Decoupling the Class Label and the Target Concept in Machine Unlearning

Jianing Zhu, Bo Han, Jiangchao Yao, Jianliang Xu, Gang Niu, Masashi Sugiyama

TMLR
TRUSTWORTHY MACHINE LEARNING AND REASONING

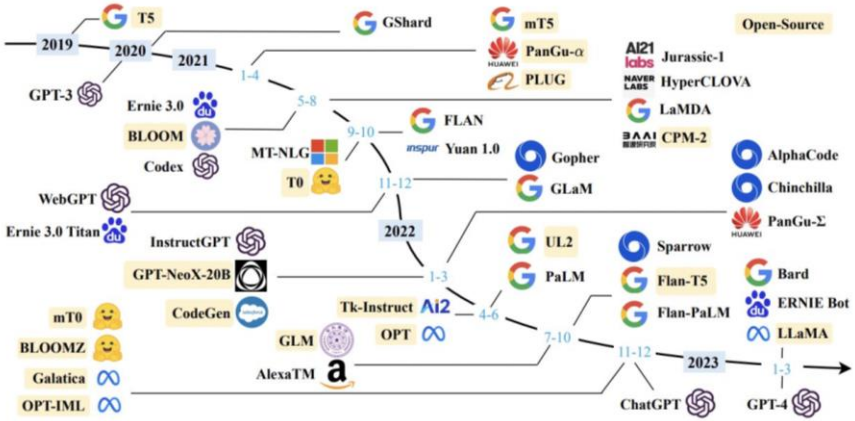
香港浸會大學
HONG KONG BAPTIST UNIVERSITY

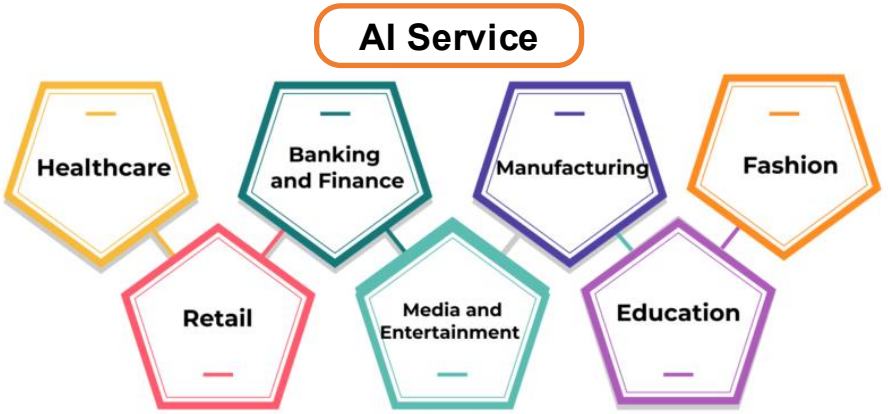
RIKEN AIP

上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

東京大学
THE UNIVERSITY OF TOKYO

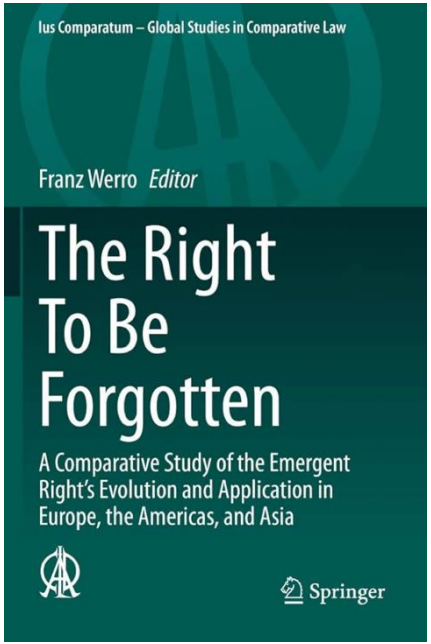
Background | Why Unlearning

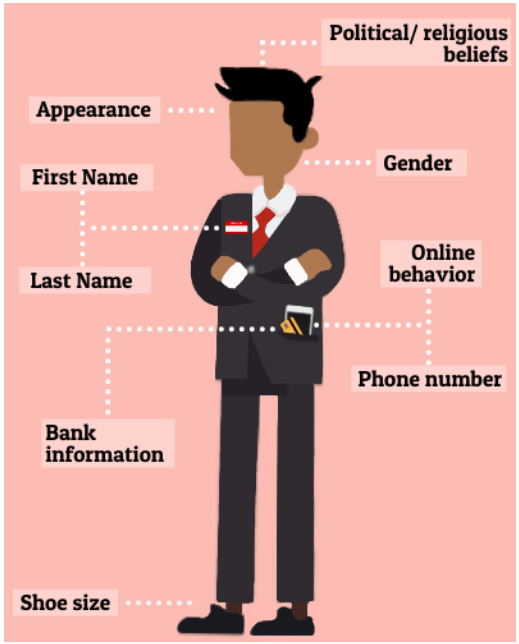




We often ask what AI can learn, but we rarely ask what it should forget.

Background | Why Unlearning





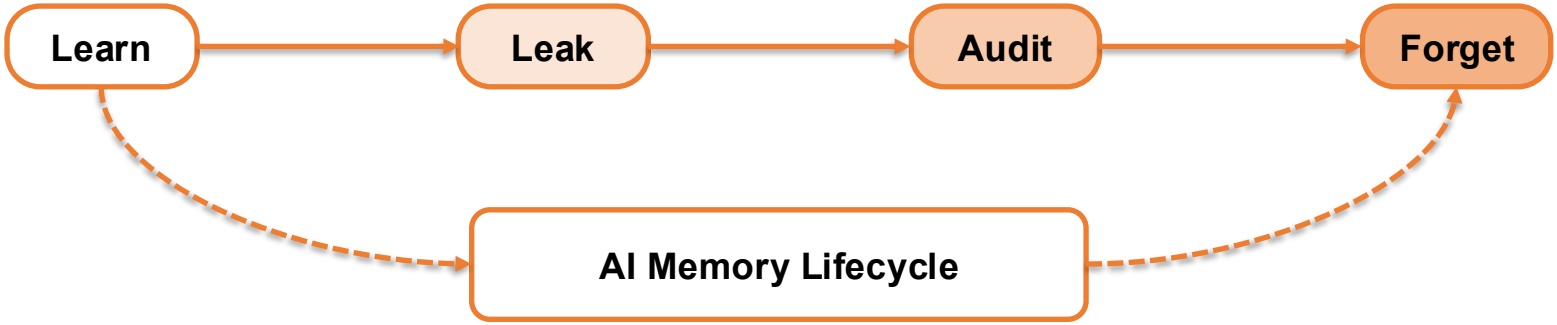
(Image credits: Dall-E and toppng.com)

