UNIVERSITÀ DEGLI STUDI
DI SALERNO

SoK: Gradient Inversion Attacks in Federated Learning

Vincenzo Carletti, Pasquale Foggia, Carlo Mazzocca, **Giuseppe Parrella*** and Mario Vento

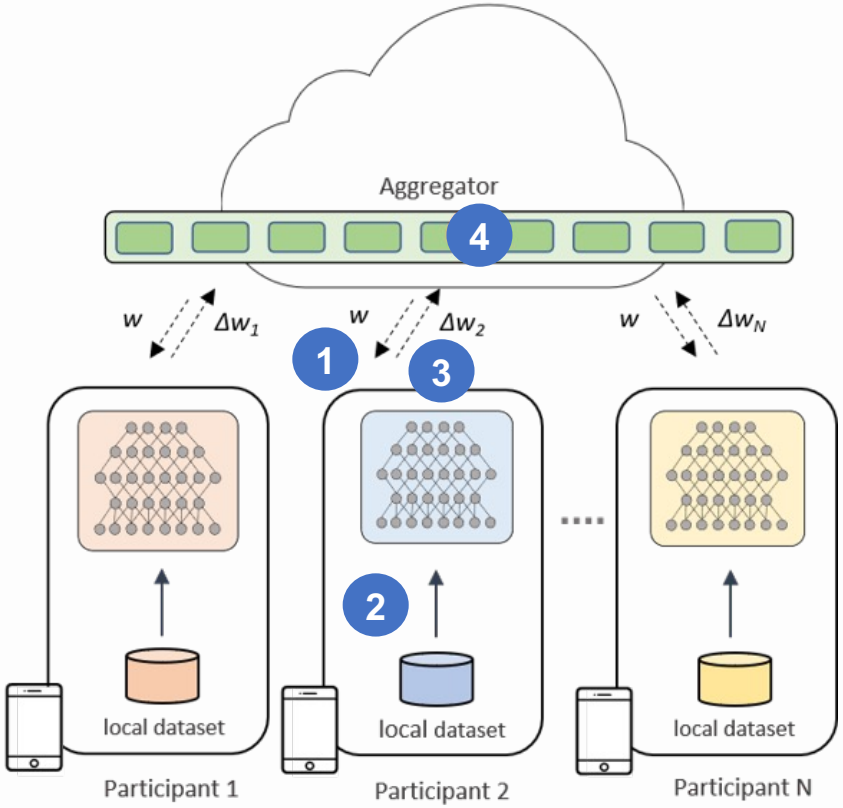
{vcarletti, pfoggia, cmazzocca, gparrella, mvento} @unisa.it*

University of Salerno

34th USENIX Security Symposium
August 2025

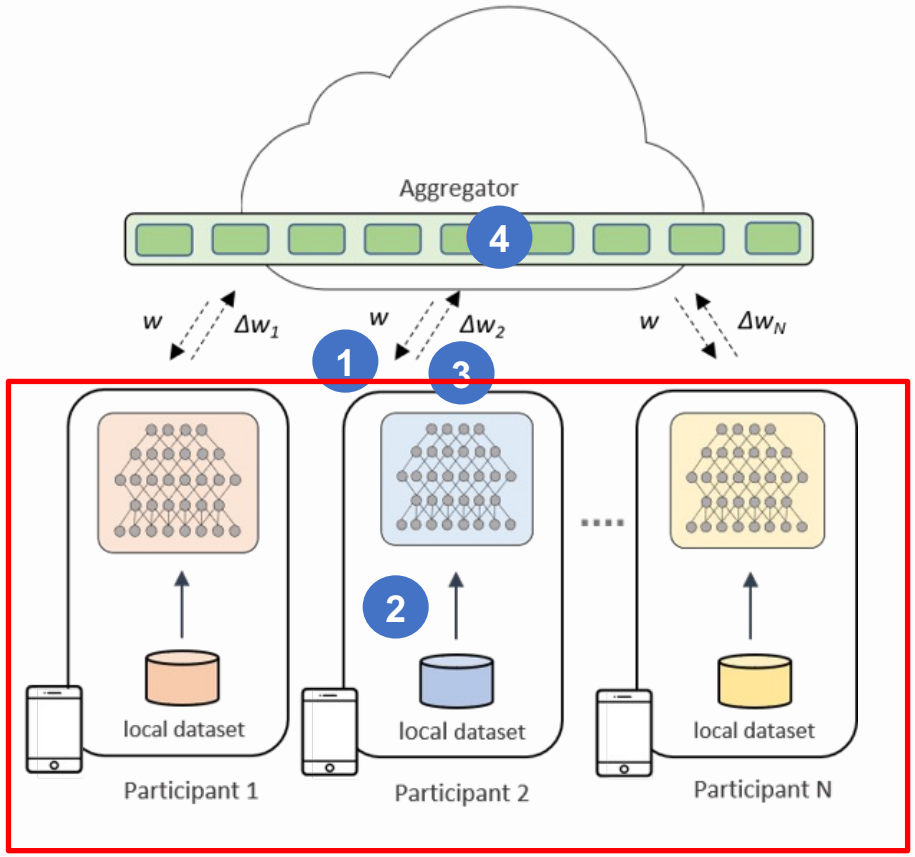
Federated Learning

Federated Learning (FL) is an emerging machine learning paradigm that allows multiple clients to train models while keeping their data private



Federated Learning

Federated Learning (FL) is an emerging machine learning paradigm that allows multiple clients to train models while keeping their data private

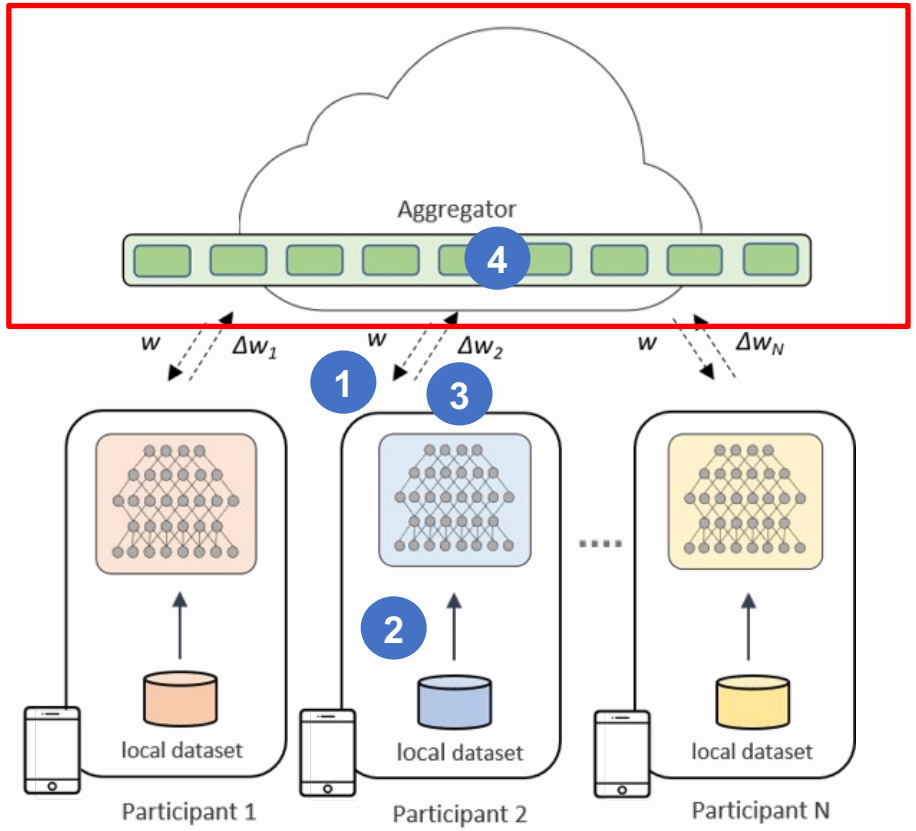


Clients

- Heterogeneous devices (e.g., smartphones, IoT, edge nodes)
- They never share their raw data during training

Federated Learning

Federated Learning (FL) is an emerging machine learning paradigm that allows multiple clients to train models while keeping their data private

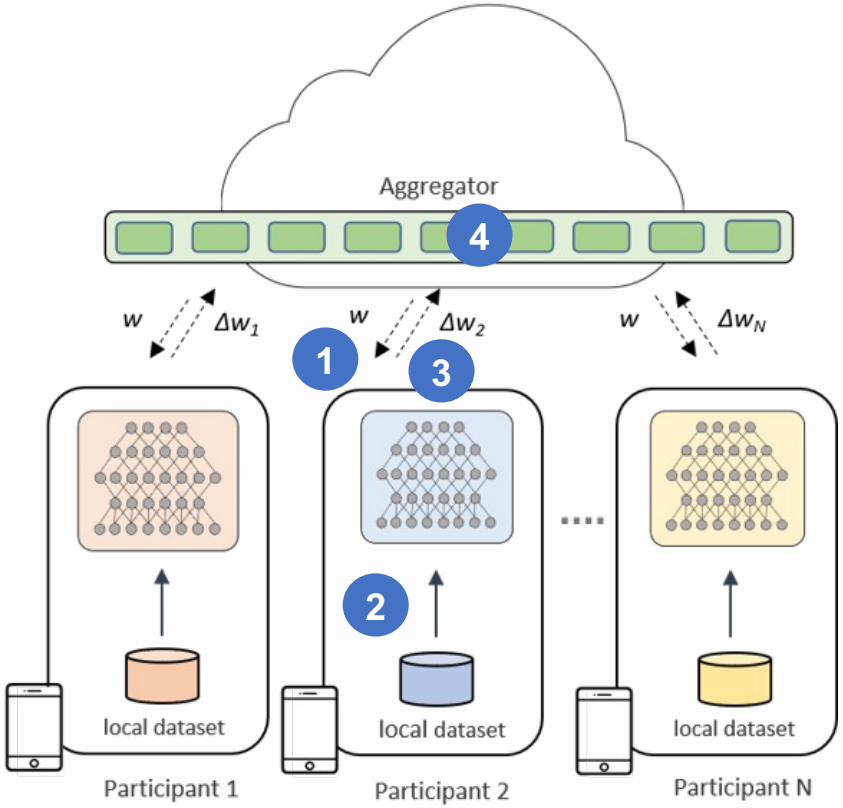


Server

- Central coordinator of the training process
- Does not have direct access to client data

