CONTROLLABLE UNLEARNING FOR IMAGE-TO-IMAGE GENERATIVE MODELS VIA ε-CONSTRAINED OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

While generative models have made significant advancements in recent years, they also raise concerns such as privacy breaches and biases. Machine unlearning has emerged as a viable solution, aiming to remove specific training data, e.g., containing private information and bias, from models. In this paper, we study the machine unlearning problem in Image-to-Image (I2I) generative models. Previous studies mainly treat it as a single objective optimization problem, offering a solitary solution, thereby neglecting the varied user expectations towards the trade-off between complete unlearning and model utility. To address this issue, we propose a controllable unlearning framework that uses a control coefficient ε to control the trade-off. We reformulate the I2I generative model unlearning problem into a ε-constrained optimization problem and solve it with a gradient-based method to find optimal solutions for unlearning boundaries. These boundaries define the valid range for the control coefficient. Within this range, every yielded solution is theoretically guaranteed with Pareto optimality. We also analyze the convergence rate of our framework under various control functions. Extensive experiments on two benchmark datasets across three mainstream I2I models demonstrate the effectiveness of our controllable unlearning framework.

1 INTRODUCTION

Generative models have recently made significant progress in fields such as image recognition (Ho et al., 2020; Dhariwal & Nichol, 2021) and natural language processing (OpenAI, 2023; Touvron et al., 2023), capturing significant academic interest due to their boundless generative potential. Typically trained on vast datasets from the Internet, generative models inevitably assimilate latent biases and expose private information (Schwarz et al., 2021). Existing studies (Kuppa et al., 2021; Tirumala et al., 2022; Carlini et al., 2023) have revealed that generative models have a strong tendency to recall specific instances encountered during training, raising concerns that the models might output biases and leak private information when put into practical situations. Machine unlearning (Nguyen et al., 2022) presents a viable solution to address this issue. It aims to eliminate the knowledge learned from specific training data (forget set) while preserving the knowledge learned from the remaining data (retain set).

Implementing unlearning for generative models serves dual objectives, i.e., fulfilling privacy requirements and enhancing model reliability. On the one hand, legislation such as the General Data Protection Regulation (Voigt & Von dem Bussche, 2017) grants individuals the *right to be forgotten*. Consequently, service providers must unlearn specific private information from the model in response to an individual's request. On the other hand, the data available on the Internet is rife with biases and inaccuracies, which compromises model performance when used for training. By proactively unlearning the biased and inaccurate data, the service providers can improve the liability of their models.

In this paper, we focus on the unlearning problem in Image-to-Image (I2I) generative models (Yang et al., 2023), where unlearning is defined by the model's incapacity to reconstruct the full image from a partially cropped one (Li et al., 2024a), as shown in Figure 1. Previous study (Li et al., 2024a) frames machine unlearning in generative models as a single-objective optimization problem,

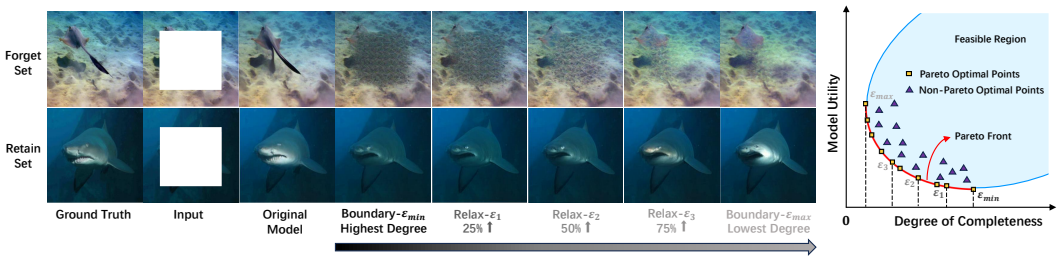

Figure 1: An overview of controllable unlearning. On the left, the first and second rows represent the forget set and the retain set, respectively. We first present the effect of unlearning in I2I generative models, followed by a collection of controllable solutions, where ε is the control coefficient. On the right, we demonstrate that for each ε, our solution is guaranteed with the Pareto optimality.

with the loss defined as a combination of performance on both the forget and retain sets. However, this approach faces **three main challenges**: i) First and foremost, this approach offers a fixed result, ignoring the real-world need for flexible trade-offs between model utility and unlearning completeness aligned with varying user expectations. Regrettably, this challenge remains overlooked in the majority of current research on unlearning. ii) This approach relies wholly on fine-tuning with manual terminating conditions, lacking a theoretical guarantee for convergence. iii) This approach integrates two optimization objectives into a single loss function, which compromises unlearning efficiency due to the competition or conflict between different objectives.

