Learning to Unlearn while Retaining: Combating Gradient Conflicts in Machine Unlearning

Gaurav Patel and Qiang Qiu

PURDUE UNIVERSITY | Elmore Family School of Electrical and Computer Engineering

Context & Problem Statement

- **Machine Unlearning (MU):**
  - Developed to comply with privacy rights like the “right to be forgotten” (GDPR).
  - *Exact unlearning* = Retraining from scratch without the unwanted data—but that is very computationally expensive.
  - *Approximate* unlearning methods aim to remove data influence more **efficiently**, by finetuning the initial model.

- **Competing Objectives**:
  - In practice, two objectives:
    - **Forget** specific data (reduce the model’s reliance or memory of it).
    - **Retain** performance on remaining data.
  - **Contradictory gradient** updates → can cause gradient conflict, slow or prevent proper unlearning.

Preliminary

- **Standard Machine Unlearning Objective:**
  - Typical approach combine the losses:

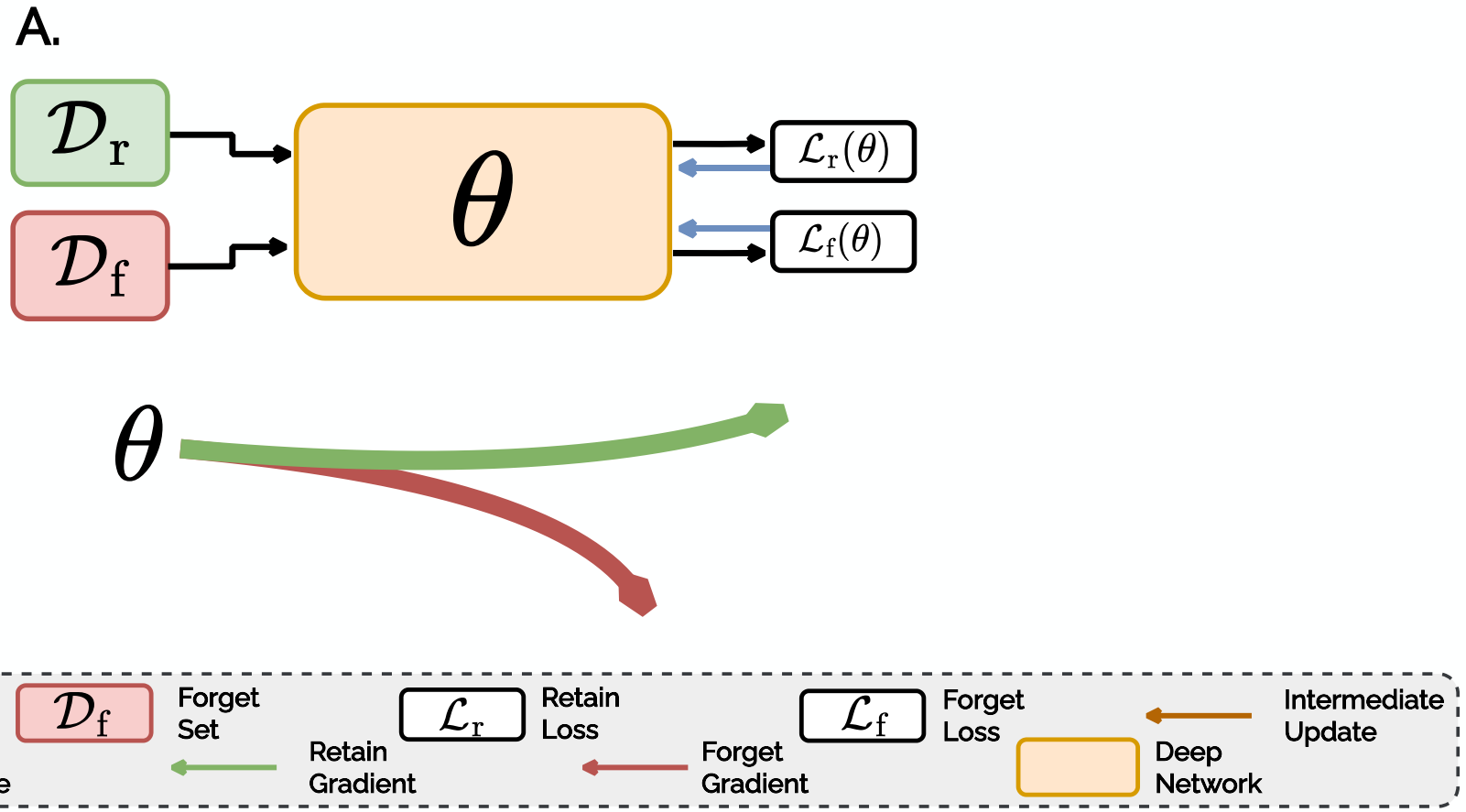
$$\min_{\theta} [\underbrace{\mathcal{L}_{\mathrm{r}}(\theta; \mathcal{D}_{\mathrm{r}})}_{\text{Retain Loss}} + \lambda \underbrace{\mathcal{L}_{\mathrm{f}}(\theta; \mathcal{D}_{\mathrm{f}})}_{\text{Forget Loss}}]$$