Federated Learning

Federated Learning (FL) is an emerging machine learning paradigm that allows multiple clients to train models while keeping their data private

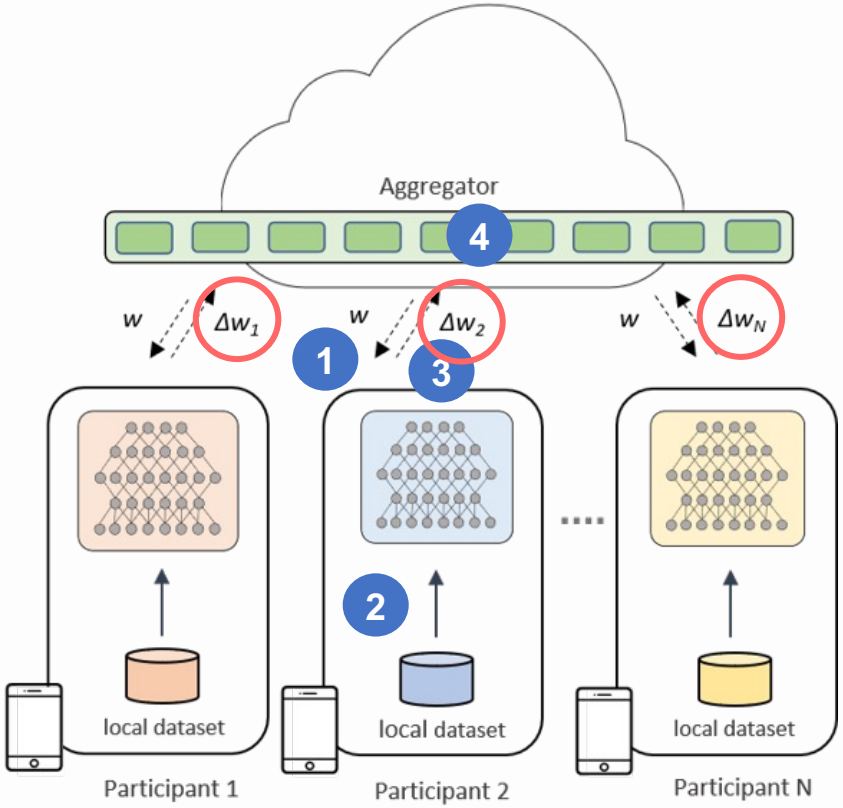


Training Process

1. Global model is broadcasted to the selected clients
2. Client perform local training
3. The clients send back a **model update**
4. The server gathers the updates and aggregate them

Gradient Inversion Attacks

Gradient Inversion Attacks (GIAs) aims to **reconstruct private training data from client updates,** thereby compromising the privacy guarantees that FL is designed to provide



A

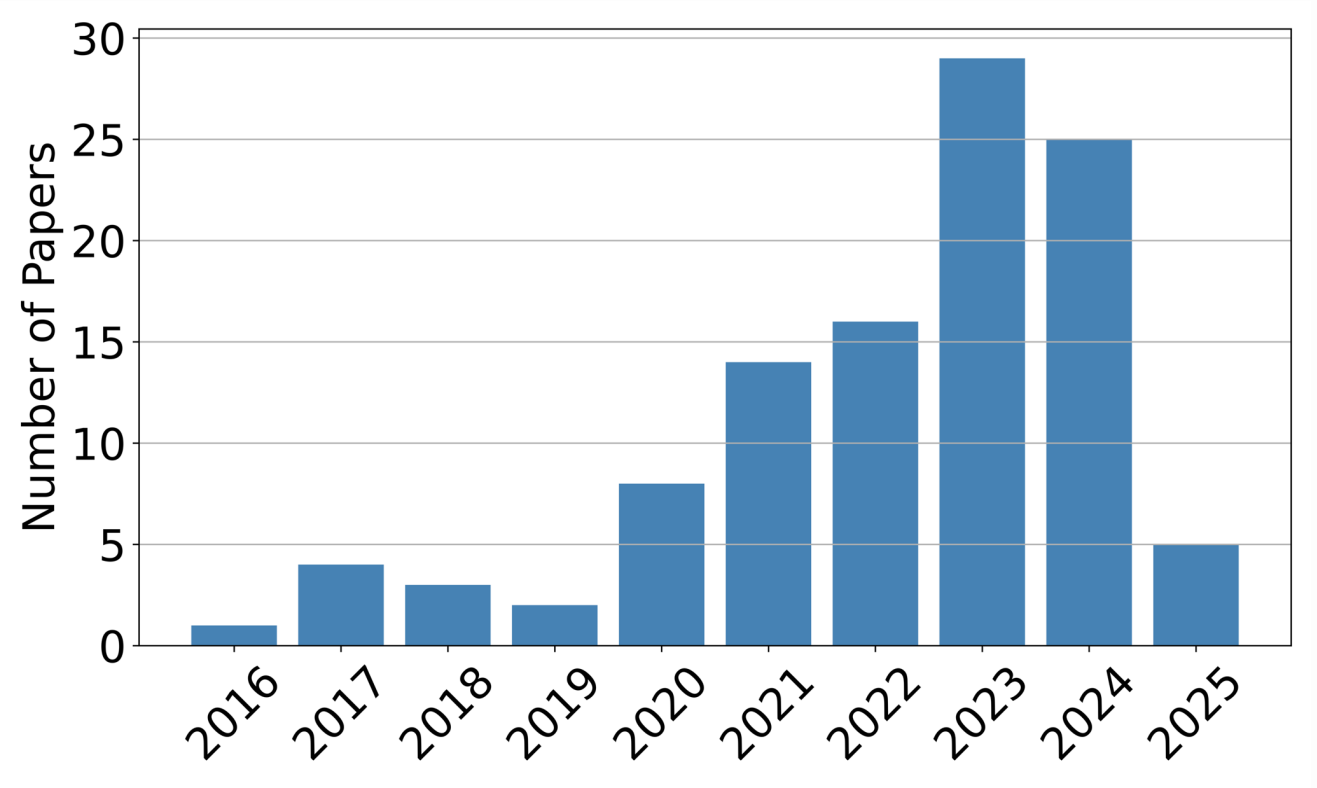
Image reconstructed with a GIA

B

Original client input

Reviewed Papers

We conduct a comprehensive Systematization of Knowledge of GIAs in FL, based on an extensive review of **107 publications** spanning from 2016 to 2025

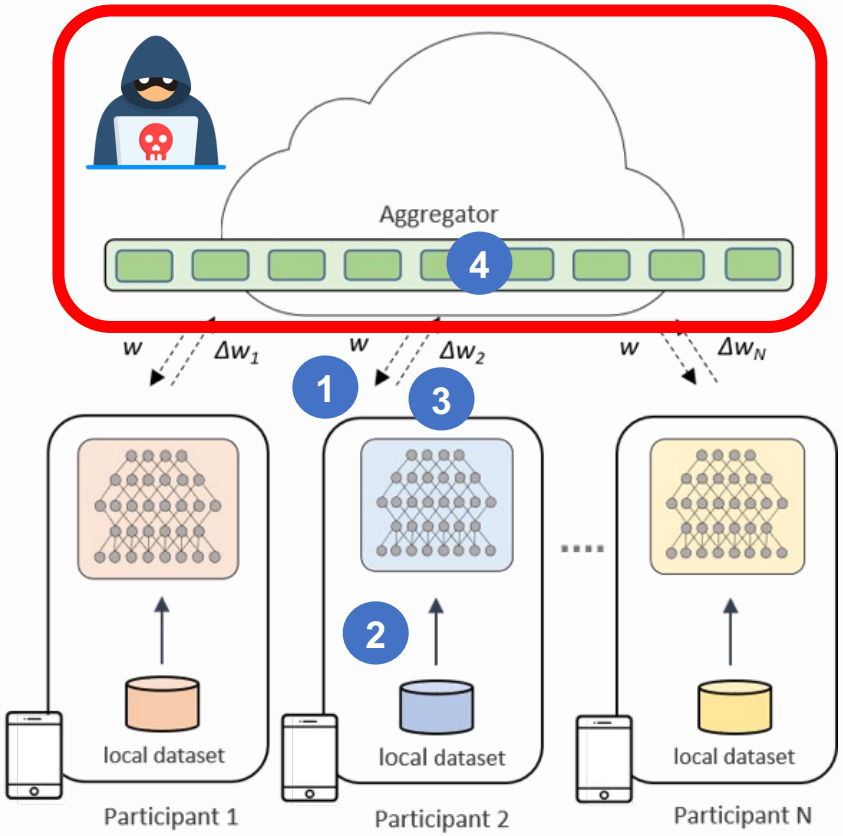


#1: Threat Model

We identify **8 threat models**, highlighting key adversary capabilities

Threat Models

The main threat actor considered in literature is the **central server**



- Previous works lack of a systematic analysis of threat models
- Only broader discussion of threat levels are provided

Threat Models

We identify **8 threat models**, highlighting key adversary capabilities

ID	Threat Model	Model Updates	Basic Knowledge	Training Details	Surrogate Data	Client Data Distribution	Active Manipulation	Client Selection	Real-World Applicability
A	Eavesdropper	✓	✗	✗	✗	✗	✗	✗	★★★
B	Informed Eavesdropper	✓	✓	✗	✗	✗	✗	✗	★★★
C	Parameter-Aware Eavesdropper	✓	✓	✓	✗	✗	✗	✗	★★★
D	Data-Enhanced Eavesdropper	✓	✓	✓	✓	✗	✗	✗	★★☆
E	Statistical-Informed Eavesdropper	✓	✓	✓	✓	✓	✗	✗	★★☆
F	Active Manipulator	✓	✓	✓	✗	✗	✓	✗	★☆☆
G	Data-Enhanced Manipulator	✓	✓	✓	✓	✗	✓	✗	★☆☆
H	Active Client Manipulator	✓	✓	✓	✓	✗	✓	✓	★☆☆

Table 1: Overview of the threat models defined in this work, categorized by adversary capabilities. Applicability in realistic settings is indicated as: ★★★: highly applicable and less detectable, ★★☆: potentially applicable (depends on specific configuration), and ★☆☆: less applicable and more detectable.