To address these challenges, we propose a **controllable unlearning** approach that *provides a set of Pareto optimal solutions to cater to varied user expectations*. Users can select a solution based on the degree of unlearning completeness through a simple control coefficient ε. Specifically, we reframe machine unlearning of I2I generative models into a bi-objective optimization problem (Kim & De Weck, 2005), i.e., unlearning the forget set (1st objective, unlearning completeness) while preserving the retain set (2nd objective, model utility). Due to legislation requirements, the first objective prioritizes the second objective, meaning that minimizing the negative impact on the retain set only arises once the unlearning objective is sufficiently optimized. Therefore, we reformulate the bi-objective optimization problem into a ε-constrained optimization problem, where the unlearning objective is treated as a constraint (primary to satisfy) and ε is the control coefficient. Utilizing gradient-based methods to solve this ε-constrained optimization, we can obtain two Pareto optimal solutions for the boundaries of unlearning with theoretical guarantee, which can be used to determine the valid range of values for ε. Subsequently, we select the value of ε within its valid range and relax the constraints on the unlearning objective by increasing ε. As a result, we obtain a set of solutions that dynamically fulfill user's varied expectations regarding the trade-off between unlearning completeness and model utility. Finally, to enhance the efficiency of unlearning, we analyze the convergence rates of our unlearning framework under various settings of the control function which is utilized to govern the direction of parameter updates. The main contributions of this paper are summarized as follows:

- We focus on I2I generative models, and propose a controllable unlearning approach that balances unlearning completeness and model utility, providing a set of solutions to fulfill varied user expectations. To the best of our knowledge, we are the first to study controllable unlearning.
- We reformulate the machine unlearning of generative models as a ε-constrained optimization problem with unlearning the forget set as the constraint, guaranteeing optimal theoretical solutions for the boundaries of unlearning. By progressively relaxing the unlearning constraint, we obtain the Pareto set and plot the corresponding Pareto front.
- We utilize gradient-based methods to solve the ε-constrained optimization problem. To enhance the efficiency of unlearning, we analyze our framework's performance across different settings of the control function and validate with multiple combinations.
- We conduct extensive experiments to evaluate our proposed method over diverse I2I generative models. The results from two large datasets demonstrate that the Pareto optimal solutions yielded by our method significantly outperform baseline methods. Additionally, the solution set achieves controllable unlearning to fulfill varied expectations regarding the trade-off between unlearning completeness and model utility.

2 RELATED WORK

2.1 I2I GENERATIVE MODELS

Many computer vision tasks can be formulated as I2I generation processes, e.g., style transfer (Zhu et al., 2017), image extension (Chang et al., 2022), restoration (Teterwak et al., 2019), and image synthesis (Yu et al., 2020). There are mainly three architectures for I2I generative models, i.e., Auto-Encoders (AEs) (Alain & Bengio, 2014), Generative Adversarial Networks (GANs) (Goodfellow et al., 2014), and diffusion models (Ho et al., 2020). AEs mainly aim to reduce the mean squared error between generated and ground truth images but often produce lower-quality outputs (Dosovitskiy et al., 2021; Esser et al., 2021). GANs, through adversarial training, significantly improve generation quality, despite their unstable training (Arjovsky et al., 2017; Gulrajani et al., 2017; Brock et al., 2019). Diffusion models, which use a diffusion-then-denoising approach, aim for stable training and high-quality generation by minimizing the distributional distance between generated images and ground truth images (Ho et al., 2020; Song & Ermon, 2020; Salimans & Ho, 2022). However, diffusion models require a greater amount of data and computational resources (Saharia et al., 2022b; Rombach et al., 2022). In this paper, we aim to design a universal unlearning method that can be applied across different I2I models.

2.2 MACHINE UNLEARNING

Machine unlearning aims to eliminate the influence of specific training data (unlearning target) from a trained model. A naive approach is to retrain the model from scratch using a modified dataset that excludes the unlearning target. However, this approach can be computationally prohibitive in practice. Based on the degree of unlearning completeness, machine unlearning can be categorized into exact unlearning and approximate unlearning (Xu et al., 2023).

Exact unlearning aims to ensure that the unlearning target is fully unlearned, i.e., as complete as retraining from scratch (Bourtoule et al., 2021; Yan et al., 2022; Li et al., 2024b). This approach, which typically relies on retraining, is limited to unlearning specific instances and cannot be readily extended to generative models with strong feature generalizations. **Approximate unlearning** aims to obtain an approximate model, whose performance closely aligns with a retrained model (Golatkar et al., 2020; Sekhari et al., 2021). This approach estimates the influence of unlearning targets, and updates the model accordingly, usually through gradient-based updates, avoiding full retraining (Basu et al., 2021; Li et al., 2023b). However, accurate influence estimation is still challenging (Graves et al., 2021), reducing the applicability of this approach to generative models.