GDPR: the right to be forgotten of personal sensitive data

Background | Why Unlearning

Grok model's disaster for "bikini" people online

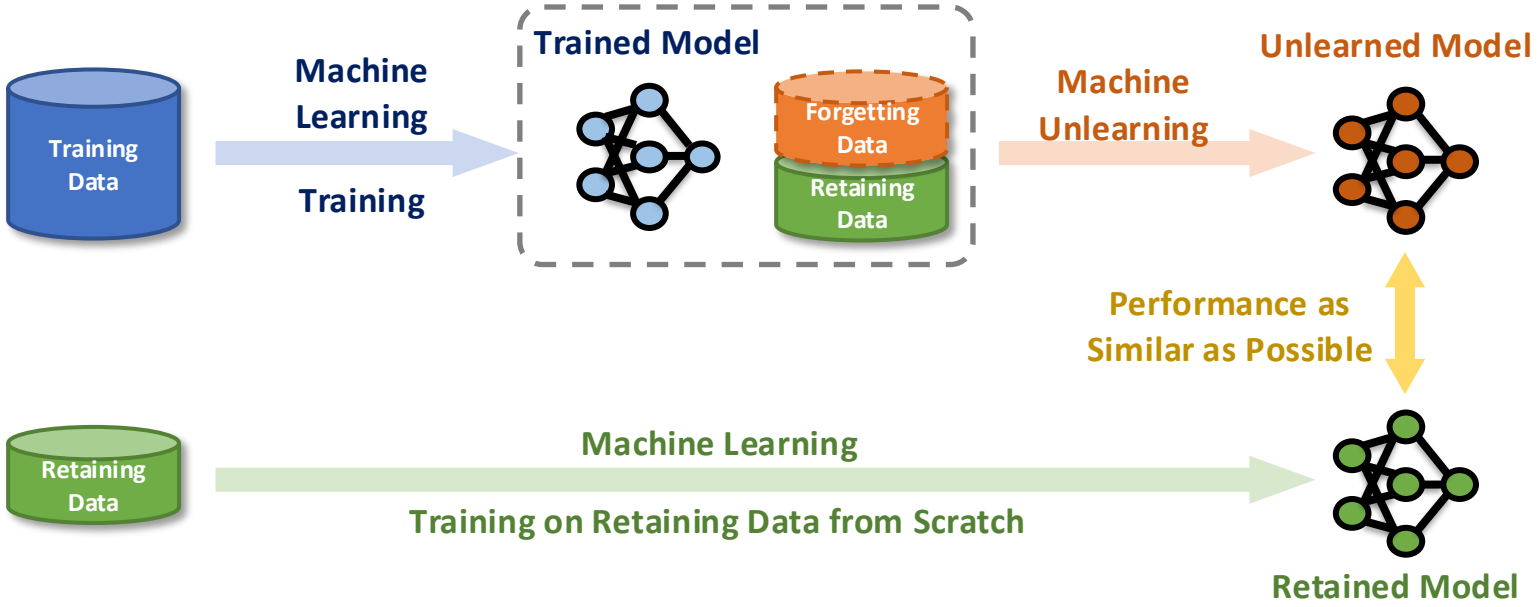
Background | Why Unlearning



Digital trust begins when AI learns not only to know — but also to forget.

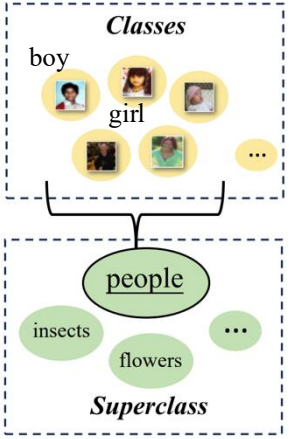
Background | Machine Unlearning

- ✓ **Machine unlearning** aims to remove the influence of the **forgetting data** from a trained model, such that it behaves similarly to a model (termed Retrained) retrained from scratch on the **retaining data**.



Decoupling the Class Label and the Target Concept in Machine Unlearning. *(ICLR 2026)*