Passive Server

A **passive server** is unable to interfere with the standard FL algorithms

ID	Threat Model	Model Updates	Basic Knowledge	Training Details	Surrogate Data	Client Data Distribution	Active Manipulation	Client Selection	Real-World Applicability
A	Eavesdropper	✓	✗	✗	✗	✗	✗	✗	★★★
B	Informed Eavesdropper	✓	✓	✗	✗	✗	✗	✗	★★★
C	Parameter-Aware Eavesdropper	✓	✓	✓	✗	✗	✗	✗	★★★
D	Data-Enhanced Eavesdropper	✓	✓	✓	✓	✗	✗	✗	★★☆
E	Statistical-Informed Eavesdropper	✓	✓	✓	✓	✓	✗	✗	★★☆
F	Active Manipulator	✓	✓	✓	✗	✗	✓	✗	★☆☆
G	Data-Enhanced Manipulator	✓	✓	✓	✓	✗	✓	✗	★☆☆
H	Active Client Manipulator	✓	✓	✓	✓	✗	✓	✓	★☆☆

Table 1: Overview of the threat models defined in this work, categorized by adversary capabilities. Applicability in realistic settings is indicated as: ★★★: highly applicable and less detectable, ★★☆: potentially applicable (depends on specific configuration), and ★☆☆: less applicable and more detectable.

Passive Server

However, it may leverage auxiliary knowledge to refine reconstructions

ID	Threat Model	Model Updates	Basic Knowledge	Training Details	Surrogate Data	Client Data Distribution	Active Manipulation	Client Selection	Real-World Applicability
A	Eavesdropper	✓	✗	✗	✗	✗	✗	✗	★★★
B	Informed Eavesdropper	✓	✓	✗	✗	✗	✗	✗	★★★
C	Parameter-Aware Eavesdropper	✓	✓	✓	✗	✗	✗	✗	★★★
D	Data-Enhanced Eavesdropper	✓	✓	✓	✓	✗	✗	✗	★★☆
E	Statistical-Informed Eavesdropper	✓	✓	✓	✓	✓	✗	✗	★★☆
F	Active Manipulator	✓	✓	✓	✗	✗	✓	✗	★☆☆
G	Data-Enhanced Manipulator	✓	✓	✓	✓	✗	✓	✗	★☆☆
H	Active Client Manipulator	✓	✓	✓	✓	✗	✓	✓	★☆☆

Table 1: Overview of the threat models defined in this work, categorized by adversary capabilities. Applicability in realistic settings is indicated as: ★★★: highly applicable and less detectable, ★★☆: potentially applicable (depends on specific configuration), and ★☆☆: less applicable and more detectable.

Active Server

An **active server** may interfere with training procedure (e.g., by manipulating the shared model)

ID	Threat Model	Model Updates	Basic Knowledge	Training Details	Surrogate Data	Client Data Distribution	Active Manipulation	Client Selection	Real-World Applicability
A	Eavesdropper	✓	✗	✗	✗	✗	✗	✗	★★★
B	Informed Eavesdropper	✓	✓	✗	✗	✗	✗	✗	★★★
C	Parameter-Aware Eavesdropper	✓	✓	✓	✗	✗	✗	✗	★★★
D	Data-Enhanced Eavesdropper	✓	✓	✓	✓	✗	✗	✗	★★☆
E	Statistical-Informed Eavesdropper	✓	✓	✓	✓	✓	✗	✗	★★☆
F	Active Manipulator	✓	✓	✓	✗	✗	✓	✗	★☆☆
G	Data-Enhanced Manipulator	✓	✓	✓	✓	✗	✓	✗	★☆☆
H	Active Client Manipulator	✓	✓	✓	✓	✗	✓	✓	★☆☆

Table 1: Overview of the threat models defined in this work, categorized by adversary capabilities. Applicability in realistic settings is indicated as: ★★★: highly applicable and less detectable, ★★☆: potentially applicable (depends on specific configuration), and ★☆☆: less applicable and more detectable.

Active Server

An active server with surrogate data can deliver even more powerful GIAs

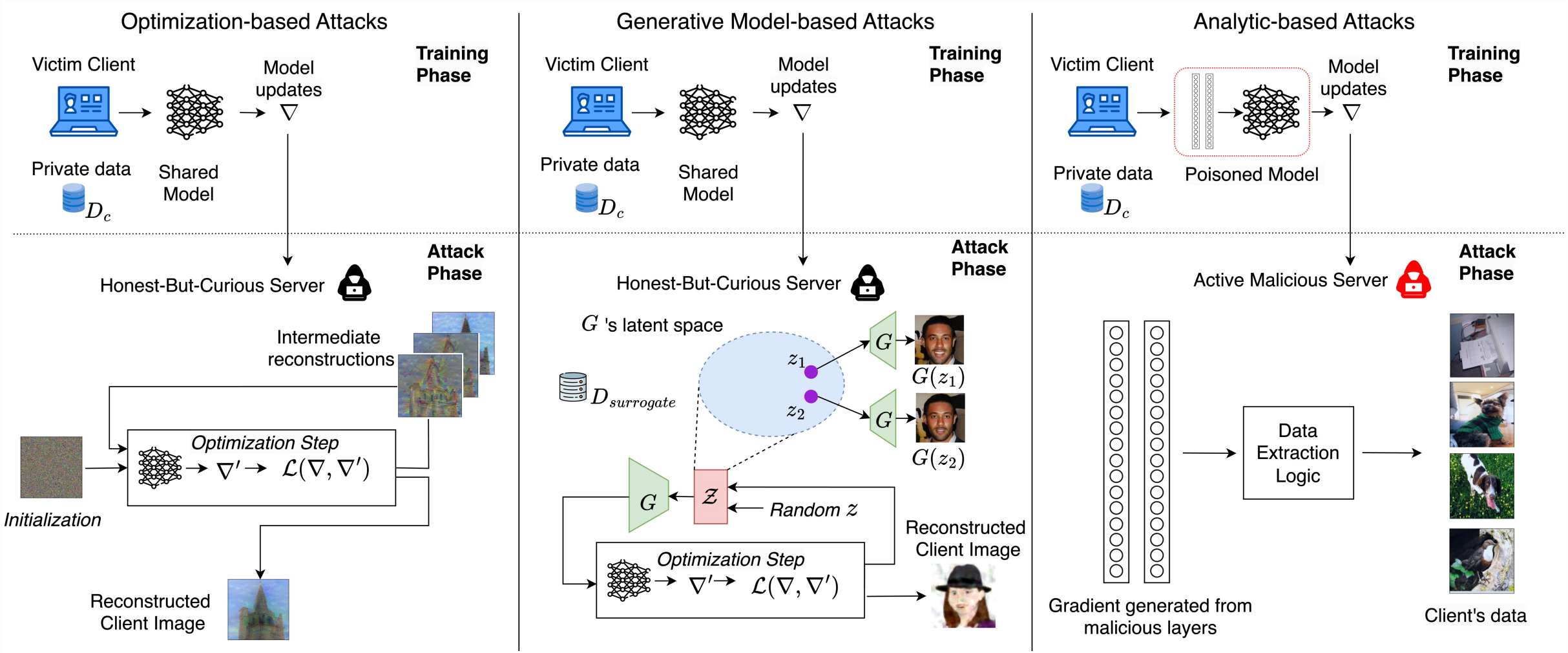
ID	Threat Model	Model Updates	Basic Knowledge	Training Details	Surrogate Data	Client Data Distribution	Active Manipulation	Client Selection	Real-World Applicability
A	Eavesdropper	✓	✗	✗	✗	✗	✗	✗	★★★
B	Informed Eavesdropper	✓	✓	✗	✗	✗	✗	✗	★★★
C	Parameter-Aware Eavesdropper	✓	✓	✓	✗	✗	✗	✗	★★★
D	Data-Enhanced Eavesdropper	✓	✓	✓	✓	✗	✗	✗	★★☆
E	Statistical-Informed Eavesdropper	✓	✓	✓	✓	✓	✗	✗	★★☆
F	Active Manipulator	✓	✓	✓	✗	✗	✓	✗	★☆☆
G	Data-Enhanced Manipulator	✓	✓	✓	✓	✗	✓	✗	★☆☆
H	Active Client Manipulator	✓	✓	✓	✓	✗	✓	✓	★☆☆

Table 1: Overview of the threat models defined in this work, categorized by adversary capabilities. Applicability in realistic settings is indicated as: ★★★: highly applicable and less detectable, ★★☆: potentially applicable (depends on specific configuration), and ★☆☆: less applicable and more detectable.