In text-to-image generative models, unlearning typically refers to concept erasure Gandikota et al. (2023); Pham et al. (2023); Lu et al. (2024), such as removing an abstract concept like the artistic style represented by "Van Gogh" or the entity concept represented by "Musk". In contrast, in I2I generative models, the objective of unlearning is to remove the knowledge the model has acquired from a specific set of samples. In this paper, We focus on I2I generative models. The exploration of unlearning is accomplished by minimizing a composite loss, which is a combination of training loss on the retain and the forget sets (Li et al., 2024a). This approach is highly dependent on manual parameter tuning and cannot guarantee unlearning completeness. As for comparison, the solutions yielded by our proposed controllable unlearning framework are theoretically guaranteed with Pareto optimality.

3 PRELIMINARY

3.1 UNLEARNING PRINCIPLES

As outlined in (Chen et al., 2022; Li et al., 2024c), an unlearning task typically has three main principles: i) unlearning completeness, which involves eliminating the influence of specific data from an already trained model; ii) unlearning efficiency, which focuses on enhancing the speed of the unlearning process; and iii) model utility, which aims to ensure that the performance of the unlearned model remains comparable to that of a model retrained from scratch.

3.2 PARETO OPTIMALITY

Consider a multi-objective optimization problem formulated as: $\min_\theta f(\theta) = (f_1(\theta), f_2(\theta), \cdots, f_m(\theta))^\top$, where $f_i(\theta)$ denotes the loss for the i-th objective.

Pareto dominance. Let θ^a, θ^b be two points in feasible set Ω, θ^a is said to dominate θ^b $\left(\theta^a \prec \theta^b\right)$ if and only if $f_i\left(\theta^a\right) \leq f_i\left(\theta^b\right), \forall i \in \{1, \ldots, m\}$ and $f_j\left(\theta^a\right) < f_j\left(\theta^b\right), \exists j \in \{1, \ldots, m\}$.

Pareto optimality (Lin et al., 2019). A point θ^* is Pareto optimal if there is no $\hat{\theta} \in \Omega$ for which $\hat{\theta} \prec \theta^*$. The collection of all such Pareto optimal points forms the Pareto set, and the surface of this set in the loss space is called the Pareto front.

3.3 I2I GENERATIVE MODEL UNLEARNING

Model architecture. Encoder-decoder structures are widely used in I2I models, with: i) an encoder E_γ reducing images to the latent space, and ii) a decoder D_ϕ reconstructing images from the latent space. For model I_θ with input image x, the output is:

$$I_\theta(x) = D_\phi(E_\gamma(x)), \tag{1}$$

where $\mathcal{T}(x)$ denotes the cropping operation (such as center cropping or random cropping), and $\theta = \{\gamma, \phi\}$ denotes the full parameter set.

Unlearning objective. Define the unlearning task for an I2I generative model I_{θ_0} involving data partitions D_f (forget set) and D_r (retain set). Consider an I_{θ_0}, i.e., the original model, with training data $D = D_f \cup D_r$. Assume that I_{θ_0} is proficiently trained to generate satisfactory results on both D_f and D_r. The objective of unlearning is to obtain an unlearned model I_θ that cannot generate satisfactory results on D_f (1st objective, unlearning completeness) while maintaining comparable performance on D_R (2nd objective, model utility). Formally,

$$\max_\theta \Big(Div\big(\mathbb{P}(X_f) \| \mathbb{P}(I_\theta(\mathcal{T}(X_f)))\big)\Big), \text{ and } \min_\theta \Big(Div\big(\mathbb{P}(X_r) \| \mathbb{P}(I_\theta(\mathcal{T}(X_r)))\big)\Big), \tag{2}$$

where X_f and X_r are the variables for ground truth images in D_f and D_r, $\mathbb{P}(I_\theta(X))$ is the model output distribution for input variable X, and $Div(\cdot||\cdot)$ represents distributional distance, measured by Kullback-Leibler (KL) divergence in this paper.

Following prior work (Kingma et al., 2019; Xia et al., 2022; Wallace et al., 2023), as the model is proficiently trained, we hypothesize that I_{θ_0} can approximately replicate the distributions over both forget and retain sets (Kingma et al., 2019; Xia et al., 2022; Wallace et al., 2023), i.e., $\mathbb{P}(X_f) \approx \mathbb{P}(I_{\theta_0}(\mathcal{T}(X_f)))$, and $\mathbb{P}(X_r) \approx \mathbb{P}(I_{\theta_0}(\mathcal{T}(X_r)))$. Let $\mathbb{P}_X := \mathbb{P}(I_{\theta_0}(\mathcal{T}(X)))$ and $\mathbb{P}_{\hat{X}} := \mathbb{P}(I_\theta(\mathcal{T}(X)))$. Then, Eq. (2) can be simplified to:

$$\max_\theta Div(\mathbb{P}_{X_f} || \mathbb{P}_{\hat{X}_f}), \text{ and } \min_\theta Div(\mathbb{P}_{X_r} || \mathbb{P}_{\hat{X}_r}), \tag{3}$$

where $\mathbb{P}_{X_f}$ and $\mathbb{P}_{\hat{X}_f}$ represent the output distributions of the forget set before and after unlearning respectively. Similarly, $\mathbb{P}_{X_r}$ and $\mathbb{P}_{\hat{X}_r}$ represent those for the retain set.

