MUNBa: Machine Unlearning via Nash Bargaining

Jing Wu[1], Mehrtash Harandi[2]
[1]Department of Data Science & AI, [2]Department of Electrical and Computer Systems Engineering
Monash University, Melbourne, VIC, Australia
{jing.wu1, mehrtash.harandi}@monash.edu

Abstract

Machine Unlearning (MU) aims to selectively erase harmful behaviors from models while retaining the overall utility of the model. As a multi-task learning problem, MU involves balancing objectives related to forgetting specific concepts/data and preserving general performance. A naive integration of these forgetting and preserving objectives can lead to gradient conflicts and dominance, impeding MU algorithms from reaching optimal solutions. To address the gradient conflict and dominance issue, we reformulate MU as a two-player cooperative game, where the two players, namely, the forgetting player and the preservation player, contribute via their gradient proposals to maximize their overall gain and balance their contributions. To this end, inspired by the Nash bargaining theory, we derive a closed-form solution to guide the model toward the Pareto stationary point. Our formulation of MU guarantees an equilibrium solution, where any deviation from the final state would lead to a reduction in the overall objectives for both players, ensuring optimality in each objective. We evaluate our algorithm's effectiveness on a diverse set of tasks across image classification and image generation. Extensive experiments with ResNet, vision-language model CLIP, and text-to-image diffusion models demonstrate that our method outperforms state-of-the-art MU algorithms, achieving a better trade-off between forgetting and preserving. Our results also highlight improvements in forgetting precision, preservation of generalization, and robustness against adversarial attacks.

WARNING: This paper contains sexually explicit imagery that may be offensive in nature.

1. Introduction

In this paper, we propose to model Machine Unlearning (MU) as a bargaining problem between two players: one seeking to forget purposefully and the other aiming to preserve the model utility. Driven by growing concerns around safety, data privacy, and data ownership, MU has seen rapid developments recently. Data protection regulations like GDPR [68] and CCPA [22] grant users the *right to be forgotten*, obligating companies to expunge data pertaining to a user upon receiving a deletion request. The goal of MU is to remove the influence of specific data points from machine learning models as if the models had never met these points during training [23], thereby ensuring compliance with intellectual property and copyright laws.

Retraining the model from scratch without forgetting data is often considered the gold standard baseline for MU [66, 67]. However, retraining is usually impractical. Consequently, a range of studies thereafter [8, 14, 16, 19–21, 27, 63, 64, 79] propose approximate MU algorithms, sought to improve the efficiency of MU without necessitating full retraining. **Despite the success of MU algorithms, little attention has been paid to the issue of gradient conflict and gradient dominance in MU.**

Roughly speaking, current MU methods involve two subgoals: erasing the influence of particular data points from the model while preserving its performance, *i.e.*, forgetting and preserving. Consider a model with parameters $\boldsymbol{\theta}$ and assume we would like to remove the influence of a set of data points $\mathcal{D}_f$ (*i.e.*, forgetting data). Let $\mathcal{D}_r$ represent the remaining data that is intended to be retained. MU is often formulated [14, 27] as minimizing a weighted sum of two objectives as: $\min_{\boldsymbol{\theta}} \alpha_r \mathcal{L}_r(\boldsymbol{\theta}; \mathcal{D}_r) + \alpha_f \mathcal{L}_f(\boldsymbol{\theta}; \mathcal{D}_f)$ where α_r and α_f are coefficients for balancing two objectives. Here, **1)** $\mathcal{L}_r(\boldsymbol{\theta}; \mathcal{D}_r)$ fine-tunes the model with the remaining data $\mathcal{D}_r$ to preserve the utility and **2)** $\mathcal{L}_f(\boldsymbol{\theta}; \mathcal{D}_f)$ directs the model to forget knowledge associated with $\mathcal{D}_f$ (by maximizing the loss on $\mathcal{D}_f$).

However, the forgetting task gradient (*i.e.*, $\nabla_{\boldsymbol{\theta}} \mathcal{L}_f$) may have conflicting directions with the preservation task gradient (*i.e.*, $\nabla_{\boldsymbol{\theta}} \mathcal{L}_r$). Moreover, the magnitudes of these gradients may differ significantly, potentially causing the joint gradient to be dominated by one of the objectives. Fig. 1 illustrates the histogram of cosine similarity between the joint update vector and both the forgetting task gradient and the preservation task gradient during the MU process, as well as the ratio of their gradient norms, highlight-

ing the frequent occurrence of gradient conflicts and dominance, which are known to cause performance degradation as studied in the literature on Multi-Objective Optimization (MOO) [38, 39, 60, 75]. Addressing these issues can improve the performance of MU algorithm across both forgetting and preserving objectives.

In this paper, we propose to Machine Unlearning via Nash Bargaining (*MUNBa*), to simultaneously resolve the gradient conflict and dominance issue using game theory concepts [45, 65]. Specifically, we frame MU as a cooperative bargaining game, where two players, *i.e.*, forgetting and preservation, offer gradient proposals and negotiate to find a mutually beneficial direction that maximizes the overall gain for both players. Inspired by the study [46], we define the utility function of each player based on the gradient information and derive a closed-form updating direction to steer the scrubbed model towards the Pareto stationary point. With our proposed method *MUNBa*, illustrated in Fig. 1, the gradient conflict and dominance issue between two players is alleviated through the bargaining process. Extensive experiments on classification and generation tasks demonstrate the effectiveness of *MUNBa* in forgetting, preserving model utility, generalization, and robustness against adversarial attacks.

Our contributions are summarized as:

- We examine and empirically demonstrate the gradient conflict and gradient dominance issue in MU. Based on the observations, we propose *MUNBa*, a straightforward optimization method using game theory to simultaneously resolve gradient conflicts and dominance in MU, approaching an equilibrium solution and thus achieving an optimal balance between forgetting and preservation.
- We further provide a theoretical analysis of the convergence, demonstrating that the solution is achieved at Pareto stationary point. Furthermore, through extensive experiments with ResNet [25], the vision-language model CLIP [53], and diffusion models [55], we empirically show that *MUNBa* consistently achieves a superior trade-off between forgetting and preservation compared with other MU methods across several MU benchmarks.

2. Methodology

In this section, we propose *MUNBa*, our unlearning framework that scrubs data from the pre-trained model while maintaining model utility via game theory. Throughout the paper, we denote scalars and vectors/matrices by lowercase and bold symbols, respectively (*e.g.*, a, $\boldsymbol{a}$, and $\boldsymbol{A}$).

2.1. Problem setup

Given a model that trains on the dataset $\mathcal{D}$ with pre-trained weights $\boldsymbol{\theta} \in \mathbb{R}^d$, our objective is

$$\min_{\boldsymbol{\theta}} \quad \alpha_r \mathcal{L}_r(\boldsymbol{\theta}; \mathcal{D}_r) + \alpha_f \mathcal{L}_f(\boldsymbol{\theta}; \mathcal{D}_f), \tag{1}$$

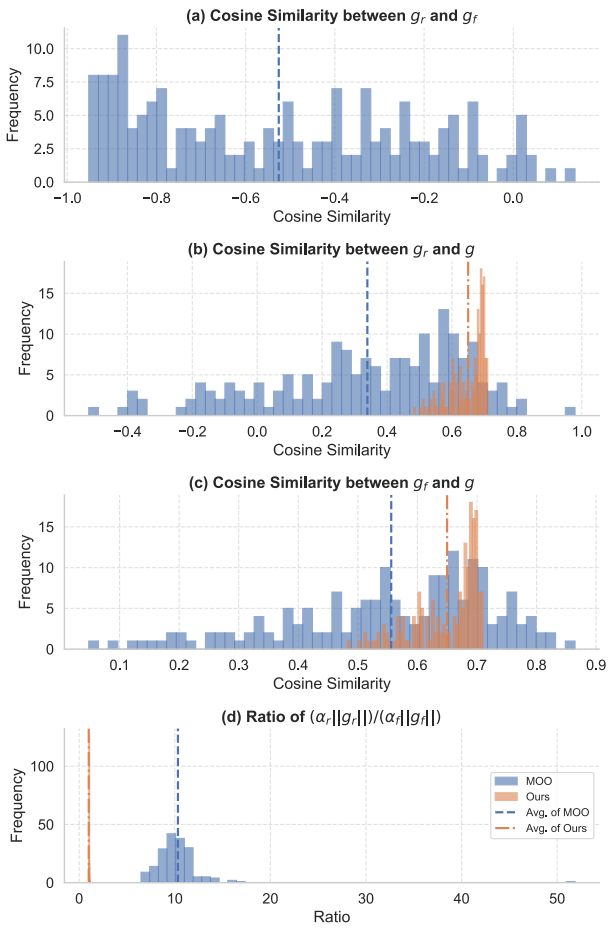

Figure 1. Gradient conflict and dominance happen across the MU process. Instead, our approach alleviates this issue, verified by the higher cosine similarity between the joint update gradient $\tilde{g}$ and both the preservation task gradient $\boldsymbol{g}_r$ and the forgetting task gradient $\boldsymbol{g}_f$. Ours achieves balanced contributions from two objectives (the ratio of gradient norms is 1.0, and the width of "Ours" bar is increased for better visibility). More examples are in §8.

