ICLR 2026

IMPERIAL

WARP: Weight Teleportation for Attack-Resilient Unlearning Protocols

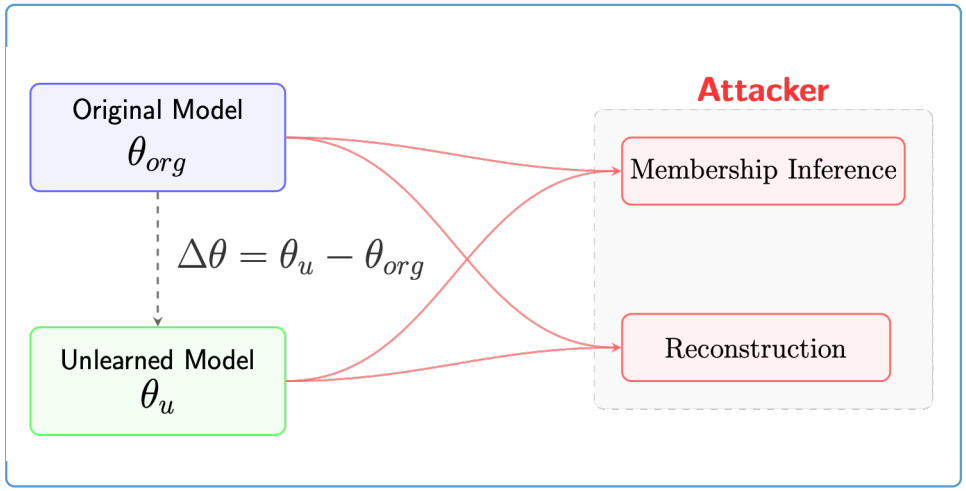
Mohammad M Maheri [1] Xavier Cadet [2] Peter Chin [2] Hamed Haddadi [1]

[1] Imperial College London [2] Dartmouth College

Problem / Motivation

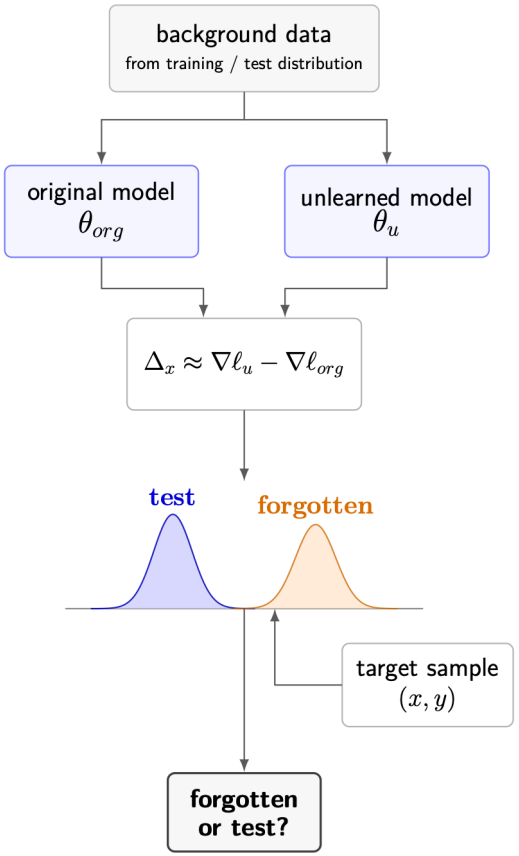
Approximate Unlearning has a privacy gap

- Original model + Unlearned model
- Attacker sees the difference
- Forgotten sample may be inferred or reconstructed

