Challenging Forgets: Unveiling the Worst-Case Forget Sets in Machine Unlearning

Supplementary Material - Appendix

Chongyu Fan*,1, Jiancheng Liu*,1, Alfred Hero[2], Sijia Liu[1,3]

[1] OPTML@CSE, Michigan State University
[2] EECS, University of Michigan, Ann Arbor
[3] MIT-IBM Watson AI Lab, IBM Research
*Equal contribution
{fanchon2,liujia45,liusiji5}@msu.edu, hero@eecs.umich.edu

A Additional Derivations

A.1 The closed-form projection operation

Recall from (7) that $\text{Proj}_{\mathbf{w}\in\mathcal{S}}(\mathbf{a})$ indicates the projection of a constant $\mathbf{a}$ onto the constraint set $\mathcal{S}$. This projection operation is defined as solving the auxiliary minimization problem $\text{Proj}_{\mathbf{w}\in\mathcal{S}}(\mathbf{a}) = \arg\min_{\mathbf{w}\in\mathcal{S}} \|\mathbf{w}-\mathbf{a}\|_2^2$, where $\mathcal{S} = \{\mathbf{w}|\mathbf{w}\in[\mathbf{0},\mathbf{1}], \mathbf{1}^\top\mathbf{w} = m\}$. It is worth noting that we have relaxed the original binary constraint into its continuous counterpart, with $\mathbf{w}\in[\mathbf{0},\mathbf{1}]$. The relaxed constraint is given by the intersection of the box constraint $\mathbf{w}\in[\mathbf{0},\mathbf{1}]$ and the hyperplane $\mathbf{1}^\top\mathbf{w} = m$.

According to [15, Proposition 1], the solution of the above projection problem yields

$$\text{Proj}_{\mathbf{w}\in\mathcal{S}}(\mathbf{a}) = P_{[\mathbf{0},\mathbf{1}]}[\mathbf{a}-\lambda\mathbf{1}], \tag{9}$$

where the variable λ is given by the root of the equation $\mathbf{1}^\top P_{[\mathbf{0},\mathbf{1}]}[\mathbf{a}-\lambda\mathbf{1}] = m$, and $P_{[\mathbf{0},\mathbf{1}]}$ is an element-wise thresholding function

$$P_{[\mathbf{0},\mathbf{1}]}[x_i] = \begin{cases} 0, & x < 0 \\ x, & x\in[0,1] \\ 1, & x > 1 \end{cases} \tag{10}$$

for the ith entry of a vector $\mathbf{x}$. We also remark that finding the root of the non-linear equation with respect to λ can be achieved using the bi-section method [1].

A.2 Worst-case forget set identification in class-wise and prompt-wise forgetting

For class-wise forgetting, the data selection variables $\mathbf{w}$ in (2)-(3) can be reinterpreted as class selection variables. Here, $w_i = 1$ indicates the selection of the ith

class for targeted worst-case unlearning. Accordingly, the *upper- and lower-level objectives* of the BLO problem (3) can be modified to

$$f(\mathbf{w}, \boldsymbol{\theta}_{\mathrm{u}}(\mathbf{w})) = \sum_i \left(w_i \mathbb{E}_{\mathbf{z}\in\mathcal{D}_i}[\ell(\boldsymbol{\theta}_{\mathrm{u}}(\mathbf{w}); \mathbf{z})]\right) + \gamma\|\mathbf{w}\|_2^2, \tag{11}$$

$$\ell_{\mathrm{MU}}(\boldsymbol{\theta}; \mathbf{w}) = \sum_i \left(w_i \mathbb{E}_{\mathbf{z}\in\mathcal{D}_i}[\ell_{\mathrm{f}}(\boldsymbol{\theta}; \mathbf{z})] + (1 - w_i)\mathbb{E}_{\mathbf{z}\in\mathcal{D}_i}[\ell_{\mathrm{r}}(\boldsymbol{\theta}; \mathbf{z})]\right), \tag{12}$$

where $\mathcal{D}_i$ represents the dataset corresponding to class i, $\mathbb{E}_{\mathbf{z}\in\mathcal{D}_i}[\ell(\boldsymbol{\theta}; \mathbf{z})]$ denotes the training loss over $\mathcal{D}_i$, and recalling that $\ell_{\mathrm{r}} = -\ell_{\mathrm{f}} = \ell$. With these specifications in place, the task of identifying the worst-case class-wise forget set can be similarly addressed by resolving the BLO problem (3).

In the context of prompt-wise forgetting, we interpret the data selection variables $\mathbf{w}$ as prompt selection variables. Here a prompt refers to a text condition used for text-to-image generation, and is known as a 'concept' within MU for generative models [5]. Thus, the unlearned generative model, when $w_i = 1$, corresponds to the scenario of removing the influence of the ith concept from the generative modeling process. Extended from the concept erasing framework for diffusion models [5], identifying the worst-case prompt-wise forget set can be formulated under the same BLO structure (3). The *upper-level* objective function can be written as

$$f(\mathbf{w}, \boldsymbol{\theta}_{\mathrm{u}}(\mathbf{w})) = \sum_i \left(w_i \mathbb{E}_{t,\boldsymbol{\epsilon}}\left[\|\boldsymbol{\epsilon}(\mathbf{x}_t|c_i, \boldsymbol{\theta}_{\mathrm{u}}(\mathbf{w})) - \boldsymbol{\epsilon}(\mathbf{x}_t|c_i, \boldsymbol{\theta}_{\mathrm{o}})\|_2^2\right]\right) + \gamma\|\mathbf{w}\|_2^2, \tag{13}$$

where $\mathbf{x}_t$ represents the latent feature subject to standard Gaussian noise injection, $\boldsymbol{\epsilon}$, during the diffusion step t through a forward diffusion process, and $\boldsymbol{\epsilon}(\mathbf{x}_t|c, \boldsymbol{\theta})$ denotes the noise estimator for $\mathbf{x}_t$ within a diffusion model parameterized by $\boldsymbol{\theta}$ and conditioned on the text prompt c. The loss term $\|\boldsymbol{\epsilon}(\mathbf{x}_t|c_i, \boldsymbol{\theta}_{\mathrm{u}}(\mathbf{w})) - \boldsymbol{\epsilon}(\mathbf{x}_t|c_i, \boldsymbol{\theta}_{\mathrm{o}})\|_2^2$ penalizes the mean squared error of image generation using the unlearned model $\boldsymbol{\theta}_{\mathrm{u}}(\mathbf{w})$ and the original diffusion model $\boldsymbol{\theta}_{\mathrm{o}}$, respectively. Therefore, minimizing (13) challenges the unlearning efficacy of $\boldsymbol{\theta}_{\mathrm{u}}(\mathbf{w})$ regarding the concept c_i to be erased (when $w_i = 1$), by steering its noise estimation accuracy towards that of the original model prior to unlearning. Furthermore, we specify the *lower-level* objective function of (3) as the Erasing Stable Diffusion (ESD) loss, $\ell_{\mathrm{ESD}}(\boldsymbol{\theta}; \mathbf{c}_i)$, developed in [5]. This loss function is designed to remove the influence of the concept c_i from the diffusion model $\boldsymbol{\theta}$. Consequently, this lower-level objective function is given by $\ell_{\mathrm{MU}}(\boldsymbol{\theta}; \mathbf{w}) = \sum_i (w_i \ell_{\mathrm{ESD}}(\boldsymbol{\theta}; \mathbf{c}_i))$.

B Additional Implementation Details

B.1 Worst-case forget set in data-wise unlearning

For the exact unlearning method Retrain, the training process comprises 182 epochs, utilizing the SGD optimizer with a cosine-scheduled learning rate initially set to 0.1. For FT [14], RL [7], EU-k [6], CF-k [6], and SCRUB [11],

the unlearning process takes 10 epochs, during which the optimal learning rate is searched within the range of $[10^{-4}, 10^{-1}]$, and $k = 1$ is set for EU-k and CF-k. For ℓ_1-sparse [9], the unlearning-enabled model updating process also takes 10 epochs, searching for the optimal sparse ratio in the range $[10^{-6}, 10^{-4}]$ and exploring the most appropriate learning rate within $[10^{-3}, 10^{-1}]$. Regarding the method BS [2], the step size of fast gradient sign method (FGSM) is fixed at 0.1. Both BS and BE [2] undergo a 10-epoch fine-tuning process, during which the optimal learning rate is sought within the interval $[10^{-8}, 10^{-4}]$. Finally, for SalUn [4], we conducted a 10-epoch fine-tuning phase, exploring learning rates within the range $[10^{-4}, 10^{-2}]$, and investigating sparsity ratios in the range $[0.1, 0.9]$.