where $\mathcal{D}_f \subset \mathcal{D}$ and $\mathcal{D}_r := \mathcal{D} \setminus \mathcal{D}_f$ represent the forgetting and remaining data, respectively; $\boldsymbol{\alpha} = \begin{bmatrix} \alpha_r & \alpha_f \end{bmatrix}$ denote the coefficient for balancing terms forgetting and preservation; the loss terms $\mathcal{L}_r(\boldsymbol{\theta}; \mathcal{D}_r) = \mathbb{E}_{\boldsymbol{x} \sim \mathcal{D}_r} \ell_r(\boldsymbol{x}; \boldsymbol{\theta}), \mathcal{L}_f(\boldsymbol{\theta}; \mathcal{D}_f) = \mathbb{E}_{\boldsymbol{x} \sim \mathcal{D}_f} \ell_f(\boldsymbol{x}; \boldsymbol{\theta})$ where $\ell_r, \ell_f : \mathcal{X} \times \Theta \to \mathbb{R}_+$ defined on the input space $\mathcal{X}$ and the parameter space Θ.

Gradient conflict and dominance. Eq. (1) involves two subgoals, *i.e.*, forgetting and preservation. Let $\boldsymbol{g}_r = \nabla_{\boldsymbol{\theta}} \mathcal{L}_r(\boldsymbol{\theta}; \mathcal{D}_r)$ and $\boldsymbol{g}_f = \nabla_{\boldsymbol{\theta}} \mathcal{L}_f(\boldsymbol{\theta}; \mathcal{D}_f)$ denote the gradient for updating these two subgoals. We first analyze the alignment between $\boldsymbol{g}_r$ and $\boldsymbol{g}_f$, as well as the alignment between the joint update direction $\tilde{\boldsymbol{g}} := \alpha_r \boldsymbol{g}_r + \alpha_f \boldsymbol{g}_f$ and $\boldsymbol{g}_r$, and the joint update direction $\tilde{\boldsymbol{g}}$ and $\boldsymbol{g}_f$ during the unlearning

process. As illustrated in Fig. 1, the cosine similarity distributions indicate a clear difference in gradient alignment between our method and MOO with $\alpha_r = 1.0, \alpha_f = 0.1$. Under the challenge scenario sample-wise forgetting on CIFAR-10 [33], we observe that there exhibit considerable gradient conflicts, as indicated by the high frequency of negative values of cosine similarity, this means that, the gradients of the preservation task $\boldsymbol{g}_r$ and that of the forgetting task $\boldsymbol{g}_f$ are often misaligned. Additionally, MOO still exhibits considerable gradient conflicts, the gradients of the preservation task and the joint update direction are often misaligned, potentially hindering effective preservation.

In contrast, our method has a much higher average cosine similarity compared to MOO, with the histogram peak shifted closer to positive values, suggesting that our method is more effective at preserving the information about the remaining data, as indicated by the closer alignment with the preservation task gradient $\boldsymbol{g}_r$. Similarly, the cosine similarity between $\tilde{\boldsymbol{g}}$ and the forgetting task gradient $\boldsymbol{g}_f$ for our method again is also positive. This alignment suggests that our method also aligns with the forgetting task, possibly leading to more effective forgetting of targeted information.

Furthermore, we examine the ratio of gradient norms for the two objectives, *i.e.*, $\frac{\alpha_r \|\boldsymbol{g}_r\|}{\alpha_f \|\boldsymbol{g}_f\|}$. We observe that MOO often exhibits an imbalance in gradient magnitudes, potentially with one task dominating the joint update direction. In contrast, our method achieves a balanced ratio of gradient norms (close to 1.0), ensuring that both tasks contribute proportionally to the unlearning process.

Overall, the comparison between the distributions suggests that our method promotes better alignment and balance contributions between the forgetting and preservation tasks, thus effectively reducing gradient conflict and supporting the model's ability to unlearn specific data influence without significantly compromising the preservation of other information.

2.2. MUNBa

2.2.1. Objective

We now describe the proposed method *MUNBa* in detail. We have two players, *i.e.*, forgetting and preservation, aiming to offer gradients to maximize the overall gain. Inspired by [46, 76], we define the utility function $u_f(\tilde{\boldsymbol{g}})$ for the player forgetting and $u_r(\tilde{\boldsymbol{g}})$ for the player preservation as

$$u_r(\tilde{\boldsymbol{g}}) \coloneqq \boldsymbol{g}_r^\top \tilde{\boldsymbol{g}}, \tag{2}$$

$$u_f(\tilde{\boldsymbol{g}}) \coloneqq \boldsymbol{g}_f^\top \tilde{\boldsymbol{g}}, \tag{3}$$

where $\tilde{\boldsymbol{g}}$ denotes the resulting joint direction for updating the model. For preservation, Eq. (2) estimates the alignment between the update direction $\tilde{\boldsymbol{g}}$ and the gradient that decreases the loss over the remaining data $\mathcal{D}_r$; while for forgetting, Eq. (3) measures the alignment between the update direction $\tilde{\boldsymbol{g}}$ and the gradient that increases the loss over the forgetting data $\mathcal{D}_f$. Consequently, if the final update direction $\tilde{\boldsymbol{g}}$ deviates significantly from the gradient $\boldsymbol{g}_r$, the payoff would decrease; and if the final update direction $\tilde{\boldsymbol{g}}$ strays far from the gradient $\boldsymbol{g}_f$, the payoff would decrease. Given that this is a cooperative game, it is reasonable to expect that players will not undermine one another without personal benefit [46]. Therefore, the agreed solution should not be dominated by any alternative, meaning the solution is considered to converge to the Pareto stationary point.

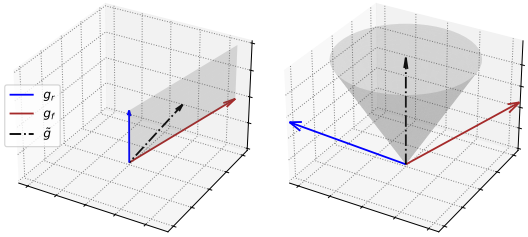

Figure 2. Visualization of update vector. There exists a solution within the convex cone where both utility functions are positive.

Lemma 2.1 (Feasibility). *Let $u_r, u_f : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ be the utility functions defined in Eqs. (2) and (3). Assume $-1 < \frac{\boldsymbol{g}_r^\top \boldsymbol{g}_f}{\|\boldsymbol{g}_r\| \|\boldsymbol{g}_f\|} < 0$. Define the feasible set C as $C = \{\tilde{\boldsymbol{g}} \mid u_r(\tilde{\boldsymbol{g}}) > 0,\ u_f(\tilde{\boldsymbol{g}}) > 0\}$. Then C is non-empty.*

Lemma 2.2 (Cone property). . *The feasible set $C \coloneqq \{\tilde{\boldsymbol{g}} \mid u_r(\tilde{\boldsymbol{g}}) > 0,\ u_f(\tilde{\boldsymbol{g}}) > 0\}$ forms a cone in $\mathbb{R}^n$.*

These lemmas ensure that, as long as the two gradients are not completely contradictory, there exists an update $\tilde{\boldsymbol{g}}$ that can improve both objectives simultaneously. Obviously, our aim is to determine a $\tilde{\boldsymbol{g}}$ that maximizes improvement across both objectives. Please see the proof in §6. We hence rewrite the objective in Eq. (1) as

$$\max_{\tilde{\boldsymbol{g}} \in \mathcal{B}_\epsilon} \quad \log\big(u_r(\tilde{\boldsymbol{g}})\big) + \log\big(u_f(\tilde{\boldsymbol{g}})\big), \tag{4}$$

where the update vector $\tilde{\boldsymbol{g}}$ is constrained to lie within a ball $\mathcal{B}_\epsilon$ of radius ϵ centered at 0. Here, the logarithm is adopted to help balance and align with the property that utility gains less benefit as it continues to improve. With this objective, the Pareto stationary point would be received. Note that in this paper, we show that the MU algorithms suffers from gradient conflict and dominance. To address these issues, we adopt the objective proposed in [46] which produces an update direction that balances the contributions of multiple tasks by leveraging principles from game theory.

