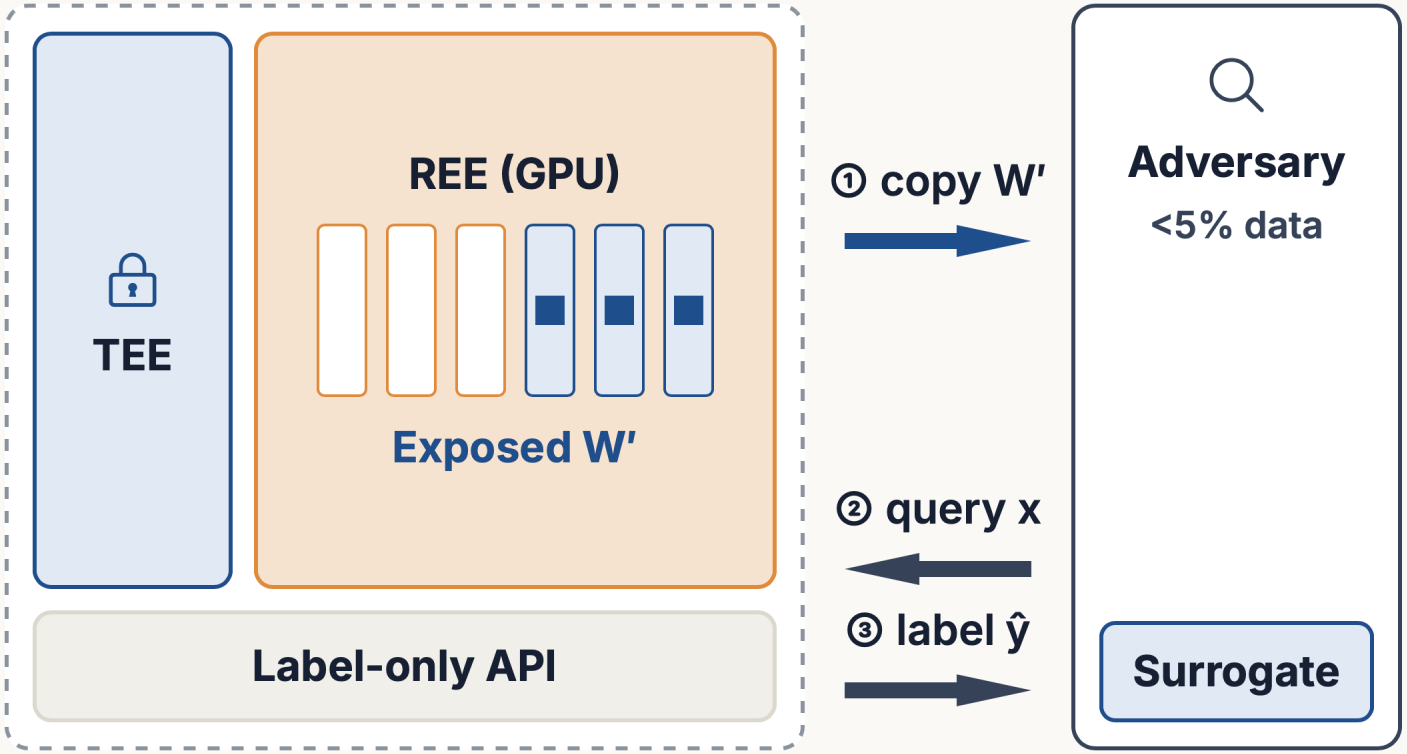
NEURIPS 2026

FILOsofer

A TEE-Shielded Model Partitioning Framework based on Fisher Information-Guided LoRA Obfuscation

Fan Zhang, Ziqi Zhang, Hossein Khalili, Neiro Cabrera, Jonathan Xue, Nader Sehatbakhsh

Model stealing and prior TSDP defenses



Layer partitioning

Selected layers run in the TEE



KEY QUESTION

Most weights still sit on the GPU.
Is partial protection enough?