B.2 Worst-case forget set in prompt-wise unlearning

In the UnlearnCanvas benchmark dataset [16] for image generation, we select 10 objects and 10 styles, leading to 100 prompt combinations. The objects include Horses, Towers, Humans, Flowers, Birds, Trees, Waterfalls, Jellyfish, Sandwiches, and Dogs, while the styles feature Crayon, Ukiyoe, Mosaic, Sketch, Dadaism, Winter, Van Gogh, Rust, Glowing Sunset, and Red Blue Ink. We target 10% of these combinations for the unlearning task.

For prompt-wise worst-case forget set identification, we utilize the Erased Stable Diffusion (ESD) method combined with SignSGD, setting a learning rate of 10^{-5} for 1000 iterations when specifying (8). After identifying the worst-case forget prompts, we apply ESD again, this time with a learning rate of 3×10^{-7} for 1000 iterations to unlearn these prompts. During image generation, DDIM is specified using 100 time steps and a conditional scale of 7.5.

C Additional Experiment Results

C.1 Additional results of Table 3

As an expansion of Table 3, **Table A1** details the performance of various approximate unlearning methods for both random and worst-case forget sets, with data forgetting ratios of 1%, 5%, 10%, and 20% on CIFAR-10. For worst-case forget sets, relabeling-free unlearning methods often follow a performance trend similar to Retrain. On the other hand, relabeling-based unlearning methods display a significant performance discrepancy from Retrain, highlighting the impact of the unlearning strategy on method efficacy.

C.2 Additional results on CIFAR-100 and Tiny ImageNet

In **Table A2**, we present the performance of various unlearning methods under random and worst-case forget sets at a 10% forgetting data ratio on the additional datasets, CIFAR-100 [10] and Tiny ImageNet [12]. When subjected to evaluation on worst-case forget sets, relabeling-free approximate unlearning

Table A1: Performance of various unlearning methods under random forget sets and worst-case forget sets on CIFAR-10 using ResNet-18 for different forgetting ratio (including 1%, 5%, 10% and 20%). The result format follows Table 3.

