CENSOR: Defense Against Gradient Inversion via Orthogonal Subspace Bayesian Sampling

Kaiyuan Zhang, Siyuan Cheng, Guangyu Shen, Bruno Ribeiro,

Shengwei An, Pin-Yu Chen†, Xiangyu Zhang, Ninghui Li

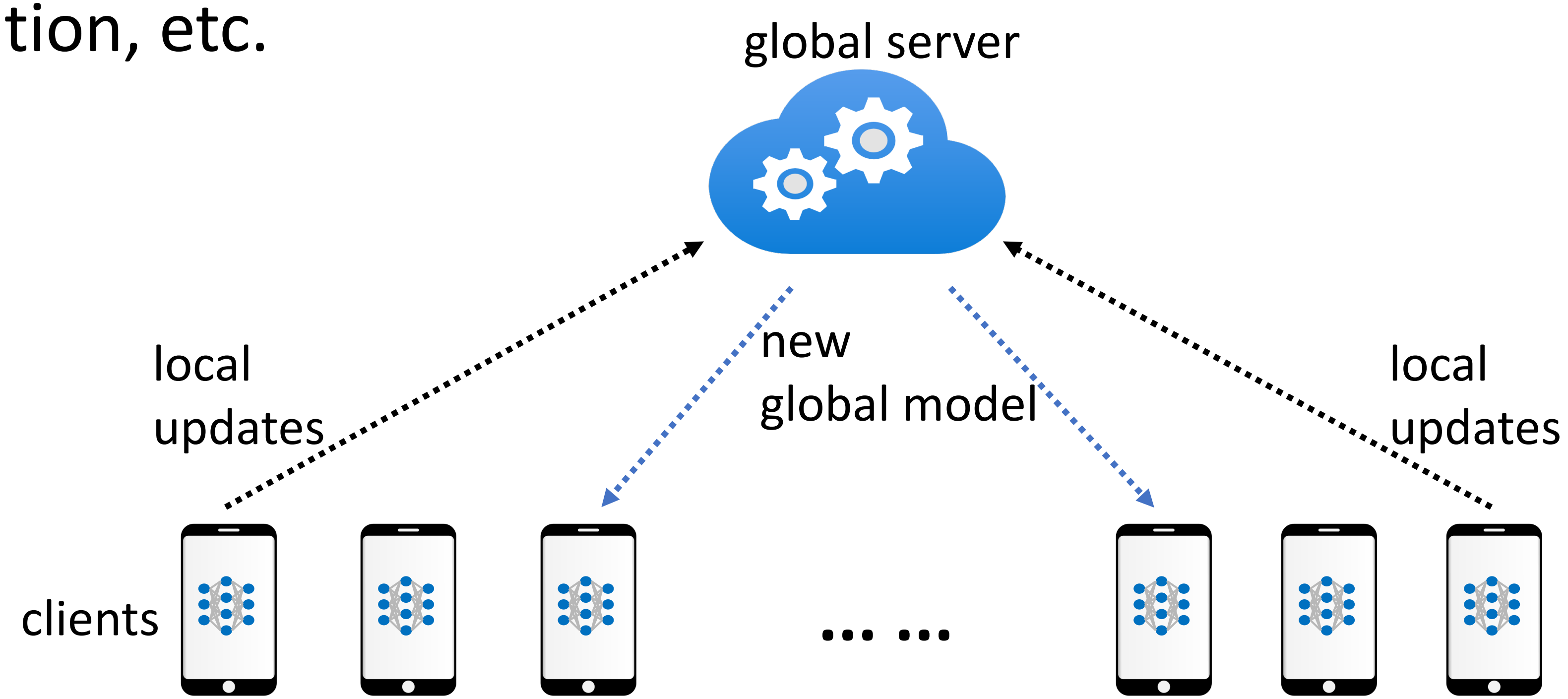
NDSS 2025

PURDUE UNIVERSITY

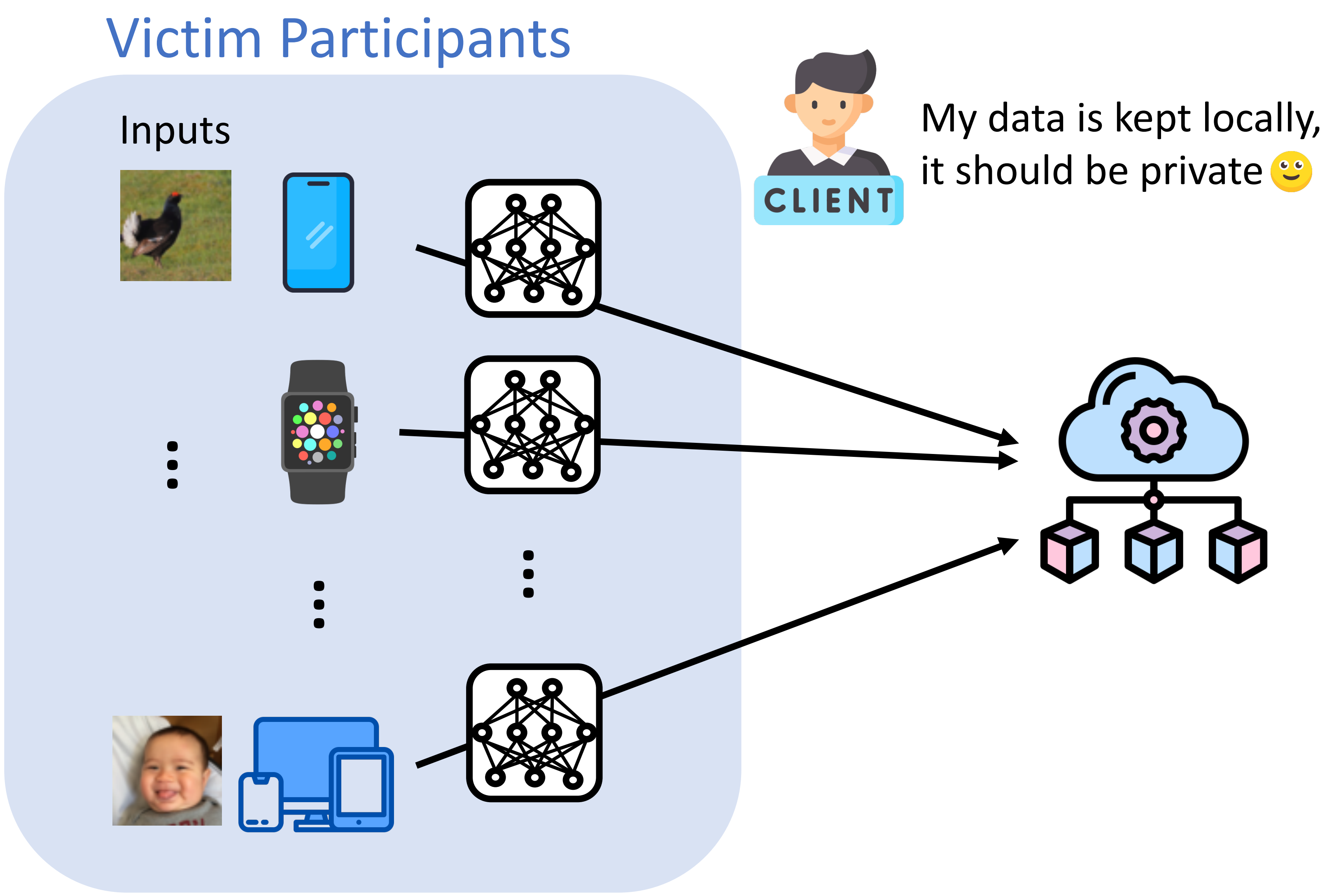
† IBM

Federated Learning