    - $\mathcal{L}_{\mathrm{r}}$ (retain loss) ensures performance (utility) on the *retain set* $\mathcal{D}_{\mathrm{r}}$.
    - $\mathcal{L}_{\mathrm{f}}$ *(*forget loss) pushes the model to "unlearn" the forget set $\mathcal{D}_{\mathrm{f}}$.
    - λ balances the two objectives.

  - **Issue**: When $\nabla \mathcal{L}_{\mathrm{r}}$ and $\nabla \mathcal{L}_{\mathrm{f}}$ point in opposing directions, updates can *cancel* each other's effect.

Our Proposition

Naïve Unlearning



Our Proposition

Learning to Unlearn while Retaining (LUR)

- **LUR Formulation:**
  - **Key Idea**: Treat forgetting as the higher-level objective and retaining as the lower-level objective in a two-step update.
  - **Implementation**:
    - **Intermediate step**: Update θ a little (with step α) using $\mathcal{L}_{\mathrm{r}}$.

$$\theta' = \theta - \alpha \nabla \mathcal{L}_{\mathrm{r}}(\theta)$$

    - **Final Update**: Evaluate $\mathcal{L}_{\mathrm{f}}$ at θ' (which accounts for the "retain" direction) and update θ.

    Formally:

$$\min_{\theta} \left[\mathcal{L}_{\mathrm{r}}(\theta; \mathcal{D}_{\mathrm{r}}) + \mathcal{L}_{\mathrm{f}}(\theta'; \mathcal{D}_{\mathrm{f}}) \right]$$

Our Proposition

Gradient Analysis

$$\mathcal{L}_{\mathrm{r}}(\theta) + \mathcal{L}_{\mathrm{f}}(\theta') = \mathcal{L}_{\mathrm{r}}(\theta) + \mathcal{L}_{\mathrm{f}}\left(\theta - \alpha \nabla \mathcal{L}_{\mathrm{r}}(\theta)\right)$$

- Gradient with respect to θ

$$\begin{aligned}
\mathbf{g}_\theta &= \nabla \mathcal{L}_{\mathrm{r}}(\theta) + \nabla \mathcal{L}_{\mathrm{f}}(\theta') \cdot \frac{\partial \theta'}{\partial \theta} \\
&= \nabla \mathcal{L}_{\mathrm{r}}(\theta) + \nabla \mathcal{L}_{\mathrm{f}}(\theta') \cdot \frac{\partial \left(\theta - \alpha \nabla \mathcal{L}_{\mathrm{r}}(\theta)\right)}{\partial \theta} \\
&= \nabla \mathcal{L}_{\mathrm{r}}(\theta) + \nabla \mathcal{L}_{\mathrm{f}}(\theta') \cdot \left(I - \alpha \nabla^2 \mathcal{L}_{\mathrm{r}}(\theta)\right).
\end{aligned}$$