2.2.2. Solution

We now present the Nash bargaining solution to Eq. (4) by the following three theorems. We provide the proofs in §6.

Theorem 2.3 (Optimality condition). *Let $f(\tilde{\boldsymbol{g}}) \coloneqq \log\big(u_r(\tilde{\boldsymbol{g}})\big) + \log\big(u_f(\tilde{\boldsymbol{g}})\big)$ and some scalar λ. The optimal*

solution $\tilde{\boldsymbol{g}}^*$ to Eq. (4) must satisfy

$$\nabla f(\tilde{\boldsymbol{g}}^*) = \lambda \tilde{\boldsymbol{g}}^*, \textit{where} \quad \tilde{\boldsymbol{g}}^* = \alpha_r \boldsymbol{g}_r + \alpha_f \boldsymbol{g}_f. \tag{5}$$

where $\alpha_r > 0$ and $\alpha_f > 0$.

Lemma 2.4 (Linear dependence)**.** *$\boldsymbol{g}_r$ and $\boldsymbol{g}_f$ are linear dependent at the Pareto stationary point.*

Theorem 2.5 (Solution characterization)**.** *Denote $\boldsymbol{G} = \begin{bmatrix}\boldsymbol{g}_r & \boldsymbol{g}_f\end{bmatrix} \in \mathbb{R}^{d\times 2}$, then the solution to Eq. (5), up to scaling, is $\tilde{\boldsymbol{g}}^* = (\alpha_r \boldsymbol{g}_r + \alpha_f \boldsymbol{g}_f)$ where $\boldsymbol{\alpha}$ is the solution to*

$$\boldsymbol{G}^\top \boldsymbol{G}\boldsymbol{\alpha} = 1/\boldsymbol{\alpha}. \tag{6}$$

We employ the same proofs as in Theorem 3.2 of [76] for Theorem 2.5, without the need to assume the linear independence of gradients. This also gives us the form of solution where $\boldsymbol{\alpha} = \begin{bmatrix}\alpha_r & \alpha_f\end{bmatrix}^\top$ solves

$$\begin{aligned} \alpha_r \|\boldsymbol{g}_r\|_2^2 + \alpha_f(\boldsymbol{g}_f^\top \boldsymbol{g}_r) &= 1/\alpha_r, \\ \alpha_f \|\boldsymbol{g}_f\|_2^2 + \alpha_r(\boldsymbol{g}_f^\top \boldsymbol{g}_r) &= 1/\alpha_f, \end{aligned} \tag{7}$$

where the relative coefficients α_r and α_f emerge from the forgetting and preservation player's impact and interactions with each other. If the interaction between two players is positive, *i.e.*, $\boldsymbol{g}_f^\top \boldsymbol{g}_r > 0$, the per-task gradient can aid each other and the relative coefficients will decrease. Conversely, the relative coefficients will increase in priority towards individual objectives.

Now, we only need to solve $\boldsymbol{\alpha}$ in Eq. (6) to obtain the bargaining solution to Eq. (4). Different from the general framework [46] that approximates $\boldsymbol{\alpha}$, we can get a closed-form solution for $\boldsymbol{\alpha}$ in this case.

Theorem 2.6 (Closed-form solution)**.** *Denote the Gram matrix $\mathbb{R}^{2\times 2} \ni \boldsymbol{K} := \boldsymbol{G}^\top \boldsymbol{G} = \begin{bmatrix}\boldsymbol{g}_r^\top \boldsymbol{g}_r & \boldsymbol{g}_r^\top \boldsymbol{g}_f \\ \boldsymbol{g}_r^\top \boldsymbol{g}_f & \boldsymbol{g}_f^\top \boldsymbol{g}_f\end{bmatrix} = \begin{bmatrix} g_1 & g_2 \\ g_2 & g_3 \end{bmatrix}$, and denote ϕ as the angle between $\boldsymbol{g}_r$ and $\boldsymbol{g}_f$. Then, closed-form solution for $\boldsymbol{\alpha}$ in $\tilde{\boldsymbol{g}}^* = \alpha_r \boldsymbol{g}_r + \alpha_f \boldsymbol{g}_f$ is*

$$\begin{cases} \alpha_r = \frac{1}{\|\boldsymbol{g}_r\|}\sqrt{\frac{1-\cos(\phi)}{\sin^2(\phi)+\xi}}, \\ \alpha_f = \frac{\sqrt{\sin^2(\phi)(1-\cos(\phi))}}{\|\boldsymbol{g}_f\|}. \end{cases} \tag{8}$$

where ξ is a very small value to avoid division by zero.

Remark 2.7. *If $\boldsymbol{g}_r$ and $\boldsymbol{g}_f$ are linearly dependent,* i.e.*, for some scalar ζ, $\boldsymbol{g}_r = \zeta \boldsymbol{g}_f$, the determinant of the Gram matrix $\boldsymbol{K}$ will become zero. To address this, we can add some noise to the gradient with a smaller norm to break dependence for $\zeta < 0$, enabling a well-defined solution for $\boldsymbol{\alpha}$. When $\zeta \geq 0$, $\boldsymbol{g}_r$ and $\boldsymbol{g}_f$ is aligned, we can simply choose $\boldsymbol{\alpha} = \begin{bmatrix}0.5 & 0.5\end{bmatrix}^\top$.*

Algorithm 1 Machine Unlearning via Nash Bargaining.

Input: Model with parameters $\boldsymbol{\theta}$, forgetting and remaining data $\mathcal{D}_f, \mathcal{D}_r$, number of iterations T, learning rate η.
Output: Parameters $\boldsymbol{\theta}^*$ for the scrubbed model.

1: Initalize $\boldsymbol{\alpha} = \begin{bmatrix}\alpha_r & \alpha_f\end{bmatrix}^\top$.
2: **for** iteration t in T **do**
3: Mini-batch $\boldsymbol{X}_f^{(t)} \sim \mathcal{D}_f$ and $\boldsymbol{X}_r^{(t)} \sim \mathcal{D}_r$.
4: $\boldsymbol{g}_r = \nabla \mathcal{L}_r(\boldsymbol{\theta}^{(t)}; \boldsymbol{X}_r^{(t)}), \boldsymbol{g}_f = \nabla \mathcal{L}_f(\boldsymbol{\theta}^{(t)}; \boldsymbol{X}_f^{(t)})$.
5: Set $\boldsymbol{G} = \begin{bmatrix}\boldsymbol{g}_r & \boldsymbol{g}_f\end{bmatrix}$ and $\boldsymbol{K} = \boldsymbol{G}^\top \boldsymbol{G} = \begin{bmatrix} g_1 & g_2 \\ g_2 & g_3 \end{bmatrix}$.
6: Solve Eq. (6) with Eq. (8) to obtain $\boldsymbol{\alpha}$:
$\alpha_r = \frac{1}{\|\boldsymbol{g}_r\|}\sqrt{\frac{1-\cos(\phi)}{\sin^2(\phi)+\xi}}, \alpha_f = \frac{\sqrt{\sin^2(\phi)(1-\cos(\phi))}}{\|\boldsymbol{g}_f\|}$.
7: Updating: $\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} - \eta \boldsymbol{G}\boldsymbol{\alpha}$.
8: **end for**
9: **return** $\boldsymbol{\theta}^{(T)}$

Algorithm 1 describes the procedure of our algorithm *MUNBa* in detail. We first calculate the gradient for each player, then solve Eq. (6) to obtain the coefficient $\boldsymbol{\alpha}$, and finally, update the model parameters with the new coefficient, which seeks the maximum gain in terms of the utilities.

2.2.3. Theoretical Properties

We now examine key theoretical properties of *MUNBa*. In particular, we show that the solution enjoys Pareto optimality, the norms of $\boldsymbol{\alpha}$ are bounded under mild conditions, and for Lipschitz functions, the updates guarantee a monotonically decreasing loss, leading to convergence. We present the following two theorems based on [46, 76] but with a slight difference in the proofs (See §6).