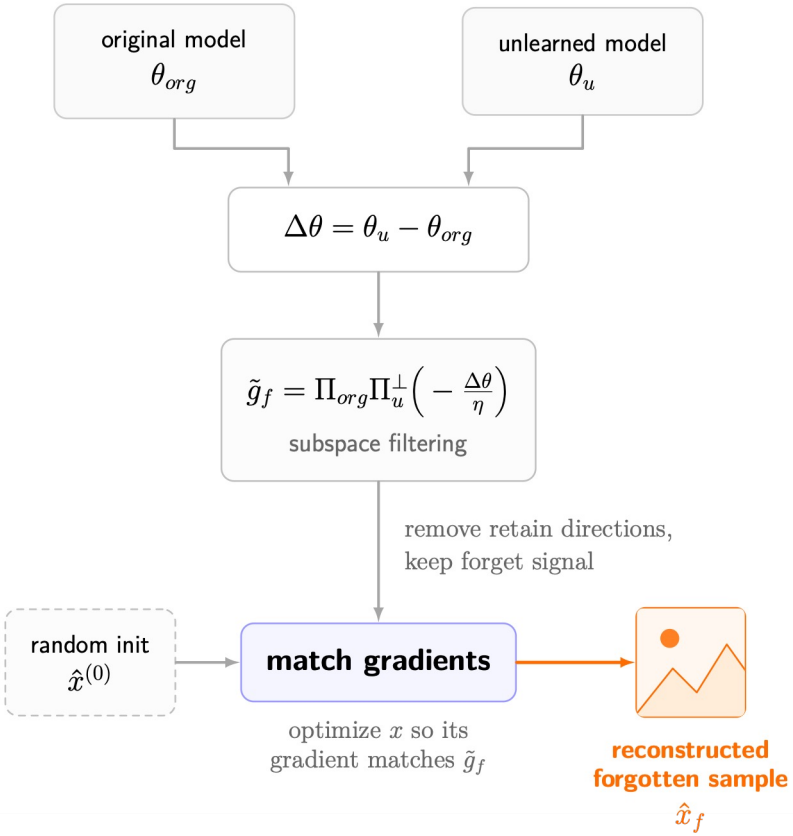


How we evaluate the risk

White Box MIA

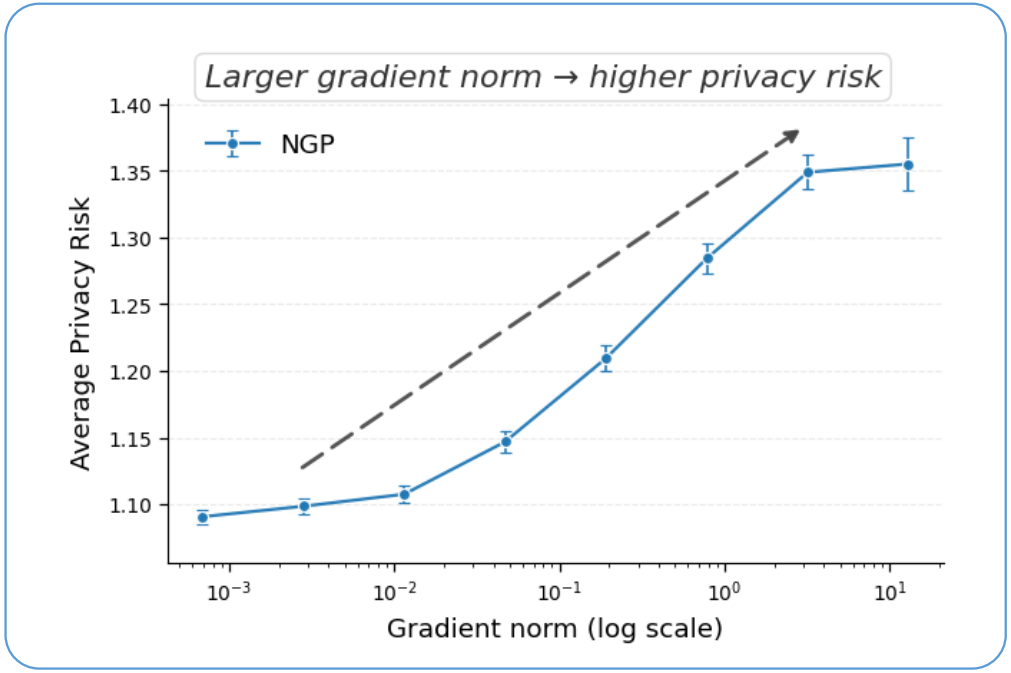


Data Reconstruction Attack

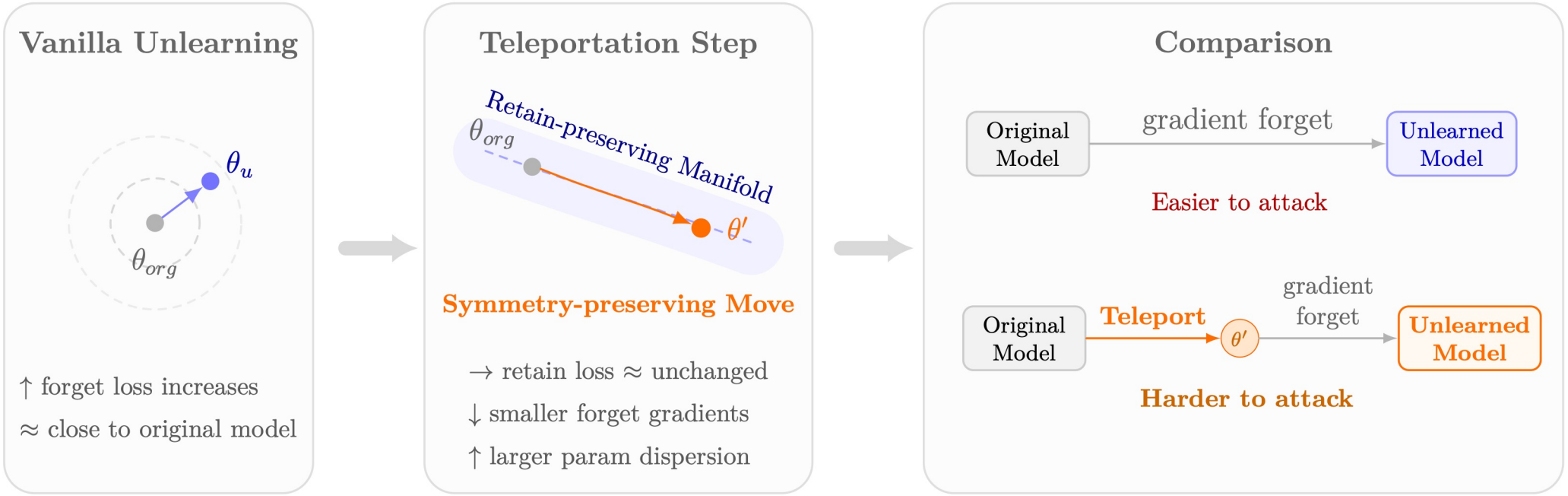


Key Observation

- Parameter closeness increases privacy leakage.
- Higher gradient norm among forget-set samples → higher privacy risk.



Methodology



Main Results

Black Box MIA

Method	All samples (BB) AUC	TPR@0.1	TPR@1	TPR@5	Most–memorized (top 1%) AUC	TPR@0.1	TPR@1	TPR@5	Acc. Test
NGP (base)	0.545	0.012	0.030	0.077	0.649	0.058	0.157	0.277	**0.808**
+ WARP	**0.516**	**0.003**	**0.014**	**0.055**	**0.598**	**0.015**	**0.082**	**0.206**	0.797
Improvement (%)	64.4	81.8	80.0	81.5	34.2	75.4	51.0	31.3	-5.7
SCRUB (base)	0.543	0.020	0.047	0.092	0.710	0.086	0.227	0.397	**0.815**
+ WARP	**0.526**	**0.015**	**0.036**	**0.078**	**0.610**	**0.041**	**0.119**	**0.213**	0.813
Improvement (%)	39.5	26.3	29.7	33.3	47.6	52.9	49.8	53.0	-1.1
PGU (base)	0.636	0.024	0.040	**0.098**	0.910	0.201	0.511	0.706	0.804
+ WARP	**0.631**	**0.018**	**0.036**	0.104	**0.875**	**0.160**	**0.431**	**0.663**	**0.808**
Improvement (%)	3.7	26.1	13.3	-12.5	8.5	20.5	16.0	6.6	+2.0
Salun (base)	0.572	0.020	0.062	0.121	0.910	0.129	0.321	0.520	0.802
+ WARP	**0.565**	**0.019**	**0.059**	**0.113**	**0.826**	**0.107**	**0.264**	**0.487**	**0.803**