#2: Gradient Inversion Attacks

We categorize GIAs into three categories. For each attack, we identify the threat model, practical insights, and applicability

GIAs: Our Categorization



Optimization-based GIAs

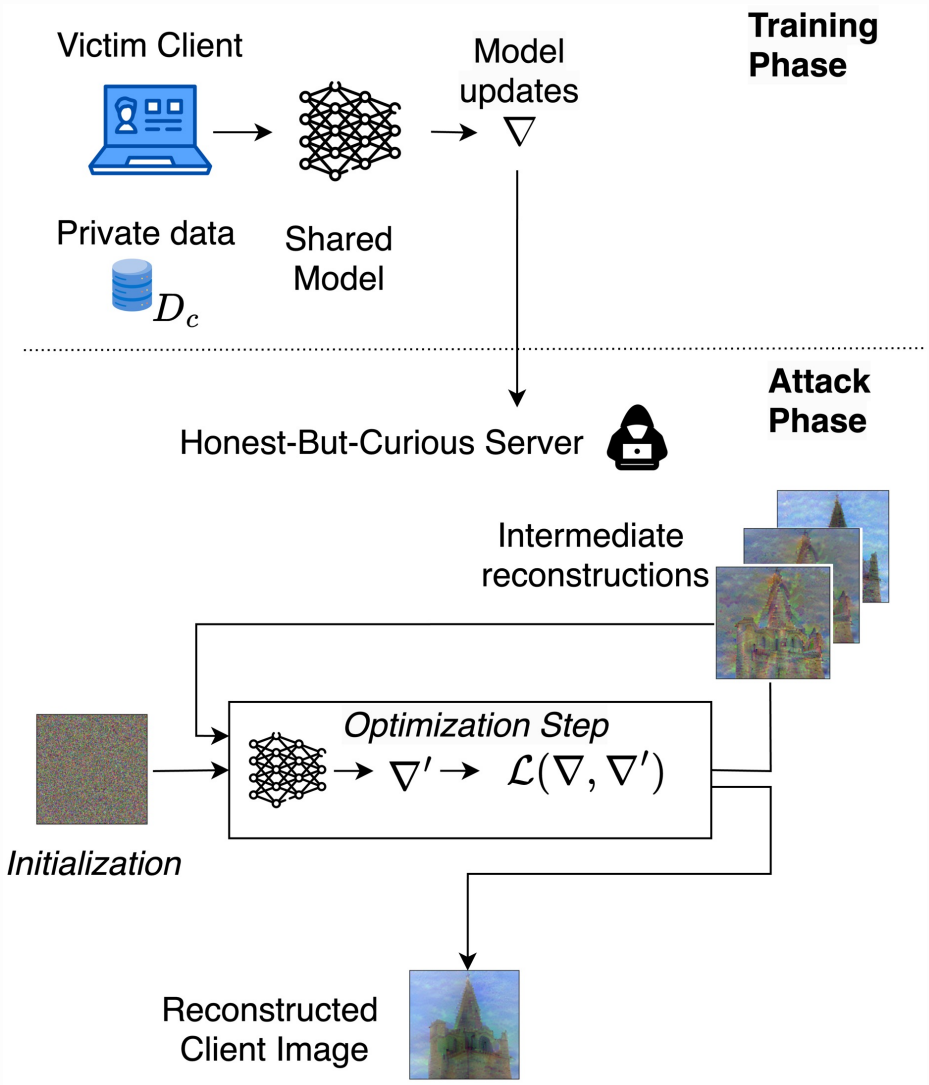
Input reconstruction is framed as an optimization problem with a **gradient-matching** objective:

$$X^*, Y^* = \arg\min_{X', Y'} \mathcal{L}_{\text{grad}}\left(\nabla_\theta \mathcal{L}(X, Y), \nabla_\theta \mathcal{L}(X', Y')\right) + \mathcal{R}_{\text{aux}}(X')$$

Recovered input and labels

Gradient-matching term

Auxiliary regularization term

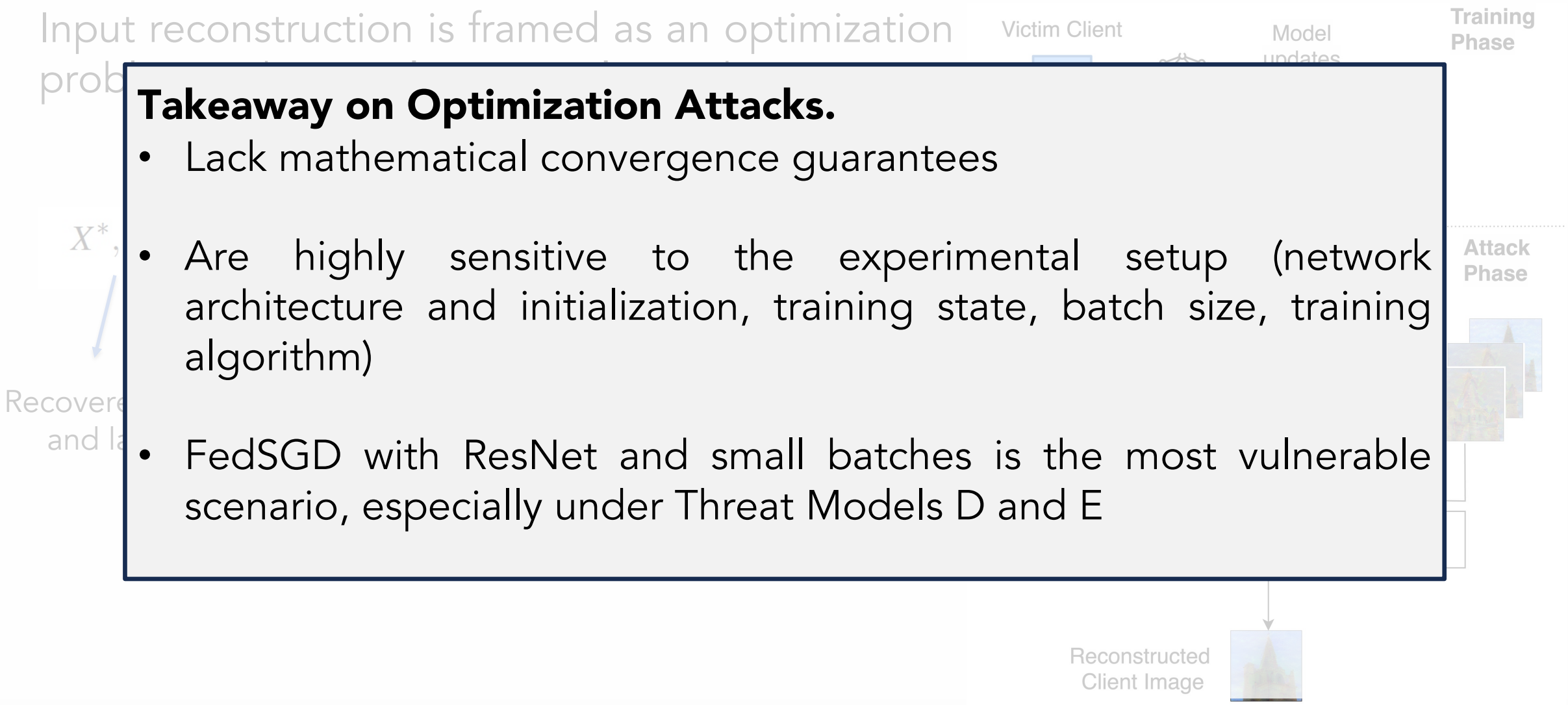


Optimization-based GIAs

Input reconstruction is framed as an optimization prob

Takeaway on Optimization Attacks.

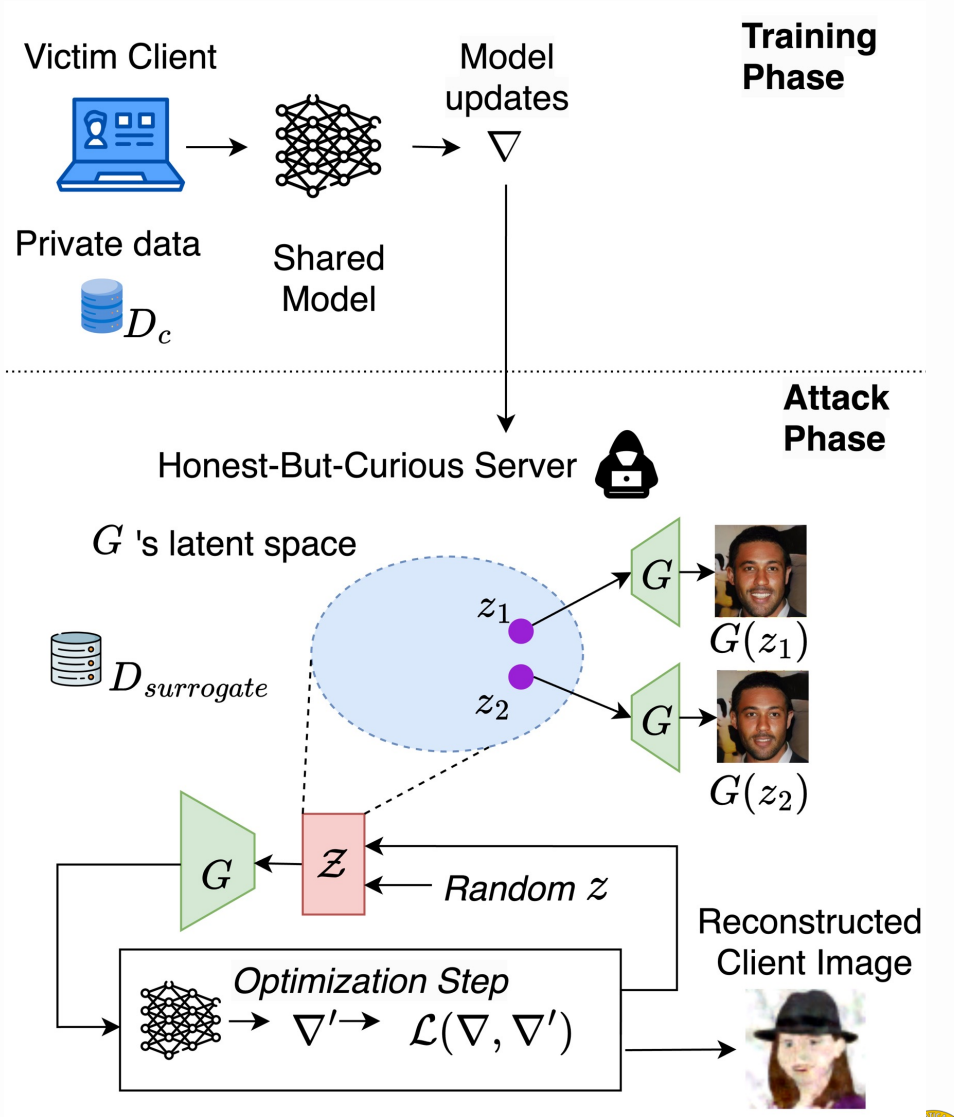
- Lack mathematical convergence guarantees
- Are highly sensitive to the experimental setup (network architecture and initialization, training state, batch size, training algorithm)
- FedSGD with ResNet and small batches is the most vulnerable scenario, especially under Threat Models D and E