Lemma 2.8 (Boundedness)**.** *For player $i \in \{r, f\}$, assume $\|\boldsymbol{g}_i\|$ is bounded by $M < \infty$, then $\frac{1}{\sqrt{2}M} \leq \|\alpha_i\| \leq \frac{\sqrt{2}}{M}$.*

Theorem 2.9 (Pareto improvement)**.** *Let $\mathcal{L}_i(\boldsymbol{\theta}^{(t)})$ denote the loss function for player $i \in \{r, f\}$ at step t, where r and f represent the preservation player and the forgetting player, respectively. Assume $\mathcal{L}_i(\boldsymbol{\theta}^{(t)})$ is differential and Lipschitz-smooth with constant $L > 0$, if the learning rate at step t is set to $\eta^{(t)} = \min \frac{1}{L\alpha_i^{(t)}}$, then the update ensures $\mathcal{L}_i(\boldsymbol{\theta}^{(t+1)}) \leq \mathcal{L}_i(\boldsymbol{\theta}^{(t)})$ for both players.*

Theorem 2.10 (Convergence)**.** *Since each player's loss $\mathcal{L}_i(\boldsymbol{\theta}^{(t)})$ is monotonically decreasing and bounded below, the combined loss $\mathcal{L}(\boldsymbol{\theta})$ converges to $\mathcal{L}(\boldsymbol{\theta}^*)$ and $\boldsymbol{\theta}^*$ is the stationary point of $\mathcal{L}(\boldsymbol{\theta})$.*

This shows that the loss value is decreasing for both players using the Nash bargaining solution, enabling them to approach an equilibrium solution without either player's loss increasing along the way, thus achieving an optimal balance between the forgetting and preservation objectives.

3. Related work

MU has applications across a wide range of domains, including classifications [8, 18, 19, 30, 43] and regression [63], diffusion models [1, 14, 16, 17, 27, 73, 79], federated learning [24, 40, 41, 69, 71, 83], graph neural networks [7, 9], as well as language models [51] and vision-language models [52]. Several benchmarks [54, 81] have been proposed for improving the quality of unlearning measurement. Retraining the model from scratch without forgetting data is often considered the gold standard for unlearning algorithms. However, this approach is impractical for most production models, which require significant training time and computational resources. As a result, approximate unlearning methods [3, 5, 6, 8, 15, 19–21, 26, 34, 35, 42, 61, 62] have gained traction as practical alternatives.

Most MU methods rely on techniques such as influence functions [32, 43, 47, 58, 72, 73] or probabilistic methods [19–21]. Tarun et al. [63] employ knowledge distillation to train a student model that mimics the behavior of the original model while filtering out the influence of the forgetting data, thereby preserving model utility. Jia et al. [30] explore the role of sparsity in enhancing the effectiveness of unlearning. Fan et al. [14], and Wu and Harandi [73] identify important parameters w.r.t. the forgetting data to erase their influence in models. Tarun et al. [64] and Chundawat et al. [10] propose MU methods considering the scenarios where training data are not available.

While most MU methods have been developed for classification, Fan et al. [14] highlight their limitations in addressing MU for image generation, which is critical for protecting copyrights and preventing inappropriate outputs. Gandikota et al. [16] propose an energy-based method tailored to classifier-free guidance mechanisms for erasing concepts in text-to-image diffusion models. Heng and Soh [27] introduce a continual learning framework to erase concepts across various types of generative models. Fan et al. [14] propose a very potent unlearning algorithm called SalUn that shifts attention to important parameters w.r.t. the forgetting data. Poppi et al. [52] recently proposed Safe-CLIP to forget unsafe linguistic or visual items in the embedding space for the vision-and-language model CLIP. Their scrubbed model can be effectively employed with pre-trained generative models. Despite these advancements, several studies [13, 78, 80, 82] demonstrate the vulnerabilities of MU methods, highlighting that with adversarial prompts, the scrubbed models can still regenerate images containing the contents requested to be forgotten.

This work. Although most MU methods are empirically demonstrated to be promising in effective forgetting and preserving model utility, they stop short of probing the control of conflict and dominance between two objectives. As for [31, 36, 73], these methods are designed to mitigate gradient conflict but do not address gradient dominance. We aim to bridge this gap by simultaneously resolving gradient conflicts and gradient dominance via game theory [46] which provides a more principled method compared to other conflict aversion techniques [38, 59, 75]. Please refer to [38, 39, 59, 60, 75] for a comprehensive overview of alternative methods.

4. Experiment

In this section, we empirically show how *MUNBa* effectively eliminates the data influence in models while maintaining the performance across various MU benchmarks.

4.1. Setup

Datasets. For the classification task, we use SVHN [48] and CIFAR-10 [33], both with an image resolution of 32×32, as well as Celeb-HQ Facial Identity Recognition Dataset [44] (Celeb-HQ-307), scaled to 224×224 resolution. For CLIP [53], ImageNet-1K [11] and Oxford Pets [49] with 37 categories are considered. For assessing unlearning in generative models, we use I2P [57], consisting of 4703 prompts that lead to NSFW (not safe for work) content generated by SD v1.4 [55], and Imagenette [29] to perform class-wise forgetting in SD. COCO-30K prompts from the MS-COCO validation set [37] are adopted to evaluate the quality of generated images. 142 nudity-related prompts presented in [82] are used to examine the robustness of MU methods against adversarial prompt attacks.

Baselines. We include the following standard MU methods, as well as recently proposed SOTA approaches: (1) *Retrain*. (2) *Fine-tuning (FT)* [70]. (3) *Gradient Ascent (GA)* [66]. (4) *Influence Unlearning (IU)* [32]. (5) *Boundary Shrink (BS)* [8] and (6) *Boundary Expand (BE)* [8]. (7) ℓ_1*-Sparse* [30]. (8) *Saliency Unlearning (SalUn)* [14]. (9) *Scciorhands (SHs)* [73]. (10) *Erased Stable Diffusion (ESD)* [16]. (11) *Forget-Me-Not (FMN)* [79]. (12) *Selective Amnesia (SA)* [27]. Note that these MU methods are not universally designed for classification and generation simultaneously, our assessment hence is specific to the task for which they were originally developed and employed.

Metrics. To evaluate the effectiveness of MU algorithms, we use the following common metrics: (1) *Accuracy*: we assess the model's accuracy on $\mathcal{D}_f$ (denoted as $\mathbf{Acc}_{\mathcal{D}_f}$), $\mathcal{D}_r$ (denoted as $\mathbf{Acc}_{\mathcal{D}_r}$), and $\mathcal{D}_t$ (denoted as $\mathbf{Acc}_{\mathcal{D}_t}$). (2) *Membership Inference Attack (MIA)*: evaluates the difficulty of inferring whether a particular data point was part of the training data. Effective MU methods should make it challenging to identify samples from $\mathcal{D}_f$ as having been in the training data. (3) *Average Gap (Avg. Gap)* [14]: average performance difference between the scrubbed model and the retrained model across the above metrics, which is calculated as Avg. Gap $= (|\text{Acc}_{\mathcal{D}_t} - \text{Acc}^*_{\mathcal{D}_t}| + |\text{Acc}_{\mathcal{D}_f} - \text{Acc}^*_{\mathcal{D}_f}| + |\text{Acc}_{\mathcal{D}_r} - \text{Acc}^*_{\mathcal{D}_r}| + |\text{MIA} - \text{MIA}^*|)/4$, where

Table 1. Quantitative results for forgetting 10% identities and 10% randomly selected samples. *MUNBa* demonstrates superiority in balancing forgetting and preservation. The best and the second best are highlighted in orange and grey, respectively.

Method	Celeb-HQ-307					CIFAR-10				
	$\text{Acc}_{\mathcal{D}_f}(\downarrow)$	$\text{Acc}_{\mathcal{D}_t}(\uparrow)$	$\text{Acc}_{\mathcal{D}_r}(\uparrow)$	MIA($\uparrow$)	Avg. Gap	$\text{Acc}_{\mathcal{D}_f}(\downarrow)$	$\text{Acc}_{\mathcal{D}_t}(\uparrow)$	$\text{Acc}_{\mathcal{D}_r}(\uparrow)$	MIA($\uparrow$)	Avg. Gap
Retrain	0.00±0.00	87.02±0.80	99.96±0.01	100.0±0.00	-	94.81±0.53	94.26±0.14	100.0±0.00	13.05±0.64	-
FT [70]	99.94±0.12	**88.59±0.59**	**99.97±7.02**	5.28±2.03	49.06	97.82±0.59	**93.58±0.17**	**99.70±0.07**	5.92±0.72	2.78
GA [66]	87.60±8.71	81.22±2.11	99.74±0.26	51.37±5.96	35.56	96.14±0.08	90.40±0.25	96.75±0.22	7.72±2.34	3.44
IU [32]	88.92±10.3	70.24±11.8	95.27±5.07	29.59±18.6	45.20	98.08±2.10	91.91±2.73	98.01±2.26	4.01±3.44	4.16
BE [8]	69.07±2.73	44.11±2.08	95.58±1.23	46.24±5.90	42.53	98.05±1.07	92.07±0.87	98.05±1.10	**18.59±0.56**	3.23
BS [8]	98.18±1.92	81.92±0.27	99.86±0.03	45.93±5.11	39.36	97.91±0.77	92.05±0.36	97.90±0.70	16.23±1.37	2.65
ℓ_1-sparse [30]	17.84±2.51	78.92±2.19	98.78±0.64	100.0±0.00	6.78	96.72±3.54	92.81±0.07	98.48±1.64	7.44±7.21	2.19
SalUn [14]	0.94±0.32	85.69±0.42	99.82±0.09	100.0±0.00	0.60	95.83±0.55	92.10±0.30	98.27±0.31	12.99±1.23	1.24
SHs [73]	0.06±0.12	85.53±0.80	99.95±0.02	100.0±0.00	0.39	95.40±1.48	92.92±0.48	98.93±0.57	9.56±2.13	1.62
MUNBa (Ours)	**0.00±0.00**	87.24±1.09	99.80±0.08	**100.0±0.00**	**0.10**	**94.99±0.53**	93.12±0.04	98.09±0.14	13.68±0.80	**0.97**