Methods	Random Forget Set					Worst-Case Forget Set				
	UA	MIA	RA	TA	Avg. Gap	UA	MIA	RA	TA	Avg. Gap
1%-Data Forgetting										
Retrain	$5.85_{\pm 0.69}$	$12.89_{\pm 1.27}$	$99.96_{\pm 0.00}$	$93.17_{\pm 0.15}$	0.00	$0.00_{\pm 0.00}$	$0.00_{\pm 0.00}$	$99.95_{\pm 0.02}$	$93.45_{\pm 0.17}$	0.00
FT	$8.93_{\pm 2.74}$ (3.08)	$14.40_{\pm 1.98}$ (1.51)	$94.52_{\pm 1.79}$ (5.44)	$88.53_{\pm 1.69}$ (4.64)	3.67	$0.13_{\pm 0.27}$ (0.13)	$0.13_{\pm 0.18}$ (0.13)	$90.69_{\pm 6.71}$ (9.26)	$85.51_{\pm 5.76}$ (7.94)	4.36
EU-k	$1.60_{\pm 1.10}$ (4.25)	$5.60_{\pm 1.49}$ (7.29)	$97.79_{\pm 0.74}$ (2.17)	$90.58_{\pm 0.76}$ (2.59)	4.07	$0.09_{\pm 0.11}$ (0.09)	$1.69_{\pm 1.04}$ (1.69)	$97.76_{\pm 0.92}$ (2.19)	$90.49_{\pm 0.88}$ (2.96)	1.73
CF-k	$0.00_{\pm 0.00}$ (5.85)	$0.44_{\pm 0.36}$ (12.45)	$99.98_{\pm 0.00}$ (0.02)	$94.33_{\pm 0.08}$ (1.16)	4.87	$0.00_{\pm 0.00}$ (0.00)	$0.00_{\pm 0.00}$ (0.00)	$99.98_{\pm 0.00}$ (0.03)	$94.40_{\pm 0.07}$ (0.95)	0.24
ℓ_1-sparse	$5.89_{\pm 1.81}$ (0.04)	$12.33_{\pm 1.84}$ (0.56)	$96.52_{\pm 0.85}$ (3.44)	$90.57_{\pm 0.66}$ (2.60)	1.66	$0.00_{\pm 0.00}$ (0.00)	$0.09_{\pm 0.18}$ (0.09)	$92.08_{\pm 3.44}$ (7.87)	$86.91_{\pm 2.88}$ (6.54)	3.62
RL	$7.26_{\pm 1.47}$ (1.41)	$43.85_{\pm 2.67}$ (30.96)	$99.99_{\pm 0.00}$ (0.03)	$94.11_{\pm 0.04}$ (0.94)	8.33	$0.00_{\pm 0.00}$ (0.00)	$84.37_{\pm 5.61}$ (84.37)	$100.00_{\pm 0.00}$ (0.05)	$94.46_{\pm 0.09}$ (1.01)	21.36
BE	$0.00_{\pm 0.00}$ (5.85)	$0.98_{\pm 0.30}$ (11.91)	$99.97_{\pm 0.01}$ (0.01)	$94.26_{\pm 0.09}$ (1.09)	4.72	$3.47_{\pm 1.54}$ (3.47)	$16.36_{\pm 6.93}$ (16.36)	$92.37_{\pm 0.48}$ (7.58)	$85.55_{\pm 0.57}$ (7.90)	8.83
BS	$0.00_{\pm 0.00}$ (5.85)	$0.98_{\pm 0.30}$ (11.91)	$99.97_{\pm 0.01}$ (0.01)	$94.27_{\pm 0.11}$ (1.10)	4.72	$3.29_{\pm 1.44}$ (3.29)	$13.69_{\pm 5.09}$ (13.69)	$92.34_{\pm 0.52}$ (7.61)	$85.48_{\pm 0.60}$ (7.97)	8.14
SalUn	$1.26_{\pm 0.55}$ (4.59)	$17.33_{\pm 1.01}$ (4.44)	$99.99_{\pm 0.01}$ (0.03)	$94.28_{\pm 0.07}$ (1.11)	2.54	$0.00_{\pm 0.00}$ (0.00)	$72.59_{\pm 2.22}$ (72.59)	$100.00_{\pm 0.00}$ (0.05)	$94.45_{\pm 0.14}$ (1.00)	18.41
5%-Data Forgetting										
Retrain	$5.92_{\pm 0.44}$	$13.00_{\pm 0.55}$	$100.00_{\pm 0.00}$	$94.51_{\pm 0.07}$	0.00	$0.00_{\pm 0.00}$	$0.02_{\pm 0.02}$	$100.00_{\pm 0.00}$	$94.67_{\pm 0.08}$	0.00
FT	$5.17_{\pm 0.73}$ (0.75)	$11.32_{\pm 0.94}$ (1.68)	$97.08_{\pm 0.44}$ (2.92)	$90.71_{\pm 0.38}$ (3.80)	2.29	$0.01_{\pm 0.02}$ (0.01)	$0.02_{\pm 0.04}$ (0.00)	$97.39_{\pm 0.48}$ (2.61)	$91.10_{\pm 0.55}$ (3.57)	1.55
EU-k	$2.13_{\pm 0.75}$ (3.79)	$6.07_{\pm 1.10}$ (6.93)	$97.81_{\pm 0.86}$ (2.19)	$90.53_{\pm 0.82}$ (3.98)	4.22	$0.16_{\pm 0.12}$ (0.16)	$2.11_{\pm 1.17}$ (2.09)	$97.61_{\pm 0.78}$ (2.39)	$90.48_{\pm 0.74}$ (4.19)	2.21
CF-k	$0.04_{\pm 0.06}$ (5.88)	$0.74_{\pm 0.04}$ (12.26)	$99.99_{\pm 0.00}$ (0.01)	$94.47_{\pm 0.02}$ (0.04)	4.55	$0.00_{\pm 0.00}$ (0.00)	$0.00_{\pm 0.06}$ (0.02)	$99.98_{\pm 0.00}$ (0.02)	$94.40_{\pm 0.04}$ (0.27)	0.08
ℓ_1-sparse	$4.63_{\pm 1.01}$ (1.29)	$10.09_{\pm 1.00}$ (2.91)	$97.13_{\pm 0.75}$ (2.87)	$90.92_{\pm 0.65}$ (3.59)	2.66	$0.00_{\pm 0.00}$ (0.00)	$0.01_{\pm 0.02}$ (0.01)	$97.12_{\pm 0.58}$ (2.88)	$91.27_{\pm 0.59}$ (3.40)	1.57
RL	$5.47_{\pm 0.45}$ (0.45)	$35.38_{\pm 0.50}$ (22.38)	$99.97_{\pm 0.01}$ (0.03)	$93.70_{\pm 0.16}$ (0.81)	5.92	$0.15_{\pm 0.18}$ (0.15)	$95.57_{\pm 0.80}$ (95.55)	$99.98_{\pm 0.00}$ (0.02)	$94.08_{\pm 0.02}$ (0.59)	24.08
BE	$0.36_{\pm 0.10}$ (5.56)	$18.77_{\pm 1.22}$ (5.77)	$99.73_{\pm 0.04}$ (0.27)	$93.07_{\pm 0.17}$ (1.44)	3.26	$38.19_{\pm 5.76}$ (38.19)	$85.74_{\pm 1.92}$ (85.72)	$79.07_{\pm 2.41}$ (20.93)	$72.56_{\pm 2.43}$ (22.11)	41.74
BS	$1.65_{\pm 0.67}$ (4.27)	$17.59_{\pm 1.65}$ (4.59)	$98.56_{\pm 0.65}$ (1.44)	$91.93_{\pm 0.55}$ (2.58)	3.22	$39.62_{\pm 9.10}$ (39.62)	$84.07_{\pm 5.29}$ (84.05)	$72.30_{\pm 7.03}$ (27.70)	$66.74_{\pm 6.86}$ (27.93)	44.82
SalUn	$0.67_{\pm 0.04}$ (5.25)	$12.87_{\pm 1.27}$ (0.13)	$100.00_{\pm 0.00}$ (0.00)	$94.13_{\pm 0.03}$ (0.38)	1.44	$0.09_{\pm 0.07}$ (0.09)	$93.23_{\pm 0.25}$ (93.21)	$100.00_{\pm 0.00}$ (0.00)	$94.20_{\pm 0.13}$ (0.47)	23.44
10%-Data Forgetting										
Retrain	$5.28_{\pm 0.33}$	$12.86_{\pm 0.61}$	$100.00_{\pm 0.00}$	$94.38_{\pm 0.15}$	0.00	$0.00_{\pm 0.00}$	$0.00_{\pm 0.00}$	$100.00_{\pm 0.00}$	$94.66_{\pm 0.09}$	0.00
FT	$5.08_{\pm 0.39}$ (0.20)	$10.96_{\pm 0.38}$ (1.90)	$97.46_{\pm 0.52}$ (2.54)	$91.02_{\pm 0.36}$ (3.36)	2.00	$0.00_{\pm 0.00}$ (0.00)	$0.02_{\pm 0.03}$ (0.02)	$97.63_{\pm 0.46}$ (2.37)	$91.58_{\pm 0.40}$ (3.08)	1.37
EU-k	$2.34_{\pm 0.79}$ (2.94)	$6.35_{\pm 0.89}$ (6.51)	$97.52_{\pm 0.89}$ (2.48)	$90.17_{\pm 0.88}$ (4.21)	4.04	$0.68_{\pm 0.56}$ (0.68)	$5.02_{\pm 4.42}$ (5.02)	$97.17_{\pm 0.86}$ (2.83)	$90.08_{\pm 0.70}$ (4.58)	3.28
CF-k	$0.02_{\pm 0.02}$ (5.26)	$0.76_{\pm 0.02}$ (12.10)	$99.98_{\pm 0.00}$ (0.02)	$94.45_{\pm 0.02}$ (0.07)	4.36	$0.00_{\pm 0.00}$ (0.00)	$0.00_{\pm 0.00}$ (0.00)	$99.98_{\pm 0.01}$ (0.02)	$94.34_{\pm 0.05}$ (0.32)	0.08
ℓ_1-sparse	$4.34_{\pm 0.73}$ (0.94)	$9.82_{\pm 1.04}$ (3.04)	$97.70_{\pm 0.72}$ (2.30)	$91.41_{\pm 0.68}$ (2.97)	2.31	$0.02_{\pm 0.03}$ (0.02)	$0.11_{\pm 0.11}$ (0.11)	$96.93_{\pm 0.73}$ (3.07)	$90.96_{\pm 0.82}$ (3.70)	1.72
RL	$3.59_{\pm 0.24}$ (1.69)	$28.02_{\pm 2.47}$ (15.16)	$99.97_{\pm 0.01}$ (0.03)	$93.74_{\pm 0.12}$ (0.64)	4.38	$1.93_{\pm 1.5}$ (1.93)	$96.70_{\pm 0.86}$ (96.70)	$99.96_{\pm 0.01}$ (0.04)	$93.83_{\pm 0.24}$ (0.83)	24.88
BE	$1.19_{\pm 0.49}$ (4.09)	$22.06_{\pm 0.61}$ (9.20)	$98.77_{\pm 0.41}$ (1.23)	$91.79_{\pm 0.32}$ (2.59)	4.28	$19.47_{\pm 2.12}$ (19.47)	$81.45_{\pm 5.16}$ (81.45)	$81.35_{\pm 2.76}$ (18.65)	$75.41_{\pm 1.77}$ (19.25)	34.70
BS	$5.72_{\pm 1.42}$ (0.44)	$27.15_{\pm 1.41}$ (14.29)	$94.29_{\pm 1.06}$ (5.71)	$87.45_{\pm 1.06}$ (6.93)	6.84	$29.75_{\pm 6.39}$ (29.75)	$74.88_{\pm 3.13}$ (74.88)	$78.34_{\pm 0.98}$ (21.66)	$72.07_{\pm 1.25}$ (22.59)	37.22
SalUn	$1.48_{\pm 0.14}$ (3.80)	$16.19_{\pm 0.34}$ (3.33)	$99.98_{\pm 0.01}$ (0.02)	$93.95_{\pm 0.01}$ (0.43)	1.89	$0.96_{\pm 0.59}$ (0.96)	$96.43_{\pm 0.35}$ (96.43)	$99.98_{\pm 0.01}$ (0.02)	$94.03_{\pm 0.08}$ (0.63)	24.51
20%-Data Forgetting										
Retrain	$5.76_{\pm 0.20}$	$14.34_{\pm 0.40}$	$100.00_{\pm 0.00}$	$94.04_{\pm 0.08}$	0.00	$0.00_{\pm 0.00}$	$0.03_{\pm 0.01}$	$100.00_{\pm 0.00}$	$94.60_{\pm 0.08}$	0.00
FT	$5.46_{\pm 0.42}$ (0.30)	$11.40_{\pm 0.78}$ (2.94)	$97.10_{\pm 0.42}$ (2.90)	$90.32_{\pm 0.41}$ (3.72)	2.46	$0.14_{\pm 0.13}$ (0.14)	$0.27_{\pm 0.15}$ (0.24)	$96.56_{\pm 1.12}$ (3.44)	$90.62_{\pm 1.07}$ (3.98)	1.95
EU-k	$3.08_{\pm 1.19}$ (2.68)	$7.43_{\pm 1.29}$ (6.91)	$96.70_{\pm 1.31}$ (3.30)	$89.37_{\pm 1.25}$ (4.67)	4.39	$1.76_{\pm 1.25}$ (1.76)	$7.39_{\pm 4.89}$ (7.36)	$95.75_{\pm 1.26}$ (4.25)	$88.96_{\pm 1.05}$ (5.64)	4.75
CF-k	$0.03_{\pm 0.01}$ (5.73)	$0.68_{\pm 0.10}$ (13.66)	$99.99_{\pm 0.00}$ (0.01)	$94.45_{\pm 0.06}$ (0.41)	4.95	$0.00_{\pm 0.00}$ (0.00)	$0.00_{\pm 0.06}$ (0.03)	$99.97_{\pm 0.01}$ (0.03)	$94.29_{\pm 0.03}$ (0.31)	0.09
ℓ_1-sparse	$3.83_{\pm 0.59}$ (1.93)	$8.76_{\pm 0.92}$ (5.58)	$97.99_{\pm 0.55}$ (2.01)	$91.30_{\pm 0.64}$ (2.74)	3.06	$0.07_{\pm 0.07}$ (0.07)	$0.14_{\pm 0.14}$ (0.11)	$97.25_{\pm 0.93}$ (2.75)	$91.22_{\pm 0.76}$ (3.38)	1.58
RL	$2.44_{\pm 0.04}$ (3.32)	$22.21_{\pm 0.43}$ (7.87)	$99.97_{\pm 0.00}$ (0.03)	$93.44_{\pm 0.04}$ (0.60)	2.95	$3.84_{\pm 2.01}$ (3.84)	$97.54_{\pm 0.12}$ (97.51)	$99.92_{\pm 0.01}$ (0.08)	$93.07_{\pm 0.40}$ (1.53)	25.74
BE	$11.54_{\pm 1.44}$ (5.78)	$29.84_{\pm 1.08}$ (15.50)	$88.20_{\pm 1.48}$ (11.80)	$80.67_{\pm 1.35}$ (13.37)	11.61	$25.05_{\pm 2.89}$ (25.05)	$82.25_{\pm 4.88}$ (82.22)	$76.88_{\pm 2.77}$ (23.12)	$71.07_{\pm 2.46}$ (23.53)	38.48
BS	$20.69_{\pm 1.59}$ (14.93)	$32.96_{\pm 2.16}$ (18.62)	$78.87_{\pm 1.91}$ (21.13)	$72.23_{\pm 1.70}$ (21.81)	19.12	$40.61_{\pm 2.26}$ (40.61)	$75.05_{\pm 6.78}$ (75.02)	$72.17_{\pm 1.16}$ (27.83)	$64.53_{\pm 1.12}$ (30.07)	43.38
SalUn	$1.31_{\pm 0.08}$ (4.45)	$17.15_{\pm 0.85}$ (2.81)	$99.98_{\pm 0.01}$ (0.02)	$93.67_{\pm 0.17}$ (0.37)	1.91	$2.82_{\pm 1.42}$ (2.82)	$97.09_{\pm 0.27}$ (97.06)	$99.95_{\pm 0.01}$ (0.05)	$93.36_{\pm 0.38}$ (1.24)	25.29

