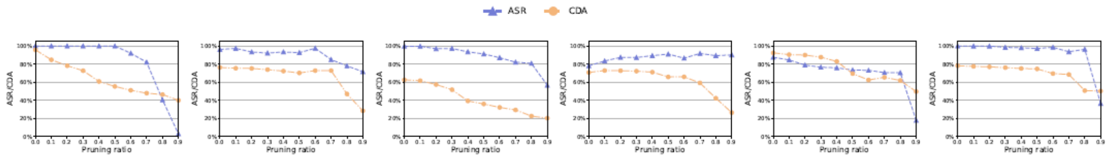
	MNIST			CIFAR-10			ImageNette	
θ	LIA	CDA	θ	LIA	CDA	θ	LIA	CDA
1	100.00%	94.51%	1	96.61%	76.92%	1	92.76%	69.91%
0.8	92.57%	92.62%	0.8	90.91%	69.05%	0.8	87.64%	70.51%
0.5	72.39%	63.14%	0.5	64.68%	53.92%	0.5	52.95%	60.59%
0.3	23.28%	12.61%	0.3	14.95%	17.61%	0.3	13.71%	13.40%
0.15	13.78%	10.26%	0.15	14.48%	11.94%	0.15	8.64%	10.04%

Villain can defeat existing label inference methods.

Possible Defenses

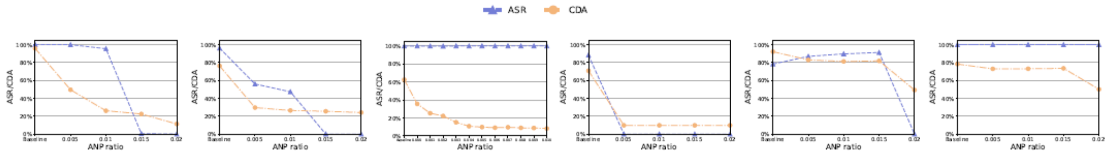
☐ *Backdoor Attack Defense*

- ***Model reconstruction***
- ***Sample preprocessing***
- ***Trigger synthesis***
- ***Poison suppression***



(a) MNIST (b) CIFAR-10 (c) CINIC-10 (d) ImageNette (e) BM (f) GM

Figure 5: Backdoor attack against defense with pruning.



(a) MNIST (b) CIFAR-10 (c) CINIC-10 (d) ImageNette (e) BM (f) GM

Figure 6: Backdoor attack against defense with ANP.

<u>Both trends prove the defense can not keep **high CDA** while reducing the ASR.</u>

Conclusion

- ➢ Design effective data poisoning strategies to strengthen the link between the trigger and the backdoor in the server model.
- ➢ Develop a new label inference algorithm to locate samples of the target label.
- ➢ Our attack is validated to be effective, robust, and efficient based on extensive experiments.

VILLAIN: Backdoor Attacks Against Vertical Split Learning

Thank you for your patience!

Contract us at:
baiyj@zju.edu.cn

USSLAB Website: www.usslab.org