Our Proposition

Gradient Analysis

- Gradient with respect to θ

$$\mathbf{g}_\theta = \nabla\mathcal{L}_\mathrm{r}(\theta) + \overbrace{\Big(\nabla\mathcal{L}_\mathrm{f}(\theta) + \nabla^2\mathcal{L}_\mathrm{f}(\theta)\ \underbrace{(\theta' - \theta)}_{=-\alpha\nabla\mathcal{L}_\mathrm{r}(\theta)} + \underbrace{\mathcal{O}(\|\theta' - \theta\|^2)}_{=\mathcal{O}(\alpha^2)}\Big)}^{=\nabla\mathcal{L}_\mathrm{f}(\theta')} \big(I - \alpha\,\nabla^2\mathcal{L}_\mathrm{r}(\theta)\big)$$

$$\Rightarrow\ \mathbf{g}_\theta = \nabla\mathcal{L}_\mathrm{r}(\theta) + \nabla\mathcal{L}_\mathrm{f}(\theta) + \nabla^2\mathcal{L}_\mathrm{f}(\theta)\,(\theta' - \theta) - \alpha\,\nabla^2\mathcal{L}_\mathrm{r}(\theta)\,\nabla\mathcal{L}_\mathrm{f}(\theta) + \mathcal{O}(\alpha^2)$$

$$(\text{using } \theta' - \theta = -\alpha\,\nabla\mathcal{L}_\mathrm{r}(\theta))$$

$$\Rightarrow\ \mathbf{g}_\theta = \nabla\mathcal{L}_\mathrm{r}(\theta) + \nabla\mathcal{L}_\mathrm{f}(\theta) - \alpha\,\nabla^2\mathcal{L}_\mathrm{f}(\theta)\,\nabla\mathcal{L}_\mathrm{r}(\theta) - \alpha\,\nabla^2\mathcal{L}_\mathrm{r}(\theta)\,\nabla\mathcal{L}_\mathrm{f}(\theta) + \mathcal{O}(\alpha^2)$$

$$\Rightarrow\ \mathbf{g}_\theta = \nabla\mathcal{L}_\mathrm{r}(\theta) + \nabla\mathcal{L}_\mathrm{f}(\theta) - \alpha\,\nabla\underbrace{\Big(\nabla\mathcal{L}_\mathrm{r}(\theta)\cdot\nabla\mathcal{L}_\mathrm{f}(\theta)\Big)}_{\text{Gradient Product}} + \mathcal{O}(\alpha^2).$$