Background | Label Domain Mismatch

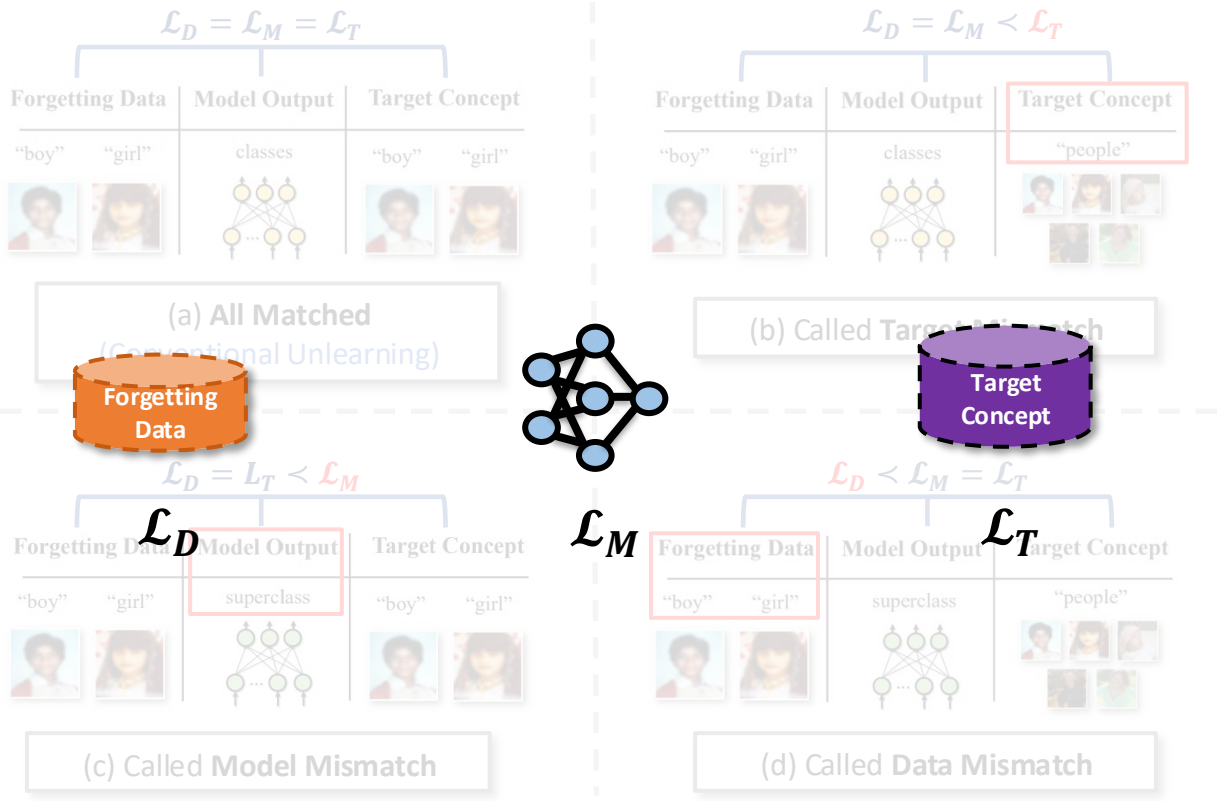


Label Domain of CIFAR-100

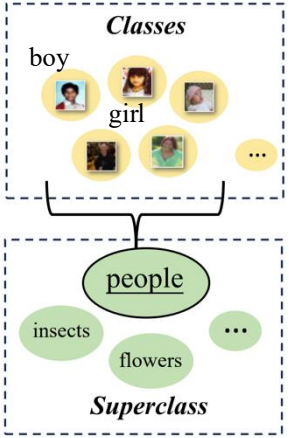
Label domains:

- ✓ $\mathcal{L}_D$: label domain of forgetting data.
- ✓ $\mathcal{L}_M$: label domain of the model output.
- ✓ $\mathcal{L}_T$: label domain of unlearning target concept.

Four Types of Unlearning Scenarios



Background | Label Domain Mismatch

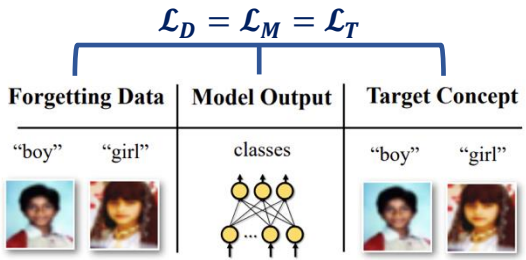


Label Domain of CIFAR-100

Label domains:

- ✓ $\mathcal{L}_D$: label domain of forgetting data.
- ✓ $\mathcal{L}_M$: label domain of the model output.
- ✓ $\mathcal{L}_T$: label domain of unlearning target concept.

Four Types of Unlearning Scenarios



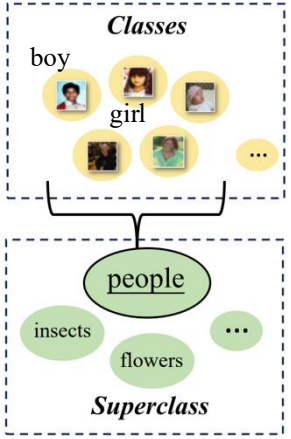
(a) **All Matched**
(Conventional Unlearning)

(b) Called **Target Mismatch**

(c) Called **Model Mismatch**

(d) Called **Data Mismatch**

Background | Label Domain Mismatch



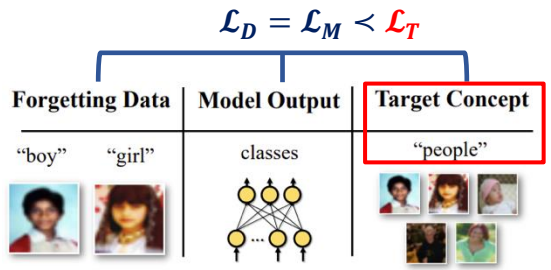
Label Domain of CIFAR-100

Label domains:

- ✓ $\mathcal{L}_D$: label domain of forgetting data.
- ✓ $\mathcal{L}_M$: label domain of the model output.
- ✓ $\mathcal{L}_T$: label domain of unlearning target concept.

Four Types of Unlearning Scenarios

(a) **All Matched** (Conventional Unlearning)

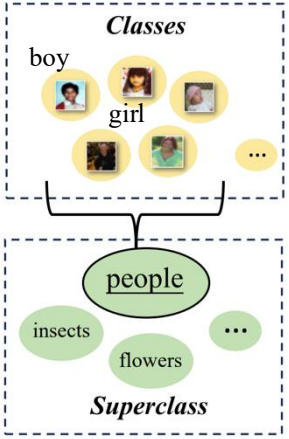


(b) Called **Target Mismatch**

(c) Called **Model Mismatch**

(d) Called **Data Mismatch**

Background | Label Domain Mismatch



Label Domain of CIFAR-100

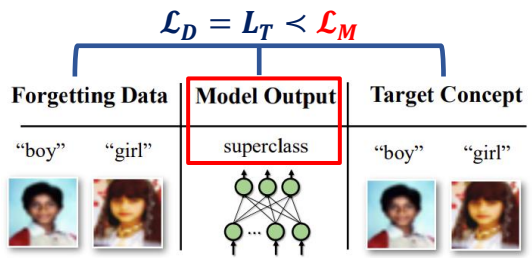
Label domains:

- ✓ $\mathcal{L}_D$: label domain of forgetting data.
- ✓ $\mathcal{L}_M$: label domain of the model output.
- ✓ $\mathcal{L}_T$: label domain of unlearning target concept.

Four Types of Unlearning Scenarios

(a) **All Matched** (Conventional Unlearning)

(b) Called **Target Mismatch**

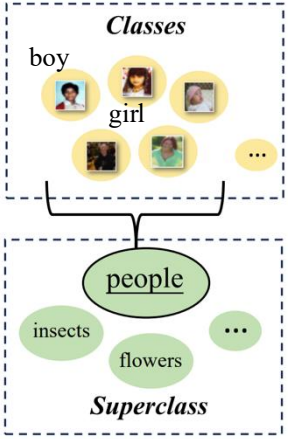


(c) Called **Model Mismatch**

(d) Called **Data Mismatch**

Decoupling the Class Label and the Target Concept in Machine Unlearning. *(ICLR 2026)*

Background | Label Domain Mismatch



Label Domain of CIFAR-100

Label domains:

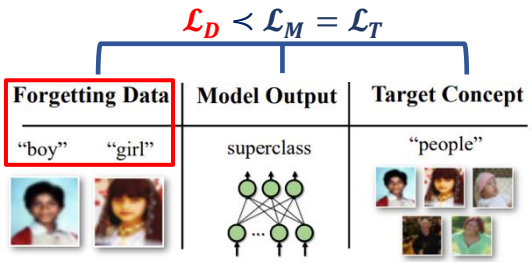
- ✓ $\mathcal{L}_D$: label domain of forgetting data.
- ✓ $\mathcal{L}_M$: label domain of the model output.
- ✓ $\mathcal{L}_T$: label domain of unlearning target concept.