$\text{Acc}^*_{\mathcal{D}_t}$, $\text{Acc}^*_{\mathcal{D}_f}$, $\text{Acc}^*_{\mathcal{D}_r}$ and MIA^* are metric values of the retrained model. A lower value implies that the unlearned model closely resembles the retrained model. (4) *Frechet Inception Distance (FID)* [28]: the widely-used metric for assessing the quality of generated images. (5) *CLIP score*: the similarity between the visual features of the generated image and its corresponding textual embedding.

4.2. Results on classification

We first evaluate MU methods on classification, trying to forget randomly selected 10% identities among 307 identities on the Celeb-HQ-307, and randomly selected 10% data on CIFAR-10. Class-wise forgetting on SVHN can be found in §8. In brief, the results suggest that *MUNBa* effectively induces forgetting for the relevant identities and samples, with minor degradation in model generalization and performance over $\mathcal{D}_r$, and *MUNBa* demonstrates the smallest average performance gap with retrained models.

In Tab. 1, among the baselines, FT exhibits high accuracies on $\mathcal{D}_r$ and $\mathcal{D}_t$ but fails to forget data traces. BE and BS are developed to perform class-wise forgetting and, as such cannot effectively forget identities and randomly selected samples. In contrast, SalUn, SHs, and *MUNBa* demonstrate superior capabilities in forgetting and preserving. SalUn achieves a forgetting accuracy of 0.94% and an accuracy of 85.69% on test data $\mathcal{D}_t$ when forgetting identities. *MUNBa* slightly surpasses SalUn in terms of the forgetting accuracy (*i.e.*, 0%) and test accuracy (*i.e.*, 87.24%) on Celeb-HQ-307, and slightly surpasses SHs and SalUn in terms of the forgetting accuracy and test accuracy on CIFAR-10. Overall, these results underscore our proposed algorithm *MUNBa* superior capabilities in balancing forgetting and preserving model utility. *MUNBa* not only minimizes privacy risks but also maintains the integrity and applicability of the model to unseen data.

4.3. Results on CLIP

We further investigate the performance of *MUNBa* when forgetting with CLIP, which plays a crucial role in tasks such as image generation. CLIP is often trained on large-scale web data, which can inadvertently introduce inappropriate content, limiting its use in sensitive or trustworthy applications and raising concerns about its suitability for widespread adoption. By effectively removing unwanted content, *MUNBa* alleviates these issues, enhancing the reliability and applicability of CLIP in these critical contexts.

We adopt a pre-trained CLIP with ViT-B/32 as the image encoder. Tab. 2 presents the performance in class-wise forgetting with CLIP on Oxford Pets. Due to CLIP's zero-shot capability, the original CLIP model demonstrates moderate performance in both erasing and retaining classes. As observed, FT achieves a good balance between forgetting and maintaining model performance, highlighting the tendency of large multimodal models to experience catastrophic forgetting when adapted to new tasks [77]. However, the generalization capability of CLIP may be damaged after fine-tuning [12], as evidenced by the performance degradation on ImageNet (here, we already exclude the classes same as those in Oxford Pets from ImageNet). While SHs excel in forgetting, it struggles to maintain a good generalization ability of CLIP, as shown by the decline in ImageNet performance after unlearning. We hypothesize that this is due to important knowledge being erased during the trimming stage in SHs. SalUn maintains relatively strong performance on ImageNet, likely because it only fine-tunes the saliency weights w.r.t. the forgetting class, thereby preserving broader generalization. Our method, *MUNBa*, outperforms existing approaches by effectively erasing and retaining class information while preserving generalization. Specifically, *MUNBa* achieves a forgetting accuracy of 2.5%, test accuracy of ∼95%, and competitive generalization performance with an ImageNet ac-

Table 2. Quantitative results for class-wise forgetting with CLIP model on Oxford Pets. Original CLIP: the zero-shot CLIP model on Oxford Pets. $\text{Acc}_{\text{ImageNet}}$: the Top-1 accuracy on ImageNet excluding the classes in forgetting data, measuring the utility of scrubbed CLIP models. SalUn excels in $\text{Acc}_{\text{ImageNet}}$ but performs less effectively than others on both $\text{Acc}_{\mathcal{D}_r}$ and $\text{Acc}_{\mathcal{D}_t}$. The best and the second best are highlighted in orange and grey, respectively.

Method	Forget one class				Forget three classes			
	$\text{Acc}_{\mathcal{D}_f}(\downarrow)$	$\text{Acc}_{\mathcal{D}_r}(\uparrow)$	$\text{Acc}_{\mathcal{D}_t}(\uparrow)$	$\text{Acc}_{\text{ImageNet}}(\uparrow)$	$\text{Acc}_{\mathcal{D}_f}(\downarrow)$	$\text{Acc}_{\mathcal{D}_r}(\uparrow)$	$\text{Acc}_{\mathcal{D}_t}(\uparrow)$	$\text{Acc}_{\text{ImageNet}}(\uparrow)$
Original CLIP	52.19±19.89	78.37±0.59	79.07±0.57	60.09±0.00	73.39±9.47	72.02±0.84	72.42±0.95	60.09±0.00
FT [70]	2.50±2.65	95.45±0.55	91.14±0.93	56.07±0.49	37.81±7.15	94.34±2.52	90.43±2.58	53.90±4.69
GA [66]	12.81±1.33	79.32±0.14	79.42±0.49	59.79±0.29	47.08±9.95	63.03±12.92	64.18±13.44	57.55±0.09
ℓ_1-sparse [30]	3.13±4.42	94.92±1.92	92.04±1.72	56.22±1.84	37.66±6.93	96.31±0.49	92.10±0.22	57.42±0.18
SalUn [14]	4.69±3.09	83.88±0.20	82.93±1.23	**59.94±0.11**	38.59±7.66	82.94±0.67	82.07±1.20	**58.92±0.02**
SHs [73]	**0.00±0.00**	98.11±0.92	91.41±1.33	37.97±1.66	**24.69±8.63**	97.61±0.32	91.00±0.59	33.38±1.20
MUNBa (Ours)	2.50±2.65	**99.66±0.16**	**94.99±0.69**	59.36±0.06	32.50±3.54	**99.81±0.12**	**94.48±0.31**	58.23±0.06

curacy of 59.36%, indicating minimal degradation in zero-shot transferability.

Furthermore, we explore the performance of scrubbed CLIP for downstream tasks such as text-to-image generation. We replace the text encoder in SD with our scrubbed CLIP text encoder, the FID score between 1K images from the training set and generated images is around 2.94, and none of the generated images are classified as the forgetting class. The SD model with our scrubbed CLIP text encoder, reduces the probabilities of generating images when using corresponding textual prompts, thus demonstrating its usefulness also in a text-to-image generation setting. Examples can be found in §8 in the Appendix. We notice that SD with our scrubbed CLIP text encoder can even learn new information. For instance, in Fig. 15, with the prompt 'A photo of Persian', original SD v1.4 generates rug, while the SD with our scrubbed CLIP text encoder successfully generates corresponding images.