Generative Model-based GIAs

In these attacks, generative model directly guide the reconstruction process

- **Online Optimization**
- **Latent-Space Optimization**
- **Direct Reconstruction**

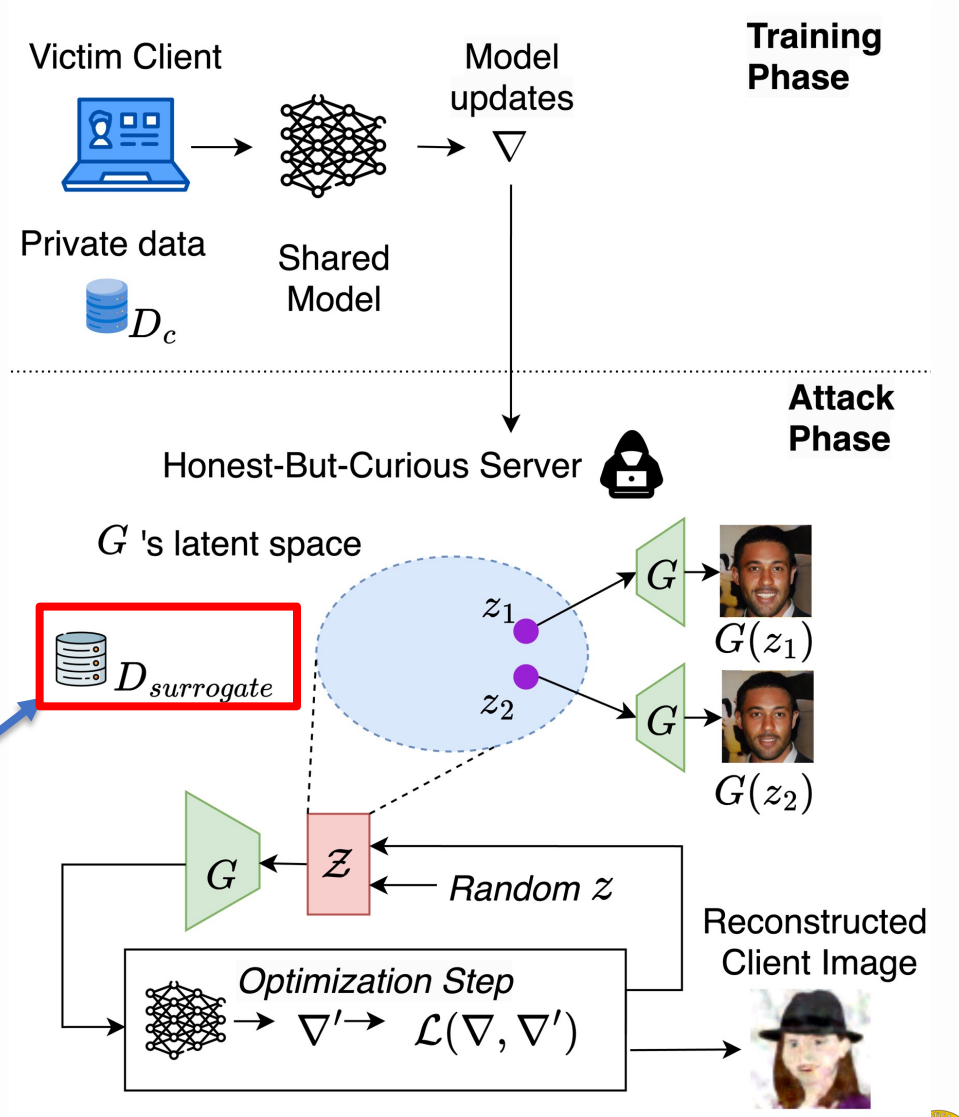


Generative Model-based GIAs

In these attacks, generative model directly guide the reconstruction process

- **Online Optimization**
- **Latent-Space Optimization**
- **Direct Reconstruction**

A surrogate dataset must be available to the attacker to train the generative model!



Generative Model-based GIAs

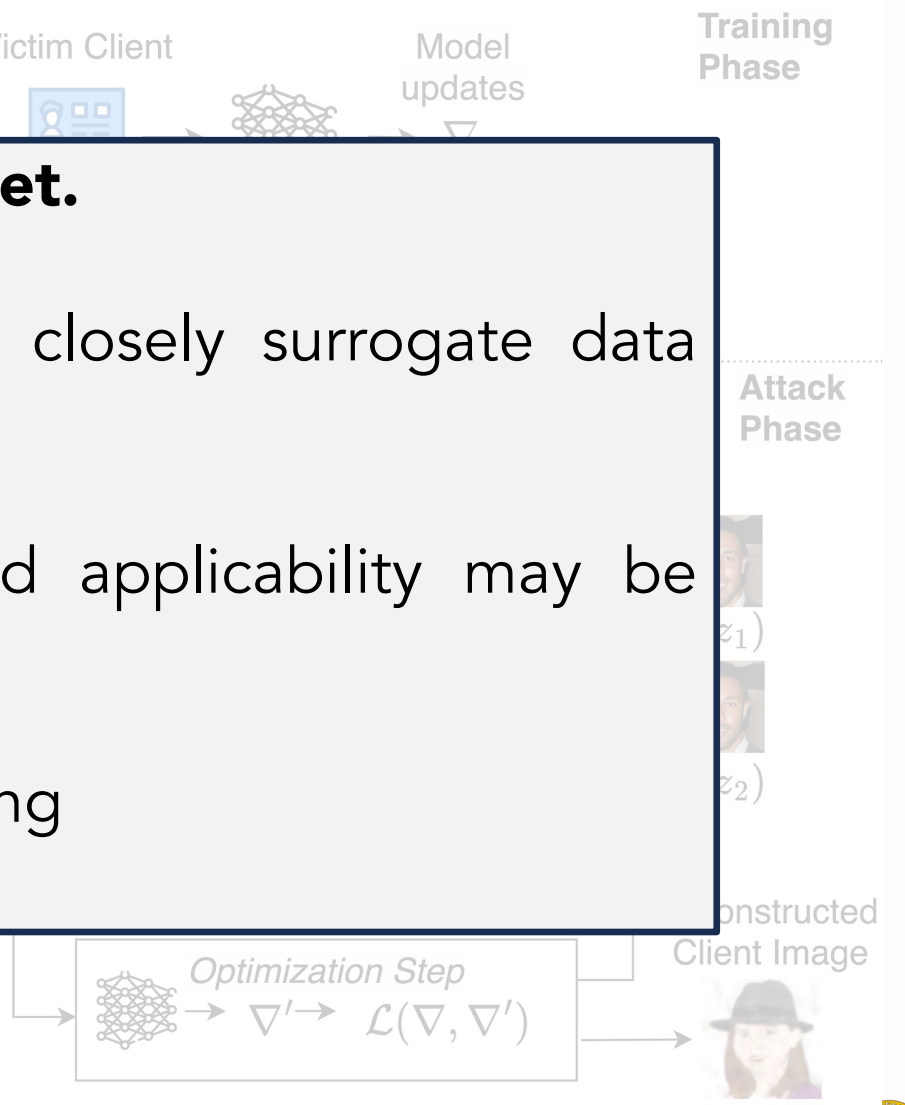
In these attacks, generative model directly guide the reconstruction process

- On
- Lat
- Dir

A sur
attacker to train the generative model!

Takeaway on Attacks with surrogate dataset.

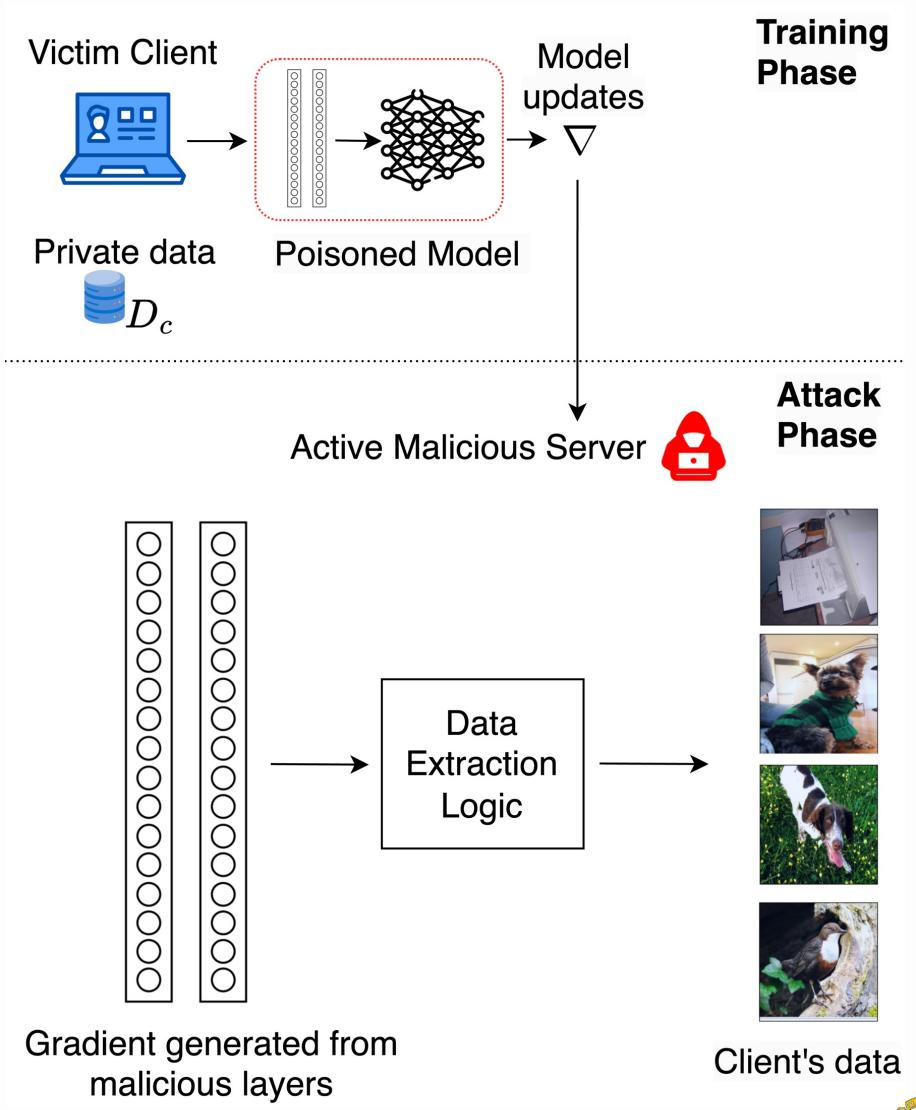
- Their real applicability depends on how closely surrogate data *aligns* with client data
- If strong similarity is required, real-world applicability may be limited
- A systematic analysis of this aspect is missing