Four Types of Unlearning Scenarios

(a) **All Matched** (Conventional Unlearning)

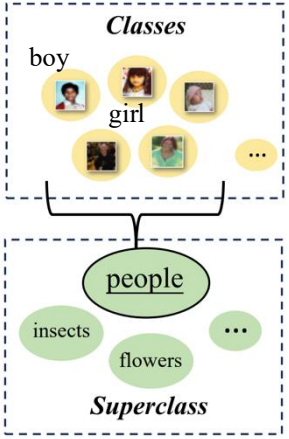
(b) Called **Target Mismatch**

(c) Called **Model Mismatch**



(d) Called **Data Mismatch**

Background | Label Domain Mismatch

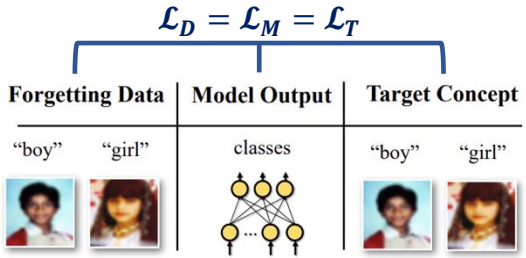


Label Domain of CIFAR-100

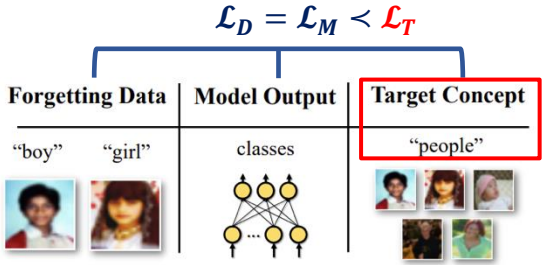
Label domains:

- ✓ $\mathcal{L}_D$: label domain of forgetting data.
- ✓ $\mathcal{L}_M$: label domain of the model output.
- ✓ $\mathcal{L}_T$: label domain of unlearning target concept.

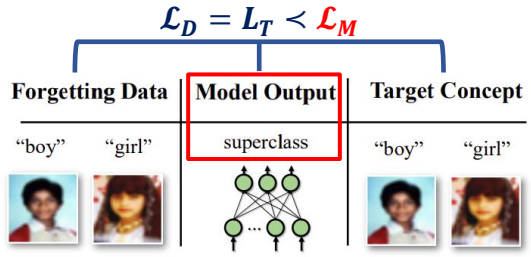
Four Types of Unlearning Scenarios



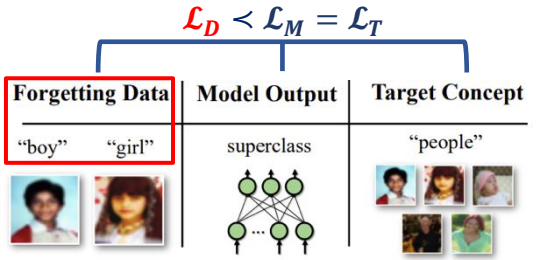
(a) **All Matched**
(Conventional Unlearning)



(b) Called **Target Mismatch**



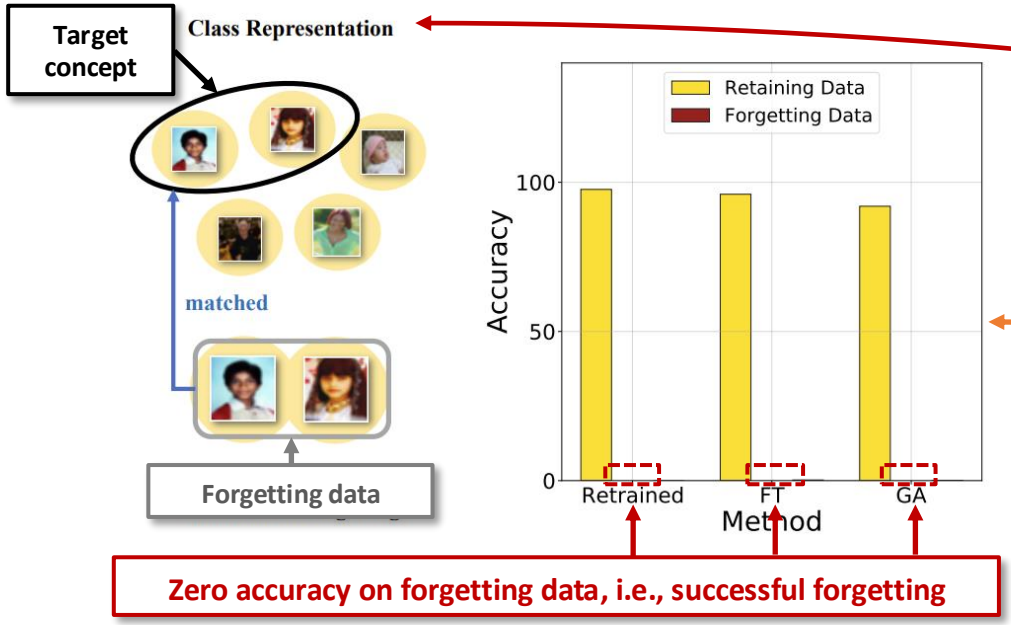
(c) Called **Model Mismatch**



(d) Called **Data Mismatch**

Decoupling the Class Label and the Target Concept in Machine Unlearning. *(ICLR 2026)*

Conventional Scenario | All Matched Forgetting



(a) **All Matched**
(Conventional Unlearning)

Observation 1: Representations are consistent.

Observation 2: FT and **GA** can achieve **similar performance** on retraining and forgetting data like **Retrained**.