methods consistently display a performance trend akin to that of Retrain. However, in sharp contrast to their relabeling-free counterparts, relabeling-based approximate unlearning methods manifest a discernible performance gap when compared to Retrain.

C.3 Additional results on different model architectures

In **Table A3**, we comprehensively assess the performance of diverse unlearning techniques under both random and worst-case forget sets scenarios, employing a 10% forgetting data ratio on CIFAR-10. This evaluation encompasses a broadened range of model architectures, including ResNet-50 [8] and VGG-16 [13]. When evaluating the methods on worst-case forget set, the relabeling-free approximate unlearning methods consistently exhibit a performance trend that closely resembles that of Retrain. Conversely, relabeling-based approximate unlearning methods demonstrate a notable performance discrepancy when compared to Retrain.

C.4 Transferability of worst-case forget sets between different models and methods

Table A2: Performance of various unlearning methods under random forget sets and worst-case forget sets on CIFAR-100 and Tiny ImageNet using ResNet-18 for forgetting ratio 10%. The result format follows Table 3.

Methods	Random Forget Set					Worst-Case Forget Set				
	UA	MIA	RA	TA	Avg. Gap	UA	MIA	RA	TA	Avg. Gap
CIFAR-100										
Retrain	$25.06_{\pm 0.25}$	$49.98_{\pm 0.91}$	$99.98_{\pm 0.00}$	$74.54_{\pm 0.07}$	0.00	$0.13_{\pm 0.04}$	$1.11_{\pm 0.19}$	$99.97_{\pm 0.00}$	$75.36_{\pm 0.31}$	0.00
FT	$23.10_{\pm 0.97}$ (1.96)	$30.47_{\pm 0.87}$ (19.51)	$90.44_{\pm 1.19}$ (9.54)	$64.42_{\pm 0.71}$ (10.12)	10.28	$0.66_{\pm 0.25}$ (0.53)	$1.08_{\pm 0.31}$ (0.03)	$90.74_{\pm 0.54}$ (9.23)	$65.77_{\pm 0.51}$ (9.59)	4.85
EU-k	$12.55_{\pm 0.63}$ (12.51)	$15.04_{\pm 0.84}$ (34.94)	$87.61_{\pm 0.35}$ (12.37)	$58.77_{\pm 0.58}$ (15.77)	18.90	$4.23_{\pm 1.45}$ (4.10)	$4.94_{\pm 1.64}$ (3.83)	$86.87_{\pm 0.14}$ (13.10)	$58.65_{\pm 0.19}$ (16.71)	9.44
CF-k	$0.02_{\pm 0.02}$ (25.04)	$2.36_{\pm 0.27}$ (47.62)	$99.98_{\pm 0.00}$ (0.00)	$75.34_{\pm 0.16}$ (0.80)	18.36	$0.00_{\pm 0.00}$ (0.13)	$0.12_{\pm 0.01}$ (0.99)	$99.98_{\pm 0.00}$ (0.01)	$75.22_{\pm 0.06}$ (0.14)	0.32
ℓ_1-sparse	$27.30_{\pm 1.49}$ (2.24)	$33.11_{\pm 1.06}$ (16.87)	$87.17_{\pm 2.03}$ (12.81)	$63.12_{\pm 1.49}$ (11.42)	10.84	$1.53_{\pm 0.75}$ (1.40)	$1.64_{\pm 0.82}$ (0.53)	$87.19_{\pm 0.64}$ (12.78)	$64.45_{\pm 0.45}$ (10.91)	6.40
RL	$47.18_{\pm 0.00}$ (22.12)	$91.71_{\pm 0.00}$ (41.73)	$99.88_{\pm 0.00}$ (0.10)	$67.31_{\pm 0.00}$ (7.23)	17.79	$62.64_{\pm 0.00}$ (62.51)	$97.18_{\pm 0.00}$ (96.07)	$99.67_{\pm 0.00}$ (0.30)	$66.22_{\pm 0.00}$ (9.14)	42.00
BE	$26.39_{\pm 2.38}$ (1.33)	$24.43_{\pm 1.51}$ (25.55)	$76.04_{\pm 2.25}$ (23.94)	$42.31_{\pm 1.34}$ (32.23)	20.76	$32.10_{\pm 1.10}$ (31.97)	$31.18_{\pm 0.72}$ (30.07)	$78.42_{\pm 2.00}$ (21.55)	$47.03_{\pm 1.46}$ (28.33)	27.98
BS	$8.44_{\pm 0.65}$ (16.62)	$19.24_{\pm 1.45}$ (30.74)	$93.17_{\pm 0.59}$ (6.81)	$62.75_{\pm 0.27}$ (11.79)	16.49	$20.64_{\pm 0.57}$ (20.51)	$27.10_{\pm 1.48}$ (25.99)	$80.81_{\pm 0.31}$ (19.16)	$52.69_{\pm 0.24}$ (22.67)	22.08
SalUn	$24.22_{\pm 0.00}$ (0.84)	$77.76_{\pm 0.00}$ (27.78)	$99.84_{\pm 0.00}$ (0.14)	$67.64_{\pm 0.00}$ (6.90)	8.92	$44.76_{\pm 0.00}$ (44.63)	$89.40_{\pm 0.00}$ (88.29)	$99.52_{\pm 0.00}$ (0.45)	$67.20_{\pm 0.00}$ (8.16)	35.38
Tiny ImageNet										
Retrain	$36.40_{\pm 0.25}$	$63.77_{\pm 0.02}$	$99.98_{\pm 0.00}$	$63.67_{\pm 0.34}$	0.00	$0.78_{\pm 0.06}$	$4.80_{\pm 0.25}$	$99.98_{\pm 0.00}$	$64.87_{\pm 0.19}$	0.00
FT	$14.41_{\pm 0.24}$ (21.99)	$25.48_{\pm 0.51}$ (38.29)	$98.72_{\pm 0.03}$ (1.26)	$62.01_{\pm 0.26}$ (1.66)	15.80	$0.05_{\pm 0.00}$ (0.73)	$0.14_{\pm 0.00}$ (4.66)	$98.36_{\pm 0.00}$ (1.62)	$61.87_{\pm 0.00}$ (3.00)	2.50
EU-k	$16.77_{\pm 0.08}$ (19.63)	$23.66_{\pm 2.70}$ (40.11)	$84.48_{\pm 0.48}$ (15.50)	$57.70_{\pm 0.42}$ (5.97)	20.30	$0.20_{\pm 0.02}$ (0.58)	$0.23_{\pm 0.03}$ (4.57)	$83.59_{\pm 0.27}$ (16.39)	$58.51_{\pm 0.34}$ (6.36)	6.98
CF-k	$13.48_{\pm 0.30}$ (22.92)	$22.49_{\pm 1.80}$ (41.28)	$87.98_{\pm 0.59}$ (12.00)	$60.29_{\pm 0.31}$ (3.38)	19.90	$0.10_{\pm 0.03}$ (0.68)	$0.11_{\pm 0.04}$ (4.69)	$86.85_{\pm 0.21}$ (13.13)	$60.37_{\pm 0.11}$ (4.50)	5.75
ℓ_1-sparse	$15.19_{\pm 0.24}$ (21.21)	$26.39_{\pm 0.54}$ (37.38)	$98.61_{\pm 0.04}$ (1.37)	$61.78_{\pm 0.21}$ (1.89)	15.46	$0.11_{\pm 0.03}$ (0.67)	$0.17_{\pm 0.05}$ (4.63)	$98.15_{\pm 0.04}$ (1.83)	$61.35_{\pm 0.12}$ (3.52)	2.66
RL	$29.13_{\pm 0.49}$ (7.27)	$42.62_{\pm 1.89}$ (21.15)	$96.25_{\pm 0.09}$ (3.73)	$58.99_{\pm 0.21}$ (4.68)	9.21	$37.04_{\pm 0.51}$ (36.26)	$47.84_{\pm 1.21}$ (43.04)	$95.39_{\pm 0.08}$ (4.59)	$56.64_{\pm 0.29}$ (8.23)	23.03
BE	$47.41_{\pm 0.44}$ (11.01)	$29.65_{\pm 0.23}$ (34.12)	$53.14_{\pm 1.00}$ (46.84)	$36.07_{\pm 0.41}$ (27.60)	29.89	$29.94_{\pm 0.77}$ (29.16)	$39.67_{\pm 7.84}$ (34.87)	$34.66_{\pm 0.26}$ (65.32)	$26.63_{\pm 0.27}$ (38.24)	41.90
BS	$30.32_{\pm 0.91}$ (6.08)	$25.45_{\pm 1.02}$ (38.32)	$70.48_{\pm 0.80}$ (29.50)	$47.00_{\pm 0.53}$ (16.67)	22.64	$24.32_{\pm 0.63}$ (23.54)	$14.40_{\pm 0.55}$ (9.60)	$54.39_{\pm 0.55}$ (45.59)	$37.54_{\pm 0.20}$ (27.33)	26.52
SalUn	$26.18_{\pm 0.50}$ (10.22)	$38.02_{\pm 1.42}$ (25.75)	$95.90_{\pm 0.07}$ (4.08)	$59.20_{\pm 0.16}$ (4.47)	11.13	$25.84_{\pm 0.10}$ (25.06)	$37.03_{\pm 0.25}$ (32.23)	$95.94_{\pm 0.06}$ (4.04)	$59.12_{\pm 0.15}$ (5.75)	16.77