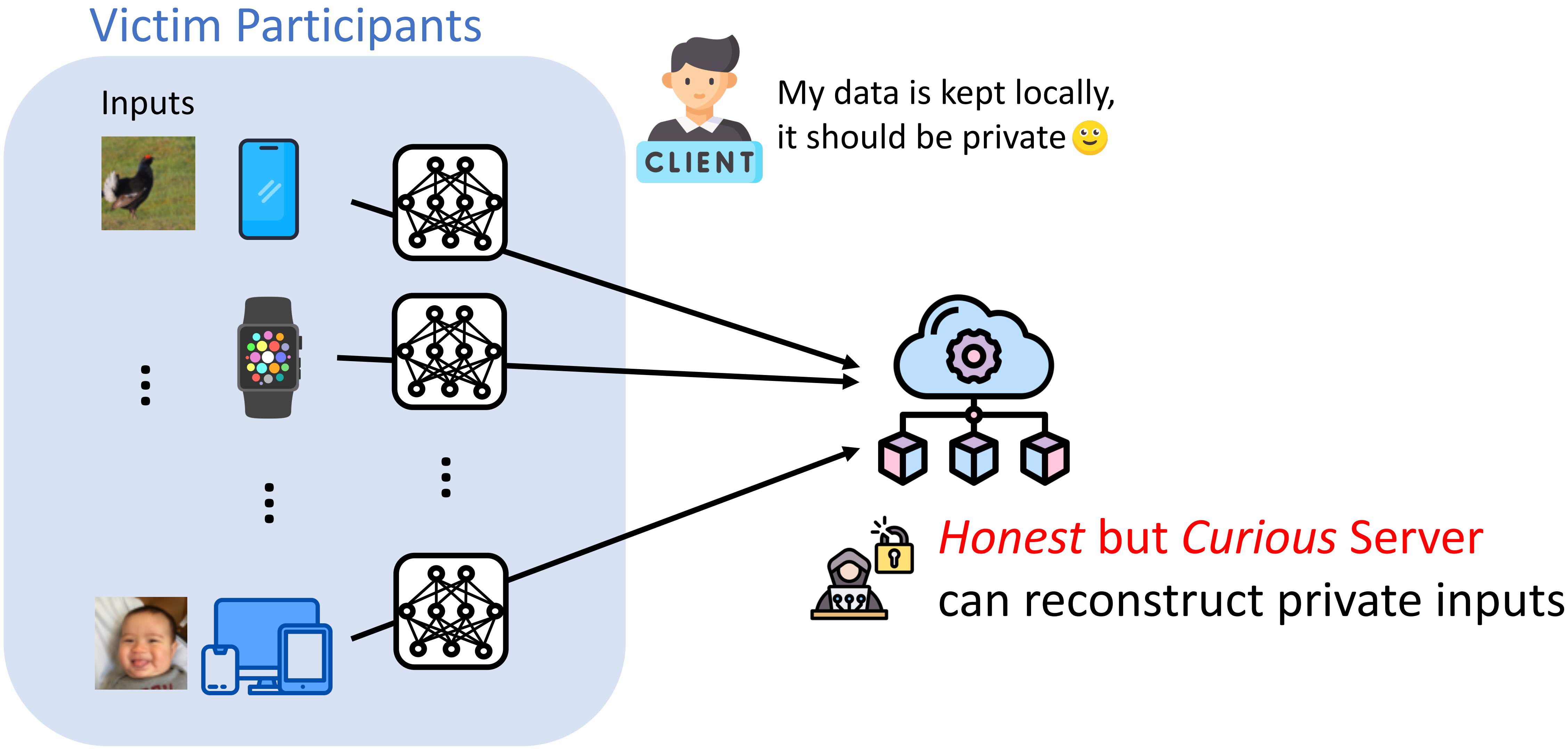
- A distributed learning paradigm that enables different parties to train a model together for high *quality* and strong *privacy protection*.
- Applications: next word prediction, credit prediction, and IoT device aggregation, etc.



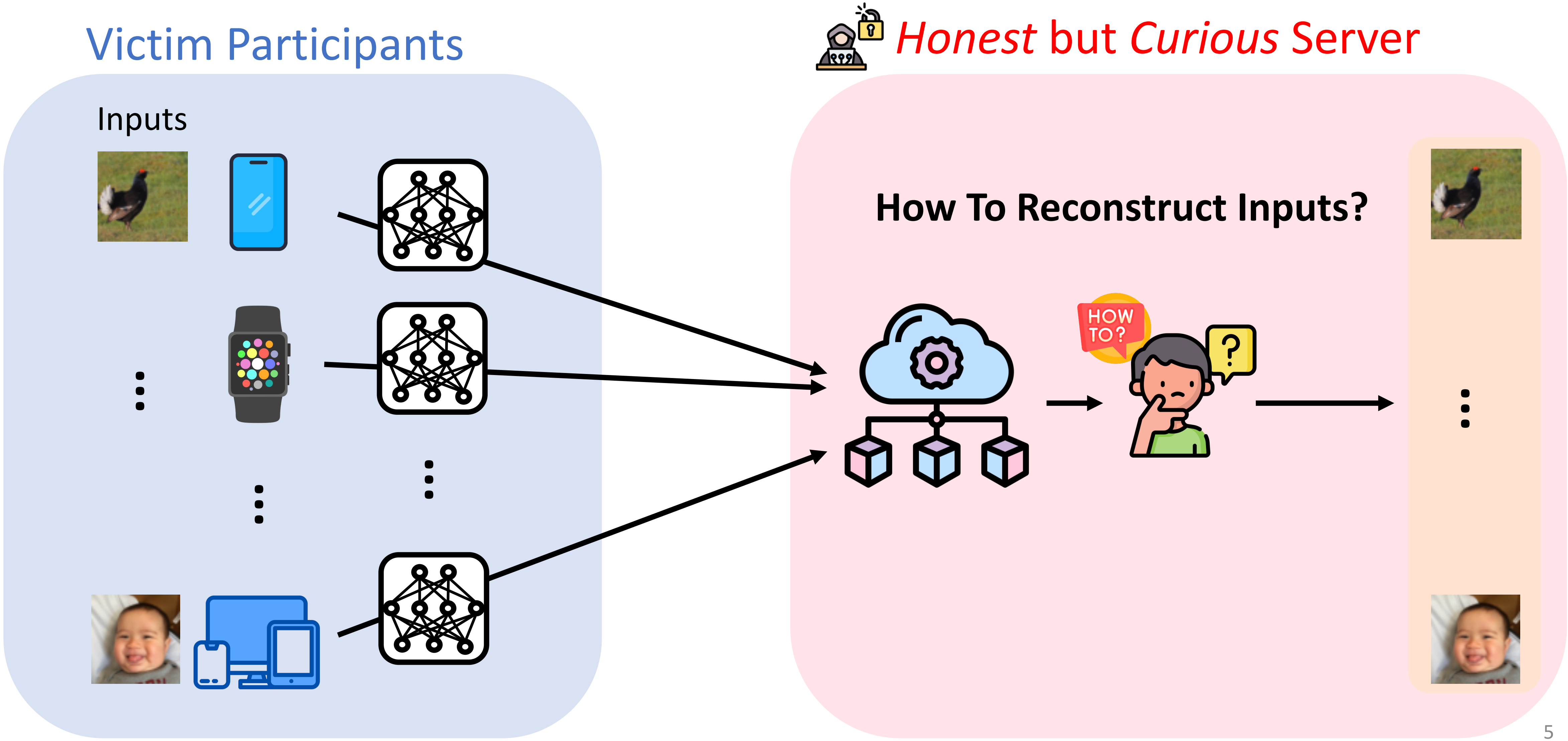
Is Your Data Really Private?