$$\theta \leftarrow \theta - \eta\mathbf{g}_\theta$$

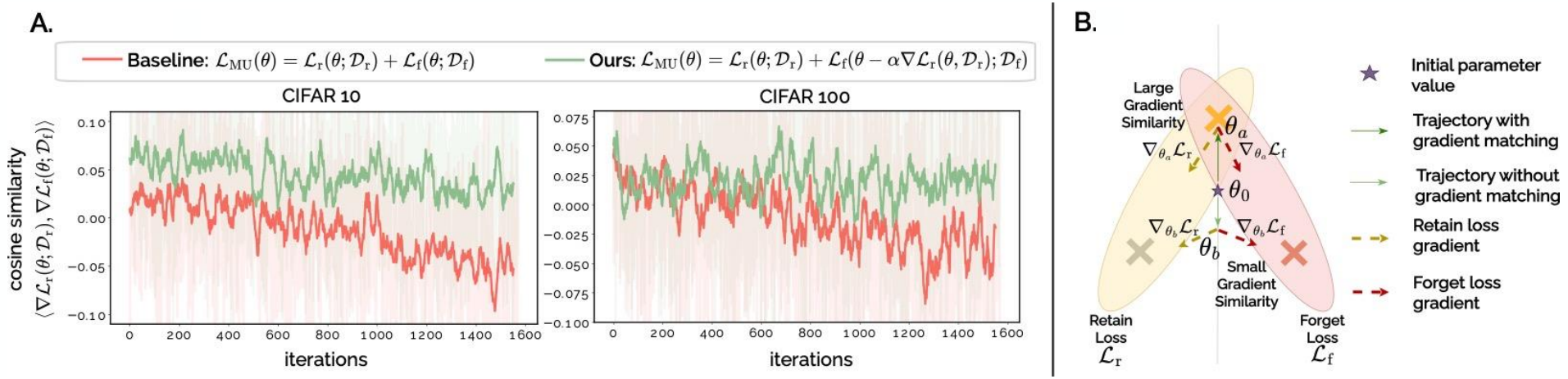
Our Proposition

Insight

- **Why does it help?**
  - The model "anticipates" how forgetting will affect retention and *implicitly* promotes alignment of $\nabla \mathcal{L}_{\mathrm{r}}$ with $\nabla \mathcal{L}_{\mathrm{f}}$.
  - Mathematically shown via **implicit gradient product regularization**:

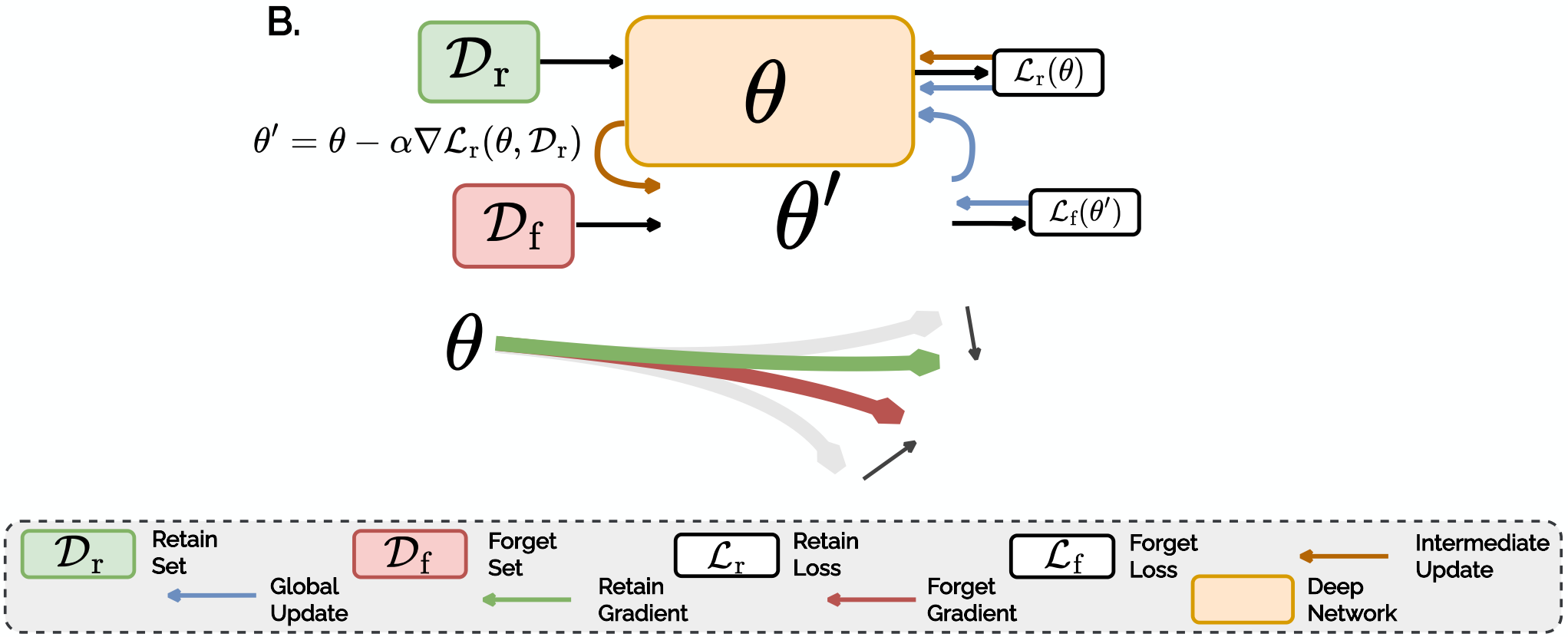
$$-\alpha \nabla \left(\nabla \mathcal{L}_{\mathrm{f}}(\theta) \cdot \nabla \mathcal{L}_{\mathrm{r}}(\theta) \right)$$

  - Encourages *maximizing* the inner product $\langle \nabla \mathcal{L}_{\mathrm{f}}(\theta) \cdot \nabla \mathcal{L}_{\mathrm{r}}(\theta) \rangle$, reducing gradient conflicts.