Zero accuracy on forgetting data, i.e., successful forgetting

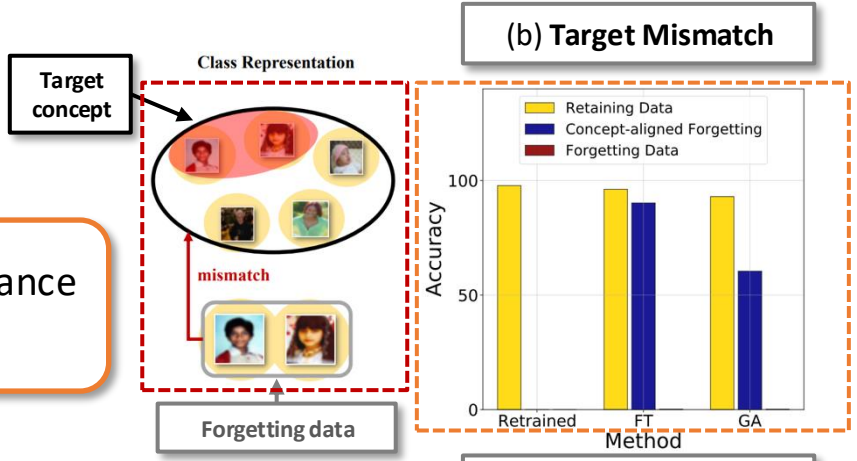
Note:
- ✓ **Retrained:** Retrain Model on Retaining Data.
- ✓ **FT:** Fine-tuning (Unlearning Method).
- ✓ **GA:** Gradient Ascent (Unlearning Method).

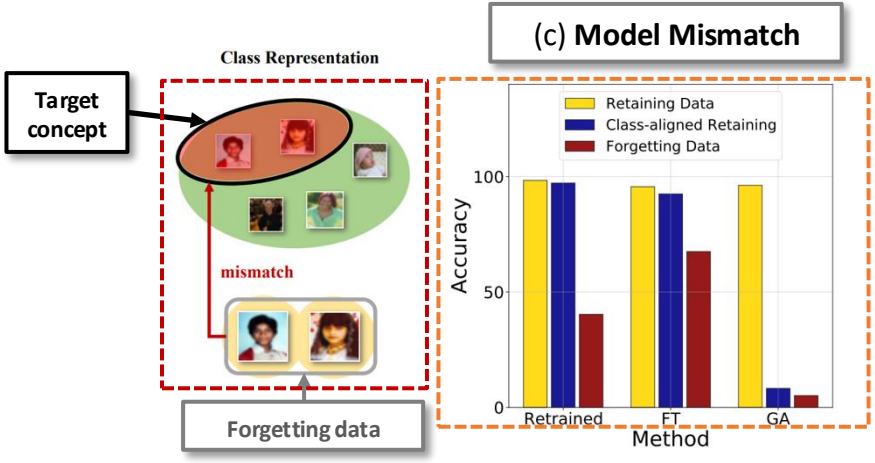
This is exactly what we aim for.

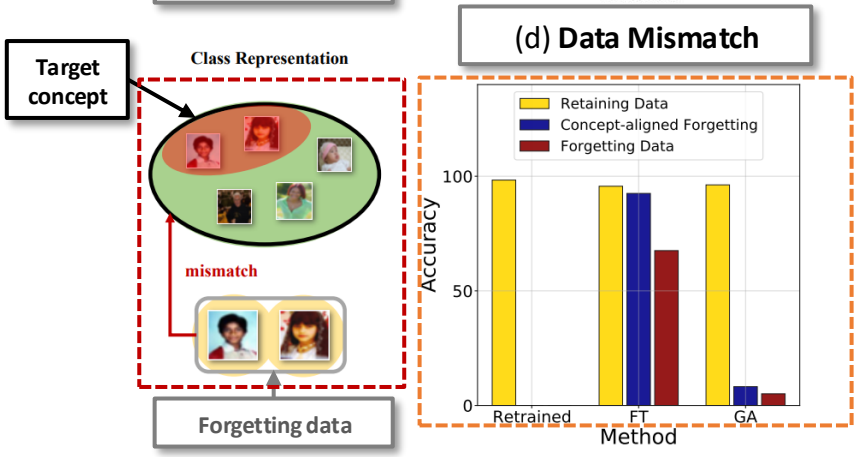
Challenge | Three Types of Mismatched Scenarios

Observation 3: Class representation mismatch issue.

Observation 4: FT and **GA** show different performance gaps compared with the **Retrained** models.





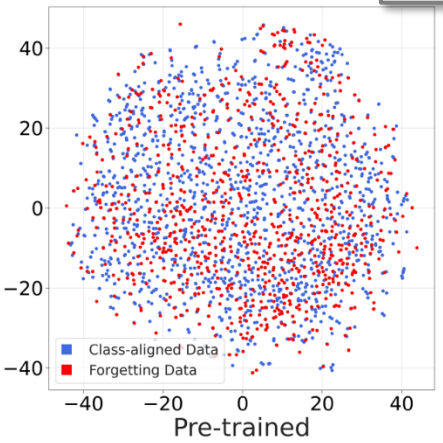


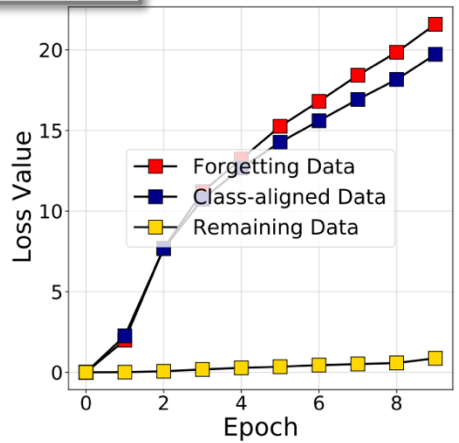
Decoupling the Class Label and the Target Concept in Machine Unlearning. *(ICLR 2026)*

Observation | Representation Entangled

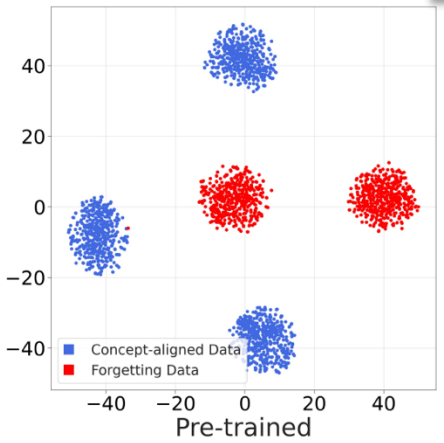
- **Visualization of the learned features** from the model trained by (left) superclass and (right) classes.
- **Loss value of** forgetting data, concept/class-aligned data, and the remaining data during GA.