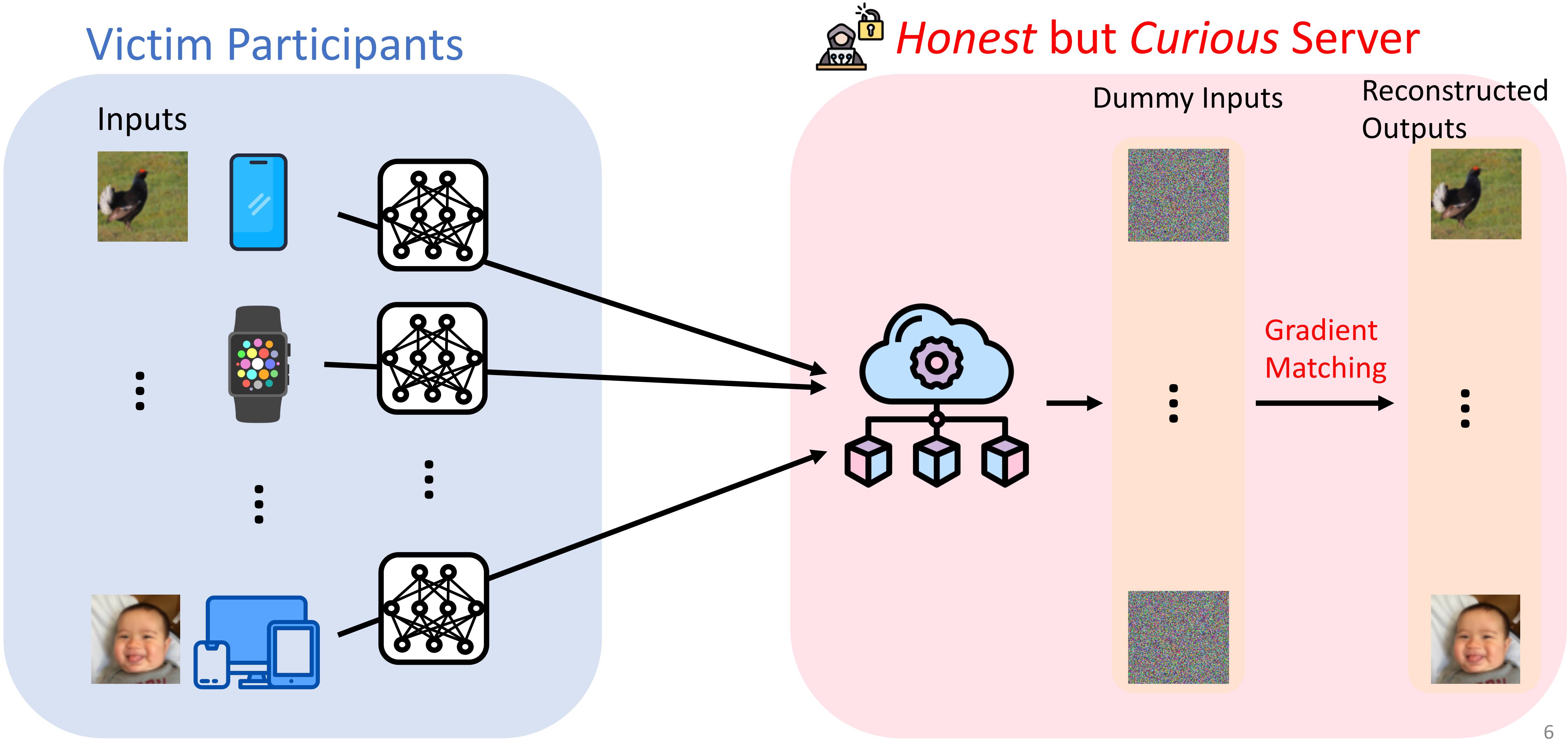
Is Your Data Really Private?



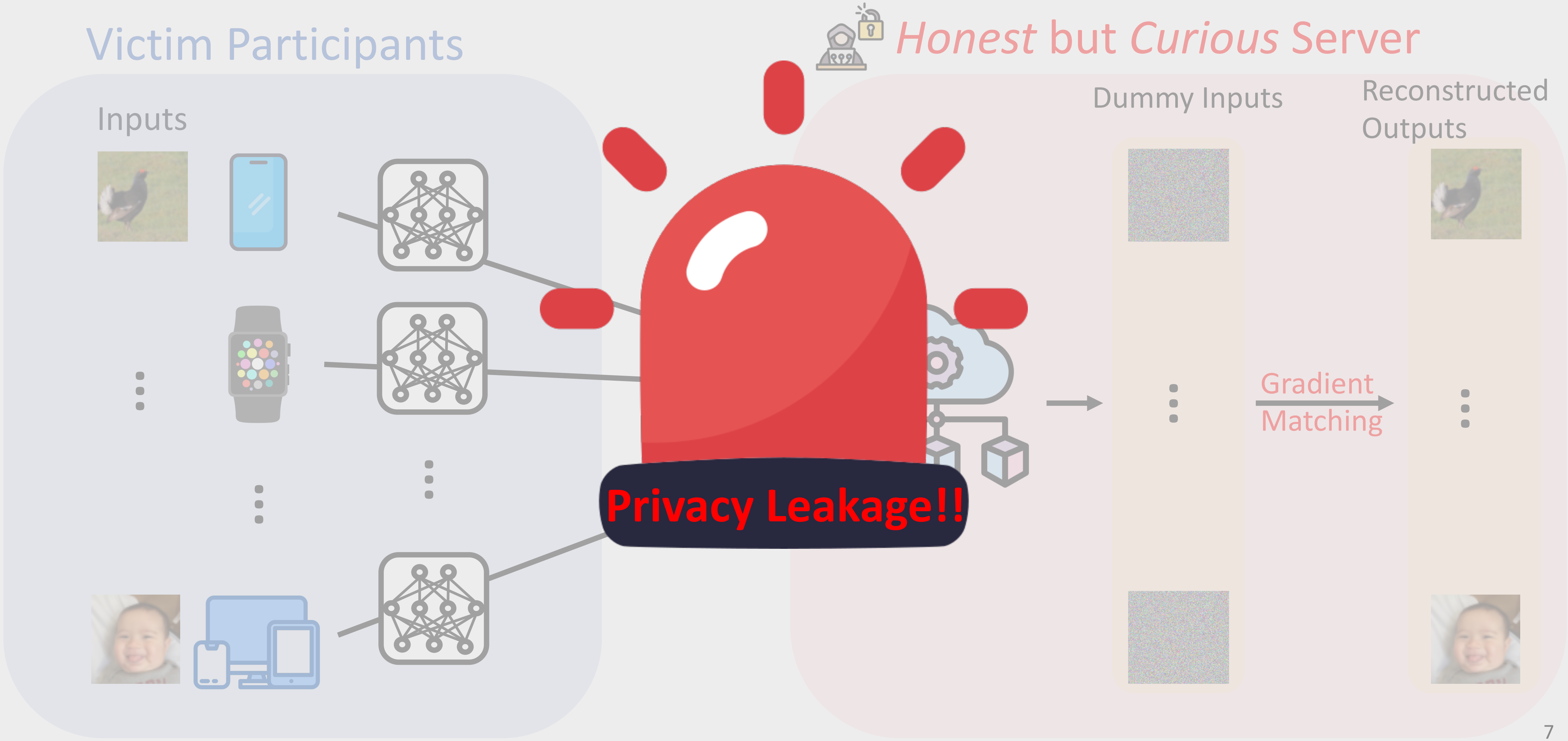
What is Gradient Inversion?