Analytic-based GIAs

The active server tries to recover input data analitically through different strategies

- **Closed Form**
- **Gradient Sparsification**
- **Gradient Isolation**



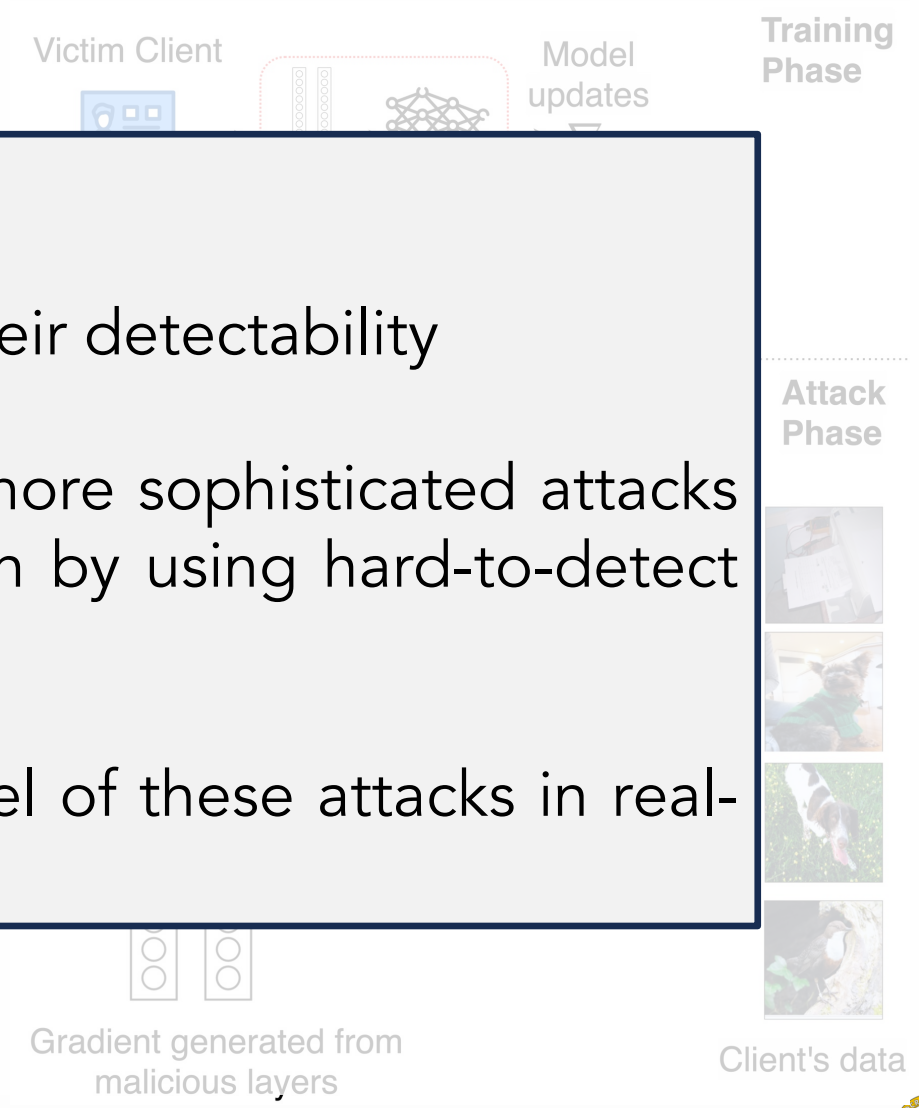
Analytic-based GIAs

The active server tries to recover input data analitically through different strategies

- Clo
- Gra
- Gra

Takeaway on Analytic Attacks.

- Historically deemed impractical due to their detectability
- Recent advancements have introduced more sophisticated attacks that achieve effective data reconstruction by using hard-to-detect model modifications
- This can increase the practical threat level of these attacks in real-world FL systems



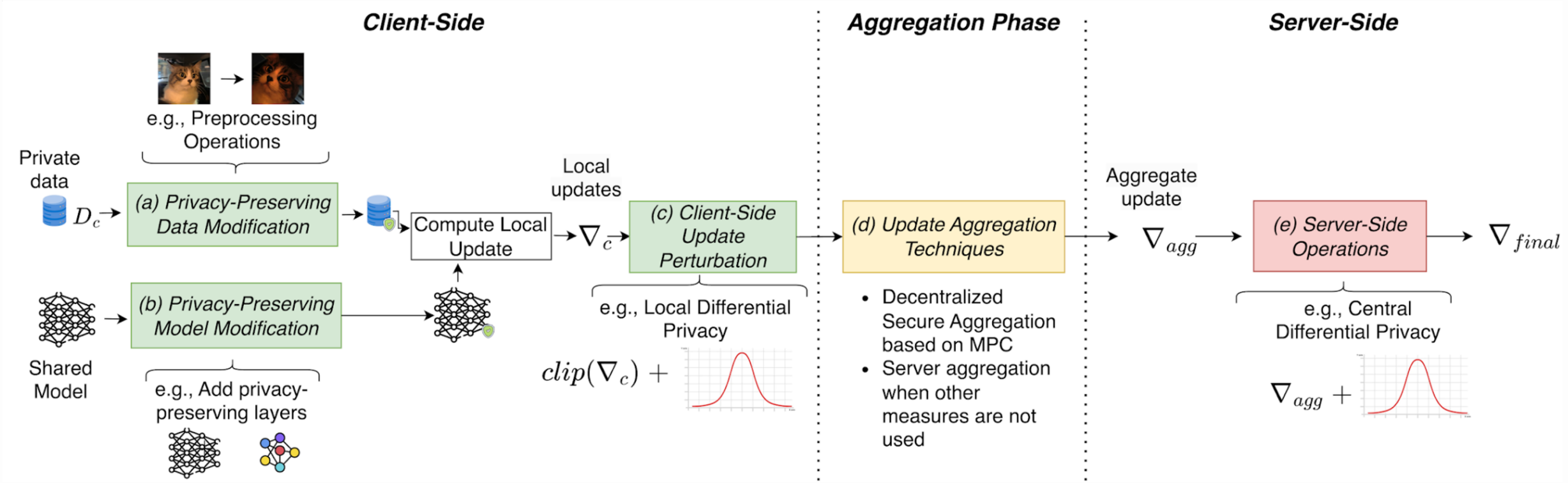
Categorization of GIAs

Category	Technique	Work	Year	Threat Model	Learning Algorithm	Image Resolution	Batch Size	Shared Model	Label Recovery	Open Source
Optimization-based	Basic Optimization	[109]	2019	A	FedSGD	64×64	8	LeNet	⋆	✓
		[105]	2020	A	FedSGD	32×32	1	LeNet	♦	✓
		[21]	2020	B	FedSGD	32×32	100	ResNets	○	✓
		[98]	2021	E	FedSGD	224×224	48	ResNet-50	♦	✗
		[27]	2022	E	FedSGD	224×224	30	ViT	[98]	✗
		[13]	2022	C	FedAVG	32×32	10×5	CNNs	[22]	✓
		[26]	2022	E	FedAVG	224×224	512×8	ResNet-18	⋆	✗
		[37]	2023	B	FedSGD	224×224	1024	VGG-16	⋆	✓
		[71]	2023	D	FedSGD	224×224	1	VGG, ResNet	[105]	✗
		[45]	2023	E	FedSGD	224×224	256	ResNet-50	♦	✓
		[96]	2024	E	FedSGD	224×224	8	ResNet-34	♦	✓
		[41]	2025	B	FedSGD	224×224	128	ResNet-18	[105]	✗
	Augmented Optimization	[94]	2023	D	FedSGD	32×32	1	LeNet, ResNet-18	[105]	✗
		[99]	2023	D	FedAVG	128×128	5×16	LeNet, ResNet-18	[105]	✓
		[70]	2024	D	FedSGD	64×64	4	ResNet-18	○	✓
		[50]	2025	D	FedSGD	N/A	1	N/A	○	✗
Generative Model-based	Online Optimization	[79]	2019	D	FedSGD	64×64	1	LeNet, ResNet-18	[105]	✗
	Direct Reconstruction	[61]	2022	B	FedSGD	64×64	256	LeNet, ResNet-18	⋆	✓
		[91]	2023	D	FedSGD	224×224	8	ResNet-50	○	✗
	Latent-Space Optimization	[34]	2021	D	FedSGD	64×64	4	ResNet-18	○	✓
		[46]	2022	D	FedSGD	224×224	1	ResNet-18	[105]	✓
		[17]	2023	D	FedSGD	64×64	1	ResNet-18	[105]	✓
		[90]	2023	D	FedSGD	128×128	1	LeNet, ResNet-18	[105]	✗
		[25]	2024	D	FedSGD	256×256	1	LeNet-7, ResNet-18	-	✗
Analytic-based	Closed Form	[108]	2021	B	FedSGD	64×64	1	6 layer CNN	♦	✓
		[51]	2022	B	FedSGD	224×224	1	ViT	[105]	✗
		[12]	2024	B	FedSGD	256×256	25	6 layer FC-NN	-	✗
	Gradient Sparsification	[18]	2022	G	Both	Input Size	256	Model Agnostic†	-	✓
		[84]	2022	F	FedSGD	Input Size	1	Model Agnostic	-	✗
		[8]	2023	G	Both	Input Size	100	FC Networks‡	-	∞
	Gradient Isolation	[7]	2023	H	FedSGD	Input Size	100	FC Networks‡	-	✗
		[106]	2023	G	FedAVG	Input Size	1×64	Model Agnostic†	-	✗
		[107]	2024	G	FedAVG	Input Size	5×8	Model Agnostic†	-	✓
		[75]	2024	F	FedSGD	Input Size	100	LeNet, VGG-16	-	✓
		[20]	2024	F	Both	Input Size	512	Model Agnostic†	-	✓
		[65]	2025	G	Both	Input Size	1024	Various	-	✓