Our Proposition

Learning to Unlearn while Retaining



Quantitative Results

Experiments: Classification & Generative Models

- **Classification: Random data forgetting on CIFAR-10 and CIFAR-100**

Table 1. Performance comparison of different MU methods for image classification under 10% (*left*) and 50% (*right*) *random data forgetting* scenarios on CIFAR-10 [31] (*top*) and CIFAR-100 [31] (*bottom*) using ResNet-18 [18]. Results are reported in the format $a \pm b$, where a denotes the mean and b represents the standard deviation over 10 independent trials. A smaller performance gap relative to Retrain indicates better MU method performance. The metric **Avg. Gap** quantifies this gap by computing the average absolute performance differences across the considered evaluation metrics (see Section 5). Best results highlighted in **Maroon** and second best in **Navy**.

Method	Random Data Forgetting (10%)					Random Data Forgetting (50%)				
	UA (↑)	TA (↑)	RA (↑)	MIA (↑)	Avg. Gap (↓)	UA (↑)	TA (↑)	RA (↑)	MIA (↑)	Avg. Gap (↓)
CIFAR 10										
Retrain	5.19 ± 0.53	94.26 ± 0.14	100.00 ± 0.00	13.05 ± 0.64	0	7.83 ± 0.26	91.71 ± 0.30	100.00 ± 0.00	19.13 ± 0.55	0
FT [63]	0.85 ± 0.46	93.83 ± 0.45	99.84 ± 0.11	3.01 ± 0.93	3.74	0.50 ± 0.33	94.32 ± 0.07	99.96 ± 0.03	2.31 ± 1.08	6.70
GA [59]	0.34 ± 0.23	94.57 ± 0.01	99.62 ± 0.25	0.91 ± 0.29	4.42	0.40 ± 0.27	94.55 ± 0.06	99.62 ± 0.26	0.96 ± 0.40	7.20
IU [30]	1.92 ± 2.10	91.91 ± 2.73	98.01 ± 2.26	4.01 ± 3.44	4.16	2.46 ± 1.99	91.10 ± 5.25	97.62 ± 1.98	5.25 ± 3.01	5.56
BE [5]	0.59 ± 0.38	93.79 ± 0.15	99.41 ± 0.38	16.16 ± 0.78	2.19	0.43 ± 0.28	94.28 ± 0.04	99.59 ± 0.28	10.82 ± 0.89	4.67
BS [5]	0.40 ± 0.25	94.24 ± 0.07	99.56 ± 0.54	4.46 ± 0.33	3.46	0.42 ± 0.28	94.44 ± 0.03	99.60 ± 0.27	1.99 ± 0.08	6.92