Improvement (%)	9.7	5.3	5.8	11.3	20.5	17.2	18.3	7.0	+0.5
SF (base)	0.509	0.004	0.015	0.056	0.518	0.089	0.034	0.079	**0.814**
+ WARP	**0.506**	**0.002**	**0.012**	**0.051**	**0.501**	**0.006**	**0.026**	**0.068**	0.811
Improvement (%)	33.3	66.7	60.0	83.3	94.4	94.3	33.3	37.9	-1.6
BT (base)	0.725	0.000	0.177	0.287	0.902	0.119	0.295	0.582	0.816
+ WARP	**0.661**	0.000	**0.137**	**0.219**	**0.865**	**0.113**	**0.275**	**0.537**	**0.818**
Improvement (%)	28.4	–	24.0	28.7	9.2	5.1	7.0	8.5	+1.1

White Box MIA

Method	AUC	TPR@0.1	TPR@1	TPR@5
NGP (base)	0.642	0.004	0.034	0.139
+ WARP	**0.614**	**0.002**	**0.021**	**0.097**
Improvement (%)	17.0	50.0	40.6	34.2
SCRUB (base)	0.700	0.011	0.102	0.287
+ WARP	**0.657**	**0.006**	**0.061**	**0.193**
Improvement (%)	14.3	54.5	42.5	33.5
PGU (base)	0.659	0.007	0.064	0.215
+ WARP	**0.533**	**0.002**	**0.025**	**0.085**
Improvement (%)	92.9	83.3	64.5	65.5
Salun (base)	0.721	0.008	0.069	0.230
+ WARP	**0.705**	**0.006**	**0.062**	**0.214**
Improvement (%)	9.5	33.3	10.1	7.0
SF (base)	0.670	0.005	0.043	0.161
+ WARP	**0.629**	**0.003**	**0.030**	**0.124**
Improvement (%)	29.2	50.0	34.9	23.2
BT (base)	0.938	0.037	0.346	0.809
+ WARP	**0.907**	**0.028**	**0.279**	**0.684**
Improvement (%)	49.2	25.7	19.4	18.4

Main Results

Thanks!

Mohammad M Maheri
m.maheri23@imperial.ac.uk