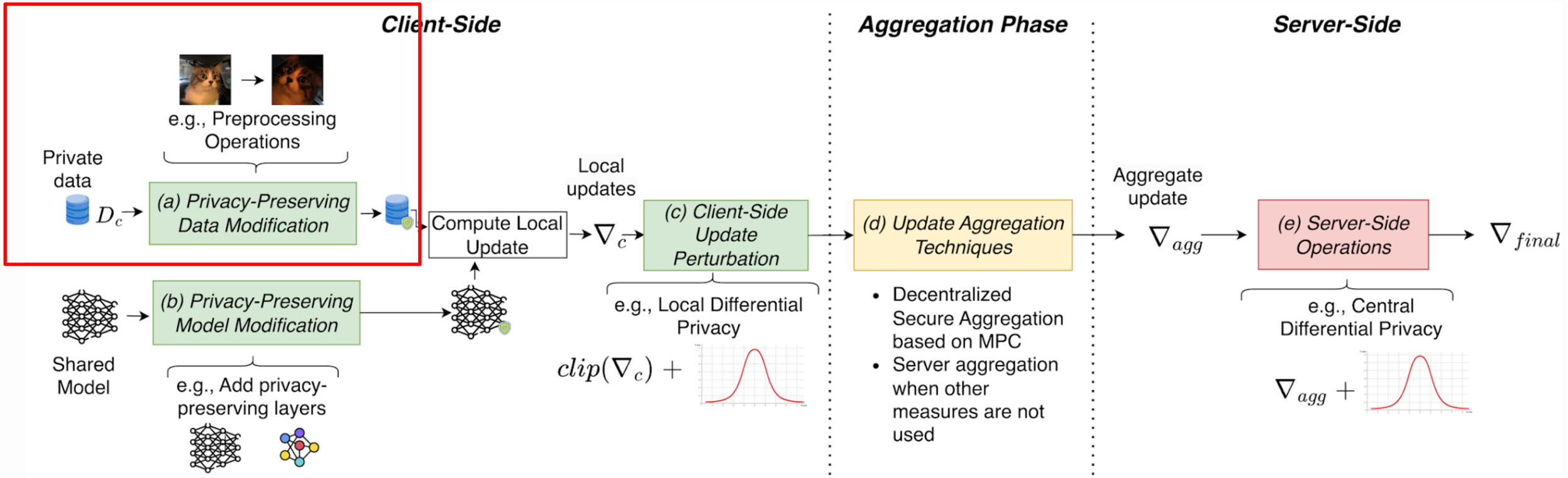
#3: Defensive Measures

We identify **two families of defenses** against GIAs, analyzing their robustness against the identified threat models