4.4. Results on generation

We also employ *MUNBa* to mitigate the generation of NSFW (not safe for work) content and perform class-wise forgetting in text-to-image Stable Diffusion (SD) models. For concept-wise forgetting, 4703 images are generated by SD v1.4 using I2P prompts and 1K images conditioned on prompts c_f ={'nudity', 'naked', 'erotic', 'sexual'} as suggested in [27] (results can be found in §8). We then evaluate on these generated images using the open-source NudeNet classifier [2], to classify the generated images into various corresponding nude body parts. For the class-wise forgetting, the forgetting class c_f is specified using the prompt 'an image of [c_f]'. The unlearning performance is measured by FID and UA (*i.e.*, $1 - P_\psi(\mathbf{y} = c_f|\mathbf{x})$) [14].

Tab. 3 presents the class-wise forgetting performance on Imagenette. More results can be found in §8 in the Appendix. Results for methods with * are from SalUn [14]. As observed, SalUn outperforms other MU methods in UA

Table 3. Performance of class-wise forgetting on Imagenette using SD. UA: the accuracy of the generated images that do not belong to the forgetting class. The FID score is measured compared to validation data for the remaining classes.

Forget. Class	ESD* [16]		SalUn* [14]		*MUNBa*	
	FID ↓	UA (%)↑	FID ↓	UA (%)↑	FID ↓	UA (%)↑
Tench	1.22	99.40	2.53	100.00	1.70	100.00
English Springer	1.02	100.00	0.79	100.00	1.17	100.00
Cassette Player	1.84	100.00	0.91	99.80	0.59	99.90
Chain Saw	1.48	96.80	1.58	100.00	1.83	99.90
Church	1.91	98.60	0.90	99.60	0.99	100.00
French Horn	1.08	99.80	0.94	100.00	0.92	99.90
Garbage Truck	2.71	100.00	0.91	100.00	1.45	100.00
Gas Pump	1.99	100.00	1.05	100.00	1.13	99.90
Golf Ball	0.80	99.60	1.45	98.80	1.04	99.90
Parachute	0.91	99.80	1.16	100.00	1.13	99.90
Average	1.49	99.40	1.22	99.82	**1.20**	**99.94**

across different forgetting classes. Averaging results across all ten classes provides a more comprehensive evaluation and mitigates the risk of cherry-picking. Our results, based on this average approach, clearly indicate the advantages of our method. Tab. 4 and Fig. 3 further present the performance of different MU methods in forgetting the concept of 'nudity'. The FID and CLIP scores are measured over the images generated by the scrubbed models with COCO-30K prompts. Here, SalUn generates the fewest harmful images across most of the nude body part classes, but *MUNBa* significantly improves the overall quality of the generated images, *i.e.*, SalUn achieves an FID of approximately 25 and *MUNBa* reaches an FID of around 15.92, while *MUNBa* slightly worse than SalUn in terms of the exposed body detected in generated images. ESD achieves a lower FID score than *MUNBa* (*i.e.*, 15.76), but *MUNBa* significantly outperforms ESD in erasing nudity, particularly on sensitive content like 'female breast' and 'male breast'.

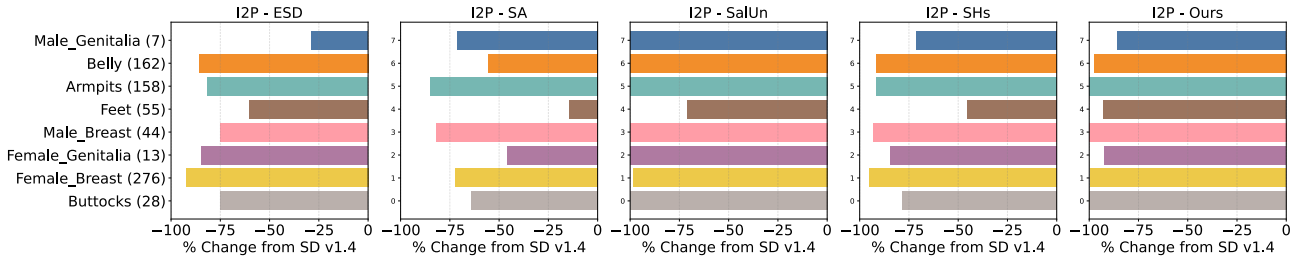

Figure 3. Quantity of nudity content detected using the NudeNet classifier from I2P data. Our method effectively erases nudity content from Stable Diffusion (SD), outperforming ESD, SA, and SHs. SalUn slightly outperforms *MUNBa* in terms of forgetting but *MUNBa* significantly improves the overall quality of the generated images as illustrated in Tab. 4.

Table 4. Evaluation of generated images by SD when forgetting 'nudity'. The FID score is measured compared to validation data, while the CLIP similarity score evaluates the alignment between generated images and the corresponding prompts. Attack success rate (ASR): the performance when adopting adversarial prompt attacks to regenerate nudity-related content.

	SD v1.4	ESD	SA	SalUn	SHs	*MUNBa*
FID ↓	15.97	**15.76**	25.58	25.06	19.45	15.92
CLIP ↑	31.32	30.33	**31.03**	28.91	30.73	30.43
ASR (%) ↓	100.00	73.24	48.59	11.27	35.92	**3.52**

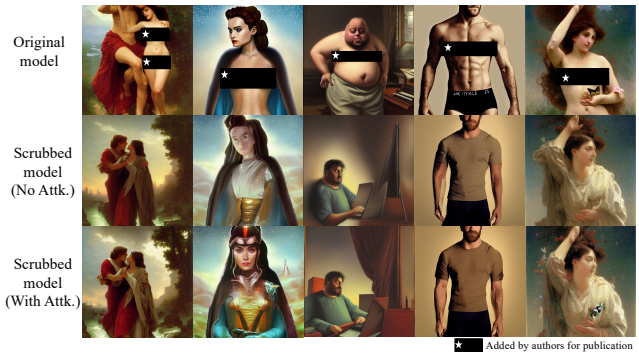

Figure 4. Top to Bottom: generated examples by SD v1.4, our scrubbed SD after erasing nudity, and our scrubbed SD conditioned on adversarial prompts generated by UnlearnDiffAtk [82], respectively. Our method *MUNBa* not only effectively erases the concept of nudity but also exhibits strong robustness against adversarial attacks.

4.5. Robustness against attacks

Finally, we investigate the robustness against adversarial attacks to analyze the safety degree of our scrubbed models. We choose the SOTA method UnlearnDiffAtk [82], and evaluate against the text-to-image SD models in erasing the concept of 'nudity'. We set the prepended prompt perturbations by $N = 5$ tokens, sample 50 diffusion time steps, and perform attack running for 40 iterations with a learning rate of 0.01 at each step. Tab. 4 presents the performance of MU methods against UnlearnDiffAtk in 'nudity' erasure. The prompts and their adversarial versions used for Fig. 4 are detailed in §7 in the Appendix. As observed, SD scrubbed by *MUNBa* exhibits stronger robustness than models scrubbed by other MU methods. Specifically, *MUNBa* achieves the lowest attack success rate of 3.52%, indicating effective resistance to adversarial prompt attacks that attempt to regenerate nudity-related content. Furthermore, *MUNBa* maintains a favorable FID score of 15.92, suggesting that *MUNBa* not only effectively erases undesired content but also preserves the image quality.

5. Conclusion, Limitations, Broader Impacts

This paper contributes *MUNBa*, erasing the influence of forgetting data in models across classification and generation. *MUNBa* resolves gradient conflicts and dominance in MU via game theory, reaching the Pareto stationary point and exhibiting superiority in balancing between forgetting and preservation compared with existing MU methods.

However, while unlearning protects privacy, it may also hinder the ability of relevant systems, potentially lead to biased outcomes, and even be adopted for malicious usage, *e.g.*, adversaries might seek to "erase" important or sensitive information to distort the model's performance, bias decision-making processes, or even obscure critical information. Therefore, ensuring that unlearning techniques are robust to malicious attempts and do not compromise model integrity is a key area for future work. In addition, although *MUNBa* is more effective than baselines, it is slower than some of them (see Section 8.1) and may fail in some cases (see Section 8.4). Although most MU methods successfully remove information about unwanted concepts/classes, they still cause an obvious impact on the remaining concepts/classes. Future works could investigate the scenario where training data are not available, as well as more efficient optimization methods targeted at resolving gradient conflicts and dominance. We hope *MUNBa* could serve as an inspiration for future research in the field of MU.

Acknowledgements

Mehrtash Harandi is supported by the Australian Research Council (ARC) Discovery Program DP250100262. The authors gratefully acknowledge the anonymous reviewers for their insightful feedback and valuable suggestions, which have significantly improved the quality of this work.