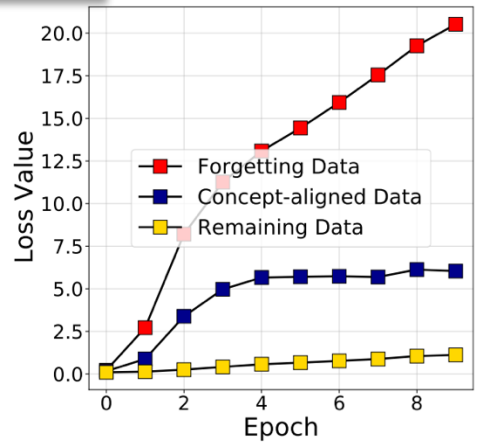
Superclass





Class



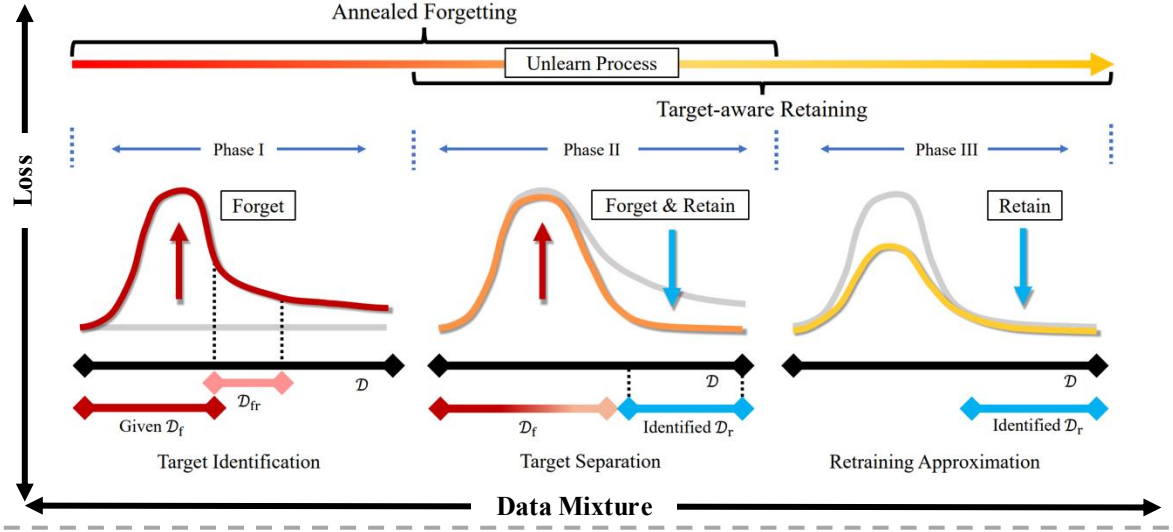


Observation 5: Representations of forgetting data and affected retaining data are closely entangled.

Observation 6: Unlearning of the forgetting data can unavoidably affect the representation of the other part.

Decoupling the Class Label and the Target Concept in Machine Unlearning. *(ICLR 2026)*

Methodology | Overview of TARF



TARget-aware Forgetting (TARF)

- ✓ Two terms consist of **Annealed Forgetting** and **Target-aware Retaining**.
- ✓ The training dynamics go through **Three Phases.**

Learning Objective of TARF

Phase II: Target Separation

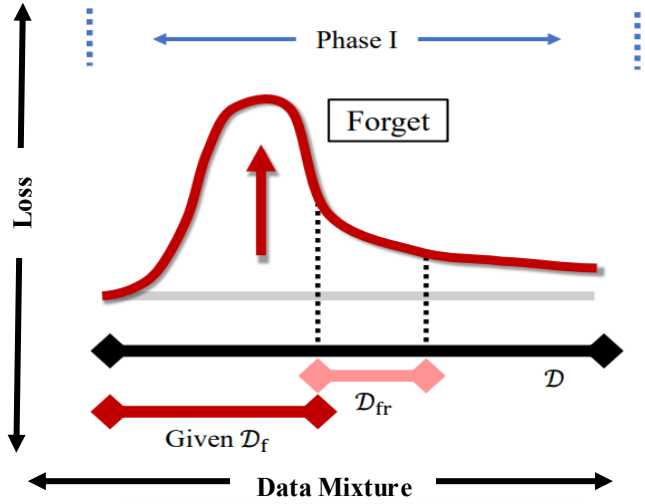
Phase I: Target Identification | **Phase III: Retraining Approximation**

$$L_{\text{TARF}} = k(t) \cdot \left(-\frac{1}{|\mathcal{D}_{\text{f}}|} \sum_{(x,y)\sim\mathcal{D}_{\text{f}}} \ell(f(x), y) \right) + \frac{1}{|\mathcal{D}_{\text{un}}|} \sum_{(x,y)\sim\mathcal{D}_{\text{un}}} \ell(f(x), y) \cdot \tau(x, y, t)$$

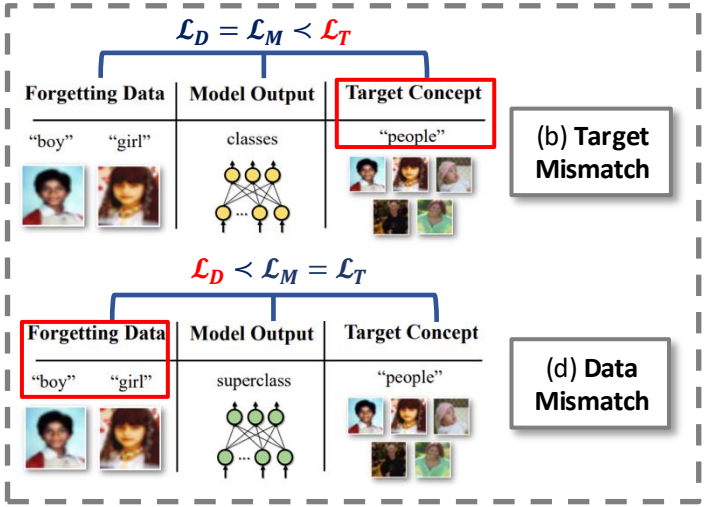
Annealed Forgetting | **Target-aware Retaining**

Training phases are controlled by t.

Methodology | Phase I: Target Identification



Aim to Address



- ✓ $\mathcal{D}_{\text{f}}$: Forgetting Data.
- ✓ $\mathcal{D}_{\text{fr}}$: False Retaining Data.