4 METHODOLOGY

In this section, we first introduce a controllable unlearning framework for I2I generative models, which formulates unlearning as a constrained optimization with the unlearning objective as a constraint. We utilize a gradient-based method to obtain the boundaries of unlearning. Then we relax the constraint within the boundaries to derive a set of Pareto optimal solutions to fulfill varied user expectations.

4.1 ε-CONSTRAINED OPTIMIZATION FORMULATION

The unlearning task for I2I models is reformulated as a bi-objective optimization (Eq. (3)), with the first objective to maximize $Div(\mathbb{P}_{X_f} || \mathbb{P}_{\hat{X}f})$. Nonetheless, the value of $Div(\cdot||\cdot)$ can theoretically be maximized to infinity, yielding an infinite number of possible $\mathbb{P}_{\hat{X}f}$ (Li et al., 2024a), consequently resulting in extremely diminished model utility. To balance unlearning completeness and model utility, we bound $Div(\mathbb{P}_{X_f} || \mathbb{P}_{\hat{X}_f})$ by Lemma 1.

Lemma 1 (Divergence Upper Bound (Cover & Thomas, 2012))**.** *Assuming the forget set with distribution $\mathbb{P}_{X_f}$ characterized by a zero-mean and covariance matrix Σ, and a signal $\mathbb{P}_{\hat{X}_f}$ with the same statistical properties, the maximal KL divergence is realized when $\mathbb{P}_{\hat{X}_f} = \mathcal{N}(0, \Sigma)$.*

$$Div(\mathbb{P}_{X_f}||\mathbb{P}_{\hat{X}_f}) \leq Div(\mathbb{P}_{X_f}||\mathcal{N}(0, \Sigma)). \tag{4}$$

As image normalization typically involves mean subtraction (Elasri et al., 2022), we can assume $\mathbb{P}_{X_f}$ and $\mathbb{P}_{\hat{X}_f}$ follow zero-mean distributions for conciseness without sacrificing generality. Lemma 1 reveals that the upper bound of $Div(\mathbb{P}_{X_f}||\mathbb{P}_{\hat{X}f})$ is achieved when $\mathbb{P}_{\hat{X}_f} \sim \mathcal{N}(0, \Sigma)$. This suggests that maximizing $Div(\mathbb{P}_{X_f}||\mathbb{P}_{\hat{X}f})$ equates to minimizing $Div(\mathbb{P}_{\hat{X}_f}||\mathcal{N}(0, \Sigma))$. Consequently, we rewrite Eq. (3) as:

$$\min_{\theta} Div(\mathcal{N}(0, \Sigma)||\mathbb{P}_{\hat{X}_f}), \text{ and } \min_{\theta} Div(\mathbb{P}_{X_r}||\mathbb{P}_{\hat{X}_r}). \tag{5}$$

As both terms in Eq. (5) depend on θ, we define $f_1(\theta) := Div(\mathcal{N}(0, \Sigma)||\mathbb{P}_{\hat{X}f})$ and $f_2(\theta) := Div(\mathbb{P}_{X_r}||\mathbb{P}_{\hat{X}_r})$ for conciseness. However, unlike classification models where their outputs are precisely univariate discrete distributions (Kurmanji et al., 2024; Zhang et al., 2023), high-dimensional KL divergence calculations in I2I generative models are intractable. Thus, following (Li et al., 2024a), we adopt the L_2 loss as a surrogate. Due to privacy legal requirements, unlearning objectives typically takes precedence. Thus, we set $f_1(\theta)$ as the primary constraint and treat Eq. (5) as a ε-constrained optimization problem:

$$\min_{\theta \in \mathbb{R}^d} f_2(\theta) \quad \text{s.t.} \quad f_1(\theta) \leq \varepsilon, \tag{6}$$

where ε is a parameter to control the completeness of unlearning. We minimize $f_2(\theta)$ inside the feasible set $\Omega = \{\theta : f_1(\theta) \leq \varepsilon\}$, which implies that our priority lies in unlearning the forget set rather than mitigating performance degradation on the retain set.

4.2 Solving the ε-Constraint Optimization

To solve the ε-constrained optimization problem in Eq. (6), approaches such as Sequential Quadratic Programming (SQP) (Nocedal & Wright; Bonnans et al., 2006), penalty function method (Yeniay, 2005), and interior point method (Renegar, 2001) are commonly employed. Given the extensive parameter set of the I2I generative model, we select a special variant of the SQP algorithm for its lower complexity and comparable convergence guarantee (Nocedal & Wright; Gill & Wong, 2011).