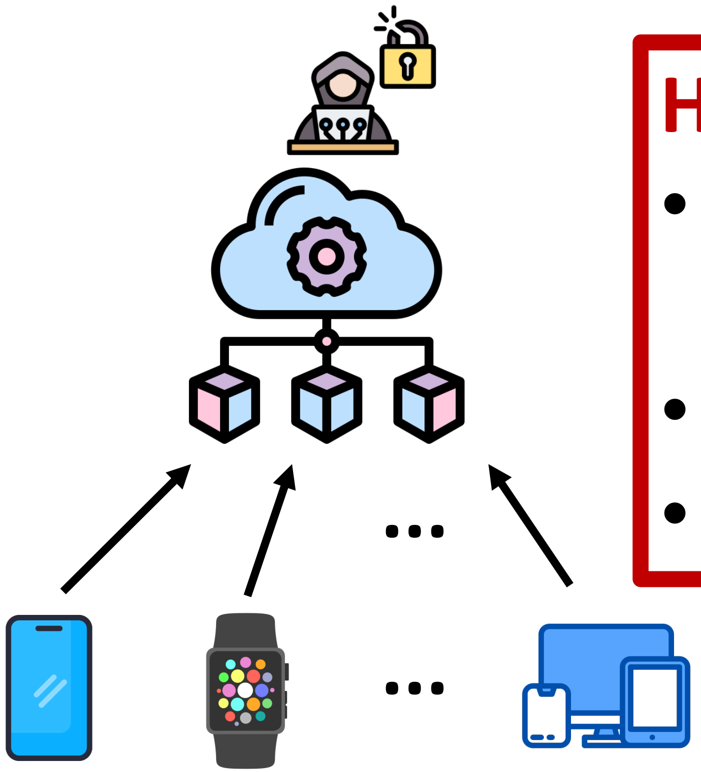
What is Gradient Inversion?



Privacy Concerns in Federated Learning



Threat Model



Honest but Curious Server

- Knows **model architecture** and **local gradients** shared by clients!
- Has access to **publicly available datasets**
- Can utilize **pre-trained models** (e.g. GANs)

Existing Gradient Inversion Attacks

- Inverting Gradients (**IG**) [1]
  - Optimizes on signed gradients with cosine similarity to refine inputs initialized from Gaussian noise.
- Grad Inversion (**GI**) [2]
  - Initializes inputs with Gaussian noise and applies Adam optimizer with regularization.
- Generative Gradient Leakage (**GGL**) [3]
  - Leverages GANs with KL-based regularization and optimizes with Bayesian or Covariance Matrix.
- Gradient Inversion in Alternative Spaces (**GIAS**) [4]
  - Employs negative cosine similarity as a gradient dissimilarity function.
- Gradient Inversion over Feature Domains (**GIFD**) [5]
  - Utilizes intermediate GAN features and optimizes with a warm-up strategy.

[1]. Geiping, Jonas, et al. "Inverting gradients-how easy is it to break privacy in federated learning?." NeurIPS 2020
[2]. Yin, Hongxu, et al. "See through gradients: Image batch recovery via gradinversion." CVPR 2021
[3]. Li, Zhuohang, et al. "Auditing privacy defenses in federated learning via generative gradient leakage." CVPR 2022
[4]. Jeon, Jinwoo, et al. "Gradient inversion with generative image prior." NeurIPS 2021
[5]. Fang, Hao, et al. "GIFD: A generative gradient inversion method with feature domain optimization." ICCV 2023

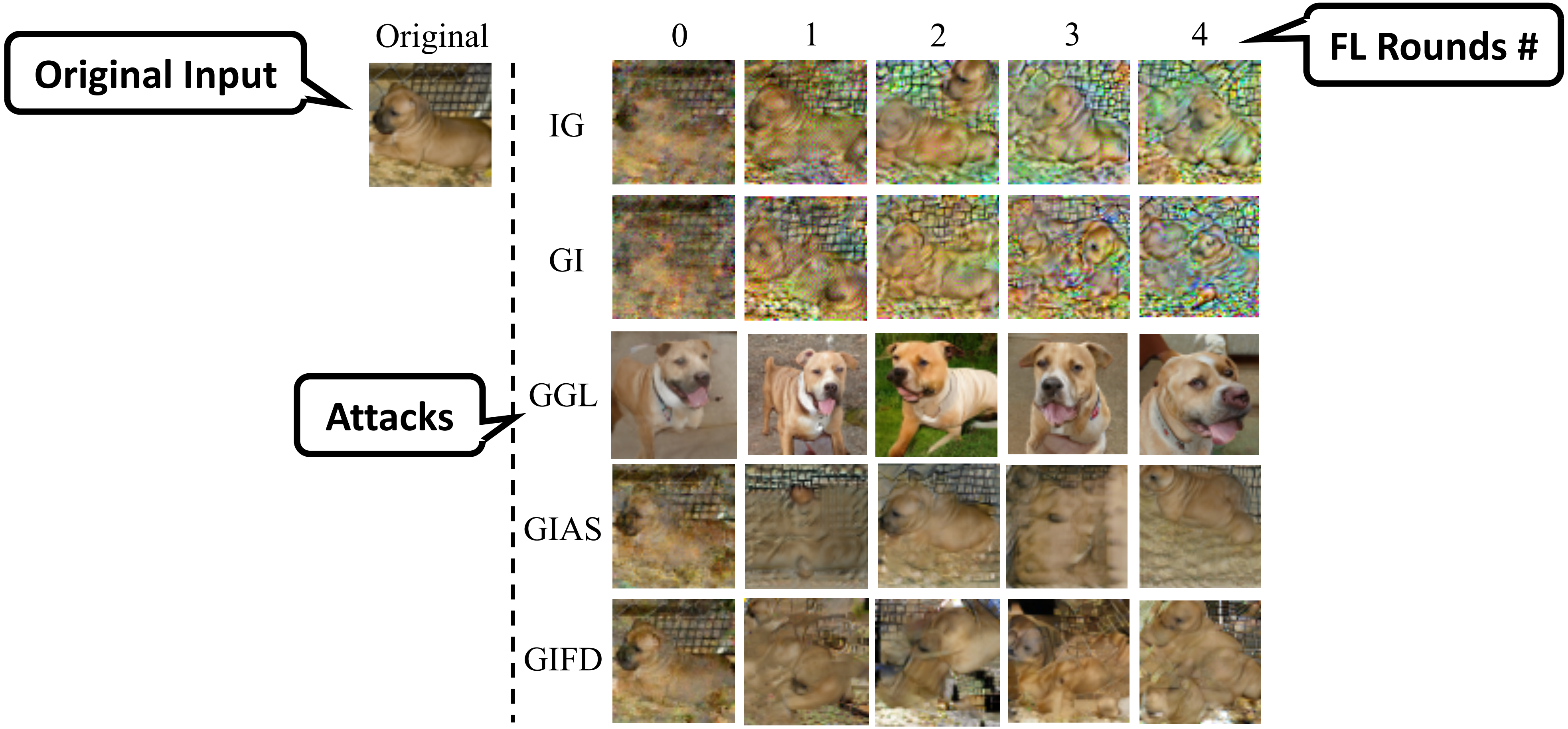
Existing Defense Methods

- Noise Gradient [1]
  - Adds Gaussian noise to gradients, reducing privacy leakage but significantly degrading utility.
- Gradient Clipping [2]
  - Bounds gradient magnitude by clipping values but fails to prevent privacy leakage.
- Gradient Sparsification [3]
  - Zeros out small gradients, transmitting only the largest values during update, yet still leaks information.
- Soteria [4]
  - Balances utility and privacy through optimization and gradient masking but is computationally expensive.