ℓ_1-sparse [39]	5.83 ± 0.49	90.64 ± 0.52	96.64 ± 0.54	11.87 ± 0.61	2.20	2.58 ± 0.60	92.10 ± 0.24	98.89 ± 0.15	6.59 ± 0.80	4.82
SalUn [10]	1.93 ± 0.42	93.92 ± 0.25	99.89 ± 0.07	17.93 ± 0.37	2.15	7.85 ± 1.18	88.15 ± 0.90	95.02 ± 0.98	19.30 ± 2.81	**2.18**
SHs [65]	4.60 ± 1.48	92.92 ± 0.48	98.93 ± 0.57	9.56 ± 2.13	**1.62**	7.98 ± 5.31	88.32 ± 4.24	94.00 ± 4.87	15.52 ± 6.43	3.29
LUR (Ours)	5.52 ± 2.16	92.95 ± 0.29	99.21 ± 0.27	11.93 ± 1.01	**0.89**	6.79 ± 0.81	90.23 ± 0.63	97.19 ± 0.72	13.98 ± 0.63	**2.62**
CIFAR 100										
Retrain	24.87 ± 0.85	74.69 ± 0.08	99.98 ± 0.01	50.22 ± 0.62	0	32.83 ± 0.14	67.27 ± 0.45	99.99 ± 0.01	60.76 ± 0.21	0
FT [63]	2.02 ± 1.36	75.28 ± 0.12	99.95 ± 0.02	9.64 ± 3.60	16.01	1.83 ± 1.20	75.36 ± 0.36	99.97 ± 0.01	9.26 ± 2.84	22.65
GA [59]	2.00 ± 1.34	75.59 ± 0.11	98.24 ± 1.16	5.00 ± 2.25	17.68	1.85 ± 1.23	75.50 ± 0.10	98.22 ± 1.17	4.94 ± 1.96	24.2
IU [30]	4.33 ± 4.82	72.13 ± 4.58	96.14 ± 4.51	9.43 ± 5.98	16.93	3.14 ± 2.19	72.08 ± 2.41	97.17 ± 2.00	8.20 ± 4.10	22.47
BE [5]	2.06 ± 1.38	74.16 ± 0.09	98.12 ± 1.24	7.60 ± 3.05	16.96	2.65 ± 1.60	67.84 ± 0.58	97.27 ± 1.62	8.62 ± 2.19	21.40
BS [5]	2.35 ± 1.48	73.20 ± 0.18	97.93 ± 1.30	8.24 ± 3.23	17.01	4.69 ± 1.47	68.12 ± 0.18	95.41 ± 1.46	10.07 ± 1.99	21.07
ℓ_1-sparse [39]	3.65 ± 0.67	70.06 ± 0.46	96.35 ± 0.67	21.33 ± 1.95	14.59	9.83 ± 2.43	69.73 ± 1.27	97.35 ± 0.89	21.72 ± 1.44	16.79
SalUn [10]	11.44 ± 1.18	71.34 ± 0.48	99.40 ± 0.35	74.66 ± 2.48	10.45	15.19 ± 0.91	64.94 ± 0.48	98.89 ± 0.48	73.86 ± 1.98	**8.54**
SHs [65]	31.24 ± 1.81	73.17 ± 0.24	99.24 ± 0.30	42.42 ± 2.06	**4.11**	20.27 ± 2.28	67.58 ± 1.76	84.64 ± 2.79	28.68 ± 2.53	15.08
LUR (Ours)	29.57 ± 0.26	73.02 ± 0.18	99.29 ± 0.06	41.44 ± 0.10	**3.96**	32.68 ± 1.75	63.02 ± 0.90	87.18 ± 0.74	45.69 ± 2.79	**8.07**