Existing TSDP leaks a warm start

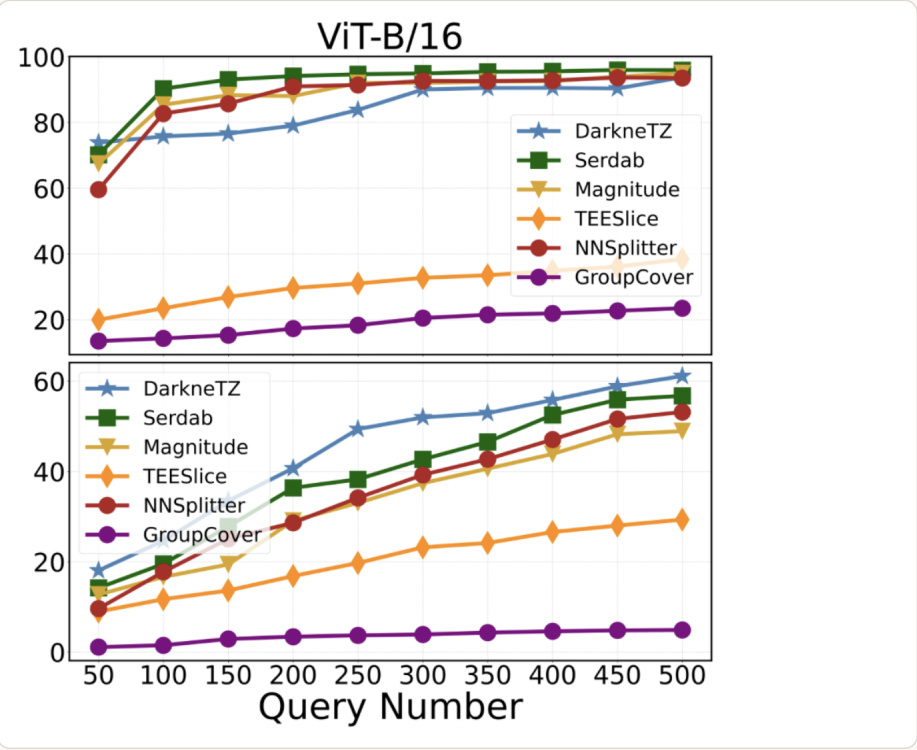


Fig. 1: surrogate accuracy vs. queries, ViT-B/16, two settings.

OBSERVATION

Every baseline's surrogate keeps improving with more queries; most reach ~90% with only 500.

Why: the weights left on the GPU are still accurate. The attacker copies them as an initialization and fine-tunes on leaked labels. At 5,000 queries on ResNet18 / CIFAR-10, even TEESlice (65.2%) and GroupCover (57.6%) exceed a pure black-box attacker (55.9%).

Insight: hiding layers is not enough. The defense must corrupt the initialization itself.