Table A3: Performance of various unlearning methods under random forget sets and worst-case forget sets on CIFAR-10 using ResNet-50 and VGG-16 for forgetting ratio 10%. The result format follows Table 3.

Methods	Random Forget Set					Worst-Case Forget Set				
	UA	MIA	RA	TA	Avg. Gap	UA	MIA	RA	TA	Avg. Gap
ResNet-50										
Retrain	$5.56_{\pm 0.35}$	$11.68_{\pm 0.66}$	$100.00_{\pm 0.00}$	$94.17_{\pm 0.01}$	0.00	$0.00_{\pm 0.00}$	$0.00_{\pm 0.00}$	$100.00_{\pm 0.00}$	$94.04_{\pm 0.30}$	0.00
FT	$4.48_{\pm 0.20}$ (1.08)	$10.05_{\pm 0.71}$ (1.63)	$98.13_{\pm 0.38}$ (1.87)	$91.47_{\pm 0.23}$ (2.70)	1.82	$0.01_{\pm 0.01}$ (0.01)	$0.04_{\pm 0.03}$ (0.04)	$97.55_{\pm 0.33}$ (2.45)	$91.38_{\pm 0.49}$ (2.66)	1.29
EU-k	$4.54_{\pm 0.16}$ (1.02)	$7.94_{\pm 0.18}$ (3.74)	$95.53_{\pm 0.21}$ (4.47)	$87.34_{\pm 0.08}$ (6.83)	4.02	$1.59_{\pm 0.67}$ (1.59)	$3.68_{\pm 1.46}$ (3.68)	$95.56_{\pm 0.47}$ (4.44)	$87.61_{\pm 0.21}$ (6.43)	4.04
CF-k	$0.01_{\pm 0.01}$ (5.55)	$0.53_{\pm 0.06}$ (11.15)	$100.00_{\pm 0.00}$ (0.00)	$94.16_{\pm 0.00}$ (0.01)	4.18	$0.00_{\pm 0.00}$ (0.00)	$0.01_{\pm 0.01}$ (0.01)	$99.99_{\pm 0.00}$ (0.01)	$94.06_{\pm 0.04}$ (0.02)	0.01
ℓ_1-sparse	$2.38_{\pm 0.12}$ (3.18)	$7.49_{\pm 0.39}$ (4.19)	$98.91_{\pm 0.01}$ (1.09)	$92.53_{\pm 0.07}$ (1.64)	2.53	$0.00_{\pm 0.00}$ (0.00)	$0.03_{\pm 0.03}$ (0.03)	$98.34_{\pm 0.15}$ (1.66)	$92.08_{\pm 0.02}$ (1.96)	0.91
RL	$7.24_{\pm 0.46}$ (1.68)	$43.07_{\pm 4.11}$ (31.39)	$99.79_{\pm 0.02}$ (0.21)	$92.52_{\pm 0.08}$ (1.65)	8.73	$1.01_{\pm 0.47}$ (1.01)	$95.76_{\pm 0.86}$ (95.76)	$99.82_{\pm 0.07}$ (0.18)	$93.25_{\pm 0.25}$ (0.79)	24.44
BE	$3.54_{\pm 0.54}$ (2.02)	$15.70_{\pm 0.71}$ (4.02)	$96.36_{\pm 0.12}$ (3.64)	$89.75_{\pm 0.10}$ (4.42)	3.52	$19.24_{\pm 1.14}$ (19.24)	$47.57_{\pm 0.61}$ (47.57)	$87.59_{\pm 0.52}$ (12.41)	$79.63_{\pm 0.39}$ (14.41)	23.41
BS	$3.62_{\pm 0.52}$ (1.94)	$10.75_{\pm 0.71}$ (0.93)	$96.27_{\pm 0.27}$ (3.73)	$89.72_{\pm 0.26}$ (4.45)	2.76	$14.27_{\pm 1.13}$ (14.27)	$37.49_{\pm 1.12}$ (37.49)	$88.56_{\pm 0.15}$ (11.44)	$81.18_{\pm 0.36}$ (12.86)	19.02
SalUn	$1.67_{\pm 0.15}$ (3.89)	$19.44_{\pm 1.95}$ (7.76)	$99.94_{\pm 0.01}$ (0.06)	$93.44_{\pm 0.07}$ (0.73)	3.11	$0.34_{\pm 0.13}$ (0.34)	$92.83_{\pm 0.59}$ (92.83)	$98.99_{\pm 0.26}$ (1.01)	$92.08_{\pm 0.20}$ (1.96)	24.04
VGG-16										
Retrain	$6.76_{\pm 0.43}$	$11.77_{\pm 0.27}$	$99.99_{\pm 0.00}$	$93.28_{\pm 0.15}$	0.00	$0.01_{\pm 0.01}$	$0.07_{\pm 0.04}$	$99.99_{\pm 0.00}$	$93.43_{\pm 0.13}$	0.00
FT	$3.91_{\pm 0.68}$ (2.85)	$8.75_{\pm 0.94}$ (3.02)	$98.31_{\pm 0.35}$ (1.68)	$90.62_{\pm 0.38}$ (2.66)	2.55	$0.07_{\pm 0.07}$ (0.06)	$0.28_{\pm 0.29}$ (0.21)	$97.36_{\pm 0.45}$ (2.63)	$90.04_{\pm 0.30}$ (3.39)	1.57
EU-k	$15.79_{\pm 7.41}$ (9.03)	$19.61_{\pm 5.29}$ (7.84)	$83.63_{\pm 8.02}$ (16.36)	$76.36_{\pm 6.99}$ (16.92)	12.54	$2.25_{\pm 2.18}$ (2.24)	$3.08_{\pm 1.86}$ (3.01)	$83.51_{\pm 0.61}$ (16.48)	$77.57_{\pm 7.75}$ (15.86)	9.40
CF-k	$0.02_{\pm 0.00}$ (6.74)	$0.33_{\pm 0.06}$ (11.44)	$99.99_{\pm 0.00}$ (0.00)	$93.59_{\pm 0.01}$ (0.31)	4.62	$0.00_{\pm 0.00}$ (0.01)	$0.00_{\pm 0.00}$ (0.07)	$99.98_{\pm 0.01}$ (0.01)	$93.54_{\pm 0.07}$ (0.11)	0.05
ℓ_1-sparse	$4.48_{\pm 0.43}$ (2.28)	$9.76_{\pm 0.32}$ (2.01)	$97.68_{\pm 0.16}$ (2.31)	$90.61_{\pm 0.13}$ (2.67)	2.32	$0.04_{\pm 0.04}$ (0.03)	$0.17_{\pm 0.19}$ (0.10)	$97.56_{\pm 0.12}$ (2.43)	$90.36_{\pm 0.15}$ (3.07)	1.41
RL	$2.71_{\pm 0.29}$ (4.05)	$14.01_{\pm 2.29}$ (2.24)	$99.97_{\pm 0.00}$ (0.02)	$92.92_{\pm 0.04}$ (0.36)	1.67	$3.50_{\pm 4.09}$ (3.49)	$95.96_{\pm 0.85}$ (95.89)	$99.89_{\pm 0.02}$ (0.10)	$92.65_{\pm 0.49}$ (0.78)	25.06
BE	$11.73_{\pm 1.40}$ (4.97)	$26.33_{\pm 10.56}$ (14.56)	$88.22_{\pm 4.15}$ (11.77)	$80.53_{\pm 3.41}$ (12.75)	11.01	$49.28_{\pm 22.21}$ (49.27)	$67.18_{\pm 8.23}$ (67.11)	$77.53_{\pm 1.63}$ (22.46)	$68.02_{\pm 3.24}$ (25.41)	41.06
BS	$7.46_{\pm 3.27}$ (0.70)	$8.61_{\pm 1.14}$ (3.16)	$92.86_{\pm 3.01}$ (7.13)	$84.23_{\pm 2.88}$ (9.05)	5.01	$52.39_{\pm 31.85}$ (52.38)	$53.78_{\pm 35.44}$ (53.71)	$75.38_{\pm 2.55}$ (24.61)	$65.64_{\pm 4.96}$ (27.79)	39.62
SalUn	$6.38_{\pm 0.49}$ (0.38)	$18.66_{\pm 1.66}$ (6.89)	$99.76_{\pm 0.14}$ (0.23)	$91.88_{\pm 0.41}$ (1.40)	2.22	$3.21_{\pm 4.48}$ (3.20)	$94.61_{\pm 0.26}$ (94.54)	$99.29_{\pm 0.17}$ (0.70)	$91.46_{\pm 0.56}$ (1.97)	25.10