References

[1] Silas Alberti, Kenan Hasanaliyev, Manav Shah, and Stefano Ermon. Data unlearning in diffusion models. *International Conference on Learning Representations (ICLR)*, 2025. 5

[2] P Bedapudi. Nudenet: Neural nets for nudity classification, detection and selective censoring, 2019. 7

[3] Jacopo Bonato, Marco Cotogni, and Luigi Sabetta. Is retain set all you need in machine unlearning? restoring performance of unlearned models with out-of-distribution images. In *European Conference on Computer Vision (ECCV)*, pages 1–19. Springer, 2024. 5, 7

[4] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004. 2

[5] Anh Bui, Khanh Doan, Trung Le, Paul Montague, Tamas Abraham, and Dinh Phung. Removing undesirable concepts in text-to-image generative models with learnable prompts. *arXiv preprint arXiv:2403.12326*, 2024. 5

[6] Sungmin Cha, Sungjun Cho, Dasol Hwang, Honglak Lee, Taesup Moon, and Moontae Lee. Learning to unlearn: Instance-wise unlearning for pre-trained classifiers. In *Proceedings of the AAAI conference on artificial intelligence*, pages 11186–11194, 2024. 5

[7] Min Chen, Zhikun Zhang, Tianhao Wang, Michael Backes, Mathias Humbert, and Yang Zhang. Graph unlearning. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, pages 499–513, 2022. 5

[8] Min Chen, Weizhuo Gao, Gaoyang Liu, Kai Peng, and Chen Wang. 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 (CVPR)*, pages 7766–7775, 2023. 1, 5, 6, 9, 10

[9] Jiali Cheng, George Dasoulas, Huan He, Chirag Agarwal, and Marinka Zitnik. Gnndelete: A general strategy for unlearning in graph neural networks. In *International Conference on Learning Representations (ICLR)*, 2023. 5

[10] Vikram S Chundawat, Ayush K Tarun, Murari Mandal, and Mohan Kankanhalli. Zero-shot machine unlearning. *IEEE Transactions on Information Forensics and Security*, 2023. 5

[11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 5

[12] Yuxuan Ding, Lingqiao Liu, Chunna Tian, Jingyuan Yang, and Haoxuan Ding. Don't stop learning: Towards continual learning for the clip model. *arXiv preprint arXiv:2207.09248*, 2022. 6

[13] Chongyu Fan, Jiancheng Liu, Alfred Hero, and Sijia Liu. Challenging forgets: Unveiling the worst-case forget sets in machine unlearning. In *European Conference on Computer Vision (ECCV)*, 2024. 5

[14] Chongyu Fan, Jiancheng Liu, Yihua Zhang, Dennis Wei, Eric Wong, and Sijia Liu. Salun: Empowering machine unlearning via gradient-based weight saliency in both image classification and generation. In *International Conference on Learning Representations (ICLR)*, 2024. 1, 5, 6, 7, 9, 10, 12

[15] Jack Foster, Stefan Schoepf, and Alexandra Brintrup. Fast machine unlearning without retraining through selective synaptic dampening. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 12043–12051, 2024. 5

[16] Rohit Gandikota, Joanna Materzynska, Jaden Fiotto-Kaufman, and David Bau. Erasing concepts from diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2426–2436, 2023. 1, 5, 7

[17] Rohit Gandikota, Hadas Orgad, Yonatan Belinkov, Joanna Materzyńska, and David Bau. Unified concept editing in diffusion models. *arXiv preprint arXiv:2308.14761*, 2023. 5

[18] Shashwat Goel, Ameya Prabhu, Amartya Sanyal, Ser-Nam Lim, Philip Torr, and Ponnurangam Kumaraguru. Towards adversarial evaluations for inexact machine unlearning. *arXiv preprint arXiv:2201.06640*, 2022. 5

[19] Aditya Golatkar, Alessandro Achille, and Stefano Soatto. Eternal sunshine of the spotless net: Selective forgetting in deep networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9301–9309, 2020. 1, 5

[20] Aditya Golatkar, Alessandro Achille, and Stefano Soatto. Forgetting outside the box: Scrubbing deep networks of information accessible from input-output observations. In *European Conference on Computer Vision (ECCV)*, pages 383–398. Springer, 2020.

[21] Aditya Golatkar, Alessandro Achille, Avinash Ravichandran, Marzia Polito, and Stefano Soatto. Mixed-privacy forgetting in deep networks. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 792–801, 2021. 1, 5

[22] Eric Goldman. An introduction to the california consumer privacy act (ccpa). *Santa Clara Univ. Legal Studies Research Paper*, 2020. 1

[23] Chuan Guo, Tom Goldstein, Awni Hannun, and Laurens Van Der Maaten. Certified data removal from machine learning models. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pages 3832–3842. PMLR, 2020. 1

[24] Anisa Halimi, Swanand Kadhe, Ambrish Rawat, and Nathalie Baracaldo. Federated unlearning: How to efficiently erase a client in fl? *arXiv preprint arXiv:2207.05521*, 2022. 5

[25] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 770–778, 2016. 2

[26] Alvin Heng and Harold Soh. Continual learning for forgetting in deep generative models. In *International Conference on Machine Learning workshop*, 2023. 5

[27] Alvin Heng and Harold Soh. Selective amnesia: A continual learning approach to forgetting in deep generative models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 1, 5, 7

[28] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems (NeurIPS)*, 30, 2017. 6

[29] Jeremy Howard and Sylvain Gugger. Fastai: A layered api for deep learning. *Information*, 11(2):108, 2020. 5

[30] Jinghan Jia, Jiancheng Liu, Parikshit Ram, Yuguang Yao, Gaowen Liu, Yang Liu, Pranay Sharma, and Sijia Liu. Model sparsification can simplify machine unlearning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 5, 6, 7, 9, 10, 12

[31] Myeongseob Ko, Henry Li, Zhun Wang, Jonathan Patsenker, Jiachen Tianhao Wang, Qinbin Li, Ming Jin, Dawn Song, and Ruoxi Jia. Boosting alignment for post-unlearning text-to-image generative models. *Advances in Neural Information Processing Systems (NeurIPS)*, 37:85131–85154, 2024. 5, 7

[32] Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *International conference on machine learning (ICML)*, pages 1885–1894. PMLR, 2017. 5, 6, 9, 10

[33] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. In *Toronto, ON, Canada*, 2009. 3, 5

[34] Nupur Kumari, Bingliang Zhang, Sheng-Yu Wang, Eli Shechtman, Richard Zhang, and Jun-Yan Zhu. Ablating concepts in text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22691–22702, 2023. 5

[35] Meghdad Kurmanji, Peter Triantafillou, Jamie Hayes, and Eleni Triantafillou. Towards unbounded machine unlearning. *Advances in neural information processing systems (NeurIPS)*, 36:1957–1987, 2023. 5

[36] Shen Lin, Xiaoyu Zhang, Willy Susilo, Xiaofeng Chen, and Jun Liu. Gdr-gma: Machine unlearning via direction-rectified and magnitude-adjusted gradients. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 9087–9095, 2024. 5

[37] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 5

[38] Bo Liu, Xingchao Liu, Xiaojie Jin, Peter Stone, and Qiang Liu. Conflict-averse gradient descent for multi-task learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 34:18878–18890, 2021. 2, 5

[39] Bo Liu, Yihao Feng, Peter Stone, and Qiang Liu. Famo: Fast adaptive multitask optimization. *Advances in Neural Information Processing Systems (NeurIPS)*, 36:57226–57243, 2023. 2, 5

[40] Gaoyang Liu, Xiaoqiang Ma, Yang Yang, Chen Wang, and Jiangchuan Liu. Federaser: Enabling efficient client-level data removal from federated learning models. In *2021 IEEE/ACM 29th International Symposium on Quality of Service (IWQOS)*, pages 1–10, 2021. 5

[41] Yi Liu, Lei Xu, Xingliang Yuan, Cong Wang, and Bo Li. The right to be forgotten in federated learning: An efficient realization with rapid retraining. In *IEEE INFOCOM 2022 - IEEE Conference on Computer Communications*, pages 1749–1758, 2022. 5

[42] Mengyao Lyu, Yuhong Yang, Haiwen Hong, Hui Chen, Xuan Jin, Yuan He, Hui Xue, Jungong Han, and Guiguang Ding. One-dimensional adapter to rule them all: Concepts diffusion models and erasing applications. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7559–7568, 2024. 5

[43] Ronak Mehta, Sourav Pal, Vikas Singh, and Sathya N. Ravi. Deep unlearning via randomized conditionally independent hessians. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10412–10421, 2022. 5

[44] Dongbin Na, Sangwoo Ji, and Jong Kim. Unrestricted black-box adversarial attack using gan with limited queries. In *European Conference on Computer Vision (ECCV)*, pages 467–482. Springer, 2022. 5