Attack: KnockoffNets (Orekondy et al., CVPR 2019) · numbers from Table 1

FILOsofer: system overview

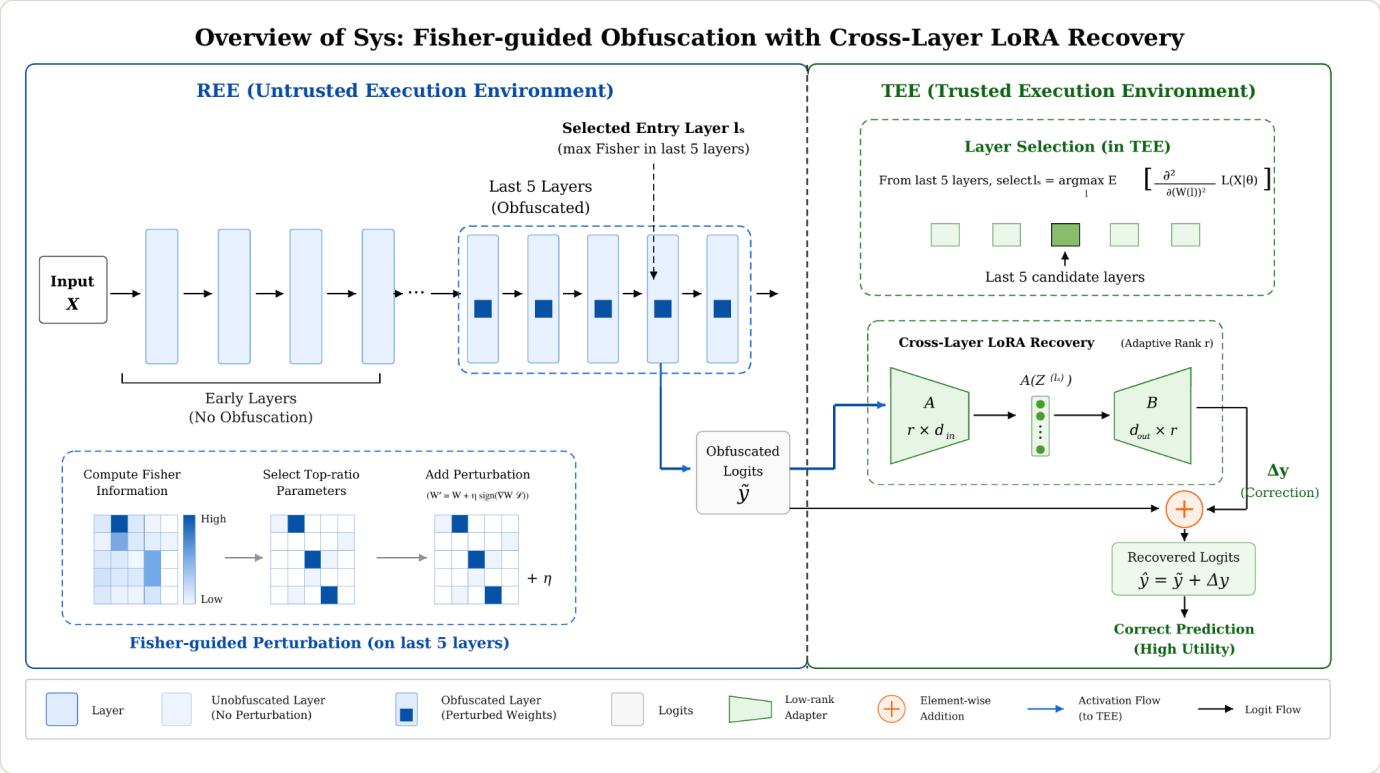


Fig. 2: the obfuscated backbone runs in the REE; the LoRA branch runs in the TEE.

1 Obfuscate

Fisher-guided perturbation of a few weights in the last layers (REE).

2 Select entry

Entry layer ℓ_s = highest Fisher among the last five.

3 Recover

Cross-layer LoRA in the TEE adds a correction: $\hat{y} = \tilde{y} + \Delta y$.

3.1 · METHOD

Fisher information-guided obfuscation

Objective. Exposed outputs should carry no information about the input (functional collapse).

$$\Delta W^* = \arg\min_{\Delta W \in \mathcal{R}} I(\, f(x; W)\, ;\, f(x; W+\Delta W)\,) \tag{1}$$

Selection. Score each weight by its target-class Fisher information; keep the top fraction $r \approx 10^{-4}$.

$$F_{Lt} = \mathbb{E}[\, (\partial \mathcal{L}(x, W) / \partial W)^2 \mid y = L_t \,] \tag{3}$$

Perturbation. Move the selected weights toward the target label; a tiny η suffices.

$$W \leftarrow W + \eta \cdot \nabla_W \mathcal{L}(x, W) \tag{4}$$

Table 5: Effectiveness of Fisher Information Guidance. ResNet18 / CIFAR-100, accuracy (%).

Strategy	Obf. Acc.	Query@50	Query@5000	LoRA Rec.
Fisher-top	1.00	1.00	2.07	80.22
Fisher-bottom	1.03	54.92	77.96	80.51
Random	1.00±0.01	54.63±0.37	77.80±0.04	80.44±0.08

Only Fisher-top (ours, highlighted) derails surrogate fine-tuning; all three keep LoRA recovery near 80%.

TAKEAWAY

High-Fisher weights shape the decision boundary, so perturbing them makes the exposed weights actively misleading.

Proof sketch (local Fisher expansion of mutual information): Appendix C.17

Cross-layer LoRA recovery in the TEE

Entry layer. The most informative of the last five layers.

$$\ell_s = \arg\max_{\ell} \mathbb{E}[\, -\partial^2 \mathcal{L} \,/\, \partial(W\ell)^2 \,] \tag{5}$$

Correction. One low-rank shortcut from ℓ_s to the logits, with $A \in \mathbb{R}^{r \times d}$ and $B \in \mathbb{R}^{d \times r}$.

$$\hat{y} = \tilde{y} + B \cdot A(\, Z_{\ell s} \,) \tag{6}$$

Joint training (Alg. 2). Raise the obfuscation strength until LoRA-recovered accuracy would fall below a threshold, then roll back.

Only A and B live in the TEE: 33.6 KB to 2.37 MB of parameters.

Table 2: The recovery accuracy of cross-layer LoRA among varying ranks.

Model	Dataset	Original	Obfuscated	LRank 4	LRank 8	LRank 16	LRank 32
AlexNet	C10	81.85	10.00	80.77	80.95	81.32	81.68
	C100	55.97	1.00	55.93	55.95	56.30	56.92
	ImageNet200	47.70	0.50	47.89	48.08	48.96	48.90
ResNet18	C10	93.07	10.00	92.92	92.98	93.04	93.42
	C100	81.50	1.00	80.57	80.63	80.50	80.92
	ImageNet200	65.68	0.50	50.03	52.13	53.82	55.20
VGG19	C10	91.42	10.00	90.85	90.94	91.48	91.65
	C100	70.39	1.00	69.01	69.14	69.67	70.03
	ImageNet200	63.23	0.50	61.10	61.77	62.66	63.32
ViT-B16	C10	97.69	10.00	97.73	97.67	97.80	97.96
	C100	87.58	1.00	88.14	86.36	87.96	88.50
	ImageNet200	81.99	0.50	83.27	83.43	83.58	84.16

Obfuscation drops every model to chance; even rank 4 restores near-original accuracy, and ViT-B16 slightly exceeds it. Harder settings (ResNet18 / ImageNet200) benefit from higher rank.