In this section, we validate the transferability of worst-case forget sets across a wider range of model architectures and methods. Concerning the transferability between models, we leverage a diverse range of models for the selection process, including ResNet-18, ResNet-50 [8], VGG-16, and VGG-19 [13]. Conversely, for the evaluation process, we employ various models and adopt Retrain as the corresponding unlearning method. The UA (unlearning accuracy) results are exhibited in **Fig. A1**. Notably, when the worst-case forget set is selected using one model and subsequently undergoes unlearning with another model, the UA shows significantly lower than that of random forget set. This observation

Table A4: Performance of various unlearning methods on CIFAR-10 using ResNet-18 with a 10% forgetting ratio under worst-case forget sets obtained using RL. The result format follows Table 3. (•) after Retrain in Worst-Case Forget Set indicates the difference from Random Forget Set.

Methods	UA	MIA	RA	TA	Avg. Gap
Random Forget Set					
Retrain	5.28	12.86	100.00	94.38	N/A
Worst-Case Forget Set					
Retrain	0.02 (5.26▼)	0.16 (12.70▼)	100.00 (0.00–)	94.67 (0.29▲)	0
FT	0.44 (0.42)	0.44 (0.29)	92.94 (7.06)	87.93 (6.74)	3.63
EU-k	1.36 (1.33)	11.93 (11.78)	97.99 (2.01)	90.75 (3.92)	4.76
SCRUB	5.04 (5.02)	20.71 (20.56)	89.92 (10.08)	85.70 (8.97)	11.16
ℓ_1-sparse	0.04 (0.02)	0.13 (0.02)	97.59 (2.41)	91.62 (3.05)	1.38
RL	4.62 (4.60)	98.36 (98.20)	99.96 (0.04)	93.59 (1.08)	25.98
BE	38.38 (38.36)	96.64 (96.49)	77.00 (23.00)	71.42 (23.25)	45.27
BS	31.02 (31.00)	96.82 (96.67)	81.88 (18.12)	75.62 (19.05)	41.21
SalUn	2.09 (2.07)	97.40 (97.24)	99.98 (0.02)	94.00 (0.67)	25.00

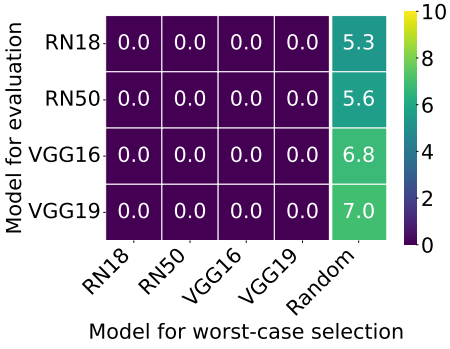


Fig. A1: UA of Retrain on CIFAR-10 using various models with 10% forgetting ratio. The rightmost column represents UA on random forget set, while other columns depict UA on worst-case forget set.

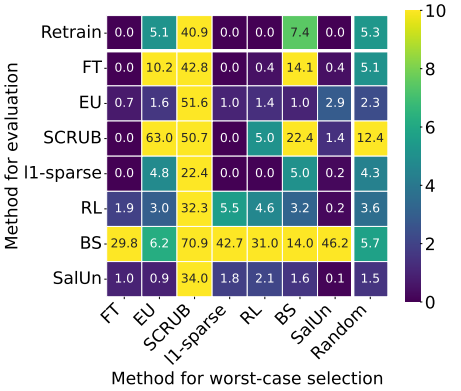


Fig. A2: UA of unlearning methods on CIFAR-10 using ResNet-18 with a 10% forgetting ratio. The rightmost column shows UA on random forget set; others show UA on worst-case forget set.

clearly demonstrates the transferability of worst-case forget set across diverse models.

Regarding the transferability between methods, we employ various approximate unlearning objectives in the selection process to specify the lower-level optimization problem (8), while utilizing different unlearning methods during evaluation. The UA results are illustrated in **Fig. A2**. As evident from the figure, the columns corresponding to the four methods, FT, ℓ_1-sparse, RL, and SalUn, exhibit deeper shades than the column for random, indicating lower UA values. Consequently, FT, ℓ_1-sparse, RL, and SalUn are more suitable for addressing the lower-level problem. In **Table A4**, we further test the unlearning methods under the worst-case forget set obtained using RL to perform the lower-level unlearning process in BLO. The results are consistent with Fig. A2.

C.5 From worst-case unlearning to easiest-case unlearning

By considering the opposite objective function of the upper-level optimization in (3), we can obtain the problem of selecting the easiest-case forget sets through BLO:

$$\min_{\mathbf{w}\in\mathcal{S}} \sum_{\mathbf{z}_i\in\mathcal{D}} -[w_i \ell(\boldsymbol{\theta}_{\mathrm{u}}(\mathbf{w}); \mathbf{z}_i)] + \gamma\|\mathbf{w}\|_2^2; \text{subject to } \boldsymbol{\theta}_{\mathrm{u}}(\mathbf{w}) = \arg\min_{\boldsymbol{\theta}} \ \ell_{\mathrm{MU}}(\boldsymbol{\theta}; \mathbf{w}), \tag{14}$$

Table A5 presents the performance of identified easiest-case forget sets. We draw two observations. *First*, for Retrain, UA and MIA on easiest-case forget sets are significantly higher than those on random forget set. *Second*, for approximate unlearning methods, Avg. Gap on easiest-case forget set is much higher than that on the random forget set. This is due to the significantly lower UA of

Table A5: Performance of approximate unlearning methods under random forget set and easiest-case forget set on CIFAR-10 using ResNet-18 with forgetting ratio 10%. The result format follows Table 3.