[1]. Geyer, Robin C., Tassilo Klein, and Moin Nabi. "Differentially private federated learning: A client level perspective." NeurIPS 2017 Workshop
[2]. Wei, Wenqi, et al. "Gradient-leakage resilient federated learning." ICDCS 2021
[3]. Aji, Alham Fikri, and Kenneth Heafield. "Sparse communication for distributed gradient descent." EMNLP 2017
[4]. Sun, Jingwei, et al. "Soteria: Provable defense against privacy leakage in federated learning from representation perspective." CVPR 2021

Observation I

Existing attacks succeed only in the early stage of training.



Overall, gradient inversion is most **effective in early training (0)**, especially when batch size = 1. Defending this stage is **critical**.

Observation II

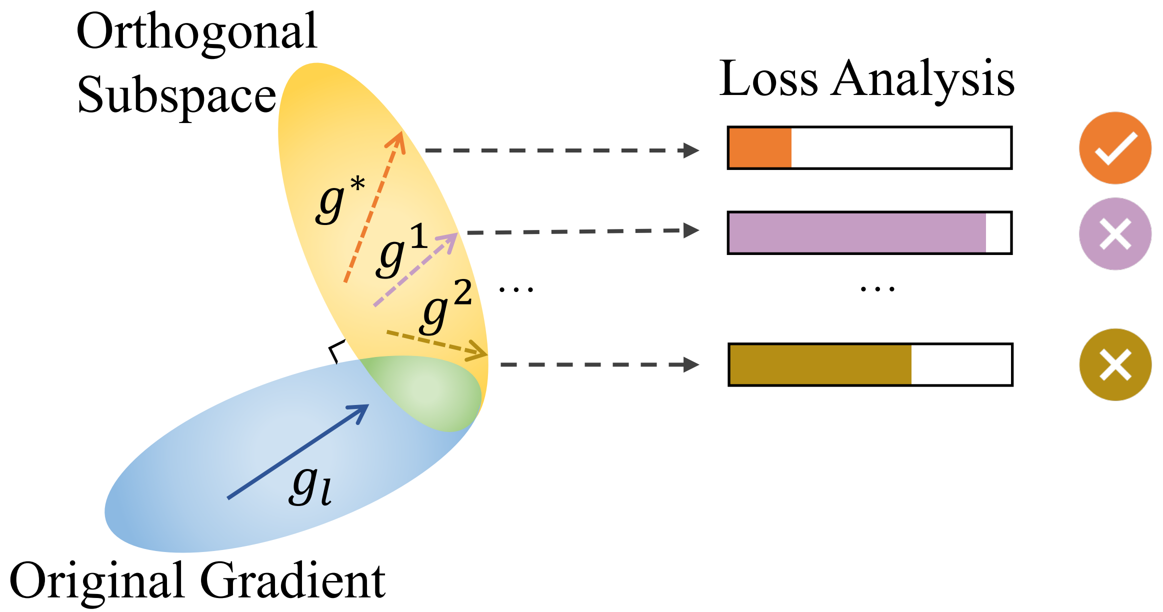
GGL [1] generates high-quality but typically low-fidelity images, and GGL leverages label information.

Even when GGL fails to reconstruct the exact input, it still leads to **privacy leakage**.

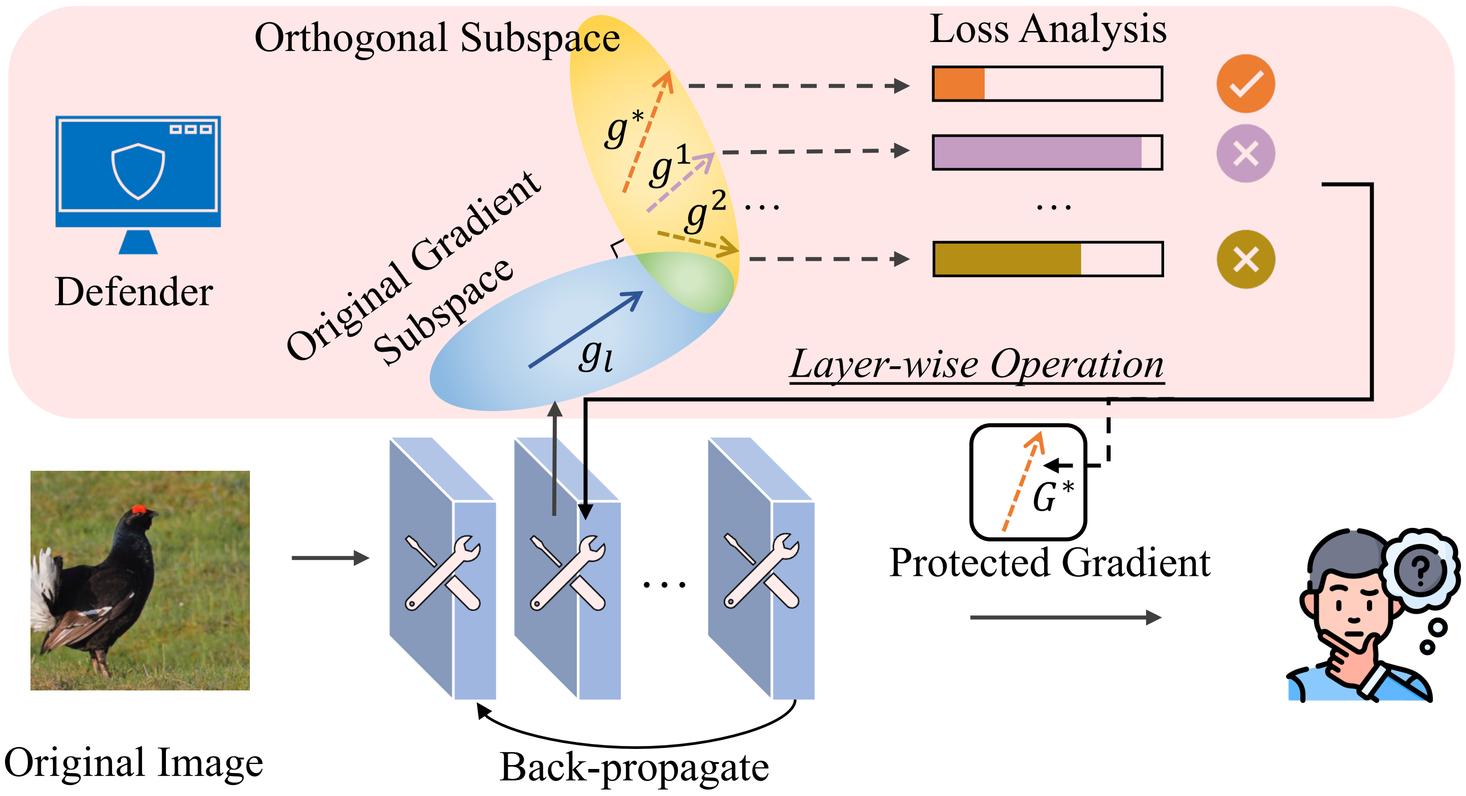
[1]. Li, Zhuohang, et al. "Auditing privacy defenses in federated learning via generative gradient leakage." CVPR 2022.