Quantitative Results

Experiments: Classification & Generative Models

- **Classification Class-wise forgetting on CelebA-HQ-FIR (face recognition benchmark)**

Table 2. Performance comparison of different MU methods for image classification under *class-wise data forgetting* on Celeb-HQ-FIR [33, 47] using ResNet-34 [18]. The content in follow the same format of Table 1. Best results highlighted in **Maroon** and second best in **Navy**.

Method	Random Class (Identity) Forgetting (10%)					Random Class (Identity) Forgetting (50%)				
	UA (↑)	TA (↑)	RA (↑)	MIA (↑)	Avg. Gap (↓)	UA (↑)	TA (↑)	RA (↑)	MIA (↑)	Avg. Gap (↓)
Retrain	100.00 ± 0.00	87.02 ± 0.80	99.96 ± 0.01	100.00 ± 0.00	0	100.00 ± 0.00	88.09 ± 1.37	99.98 ± 0.03	100.00 ± 0.00	0
FT [63]	0.06 ± 0.12	88.59 ± 0.59	99.97 ± 7.02	5.28 ± 2.03	49.06	0.02 ± 0.03	90.71 ± 1.27	99.98 ± 0.03	3.08 ± 0.24	49.46
GA [59]	12.4 ± 8.71	81.22 ± 2.11	99.74 ± 0.26	51.37 ± 5.96	35.56	0.04 ± 0.02	88.41 ± 0.40	99.98 ± 0.03	2.44 ± 0.43	49.46
IU [30]	11.08 ± 10.25	70.24 ± 11.77	95.27 ± 5.07	29.59 ± 18.59	45.20	9.63 ± 8.78	68.40 ± 7.91	94.80 ± 6.61	30.10 ± 9.65	46.29
BE [5]	30.93 ± 2.73	44.11 ± 2.08	95.58 ± 1.23	46.24 ± 5.90	42.53	0.06 ± 0.02	83.12 ± 1.68	99.97 ± 0.02	3.62 ± 0.52	50.33
BS [5]	1.82 ± 1.92	81.92 ± 0.27	99.86 ± 0.03	45.93 ± 5.11	39.36	0.02 ± 0.03	87.80 ± 0.95	99.98 ± 0.03	2.76 ± 0.35	49.38
ℓ_1-sparse [39]	1.19 ± 0.72	89.37 ± 0.70	99.97 ± 0.00	76.78 ± 5.66	31.10	23.86 ± 3.63	90.29 ± 1.05	99.92 ± 0.10	99.86 ± 0.19	19.64
SalUn [10]	100.00 ± 0.00	78.36 ± 1.34	96.90 ± 1.11	100.00 ± 0.00	2.93	45.10 ± 2.60	90.92 ± 1.66	99.98 ± 0.03	99.95 ± 0.00	14.45
SHs [65]	98.48 ± 2.73	80.18 ± 6.60	97.20 ± 3.81	99.83 ± 0.35	**2.82**	99.24 ± 0.52	81.64 ± 3.75	99.14 ± 0.95	100.00 ± 0.00	**2.01**
LUR (Ours)	100.00 ± 0.00	86.61 ± 1.01	99.97 ± 0.00	100.00 ± 0.00	**0.10**	99.75 ± 0.20	91.64 ± 0.74	99.97 ± 0.02	100.00 ± 0.00	**0.95**

Quantitative Results

Experiments: Classification & Generative Models

- **Generative Models: Diffusion (Stable Diffusion)**

Table 4. Class-wise forgetting performance on Imagenette [26] using SD [50]. Best results highlighted in **Maroon** and second best in **Navy**.

Forget Class	ESD [12]		FMN [68]		SalUn [10]		**LUR** (Ours)	
	UA (↑)	FID (↓)	UA (↑)	FID (↓)	UA (↑)	FID (↓)	UA (↑)	FID (↓)
Tench	99.40	1.22	42.40	1.63	100.00	2.53	100.00	0.74
English Springer	100.00	1.02	27.20	1.75	100.00	0.79	100.00	0.97
Cassette Player	100.00	1.84	93.80	0.80	99.80	0.91	99.80	0.99
Chain Saw	96.80	1.48	48.40	0.94	100.00	1.58	100.00	1.30
Church	98.60	1.91	23.80	1.32	99.60	0.90	100.00	1.04
French Horn	99.80	1.08	45.00	0.99	100.00	0.94	100.00	0.75
Garbage Truck	100.00	2.71	41.40	0.92	100.00	0.91	100.00	0.94
Gas Pump	100.00	1.99	53.60	1.30	100.00	1.05	100.00	0.88
Golf Ball	99.60	0.80	15.40	1.05	98.80	1.45	100.00	0.88
Parachute	99.80	0.91	34.40	2.33	100.00	1.16	99.80	1.29
Average	99.40	1.50	42.54	1.30	**99.82**	**1.22**	**99.96**	**0.98**

Qualitative Results

Figure 4. *Quantitative and qualitative evalutation on I2P [53] benchmark.* **A.** Evaluation of the amount of nudity content detected by the NudeNet classifier [1] for each unlearning method. The bars represent the percentage decrease in the number of images from each nudity class compared to SD [50]. **B.** Generated images from SD with and without MU. Unlearning methods: SalUn [10], SHs [65], and **LUR** (Ours). Each column shows images from different MU methods using the same prompt (P_i) and seed. Prompt details in Appendix Table 5 (*Supplementary Material*).