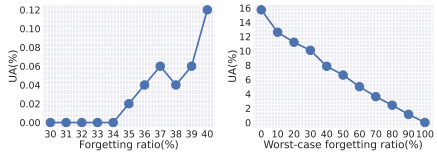
Methods	Random Forget Set					Easiest-Case Forget Set				
	UA	MIA	RA	TA	Avg. Gap	UA	MIA	RA	TA	Avg. Gap
Retrain	$5.28_{\pm 0.33}$	$12.86_{\pm 0.61}$	$100.00_{\pm 0.00}$	$94.38_{\pm 0.15}$	0.00	$43.18_{\pm 1.06}$	$67.72_{\pm 0.87}$	$100.00_{\pm 0.00}$	$93.15_{\pm 0.14}$	0.00
Relabeling-free										
FT	$5.08_{\pm 0.39}$ (0.20)	$10.96_{\pm 0.38}$ (1.90)	$97.46_{\pm 0.52}$ (2.54)	$91.02_{\pm 0.36}$ (3.36)	2.00	$28.86_{\pm 0.99}$ (14.32)	$49.39_{\pm 1.83}$ (18.33)	$98.49_{\pm 0.40}$ (1.51)	$90.90_{\pm 0.43}$ (2.25)	9.10
EU-k	$2.34_{\pm 0.79}$ (2.94)	$6.35_{\pm 0.89}$ (6.51)	$97.52_{\pm 0.89}$ (2.48)	$90.17_{\pm 0.88}$ (4.21)	4.04	$9.48_{\pm 4.75}$ (33.70)	$19.27_{\pm 6.13}$ (48.45)	$97.87_{\pm 1.01}$ (2.13)	$89.81_{\pm 1.17}$ (3.34)	21.90
CF-k	$0.02_{\pm 0.02}$ (5.26)	$0.76_{\pm 0.02}$ (12.10)	$99.98_{\pm 0.00}$ (0.02)	$94.45_{\pm 0.02}$ (0.07)	4.36	$0.10_{\pm 0.03}$ (43.08)	$3.09_{\pm 0.13}$ (64.63)	$99.99_{\pm 0.00}$ (0.01)	$94.40_{\pm 0.03}$ (1.25)	27.24
ℓ_1-sparse	$4.34_{\pm 0.73}$ (0.94)	$9.82_{\pm 1.04}$ (3.04)	$97.70_{\pm 0.72}$ (2.30)	$91.41_{\pm 0.68}$ (2.97)	2.31	$26.26_{\pm 1.24}$ (16.92)	$46.56_{\pm 1.75}$ (21.16)	$98.55_{\pm 0.14}$ (1.45)	$90.75_{\pm 0.43}$ (2.40)	10.48
Relabeling-based										
RL	$3.59_{\pm 0.24}$ (1.69)	$28.02_{\pm 2.47}$ (15.16)	$99.97_{\pm 0.01}$ (0.03)	$93.74_{\pm 0.12}$ (0.64)	4.38	$22.84_{\pm 2.76}$ (20.34)	$92.51_{\pm 0.98}$ (24.79)	$99.95_{\pm 0.01}$ (0.05)	$92.73_{\pm 0.25}$ (0.42)	11.40
BE	$1.19_{\pm 0.49}$ (4.09)	$22.06_{\pm 0.61}$ (9.20)	$98.77_{\pm 0.41}$ (1.23)	$91.79_{\pm 0.32}$ (2.59)	4.28	$13.21_{\pm 1.48}$ (29.97)	$33.93_{\pm 3.58}$ (33.79)	$97.19_{\pm 1.59}$ (2.81)	$88.65_{\pm 1.31}$ (4.50)	17.77
BS	$5.72_{\pm 1.42}$ (0.44)	$27.15_{\pm 1.41}$ (14.29)	$94.29_{\pm 1.06}$ (5.71)	$87.45_{\pm 1.06}$ (6.93)	6.84	$17.55_{\pm 1.00}$ (25.63)	$34.93_{\pm 3.03}$ (32.79)	$96.33_{\pm 0.55}$ (3.67)	$87.46_{\pm 0.48}$ (5.69)	16.94
SalUn	$1.48_{\pm 0.14}$ (3.80)	$16.19_{\pm 0.34}$ (3.33)	$99.98_{\pm 0.01}$ (0.02)	$93.95_{\pm 0.01}$ (0.43)	1.89	$16.76_{\pm 1.71}$ (26.42)	$88.71_{\pm 0.94}$ (20.99)	$99.95_{\pm 0.01}$ (0.05)	$92.85_{\pm 0.19}$ (0.30)	11.94

approximate unlearning methods on easiest-case forget set compared to that of Retrain. This suggests that the current approximate unlearning methods are not yet effective enough, even for data in easiest-case forget set, and cannot accurately forget them.

C.6 Uniqueness and mixture of worst-case forget set

To verify the uniqueness of worst-case forget set, we identified worst-case forget set for different forgetting data ratios and performed unlearning using Retrain. We found that a **maximal** set with zero UA can exist. As shown in **Fig. A3-(a)**, with appropriately defined set sizes (up to 34% of the entire dataset), our method consistently identifies a worst-case forget set with 0 UA.

Furthermore, any subset of this set will also exhibit the worst-case property. **Fig. A3-(b)** illustrates that including any part of the worst-case set complicates the unlearning process. When the forget set represents a 34% ratio comprising a mix of worst-case forget set and random forget set and unlearning is performed using Retrain, the unlearning becomes increasingly difficult as the proportion of worst-case random forget set increases, which is indicated by the decrease in UA. This highlights the importance and significance of worst-case forgetting.



(a) Uniqueness (b) Mixture

Fig. A3: UA of Retrain on CIFAR-10 using ResNet-18. (a) UA under worst-case forgetting scenarios at different forgetting data ratios. (b) UA under a mixture of random and worst-case forgetting scenarios at different mixture ratios.

C.7 Identifying worst-case forget set in class-wise forgetting.

Extended from data-wise forgetting, **Table A6** showcases the effectiveness of our proposal in class-wise forgetting for image classification on the ImageNet dataset [3]. Recall that the data selection variables are now interpreted as class selection variables. In this experiment, our objective is to eliminate the influence of 10% of the ImageNet classes on classification performance.

To avoid completely eliminating the prediction head for the forgetting class in the model (ResNet-18), we define a class removal as the elimination of 90% of its data points. Consistent with our previous observations in class-wise forgetting, we can observe from Table A6 that our identified worst-case forget set constitutes a more challenging subset for the erasure of data influence as compared to random forget set, evidenced by a significant decline in UA of Retrain from 72.92% to 45.92%. A smaller reduction in MIA performance compared to data-wise forgetting suggests that class-wise forgetting presents a relatively simpler challenge. In addition, by examining the performance of representative approximate unlearning methods (FT, ℓ_1-sparse, and RL), we observe that relabeling-free unlearning methods exhibit performances akin to Retrain under the worst-case forget set, whereas relabeling-based methods demonstrate substantial discrepancies in UA, consistent with our observations in Table 3.

Table A6: Performance of various MU methods on ImageNet, ResNet-18. The content format follows Table A4.

Methods	UA	MIA	RA	TA	Avg. Gap
Random Forget Set					
Retrain	72.92	98.78	65.90	66.03	N/A
Worst-Case Forget Set					
Retrain	45.92 (27.00▼)	98.56 (0.22▼)	66.64 (0.74▲)	66.68 (0.65▲)	0.00
FT	36.40 (9.52)	96.71 (1.85)	65.40 (1.24)	65.73 (0.95)	3.39
ℓ_1-sparse	38.50 (7.42)	95.57 (2.99)	64.16 (2.48)	65.00 (1.68)	3.64
RL	99.89 (53.97)	99.01 (0.45)	39.70 (26.94)	44.04 (22.64)	26.00

Moreover, **Fig. A4** portrays the class-wise entropy for ImageNet classes within worst-case forget set in comparison to other classes. This visualization elucidates a predilection for selecting low-entropy classes as the worst-case scenarios for unlearning, suggesting that these classes are ostensibly simpler to learn. Furthermore, the worst-case forget class is primarily composed of animals and insects. In **Fig. A5**, we use t-SNE to show the relationship between worst-case classes and other classes. As we can see, the worst-case class primarily resides on the periphery of the distribution.

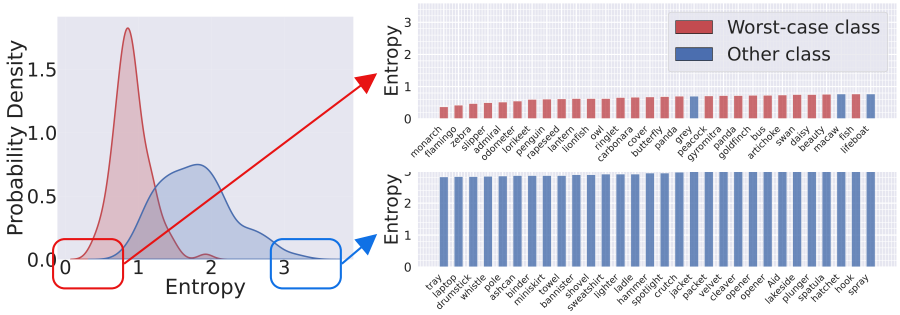


Fig. A4: Average entropy of worst-case forget classes vs. that of other classes on ImageNet using ResNet-18. The number of worst-case forget classes is 100.