CENSOR Intuition

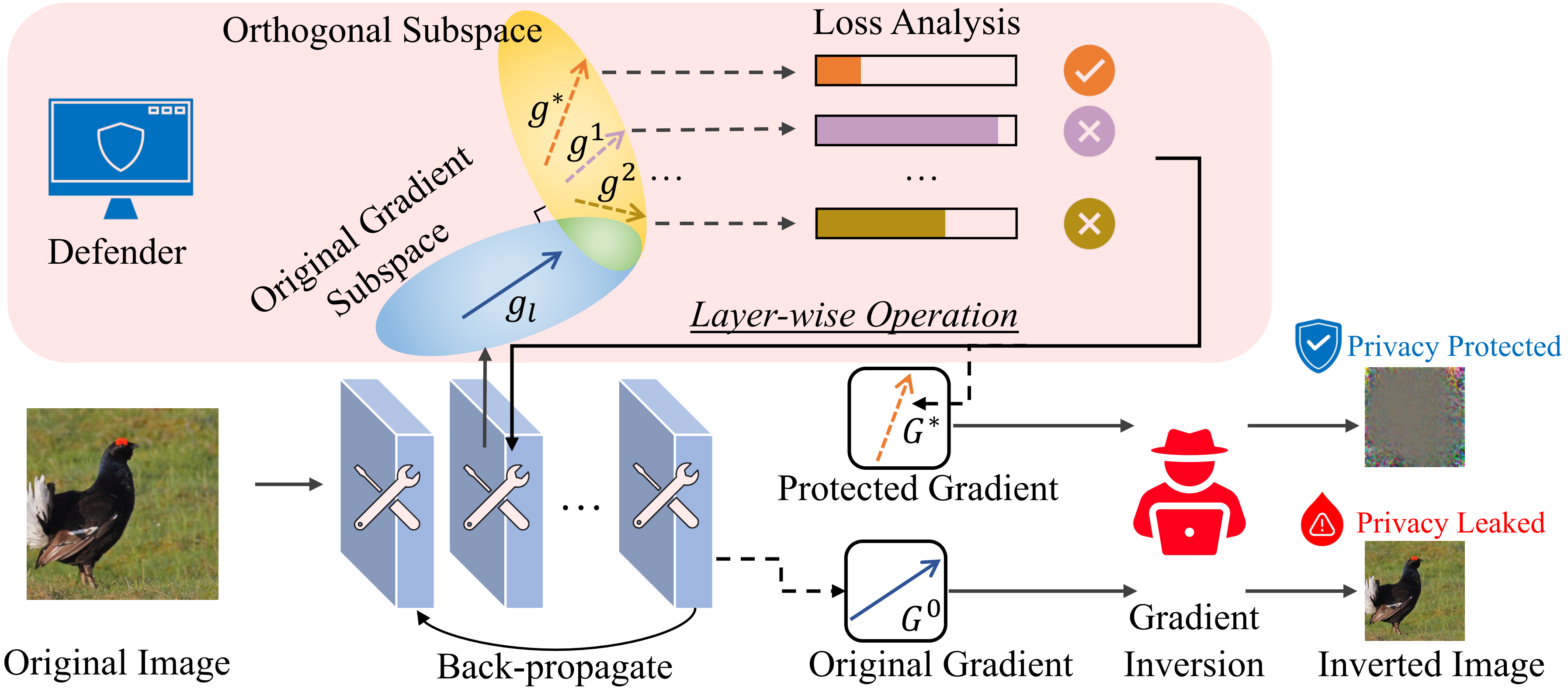
- In high-level, CENSOR samples gradients in a subspace that is orthogonal to the original gradient layer by layer and select the one that achieves the *lowest loss*.



CENSOR: Layer-wise Orthogonal Subspace Perturbation



CENSOR: Layer-wise Orthogonal Subspace Perturbation



Quantitative Experiment

Metrics
↑higher, better
↓lower, better

Table I: Quantitative evaluation of various defense methods against existing attacks. (An upward arrow denoting the higher the better, a downward arrow denoting the lower the better.)

Attacks

DA	Defense	IG [5]				GI [9]				GGL [8]				GIAS [6]				GIFD [4]			
		MSE↑	LPIPS↑	PSNR↓	SSIM↓	MSE↑	LPIPS↑	PSNR↓	SSIM↓	MSE↑	LPIPS↑	PSNR↓	SSIM↓	MSE↑	LPIPS↑	PSNR↓	SSIM↓	MSE↑	LPIPS↑	PSNR↓	SSIM↓
ImageNet	No Defense	0.0195	0.5574	17.819	0.2309	0.0191	0.5402	17.908	0.2400	0.0453	0.5952	13.873	0.0745	0.0191	0.4795	18.452	0.3099	0.0130	0.3782	21.364	0.4528
	Noise [13]	0.0246	0.6294	16.338	0.1754	0.0269	0.6300	15.883	0.1549	0.0410	0.5697	14.252	0.0817	0.0253	0.5947	16.601	0.1854	0.0196	0.5380	18.166	0.2686
	Clipping [14]	0.0167	0.5008	18.883	0.3128	0.0383	0.7302	14.844	0.0106	**0.0477**	0.5823	**13.520**	0.0749	0.0203	0.4825	18.738	0.3186	0.0150	0.4433	19.547	0.3798
	Sparsi [15]	0.0137	0.4945	19.383	0.3419	0.0157	0.4941	18.799	0.3099	0.0456	0.6080	13.743	0.0776	0.0135	0.3981	20.483	0.4182	0.0179	0.4444	19.486	0.3686
	Soteria [16]	**0.0662**	**0.7596**	**12.220**	0.0135	**0.0682**	0.7485	**12.215**	0.0134	0.0461	0.5986	13.879	0.0708	0.0245	0.4986	17.646	0.2664	0.0139	0.3967	20.602	0.4335
	CENSOR	0.0600	0.7551	12.463	**0.0067**	0.0416	**0.8615**	14.446	**0.0021**	0.0419	**0.7912**	14.262	**0.0094**	**0.0650**	**0.7591**	**12.266**	**0.0139**	**0.0507**	**0.7610**	**13.323**	**0.0094**

Defenses

[5]. Geiping, Jonas, et al. "Inverting gradients-how easy is it to break privacy in federated learning?." NeurIPS 2020
[9]. Yin, Hongxu, et al. "See through gradients: Image batch recovery via gradinversion." CVPR 2021
[8]. Li, Zhuohang, et al. "Auditing privacy defenses in federated learning via generative gradient leakage." CVPR 2022
[6]. Jeon, Jinwoo, et al. "Gradient inversion with generative image prior." NeurIPS 2021
[4]. Fang, Hao, et al. "GIFD: A generative gradient inversion method with feature domain optimization." ICCV 2023
[13]. Geyer, Robin C., Tassilo Klein, and Moin Nabi. "Differentially private federated learning: A client level perspective." NeurIPS 2017 Workshop
[14]. Wei, Wenqi, et al. "Gradient-leakage resilient federated learning." ICDCS 2021
[15]. Aji, Alham Fikri, and Kenneth Heafield. "Sparse communication for distributed gradient descent." EMNLP 2017
[16]. Sun, Jingwei, et al. "Soteria: Provable defense against privacy leakage in federated learning from representation perspective." CVPR 2021

Quantitative Experiment

Metrics
↑higher, better
↓lower, better

Table I: Quantitative evaluation of various defense methods against existing attacks. (An upward arrow denoting the higher the better, a downward arrow denoting the lower the better.)