[45] John Nash. Two-person cooperative games. *Econometrica: Journal of the Econometric Society*, pages 128–140, 1953. 2

[46] Aviv Navon, Aviv Shamsian, Idan Achituve, Haggai Maron, Kenji Kawaguchi, Gal Chechik, and Ethan Fetaya. Multi-task learning as a bargaining game. In *International Conference on Machine Learning (ICML)*, 2022. 2, 3, 4, 5

[47] Seth Neel, Aaron Roth, and Saeed Sharifi-Malvajerdi. Descent-to-delete: Gradient-based methods for machine unlearning. In *Proceedings of the 32nd International Conference on Algorithmic Learning Theory*, pages 931–962. PMLR, 2021. 5

[48] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Baolin Wu, Andrew Y Ng, et al. Reading digits in natural images with unsupervised feature learning. In *NIPS workshop on deep learning and unsupervised feature learning*, page 4. Granada, 2011. 5

[49] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition (CVPR)*, pages 3498–3505. IEEE, 2012. 5

[50] Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. On the difficulty of training recurrent neural networks. In *International conference on machine learning (ICML)*, pages 1310–1318. Pmlr, 2013. 6

[51] Martin Pawelczyk, Seth Neel, and Himabindu Lakkaraju. In-context unlearning: Language models as few-shot unlearners. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, pages 40034–40050. PMLR, 2024. 5

[52] Samuele Poppi, Tobia Poppi, Federico Cocchi, Marcella Cornia, Lorenzo Baraldi, Rita Cucchiara, et al. Safe-clip: Removing nsfw concepts from vision-and-language models. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024. 5

[53] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning (ICML)*, pages 8748–8763. PMLR, 2021. 2, 5

[54] Jie Ren, Kangrui Chen, Yingqian Cui, Shenglai Zeng, Hui Liu, Yue Xing, Jiliang Tang, and Lingjuan Lyu. Six-cd: Benchmarking concept removals for benign text-to-image diffusion models. *arXiv preprint arXiv:2406.14855*, 2024. 5

[55] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 10684–10695, 2022. 2, 5

[56] Abhishek Roy, Geelon So, and Yi-An Ma. Optimization on pareto sets: On a theory of multi-objective optimization. *arXiv preprint arXiv:2308.02145*, 2023. 2

[57] Patrick Schramowski, Manuel Brack, Björn Deiseroth, and Kristian Kersting. Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22522–22531, 2023. 5, 7

[58] Ayush Sekhari, Jayadev Acharya, Gautam Kamath, and Ananda Theertha Suresh. Remember what you want to forget: Algorithms for machine unlearning. *Advances in Neural Information Processing Systems (NeurIPS)*, 34:18075–18086, 2021. 5

[59] Ozan Sener and Vladlen Koltun. Multi-task learning as multi-objective optimization. *Advances in neural information processing systems (NeurIPS)*, 31, 2018. 5

[60] Dmitry Senushkin, Nikolay Patakin, Arseny Kuznetsov, and Anton Konushin. Independent component alignment for multi-task learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20083–20093, 2023. 2, 5

[61] Juwon Seo, Sung-Hoon Lee, Tae-Young Lee, Seungjun Moon, and Gyeong-Moon Park. Generative unlearning for any identity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9151–9161, 2024. 5

[62] Christoforos N Spartalis, Theodoros Semertzidis, Efstratios Gavves, and Petros Daras. Lotus: Large-scale machine unlearning with a taste of uncertainty. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pages 10046–10055, 2025. 5

[63] Ayush Kumar Tarun, Vikram Singh Chundawat, Murari Mandal, and Mohan Kankanhalli. Deep regression unlearning. In *International Conference on Machine Learning (ICML)*, pages 33921–33939, 2023. 1, 5

[64] Ayush K Tarun, Vikram S Chundawat, Murari Mandal, and Mohan Kankanhalli. Fast yet effective machine unlearning. *IEEE Transactions on Neural Networks and Learning Systems*, 2023. 1, 5

[65] William Thomson. Cooperative models of bargaining. *Handbook of game theory with economic applications*, 2:1237–1284, 1994. 2

[66] Anvith Thudi, Gabriel Deza, Varun Chandrasekaran, and Nicolas Papernot. Unrolling sgd: Understanding factors influencing machine unlearning. In *2022 IEEE 7th European Symposium on Security and Privacy (EuroS&P)*, pages 303–319. IEEE, 2022. 1, 5, 6, 7, 9, 10, 12

[67] Anvith Thudi, Hengrui Jia, Ilia Shumailov, and Nicolas Papernot. On the necessity of auditable algorithmic definitions for machine unlearning. In *31st USENIX Security Symposium (USENIX Security 22)*, pages 4007–4022, 2022. 1

[68] Paul Voigt and Axel Von dem Bussche. The eu general data protection regulation (gdpr). *A Practical Guide, 1st Ed., Cham: Springer International Publishing*, 10(3152676):10–5555, 2017. 1

[69] Junxiao Wang, Song Guo, Xin Xie, and Heng Qi. Federated unlearning via class-discriminative pruning. In *Proceedings of the ACM Web Conference 2022*, pages 622–632, 2022. 5

[70] Alexander Warnecke, Lukas Pirch, Christian Wressnegger, and Konrad Rieck. Machine unlearning of features and labels. *arXiv preprint arXiv:2108.11577*, 2021. 5, 6, 7, 9, 10, 12

[71] Chen Wu, Sencun Zhu, and Prasenjit Mitra. Federated unlearning with knowledge distillation. *arXiv preprint arXiv:2201.09441*, 2022. 5

[72] Ga Wu, Masoud Hashemi, and Christopher Srinivasa. Puma: Performance unchanged model augmentation for training data removal. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8675–8682, 2022. 5

[73] Jing Wu and Mehrtash Harandi. Scissorhands: Scrub data influence via connection sensitivity in networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024. 5, 6, 7, 9, 10, 12

[74] Mao Ye and Qiang Liu. Pareto navigation gradient descent: a first-order algorithm for optimization in pareto set. In *Uncertainty in artificial intelligence*, pages 2246–2255. PMLR, 2022. 2

[75] Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:5824–5836, 2020. 2, 5

[76] Yi Zeng, Xuelin Yang, Li Chen, Cristian Canton Ferrer, Ming Jin, Michael I Jordan, and Ruoxi Jia. Fairness-aware meta-learning via nash bargaining. *arXiv preprint arXiv:2406.07029*, 2024. 3, 4, 9

[77] Yuexiang Zhai, Shengbang Tong, Xiao Li, Mu Cai, Qing Qu, Yong Jae Lee, and Yi Ma. Investigating the catastrophic forgetting in multimodal large language models. *Advances in neural information processing systems workshop*, 2023. 6

[78] Binchi Zhang, Zihan Chen, Cong Shen, and Jundong Li. Verification of machine unlearning is fragile. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, pages 58717–58738. PMLR, 2024. 5

[79] Eric Zhang, Kai Wang, Xingqian Xu, Zhangyang Wang, and Humphrey Shi. Forget-me-not: Learning to forget in text-to-image diffusion models. *arXiv preprint arXiv:2303.17591*, 2023. 1, 5

[80] Yimeng Zhang, Xin Chen, Jinghan Jia, Yihua Zhang, Chongyu Fan, Jiancheng Liu, Mingyi Hong, Ke Ding, and Sijia Liu. Defensive unlearning with adversarial training for robust concept erasure in diffusion models. *arXiv preprint arXiv:2405.15234*, 2024. 5, 7

[81] Yihua Zhang, Chongyu Fan, Yimeng Zhang, Yuguang Yao, Jinghan Jia, Jiancheng Liu, Gaoyuan Zhang, Gaowen Liu, Ramana Rao Kompella, Xiaoming Liu, and Sijia Liu. Unlearncanvas: A stylized image dataset to benchmark machine unlearning for diffusion models. *arXiv preprint arXiv:2402.11846*, 2024. 5

[82] Yimeng Zhang, Jinghan Jia, Xin Chen, Aochuan Chen, Yihua Zhang, Jiancheng Liu, Ke Ding, and Sijia Liu. To generate or not? safety-driven unlearned diffusion models are still easy to generate unsafe images... for now. In *European Conference on Computer Vision (ECCV)*, 2024. 5, 8, 7, 15

[83] Yang Zhao, Jiaxi Yang, Yiling Tao, Lixu Wang, Xiaoxiao Li, and Dusit Niyato. A survey of federated unlearning: A taxonomy, challenges and future directions. *arXiv preprint arXiv:2310.19218*, 2023. 5