Objective of TARF-Phase-I

Phase I: Target Identification

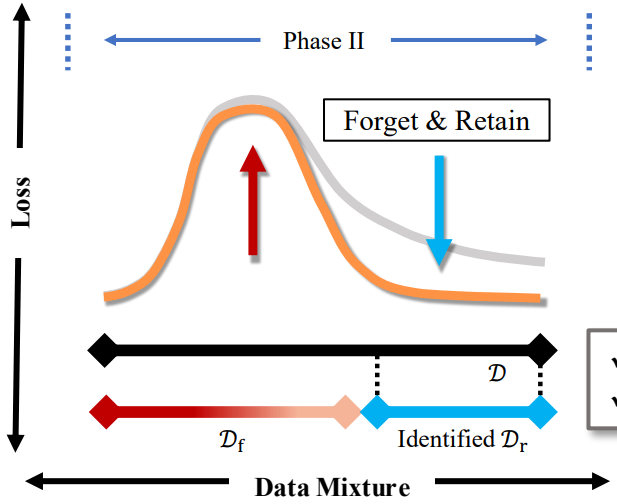
$$L_{\text{TARF-Phase-I}} = \boldsymbol{k(t)} \cdot \left(-\frac{1}{|\mathcal{D}_{\text{f}}|} \sum_{(x,y)\sim\mathcal{D}_{\text{f}}} \ell(f(x), y) \right)$$

Annealed Forgetting

Goal of TARF Phase I:

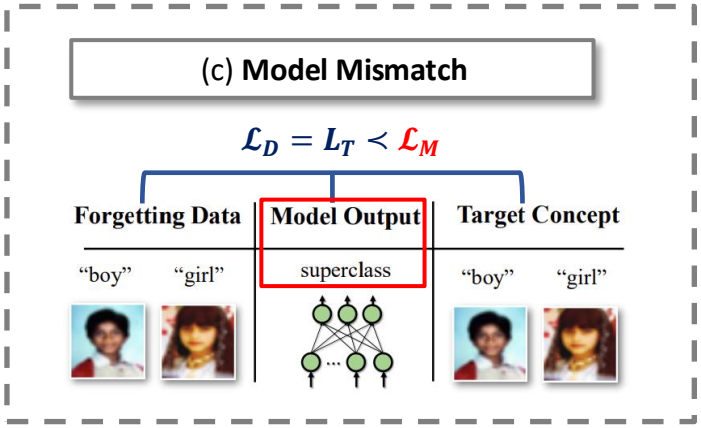
- ✓ Learn the **representations** of forgetting data.
- ✓ Identify **potential forgetting data**, i.e., false retaining data, from remaining data.

Methodology | Phase II: Target Separation



Aim to Address

- ✓ $\mathcal{D}_f$: Forgetting Data.
- ✓ $\mathcal{D}_r$: Retaining Data.



Objective of TARF-Phase-II

Phase II: Target Separation

$$L_{\text{TARF-Phase-II}} = k(t) \cdot \left(-\frac{1}{|\mathcal{D}_f|} \sum_{(x,y)\sim\mathcal{D}_f} \ell(f(x), y) \right) + \frac{1}{|\mathcal{D}_{un}|} \sum_{(x,y)\sim\mathcal{D}_{un}} \ell(f(x), y) \cdot \tau(x, y, t)$$

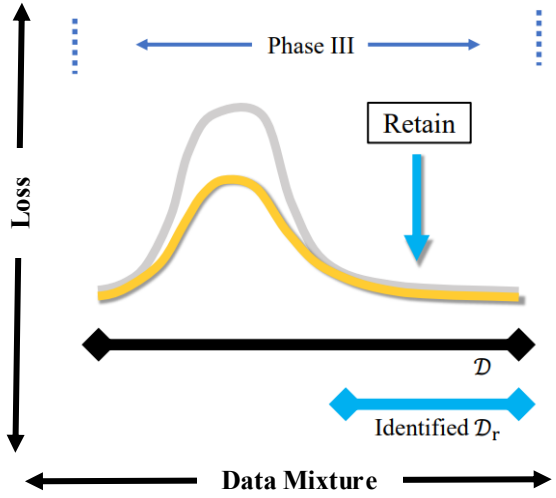
Annealed Forgetting | **Target-aware Retaining**

Goal of TARF Phase II:

- ✓ Learn the **representations** of forgetting data and retaining data.
- ✓ Encourage the model to **deconstruct the target concept** and **reconstruct representations** of the retaining part.

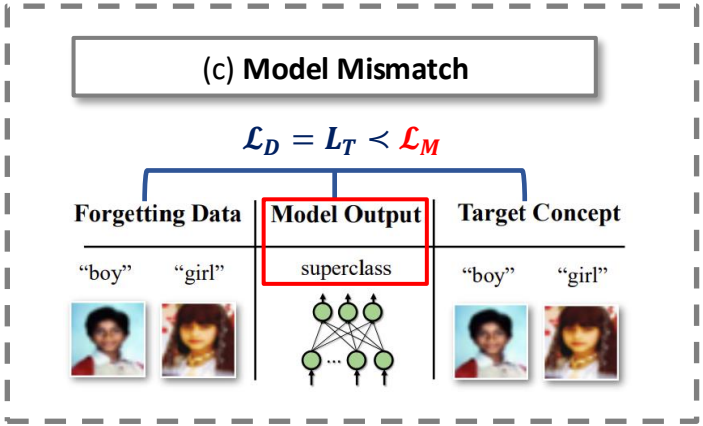
Decoupling the Class Label and the Target Concept in Machine Unlearning. *(ICLR 2026)*

Methodology | Phase III: Retraining Approximation



Aim to Address

- ✓ $\mathcal{D}_r$: Retaining Data.
- ✓ $\mathcal{D}_{un}$: Remaining Data, including True Retaining Data and False Retaining Data.



Objective of TARF-Phase-III

Phase III: Retraining Approximation

$$L_{\text{TARF-Phase-III}} = \underbrace{\frac{1}{|\mathcal{D}_{un}|} \sum_{(x,y)\sim\mathcal{D}_{un}} \ell(f(x), y) \cdot \tau(x, y, t)}_{\text{Target-aware Retaining}}$$

Goal of TARF Phase III:

- ✓ Learn to tune the **representations** of retaining data.
- ✓ **Prevent excessive forgetting**.
- ✓ **Approximate the retraining objective**.