Attacks

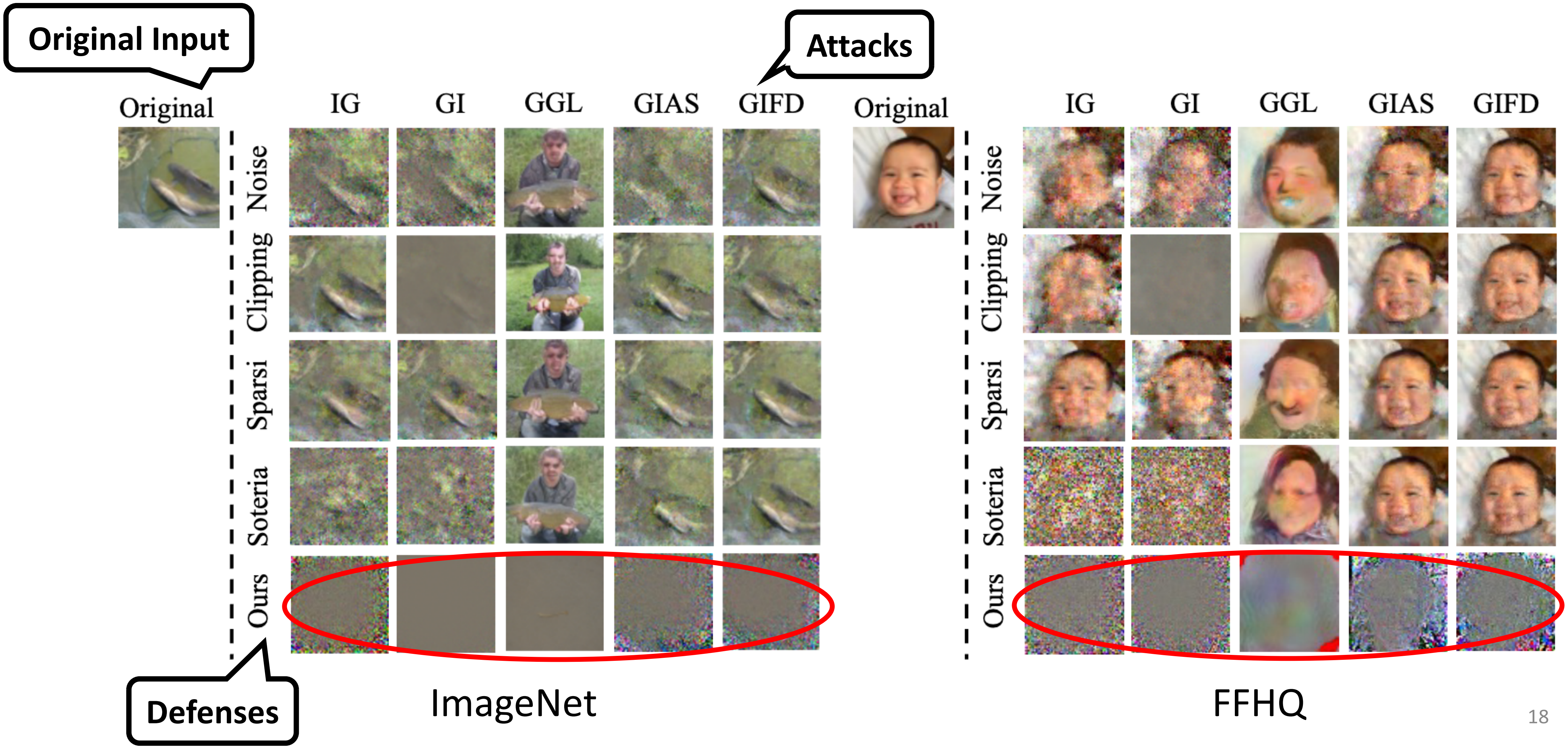
DA	Defense	IG [5]				GI [9]				GGL [8]				GIAS [6]				GIFD [4]			
		MSE↑	LPIPS↑	PSNR↓	SSIM↓	MSE↑	LPIPS↑	PSNR↓	SSIM↓	MSE↑	LPIPS↑	PSNR↓	SSIM↓	MSE↑	LPIPS↑	PSNR↓	SSIM↓	MSE↑	LPIPS↑	PSNR↓	SSIM↓
	No Defense	0.0195	0.5574	17.819	0.2309	0.0191	0.5402	17.908	0.2400	0.0453	0.5952	13.873	0.0745	0.0191	0.4795	18.452	0.3099	0.0130	0.3782	21.364	0.4528
ImageNet	Noise [13]	0.0246	0.6294	16.338	0.1754	0.0269	0.6300	15.883	0.1549	0.0410	0.5697	14.252	0.0817	0.0253	0.5947	16.601	0.1854	0.0196	0.5380	18.166	0.2686
	Clipping [14]	0.0167	0.5008	18.883	0.3128	0.0383	0.7302	14.844	0.0106	**0.0477**	0.5823	**13.520**	0.0749	0.0203	0.4825	18.738	0.3186	0.0150	0.4433	19.547	0.3798
	Sparsi [1[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.0456	0.6080	13.743	0.0776	0.0135	0.3981	20.483	0.4182	0.0179	0.4444	19.486	0.3686
	Soteria [illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.0461	0.5986	13.879	0.0708	0.0245	0.4986	17.646	0.2664	0.0139	0.3967	20.602	0.4335
	CENSO[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.0419	**0.7912**	14.262	**0.0094**	**0.0650**	**0.7591**	**12.266**	**0.0139**	**0.0507**	**0.7610**	**13.323**	**0.0094**

Defenses

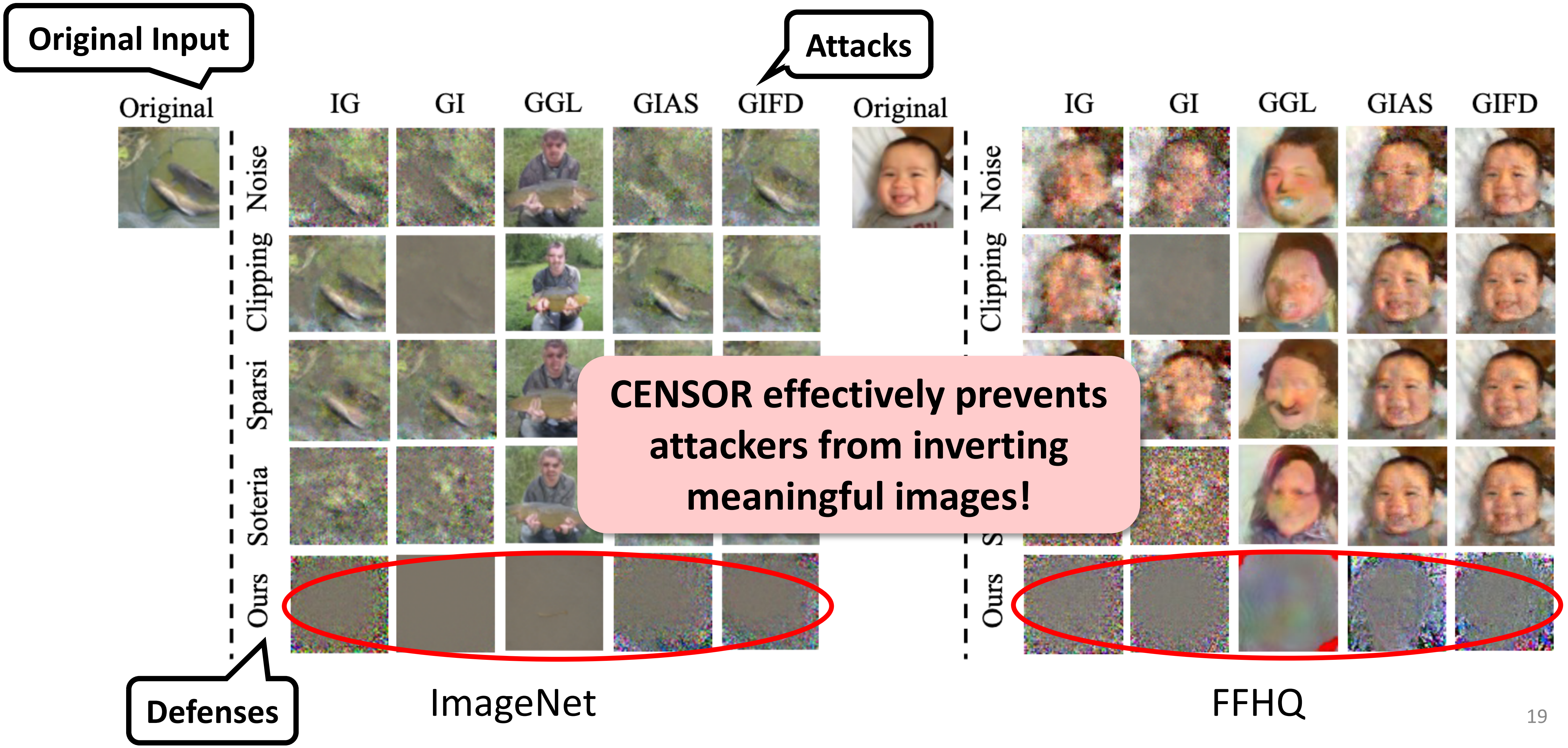
CENSOR outperforms existing defenses in almost all cases, and significantly surpasses the SOTA defense Soteria (up to 114% in the metrics)!