Qualitative Results

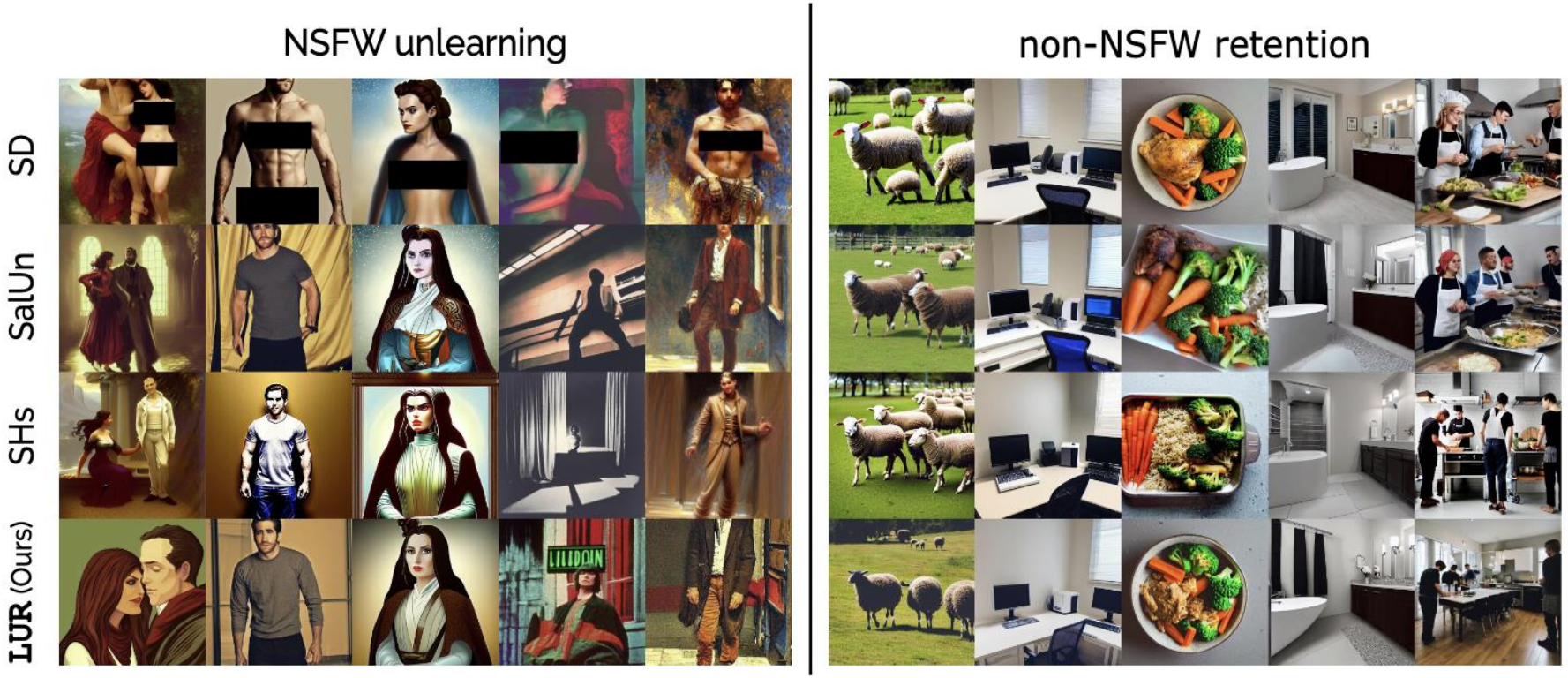


Figure 8. Example generations from prompts in I2P [53] (*left*) and COCO-10k [37, 69] (*right*) after unlearning the concept of *nudity*. Each column represents the generation from one prompts with a fixed seed. The prompts corresponding to the generated images are provided in Table 10.

Conclusion & Future Directions

- **Main Contributions:**
  - **LUR** framework aligns conflicting gradients for retain vs. forget → fosters a more robust approximate unlearning.
  - **Implicit Gradient Product Regularization** elegantly emerges from the two-step update.
  - Demonstrated **wide applicability** (discriminative and generative tasks), consistently improving over baseline and current MU methods.

- **Looking Ahead**:
  - Potential for multi-concept unlearning, language model unlearning, and other multimodal settings (VLMs).

Thank You