Experiments | Empirical Evaluations

Type / $\mathcal{D}$	Dataset	CIFAR-10						CIFAR-100					
	Method / Metrics	UA	RA	TA	MIA	Gap↓	TIME↓	UA	RA	TA	MIA	Gap↓	TIME↓
All matched	Retrained (Ref.)	0.00	99.51	94.69	100.00	-	43.3	0.00	97.85	76.03	100.00	-	43.2
	FT [58]	1.07	98.62	92.36	100.00	1.07	4.43	0.67	96.32	72.34	100.00	1.47	5.02
	RL [56]	4.13	97.65	91.23	100.00	2.36	4.88	1.00	96.09	72.00	100.00	1.70	4.96
	GA [28]	0.49	95.24	88.17	99.78	2.88	**0.25**	1.33	94.74	68.56	99.89	3.01	**0.06**
	IU [29]	0.22	88.15	82.38	99.96	5.99	0.45	0.00	37.61	29.58	100.00	26.67	0.51
	BS [6]	25.04	87.94	80.90	88.67	15.43	0.82	4.60	90.18	63.66	99.55	6.27	0.78
	L_1-sparse [30]	0.00	94.20	89.77	100.00	2.56	4.39	0.00	94.60	71.57	100.00	1.93	4.39
	SalUn [11]	0.00	91.32	86.87	100.00	4.00	5.65	0.00	75.34	62.14	100.00	9.10	5.75
	SCRUB [37]	0.00	99.94	91.00	100.00	1.03	2.88	0.00	99.98	76.75	100.00	**0.71**	3.23
	TARF (ours)	0.00	98.23	91.95	100.00	**1.01**	4.21	0.00	96.90	72.53	100.00	1.11	4.68
Model mismatch	Retrained (Ref.)	87.76	99.58	95.91	20.57	-	43.8	88.22	98.58	78.50	25.78	-	43.8
	FT [58]	94.67	98.53	93.56	9.56	5.33	4.29	92.67	95.02	79.34	16.33	4.58	4.86
	RL [56]	53.69	97.85	92.39	96.60	28.84	4.82	80.11	95.83	79.83	99.00	21.35	4.93
	GA [28]	5.76	86.99	82.20	94.98	45.68	**0.25**	6.78	94.83	76.96	97.78	39.68	**0.06**
	IU [29]	23.69	87.34	82.57	89.87	39.74	0.44	34.67	96.83	79.08	86.44	29.14	0.49
	BS [6]	10.29	50.77	49.39	95.96	62.05	0.79	18.11	95.90	72.28	95.22	37.14	0.89
	L_1-sparse [30]	93.11	94.76	91.63	14.44	5.15	4.24	90.22	94.78	78.81	18.88	3.25	5.00
	SalUn [11]	8.91	93.95	84.38	99.32	43.69	6.04	66.33	78.83	70.78	77.00	25.15	5.97
	SCRUB [37]	95.14	99.81	94.22	15.38	3.61	3.06	91.44	99.74	79.23	21.11	2.45	4.12
	TARF (ours)	91.11	97.49	92.49	17.82	**2.90**	4.31	86.67	97.05	80.07	26.00	**1.21**	4.81
Target mismatch	Retrained (Ref.)	0.00	99.38	93.85	100.00	-	52.1	0.00	97.85	73.72	100.00	-	53.2
	FT [58]	50.43	98.47	91.65	50.44	25.78	4.38	58.18	96.32	72.53	46.76	28.54	5.00
	RL [56]	51.25	97.56	90.90	56.23	24.95	4.79	58.89	96.05	72.20	46.98	28.81	4.93
	GA [28]	40.82	97.01	89.51	64.32	20.80	**0.26**	21.38	96.64	70.22	90.67	8.86	**0.05**
	IU [29]	44.51	88.07	81.80	58.73	27.29	0.44	30.62	37.19	29.58	63.69	42.93	0.50
	BS [6]	53.62	88.65	75.39	76.33	26.62	0.82	40.44	98.32	68.66	85.16	15.20	0.97
	L_1-sparse [30]	49.47	93.61	88.83	51.24	27.26	4.38	56.09	94.63	72.00	48.04	28.25	4.78
	SalUn [11]	46.63	91.08	86.31	60.94	25.38	5.90	59.64	75.52	62.37	65.96	27.35	5.81
	SCRUB [37]	49.98	99.94	92.10	50.18	25.53	2.89	59.64	99.99	75.32	44.89	29.90	3.52
	TARF (ours)	0.06	97.57	90.81	100.00	**1.23**	4.23	0.31	97.35	73.68	100.00	**0.21**	4.85
Data mismatch	Retrained (Ref.)	0.00	99.54	95.56	100.00	-	52.1	0.00	98.50	80.15	100.00	-	53.2
	FT [58]	96.79	98.49	93.26	6.48	48.41	4.32	82.62	95.66	79.77	37.24	37.15	4.93
	RL [56]	76.47	97.68	91.93	49.81	33.04	4.76	89.78	96.82	79.90	70.76	30.49	4.97
	GA [28]	8.69	96.41	90.78	93.03	5.89	**0.25**	6.00	97.65	79.23	98.04	2.43	**0.05**
	IU [29]	22.84	95.50	89.54	88.57	11.08	0.44	31.51	98.96	78.20	88.09	11.46	0.48
	BS [6]	16.70	61.21	49.76	92.24	22.37	0.82	15.38	98.50	72.28	96.22	6.76	0.96
	L_1-sparse [30]	95.76	94.31	91.08	9.52	48.99	4.78	88.31	94.91	79.02	22.49	42.64	5.03
	SalUn [11]	51.77	93.87	90.46	63.52	24.75	5.72	72.93	78.87	71.04	54.13	36.89	5.72
	SCRUB [37]	97.13	99.89	95.03	10.99	46.76	2.94	95.50	99.79	79.68	15.11	45.54	3.68
	TARF (ours)	0.00	98.17	93.09	100.00	**0.96**	4.22	0.00	95.01	78.98	100.00	**1.17**	4.78

- ✓ **Dataset:** CIFAR-10 and CIFAR-100
- ✓ **Trained Model:** ResNet-18, WideResNet-50
- ✓ **Golden Reference Method:** Retrain model using Retaining data.
- ✓ **Gap: Average performance gap** between unlearned model and retrained reference model across four metrics (UA, RA, TA and MIA) [1].

Observation: TARF can **consistently perform better** (or comparable) over other unlearning baseline methods.

See our paper for more results.

[1] Jia et al. Model sparsity can simplify machine unlearning. In *NeurIPS*, 2023.

Take Home Messages

- ✓ **New and Practical Unlearning Scenarios:** Compared to conventional label-aligned unlearning, **decoupling the class label** from the target concept reflects a more **realistic and practical** unlearning scenario.
- ✓ **Formal Formulation of Label Domain Mismatch in Unlearning:** We formally define and formulate the three types of label domain mismatch in unlearning, i.e., **target mismatch**, **model mismatch**, and **data mismatch**.
- ✓ **General Unlearning Framework:** We propose a novel unlearning method **TARF**, which assigns an annealed gradient ascent on the **identified potential forgetting data** and the normal gradient descent on the **selected retaining data**.