Security against model stealing

Table 1: **The accuracy of the surrogate model.** Green and Red boxes highlight the lowest and highest accuracy, respectively.

Model	Data	No-Shield	DarkneTZ		Serdab		Magnitude		NNSplitter		TEESlice		GroupCover		Ours		Blackbox	
		/	50	5000	50	5000	50	5000	50	5000	50	5000	50	5000	50	5000	50	5000
AlexNet	C10	81.58	66.01	80.97	74.08	80.26	68.42	78.08	58.90	75.26	24.79	66.51	13.60	47.70	10.00	12.17	19.37	64.20
	C100	55.97	13.60	52.19	42.59	54.25	30.64	50.40	19.92	42.76	3.98	35.20	1.00	1.00	1.00	1.00	2.89	30.75
	IN200	47.70	4.12	33.16	31.29	34.44	14.92	35.83	3.67	29.73	0.65	13.39	0.50	0.50	0.50	0.50	0.66	13.46
ResNet18	C10	93.07	86.84	92.79	86.86	92.25	73.26	88.63	78.30	90.66	23.87	65.21	10.00	57.60	10.00	18.51	23.12	55.92
	C100	81.50	26.36	79.23	77.93	80.79	59.55	77.02	47.86	78.78	5.31	58.01	1.00	21.90	1.00	1.00	3.25	31.53
	IN200	65.68	61.22	63.50	5.96	59.10	43.08	58.39	25.24	42.12	2.27	48.08	0.50	6.30	0.50	0.50	1.16	34.06
VGG19	C10	91.42	89.34	91.45	91.44	91.42	83.93	90.15	79.87	89.51	40.52	89.62	10.00	21.70	10.00	17.25	40.61	71.05
	C100	70.39	22.63	67.71	69.07	69.85	59.95	65.34	25.50	59.62	7.31	48.97	1.00	1.30	1.00	1.00	7.09	39.34
	IN200	63.23	60.89	61.17	5.24	52.66	24.67	45.20	13.28	49.26	2.83	43.72	0.50	2.90	0.50	0.50	2.54	32.09
ViT-B/16	C10	97.69	67.96	97.12	65.54	94.99	95.26	97.92	59.55	93.56	22.64	97.64	12.60	39.70	10.31	14.87	20.63	85.02
	C100	86.58	24.38	78.17	14.70	80.84	15.64	85.48	9.63	53.17	12.69	86.89	1.60	11.50	1.41	6.83	11.62	74.36
	IN200	81.99	12.94	78.32	9.02	80.10	72.62	85.91	14.68	82.74	10.18	81.71	0.80	4.30	0.85	2.21	8.82	70.50

Ours (green) is the lowest in every setting and below the black-box attacker at both budgets.

Efficiency and real-device deployment

Table 3 (excerpt). %FLOPs in the TEE to match black-box security, CIFAR-10.

Method	ResNet18	VGG19	ViT
DarkneTZ	100.00	98.85	91.07
Serdab	100.00	100.00	91.73
Magnitude	100.00	100.00	100.00
TEESlice	3.80	0.34	7.24
GroupCover	2.18	2.75	0.24
Ours	0.0027	0.0032	0.0069

Table 6: Memory footprint and Fisher preparation cost.

Model	REE (MB)	TEE (KB)	Infer. (MB)	Fisher (MB)	Train (MB)	Fisher (ms)
ResNet18	42.8	257.6	69.7	32.21	555.6	17.92
VGG19	76.6	33.6	103.3	9.20	320.7	10.05
ViT-B/16	327.6	2365.6	352.6	108.32	11890.8	676.05

Table 8. Per-image latency (ms) on Jetson Orin Nano; LoRA branch in TrustZone.

Model	GPU	TrustZone	Overhead
AlexNet	12.4	0.84	6.3%
ResNet18	22.1	0.86	3.7%
VGG19	48.9	0.85	1.7%
ViT-B/16	91.7	0.87	0.9%

Table 6: inference at batch size 1; Fisher preparation at batch size 100, offline only.

5 · CONCLUSION

Limitations and conclusion

Limitations

- Larger budgets still help: 70.6% at 50k queries (black-box: 90.9%).
- Assumes TEE confidentiality; side channels are out of scope.
- Image classification only; LLMs are future work.

Don't just hide weights. Make the exposed ones misleading.

FILOsofer removes the attacker's warm start with Fisher-guided obfuscation and restores accuracy with a tiny cross-layer LoRA in the TEE.

Code: github.com/ssysarch/FILOsofer

Limitations: Appendix C.8, C.9, C.11–C.13