Specifically, we employ a gradient-based method to solve Eq. (6), updating the parameter by $\theta_{t+1} \leftarrow \theta_t - \mu_t g_t$. Here, $\mu_t > 0$ denotes the step size, and g_t represents the direction of the parameter update, which is determined by solving a convex quadratic programming problem w.r.t. g (for a detailed derivation, please refer to the Appendix C.1):

$$g_t = \min_{g \in \mathbb{R}^d} \Big\{ \|\nabla f_2(\theta_t) - g\|^2 \quad \text{s.t.} \quad \nabla f_1(\theta_t)^\top g \geq f_1(\theta_t) - \varepsilon \Big\}. \tag{7}$$

Due to the inability to obtain the effective range of ε in the early stages of unlearning, direct computation of $f_1(\theta_t) - \varepsilon$ is not feasible. Consequently, we adjust the constraint of Eq. (7) by employing a control function $\psi(\theta_t)$ (i.e., $\nabla f_1(\theta_t)^\top g \geq \psi(\theta_t)$), which should satisfy $sign(\psi(\theta_t)) = sign(f_1(\theta_t) - \varepsilon)$, where $sign(x) = x/|x|$ for $x \neq 0$ and $sign(0) = 0$. This ensures that the direction of updates remains as consistent as possible before and after the substitution. Further, we provide a summary of our proposed unlearning algorithm in Algorithm 1.

Assumption 1. *Assume $f_1(\theta)$ and $f_2(\theta)$ are continuously differentiable, and the trajectory $\{\theta_t : t \in [0, +\infty)\}$ follows the continuous-time dynamics $\dot{\theta}_t = -g_t$, where g_t is defined in Eq. (7) and $\max_{t \in [0, +\infty)} \eta_t < +\infty$.*

The convergence analysis of Algorithm 1 regarding Eq. (6) utilizes the continuous-time framework given by $\dot{\theta}_t = -g_t$, as mentioned in Assumption 1. Please refer to Theorem 2 in Appendix C.2 for further details of convergence.

Algorithm 1 Gradient-based Optimization Method

Require: Original model I_{θ_0}, forget set D_f, retain set D_r, control function $\psi(\theta)$, step size μ, covariance matrix Σ, numerical stability variable $\varpi = 1e-7$.

Initial: Initialize $t = 0$, $I_{\theta_t} = I_{\theta_0}$;
for $t = 0$ to $T-1$ **do**
  Sample $\{x_f\}$, $\{x_r\}$ and $\{x_n\}$ from D_f, D_r and $\mathcal{N}(0, \varepsilon)$ respectively, ensuring that $|\{x_f\}| = |\{x_r\}| = |\{x_n\}|$;
  Compute loss:
    $f_1(\theta_t) = \|I_{\theta_t}(\mathcal{T}(D_f)) - I_{\theta_0}(\mathcal{T}(x_n))\|_2$
    $f_2(\theta_t) = \|I_{\theta_t}(\mathcal{T}(D_r)) - I_{\theta_0}(\mathcal{T}(D_r))\|_2$
  Compute gradient: $\nabla f_1(\theta_t)$, $\nabla f_2(\theta_t)$;
  Compute the solution to the dual problem of Eq. (7): $\eta_t = \max\left(\frac{\psi(\theta_t) - \nabla f_2(\theta_t)^\top \nabla f_1(\theta_t)}{\|\nabla f_1(\theta_t)\|^2 + \varpi}, 0\right)$;
  Compute parameter update direction: $g_t = \nabla f_2(\theta_t) + \eta_t \nabla f_1(\theta_t)$;
  Update the parameter of the target model $I_{\theta_{t+1}}$: $\theta_{t+1} \leftarrow \theta_t - \mu_t g_t$;
end for
Return Unlearned model I_{θ_T};

4.3 A Controllable Unlearning Framework

Our controllable unlearning framework consists of two phases. In Phase I, we reformulate Eq. (6) into a special form to obtain the solution for the boundaries of unlearning. In Phase II, we adjust the value ε within its valid range to relax the unlearning constraint and obtain the Pareto optimal solutions for controllable unlearning. This relaxation of unlearning completeness allows for a controllable trade-off between completeness and model utility, thereby catering to varied user expectations.

Phase I: Boundaries of unlearning. The boundaries of unlearning refer to the two Pareto optimal solutions with the highest and lowest degrees of unlearning completeness.

To obtain the Pareto optimal solutions with the highest degrees of unlearning completeness, we reformulate Eq. (6)into the following special form:

$$\min_{\theta \in \mathbb{R}^d} f_2(\theta) \quad \text{s.t.} \quad f_1(\theta) \leq \varepsilon, \\ \text{where} \quad \varepsilon = f_1^*, \text{ and } f_1^* := \inf_{\theta \in \mathbb{R}^d} f_1(\theta). \tag{8}$$

The solution of this optimization problem can be obtained by Algorithm 1. According to Assumption 1, we need to ensure that $\psi(\theta) \geq 0$ in Eq. (8) to guarantee the same sign with $f_1(\theta) - \varepsilon$. In this paper, we we simply define $\psi(\theta) = \alpha \|\nabla f_1(\theta)\|^\delta$ with $\alpha > 0$ and $\delta \geq 1$.