[5]. Geiping, Jonas, et al. "
[9]. Yin, Hongxu, et al. "Se
[8]. Li, Zhuohang, et al. "Au 22
[6]. Jeon, Jinwoo, et al. "Grad
[4]. Fang, Hao, et al. "GIFD: A generative gradient inversion method with feature domain optimization." ICCV 2023
[13]. Geyer, Robin C., Tassilo Klein, and Moin Nabi. "Differentially private federated learning: A client level perspective." NeurIPS 2017 Workshop
[14]. Wei, Wenqi, et al. "Gradient-leakage resilient federated learning." ICDCS 2021
[15]. Aji, Alham Fikri, and Kenneth Heafield. "Sparse communication for distributed gradient descent." EMNLP 2017
[16]. Sun, Jingwei, et al. "Soteria: Provable defense against privacy leakage in federated learning from representation perspective." CVPR 2021

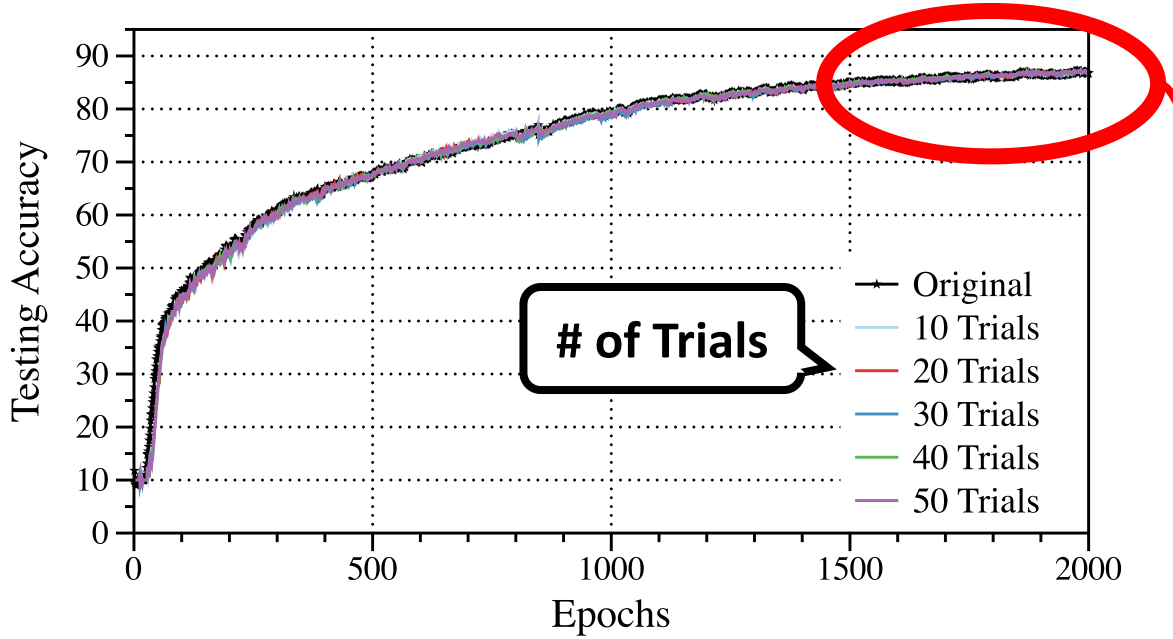
Qualitative Experiment

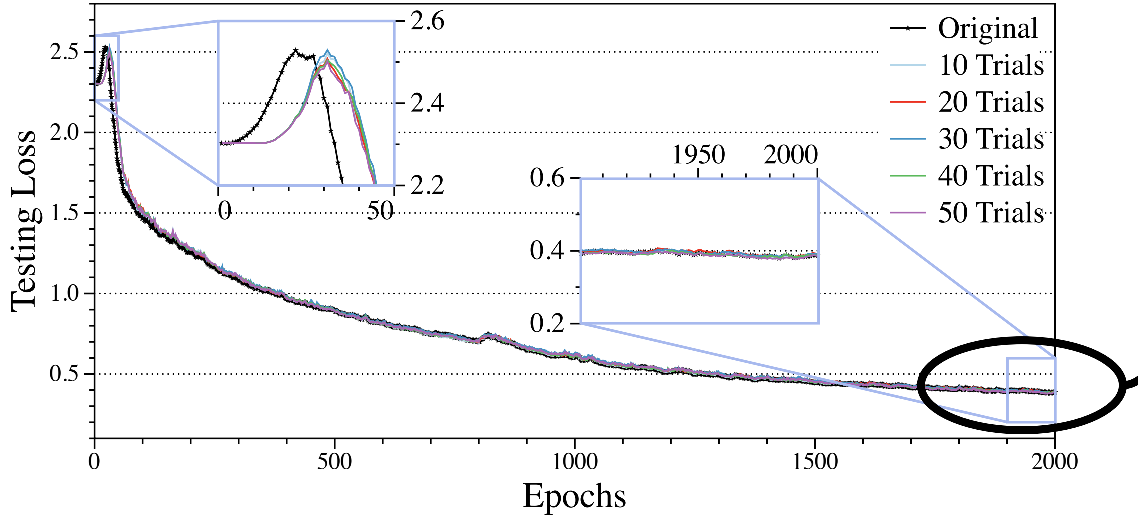


Qualitative Experiment



Convergence Study





- 100 clients in total on CIFAR-10 dataset
- *Non-i.i.d.* data distribution
- Randomly selected 10 clients each epoch

Both Original (vanilla) and CENSOR in different trials converged to the similar level.

Testing loss exhibits only slight variations at the beginning. In the end, all have settled at a relatively low level.

Adaptive Attack (EOT)

- Expectation Over Transformation (EOT) is to perform the gradient transformation multiple times, and take the *average gradient* over several runs, to approximate the gradient and *mitigate the randomization effect* as much as possible.

Table III: Adaptive attack with EOT.

Dataset	**EOT**	**MSE ↑**	**LPIPS ↑**	**PSNR ↓**	**SSIM ↓**
ImageNet	w/o	0.0507	0.7610	13.32	0.0094
	w/.	0.0518	0.7668	13.39	0.0087
FFHQ	w/o	0.1037	0.8097	9.90	0.0195
	w/.	0.1098	0.8340	9.82	0.0195

CENSOR: Defense Against Gradient Inversion via Orthogonal Subspace Bayesian Sampling

Take-aways:

1. CENSOR is designed to **mitigate gradient inversion attacks**.
2. CENSOR samples gradients within a ***subspace orthogonal*** to the original gradients.
3. CENSOR enhances the data privacy and maintains model utility.
4. Paper, code, slides: https://censor-gradient.github.io/

Thank you for listening!

On the academic job market in 2025-26 cycle!

SCAN ME

PURDUE UNIVERSITY

IBM