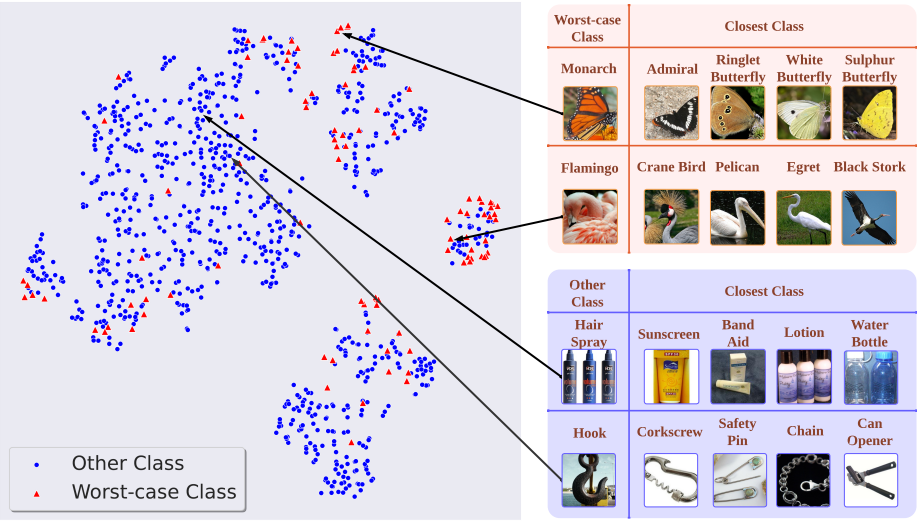


Fig. A5: T-SNE for all classes in the learned feature space, with an additional display on the right side showcasing two worst-case classes and two others, along with their four closest classes.

C.8 Additional results of Fig. 5

In **Tables** A7-A8, we present more examples using the original stable diffusion model (w/o unlearning), the unlearned diffusion model over the worst-case forgetting prompt set (Worst). For each diffusion model, images are generated based on an unlearned prompt from the worst-case forget set. It is evident that the unlearned diffusion model is still capable of generating corresponding images for prompts from the worst-case forget set.

D Broader Impacts and Limitations

Worst-Case Forget Set represents a novel perspective in evaluating data privacy and security. This set strikes a balance between data influence erasure and model utility, offering a robust assessment of the effectiveness of existing unlearning methods from an adversarial standpoint. It also provides a deeper understanding of datasets from the perspective of machine unlearning.

However, it is crucial to acknowledge the limitations of worst-case forget set. While worst-case forget set has demonstrated its effectiveness in various scenarios, including data-wise, class-wise, and prompt-wise, the effectiveness of unlearning methods for language models on worst-case forget set remains an area worthy of further exploration.

Table A7: Examples of image generation using the original stable diffusion model (w/o unlearning), the unlearned diffusion model over the worst-case forgetting prompt set (Worst). For each diffusion model, images are generated based on an unlearned prompt from the worst-case forget set.

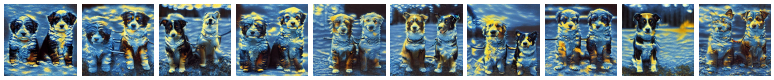
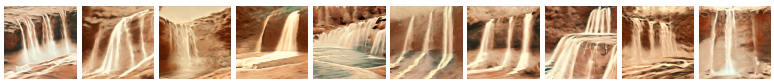
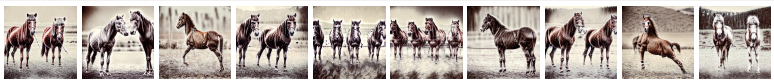
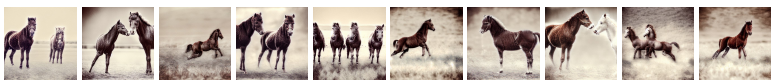
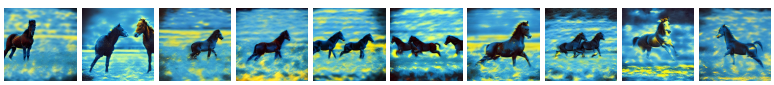
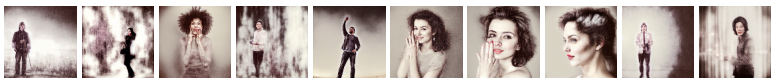
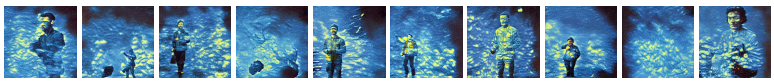
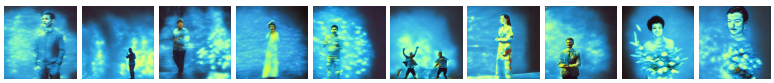
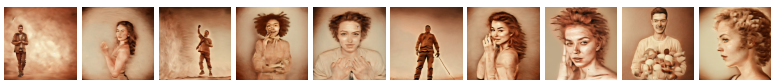
Model	Generation Condition
	$P_u^{(w)}$: *A painting of* Dogs *in* Van Gogh *Style.*
Original Diffusion Model	
Unlearned Diffusion Model (Worst)	
	$P_u^{(w)}$: *A painting of* Dogs *in* Rust *Style.*
Original Diffusion Model	
Unlearned Diffusion Model (Worst)	
	$P_u^{(w)}$: *A painting of* Waterfalls *in* Rust *Style.*
Original Diffusion Model	
Unlearned Diffusion Model (Worst)	
	$P_u^{(w)}$: *A painting of* Horses *in* Winter *Style.*
Original Diffusion Model	
Unlearned Diffusion Model (Worst)	
	$P_u^{(w)}$: *A painting of* Horses *in* Van Gogh *Style.*
Original Diffusion Model	
Unlearned Diffusion Model (Worst)	

Table A8: Examples of image generation using the original stable diffusion model (w/o unlearning), the unlearned diffusion model over the worst-case forgetting prompt set (Worst). For each diffusion model, images are generated based on an unlearned prompt from the worst-case forget set.

Model	Generation Condition
	$P_u^{(w)}$: *A painting of* Horses *in* Rust *Style.*
Original Diffusion Model	
Unlearned Diffusion Model (Worst)	
	$P_u^{(w)}$: *A painting of* Human *in* Winter *Style.*
Original Diffusion Model	
Unlearned Diffusion Model (Worst)	
	$P_u^{(w)}$: *A painting of* Human *in* Van Gogh *Style.*
Original Diffusion Model	
Unlearned Diffusion Model (Worst)	
	$P_u^{(w)}$: *A painting of* Human *in* Rust *Style.*
Original Diffusion Model	
Unlearned Diffusion Model (Worst)	
	$P_u^{(w)}$: *A painting of* Dogs *in* Winter *Style.*
Original Diffusion Model	
Unlearned Diffusion Model (Worst)	

Appendix References

1. Boyd, S.P., Vandenberghe, L.: Convex optimization. Cambridge university press (2004) 1
2. Chen, M., Gao, W., Liu, G., Peng, K., Wang, C.: Boundary unlearning: Rapid forgetting of deep networks via shifting the decision boundary. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7766–7775 (2023) 3
3. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009) 7
4. Fan, C., Liu, J., Zhang, Y., Wei, D., Wong, E., Liu, S.: Salun: Empowering machine unlearning via gradient-based weight saliency in both image classification and generation. arXiv preprint arXiv:2310.12508 (2023) 3
5. Gandikota, R., Materzynska, J., Fiotto-Kaufman, J., Bau, D.: Erasing concepts from diffusion models. arXiv preprint arXiv:2303.07345 (2023) 2
6. Goel, S., Prabhu, A., Sanyal, A., Lim, S.N., Torr, P., Kumaraguru, P.: Towards adversarial evaluations for inexact machine unlearning. arXiv preprint arXiv:2201.06640 (2022) 2
7. Golatkar, A., Achille, A., Soatto, S.: Eternal sunshine of the spotless net: Selective forgetting in deep networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9304–9312 (2020) 2
8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016) 4, 5
9. Jia, J., Liu, J., Ram, P., Yao, Y., Liu, G., Liu, Y., Sharma, P., Liu, S.: Model sparsity can simplify machine unlearning. Advances in neural information processing systems **36** (2023) 3
10. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images (2009) 3
11. Kurmanji, M., Triantafillou, P., Triantafillou, E.: Towards unbounded machine unlearning. arXiv preprint arXiv:2302.09880 (2023) 2
12. Le, Y., Yang, X.: Tiny imagenet visual recognition challenge. CS 231N **7**(7), 3 (2015) 3
13. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014) 4, 5
14. Warnecke, A., Pirch, L., Wressnegger, C., Rieck, K.: Machine unlearning of features and labels. arXiv preprint arXiv:2108.11577 (2021) 2
15. Xu, K., Chen, H., Liu, S., Chen, P.Y., Weng, T.W., Hong, M., Lin, X.: Topology attack and defense for graph neural networks: An optimization perspective. arXiv preprint arXiv:1906.04214 (2019) 1
16. Zhang, Y., Zhang, Y., Yao, Y., Jia, J., Liu, J., Liu, X., Liu, S.: Unlearncanvas: A stylized image dataset to benchmark machine unlearning for diffusion models. arXiv preprint arXiv:2402.11846 (2024) 3