Proposition 1 (Boundary of Pareto Set)**.** *Under Assumption 1, let $f_1^* > -\infty$ and $f_2^* > -\infty$ be the infimum of $f_1(\theta), f_2(\theta)$, respectively. Further, let $\psi(\theta)$ be continuous and $\nabla f_1(\theta)$ be continuously differentiable. If $\theta_t \to \theta^*$ and $g_t \to 0$ as $t \to +\infty$, with $\nabla^2 f_1(\theta)$ of constant rank near θ^* and $f_1(\theta), f_2(\theta)$ being convex near θ_t, then θ^* is a Pareto optimal solution and $f_1(\theta^*) = f_1^*$.*

Proof. The proof can be found in Appendix C.3. □

Proposition 1 ensures that the solution θ_1^* obtained by Algorithm 1 for solving Eq. (8) is on the boundary of the Pareto set, specifically refer to the highest degree of unlearning completeness. Meanwhile, $f_1(\theta_1^*)$ achieve the infimum of $f_1(\theta)$.

Obtaining the Pareto optimal solution with the lowest unlearning completeness is similar to the process mentioned above, with the difference of exchanging the positions of $f_1(\theta)$ and $f_2(\theta)$ in Eq. (8). This new problem is formulated as $\min_{\theta \in \mathbb{R}^d} f_1(\theta)$, s.t. $f_2(\theta) \leq \varepsilon$, where $\varepsilon = f_2^*$, and $f_2^* := \inf_{\theta \in \mathbb{R}^d} f_2(\theta)$. The solution θ_2^* obtained by solving this problem is another boundary the Pareto set, i.e., the Pareto optimal solution with the lowest unlearning completeness, with $f_2(\theta_2^*)$ achieving the infimum of $f_2(\theta)$.

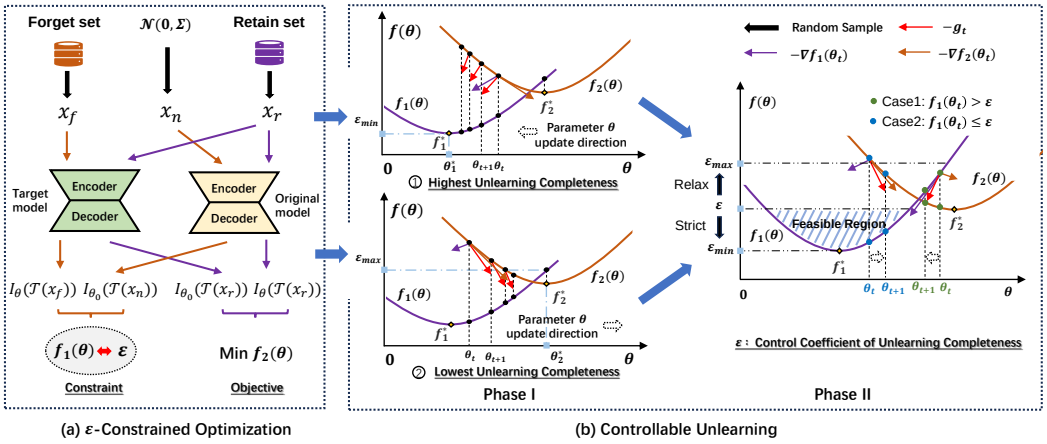

Figure 2: Pipeline of the controllable unlearning framework. (a) shows the unlearning task of the I2I generative model which is framed as a ε-constrained optimization problem. (b) shows that the implementation of controllable unlearning unfolds in two phases: i) initially identifying two boundary points of unlearning, necessitating a strict reduction in $f_1(\theta)$ (or $f_2(\theta)$) for optimality; and ii) then locating the given ε's Pareto optimal point, with strict reduction in $f_1(\theta)$ when $f_1(\theta_t) > \varepsilon$ and permitting an increase when $f_1(\theta_t) \leq \varepsilon$.

Phase II: Controllable unlearning. To adjust the trade-off between unlearning completeness and model utility, we relax the unlearning constraint by defining $f_1(\theta_1^*) < \varepsilon < f_1(\theta_2^*)$ in Eq. (6), where θ_1^* and θ_2^* have already been obtained in Phase I. Then we rewrite Eq. (8) for controllable unlearning:

$$\min_{\theta \in \mathbb{R}^d} f_2(\theta) \quad \text{s.t.} \quad f_1(\theta) \leq \varepsilon,$$
$$\text{where} \quad \varepsilon > f_1^*, \text{ and } f_1^* := \inf_{\theta \in \mathbb{R}^d} f_1(\theta), \tag{9}$$

where $\varepsilon \in \mathbb{R}$ is used to adjust the completeness of unlearning. In Phase II, according to the sign condition in Assumption 1, we simply set $\psi(\theta) = \beta(f_1(\theta) - \varepsilon)^\delta$ with $\beta > 0$, $\delta = 2n+1$ and $n \in \mathbb{N}$.