Defensive Measures



Client-Side Defensive Measures

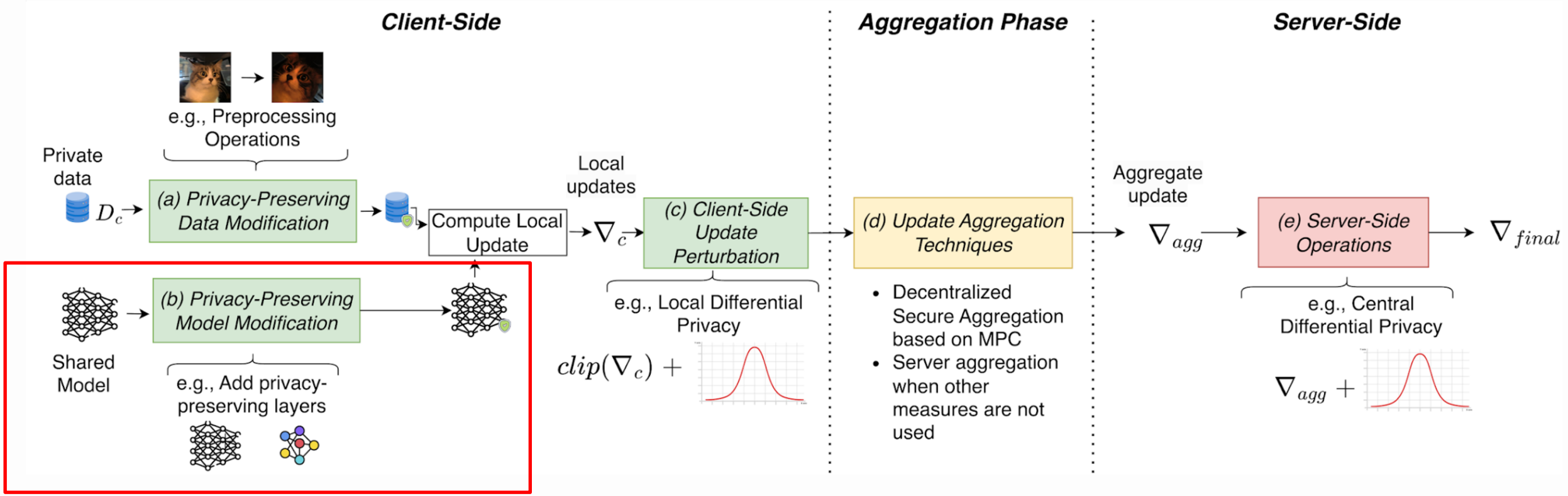


(a): Client locally modify input data

+ Does not rely on other clients or central server

– Usually impact on model performances

Client-Side Defensive Measures

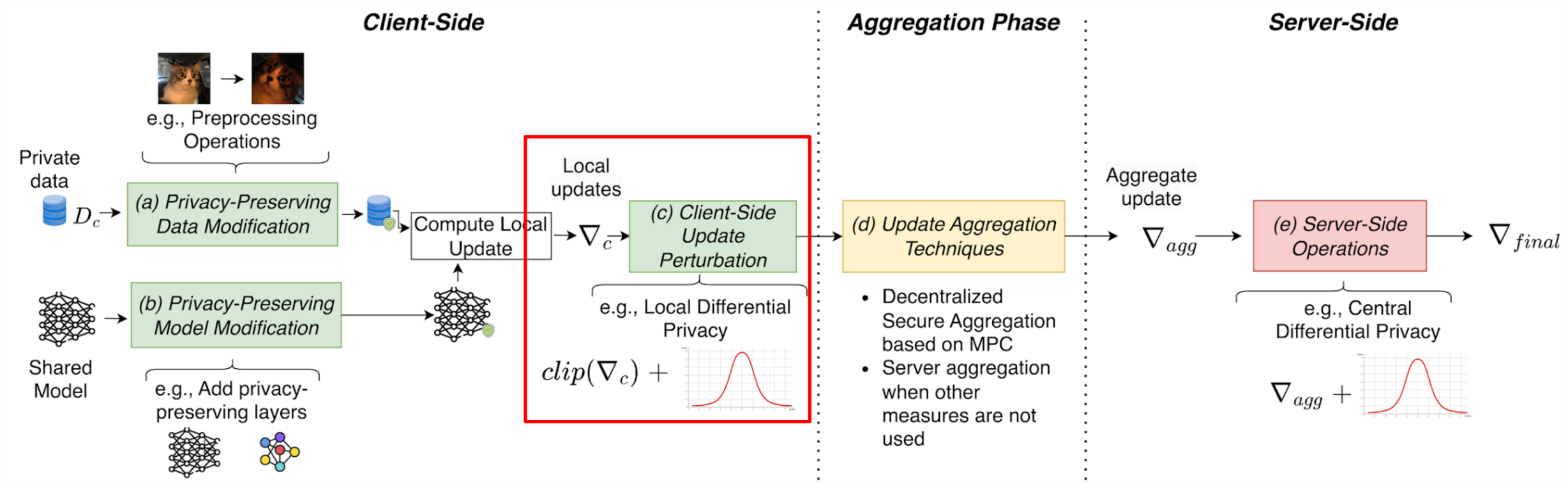


(b): Client locally modify the model

\+ No server or other trusted clients are required

\- Not evaluated against recent GIAs

Client-Side Defensive Measures

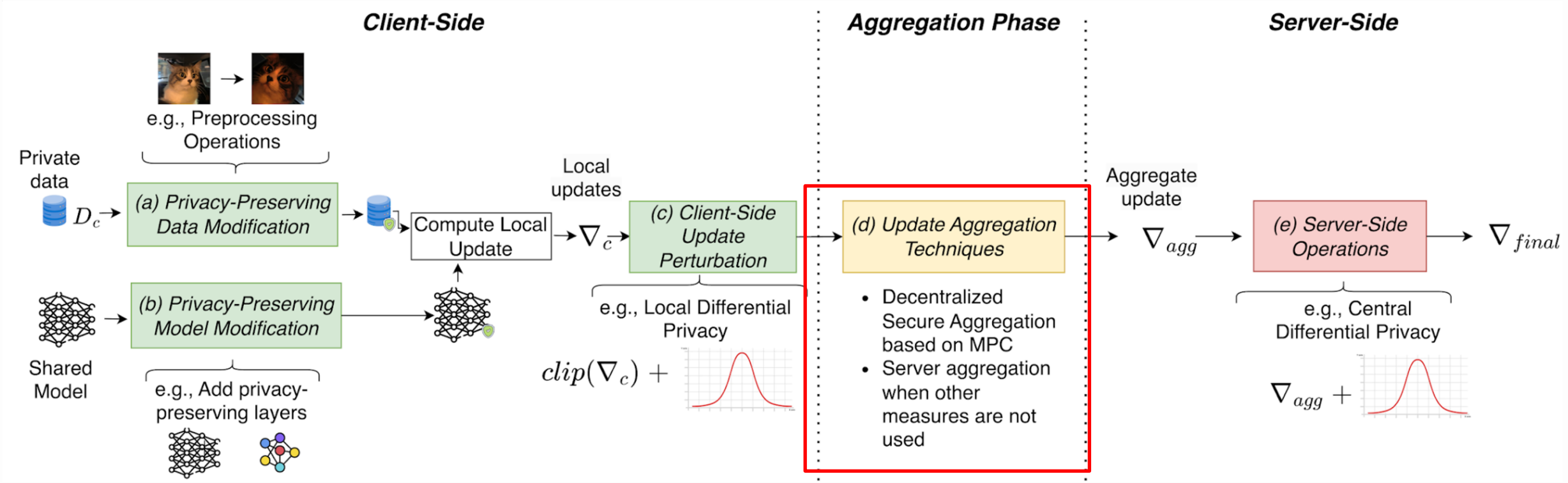


(c): Client locally modify the model update

+ No server or other trusted clients are required

– May significantly compromise model utility

Aggregation Defensive Measures

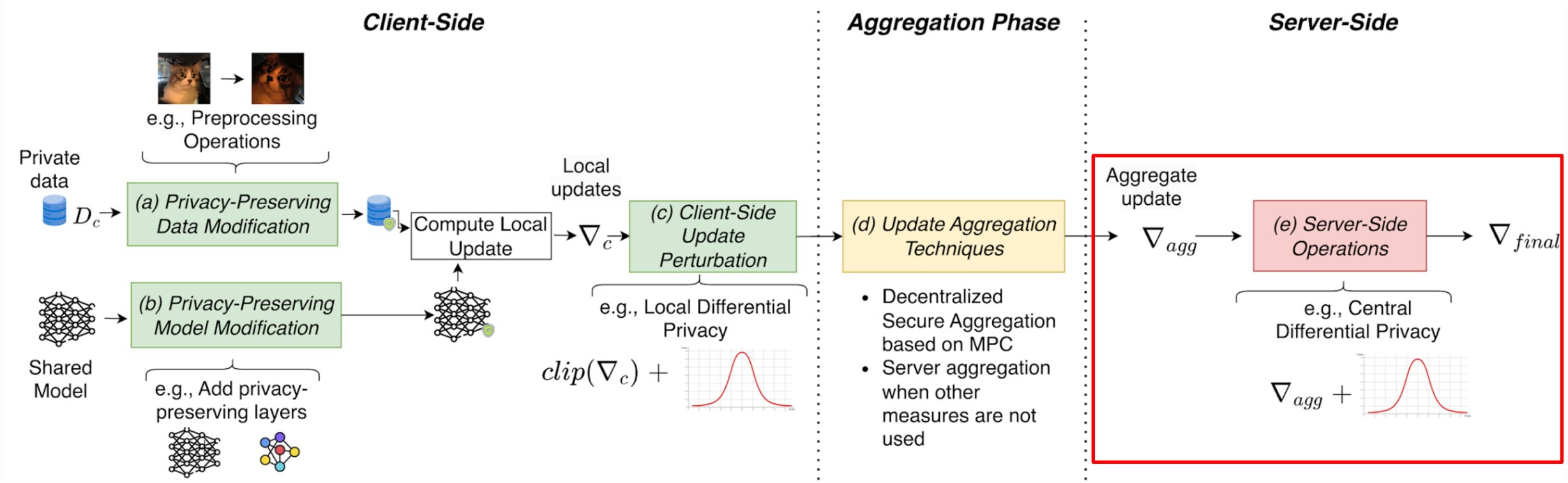


(d): Client aggregate updates to secure the *individual* updates from server

\+ No trusted server required

\- Introduce communicational overhead

Server-Side Defensive Measures



(e): Server manipulate aggregate update

+ Less impact on model utility with respect to LDP

– The server must be trusted

Categorization of Defensive Measures

Category	Technique	Work	Where	Threat Models	Intuition	Main Weakness	Open Source
Formal-Defenses	DP-based	CDP [55]	Server	A B C D E F G H	Server adds noise to clipped client contributions	Requires trusted (passive) server and ideal sampling conditions	~
		LDP [82, 83, 93]	Client	A B C D E F G H	Clients add noise to their own updates	Significantly compromises model utility; May be weakened from tailored GIAs [17, 30, 99]	~
	Cryptography-Based	SA [56]	Client	A B C D E F G H	Server has access to aggregated client contributions only	Vulnerable to active malicious servers; Adds communication overhead	~
		HE [9]	Client	A B C D E F G H	Enables computations on encrypted data without decryption	High computational and communication overhead	~
Heuristic Defenses	Gradient Perturbation	Pruning [92, 99, 104]	Client	A B C D E F G H	Transmits only the most significant gradient elements	Bypassed by modern GIAs [17, 50, 91, 99]	~
		Quantization [99]	Client	A B C D E F G H	Reduces gradient precision with fewer bits	Bypassed by modern GIAs [17, 50, 91, 99]	~
		Clipping [47]	Client	A B C D E F G H	Limits the magnitude of gradients	Bypassed by modern GIAs [17, 50, 91, 99]	~
		Sun et al. (2021) [68]	Client	A B C D E F G H	Perturbs data representation in FC layer to modify gradient pattern	Bypassed by modern GIAs [17, 99]	✓
		Wang et al. (2022) [76]	Client	A B C D E F G H	Adds Gaussian noise to high-sensitivity components of model weights	Not tested against recent generative model-based GIAs	✓
		Wang et al. (2024) [74]	Client	A B C D E F G H	Adaptive noise injection with sensitivity-informed perturbation strategy	Not tested against recent generative model-based GIAs	✗
		Zhang et al. (2025) [101]	Client	A B C D E F G H	Perturb gradients in a subspace orthogonal to the original one	Not evaluated against attack with stronger threat model	✓
	Learning Algorithm Modification	Lee et al. (2021) [40]	Client	A B C D E F G H	Transforms data into dissimilar representations	Not tested against generative model-based GIAs	✗
		Jebreel et al. (2022) [33]	Client	A B C D E F G H	Slices and encrypts gradients between clients	Not tested against generative model-based GIAs	✓
		Wan et al.(2022) [73]	Client	A B C D E F G H	Dynamically modifies learning rate for each client to make gradient estimation difficult	Uncertain impact on optimization dynamics	✗
		Gao et al. (2023) [19]	Client	A B C D E F G H	Uses augmentation to balance privacy and utility	Vulnerable during early training phases [5]	✓
		Liu et al. (2023) [49]	Client	A B C D E F G H	Decomposes weight matrices into cascading submatrices creating nonlinear mapping between gradients and raw data	Not tested against generative model-based GIAs	✗
		Ye et al. (2024) [97]	Client	A B C D E F G H	Plug-and-play defense using vicinal distribution augmentation of training data	Not tested against generative model-based GIAs	✓
		Wu et al. (2024) [85]	Client	A B C D E F G H	Use visually different synthesized concealed samples to compute model updates	Introduce computational overhead to synthesize concealed images	✓
	Model Modification	Fan et al. (2020) [16]	Client	A B C D E F G H	Parallel branch with weights hidden from server	May be vulnerable to branch simulation scenarios or recent GIAs	✗
		Scheliga et al. (2022) [64]	Client	A B C D E F G H	Variational block adding randomness	Proven ineffective against advanced GIAs [99]	✓
		Ren et al. (2023) [62]	Client	A B C D E F G H	Extends model with branch hidden from server	May be vulnerable to branch simulation scenarios or recent GIAs	✓

#4: Evaluation Metrics

We provide **a novel taxonomy for existing metrics** employed to measure privacy leakage for GIAs

Evaluation Metrics

GIAs evaluation metrics are based on **image similarity** or **image recognition** to evaluate privacy leakage risk associated to such attacks

Assessment Type	Metric	DL-based	Privacy-specific	Key Aspect Evaluated
Image Similarity	PSNR	✗	✗	Measures reconstruction quality through signal-to-noise ratio
	MSE	✗	✗	Calculates the mean squared difference between original and reconstructed image pixels
	SSIM [80]	✗	✗	Evaluates structural similarity by considering brightness, contrast, and structural patterns
	LPIPS [102]	✓	✗	Measures visual similarity as perceived by humans using pre-trained neural network representations
	Jaccard [99]	✓	✗	Evaluates correspondence between semantic attributes extracted from images using neural networks
	SemSim [69]	✓	✓	Captures the subjective dimension of privacy violations based on human-annotated judgments
	AVD [58]	✗	✓	Quantifies structural information leakage through analysis of distances in gradient space
Image Recognition	IIP [18, 98]	✓	✓	Measures the probability of correctly matching a reconstructed image to its original source in the training set
	RDLV [26]	✗	✓	Quantifies additional information extracted through an attack beyond baseline knowledge by normalizing SSIM improvement

Evaluation Metrics

GIAs evaluation metrics are based on **image similarity** or **image recognition** to evaluate privacy leakage risk associated to such attacks

Takeaway on Evaluation Metrics.

- For Threat Models A to E, where GIAs aim to approximate user inputs, privacy-centric evaluation metrics are essential for accurately evaluating privacy leakage and defense effectiveness
- Traditional image similarity metrics often underestimate privacy risks in approximate reconstructions

Image Recognition	IIP [18, 98]	✓	✓	Measures the probability of correctly matching a reconstructed image to its original source in the training set
	RDLV [26]	✗	✓	Quantifies additional information extracted through an attack beyond baseline knowledge by normalizing SSIM improvement

#5: Open Challenges

We highlight **key challenges** and **promising future research directions** specific to GIAs in FL

Open Challenges

- **Extending GIAs beyond image classification.** The literature overlooks other vision tasks
- **Client-side defenses against active adversaries.** Defenses for adversaries who manipulate models stealthily on the client side are lacking
- **Impact of distribution shifts on GIAs.** Differences between surrogate and real data affect attack success
- **Limitations of current privacy metrics.** Existing metrics do not fully capture privacy risks

UNIVERSITÀ DEGLI STUDI
DI SALERNO

Thank you for your attention!

Contact: gparrella@unisa.it

GitHub:

34th USENIX Security Symposium
August 2025