Proposition 2 (Interior of Paret Set)**.** *Under Assumption 1, let $f_2^* = \inf_{\theta \in \mathbb{R}^d} f_2(\theta) > -\infty$ and $\sup_{t \in [0,+\infty)} \eta_t = \eta_{max} < +\infty$. If θ_t is a stationary point with $g_t = 0$ and $\eta_t < +\infty$, and both $f_1(\theta)$ and $f_2(\theta)$ are convex at θ_t, then θ_t is a Pareto optimal solution w.r.t. ε.*

Proof. The proof can be found in Appendix C.4. □

From Proposition 2, Eq. (9) provides a Pareto optimal solution w.r.t. ε. By progressively increasing ε from f_1^*, which is estimated by $f_1(\theta_1^*)$ in Phase I, we can trace a path of Pareto optimal solutions for different completeness of unlearning. As a result, this path offers controllable unlearning for varied user expectations.

4.4 Enhancing the Efficiency of Unlearning

To enhance the efficiency of unlearning, we investigate the influence of the control function $\psi(\theta)$ on convergence rates across different phases, as outlined in the proposition below:

Proposition 3. *Under Assumption 1, with $f_2^* = \inf_{\theta \in \mathbb{R}^d} f_2(\theta) > -\infty$, then:*

1. *For Phase I, if $\psi(\theta) = \alpha \|\nabla f_1(\theta)\|^\delta$ with $\alpha > 0$ and $\delta \geq 1$, the convergence rates of $f_1(\theta)$ and $f_2(\theta)$ are $O\left(1/t^{\frac{1}{\delta}}\right)$ and $O\left(1/t^{\frac{1}{2}-\frac{1}{2\delta}}\right)$, respectively.*
2. *For Phase II, if $\psi(\theta) = \beta(f_1(\theta) - \varepsilon)^\delta$ with $\beta > 0$, $\delta = 2n+1$, $n \in \mathbb{N}$, and $\sup_{t \in [0,+\infty)} \eta_t = \eta_{max} < +\infty$, the convergence rate of $[f_1(\theta) - \varepsilon]_+$ is $O\left(1/t^{\frac{1}{\delta}}\right)$.*

Proof. The proof can be found in Appendix C.5. □

Proposition 3 demonstrates that the convergence rate depends on the exponent δ in $\psi(\theta)$, where higher values of δ result in a faster convergence rate of $f_1(\theta)$. However, excessively large δ can also lead to a slower convergence rate of $f_2(\theta)$ and instabilities in training. To balance convergence rate and training stability, we explore various ε in $\psi(\theta)$ in both phases with extensive empirical studies. The results can be found in Section 5.4.

5 EXPERIMENTS

5.1 EXPERIMENTAL SETTINGS

We evaluate our proposed method on three mainstream I2I generative models, i.e., Masked Autoencoder (MAE) (He et al., 2022), Vector Quantized Generative Adversarial Networks (VQ-GAN) (Li et al., 2023a), and diffusion probabilistic models (Saharia et al., 2022a). Please refer to Appendix D.1 for the settings of hyperparameters.

Datasets: Following (Li et al., 2024a), we conduct experiments on the following two large-scale datasets: i) ImageNet-1K (Deng et al., 2009), from which we randomly select 200 classes, designating 100 of these as the forget set and the remaining 100 as the retain set. Each class contains 150 images, with 100 allocated for training and the remaining for validation; and ii) Places-365 (Zhou et al., 2017), from which we randomly select 100 classes, designating 50 of these as the forget set and the remaining 50 as the retain set. Each class contains 5500 images, with 5000 allocated for training and the remaining 500 for validation.

Baselines: We first report the performance of the original model (i.e., before unlearning) as a reference. Following (Li et al., 2024a), we set the following baselines: i) Max Loss (Warnecke et al., 2023; Gandikota et al., 2023), which maximizes the training loss on the forget set; ii) Retain Label (Kong & Chaudhuri, 2023), which minimizes training loss by setting the true values of the retain samples as those of the forget set; iii) Noisy Label (Graves et al., 2021; Gandikota et al., 2023), which minimizes the training loss by introducing Gaussian noise to the ground truth images of the forget set; and iv) Composite Loss (Li et al., 2024a), the State-Of-The-Art (SOTA) method, which builds upon Noisy Label by calculating the loss on the retain set and obtaining their weighted sum, thereby minimizing this weighted training loss.

Evaluation metrics. We adopt three different types of metrics to comprehensively compare our method with other baselines: i) Inception Score (IS) of the generated images (Salimans et al., 2016); ii) the Frechét Inception Distance (FID) between the generated images and the ground truth images (Heusel et al., 2017); and iii) the cosine similarity between the CLIP embeddings of the generated images and the ground truth images (Radford et al., 2021). IS evaluates the quality of the generated images independently, while the FID further measures the similarity between the generated and ground truth images. On the other hand, the distance of CLIP embeddings assesses whether the generated images still capture similar semantics. Please refer to Appendix D.2 for more information of evaluation metrics.

5.2 UNLEARNING PERFORMANCE

We test our method on image extension, inpainting, and reconstruction tasks. We report the results for center uncropping (i.e., inpainting) in Tabel 1, and the others in Appendix H.1.

Baseline comparison: As shown in Table 1, compared to the original model, our method retains almost the same performance on the retain set or only exhibits minor degradation. Meanwhile, there is a significant reduction in the three metrics on the forget set. In contrast, these baselines generally cannot perform well simultaneously on both the forget set and the retain set. For instance, in MAE, Composite Loss has the least performance degradation on the retain set, but its performance on the forget set is also the worst. We also observe similar findings for Max Loss in VQ-GAN. Furthermore, we provide some examples of generated images in Figure 3, and more images in Appendix F.

T-SNE analysis: Following (Li et al., 2024a), we conduct a T-SNE analysis (Van der Maaten & Hinton, 2008) to further analyze our method's effectiveness. Using our unlearned model, we generate 50 images for both the retain set and the forget set. We then calculate the CLIP embedding vectors for these images and their corresponding ground truth images. As illustrated in Figure 4,

Table 1: Results of center cropping 50% of the images. ‘F’ and ‘R’ stand for the forget set and retain set, respectively. Here, ”Ours” refers to the boundary points of unlearning obtained in Phase I, that is, the solution with the highest degree of unlearning completeness. The best results are highlighted in **bold**, and secondary results are highlighted with underline.

	MAE						VQ-GAN						Diffusion Models					
	IS		FID		CLIP		IS		FID		CLIP		IS		FID		CLIP	
	F ↓	R ↑	F ↑	R ↓	F ↓	R ↑	F ↓	R ↑	F ↑	R ↓	F ↓	R ↑	F ↓	R ↑	F ↑	R ↓	F ↓	R ↑
Original	21.59	21.83	16.28	14.87	0.88	0.88	23.74	24.06	21.80	18.17	0.78	0.85	16.90	19.65	82.12	81.51	0.89	0.91
Max Loss	15.42	16.55	129.54	87.13	0.72	0.72	19.20	21.23	23.52	43.88	0.77	0.75	17.27	18.10	95.93	108.70	0.83	0.79
Retain Label	20.74	14.14	90.62	103.72	0.71	0.73	14.44	19.24	106.01	46.25	0.47	0.75	17.02	**19.08**	86.10	**89.18**	0.87	**0.83**
Noisy Label	15.38	**17.97**	135.47	**63.89**	0.71	**0.77**	15.95	20.63	93.55	47.03	0.49	0.74	17.15	18.36	125.99	121.55	0.72	0.76
Composite Loss	13.96	15.71	149.78	74.14	0.70	0.72	14.34	21.60	103.17	37.92	0.48	0.77	14.33	17.80	149.22	98.82	0.64	0.80
Ours	**12.33**	17.47	**154.60**	68.453	**0.69**	0.75	**13.23**	**22.55**	**139.21**	**26.39**	**0.46**	**0.82**	**11.84**	18.47	**165.05**	95.42	**0.55**	0.81

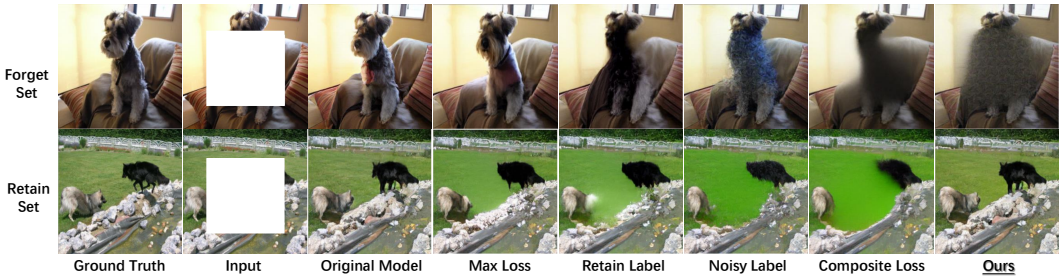

Figure 3: Generated images of cropping 50% at the center of the image on VQ-GAN. From left to right, the images generated by baselines are presented. Our method results in the highest degree of unlearning completeness while maintaining a minimal reduction in model utility.

after unlearning, the embeddings of retain set are close to that of the ground truth images, while most of the generated images on the forget set diverge significantly from the ground truth one.

Unlearning robustness: We validate the performance of our controllable unlearning framework in different image generation tasks by changing the cropping patterns. The results indicate that our framework is robust to various image generation tasks and generally outperforms baselines, with detailed results provided in Appendix H.1. Moreover, we examine the unlearning effects of our controllable unlearning framework under different crop ratios. The results in Appendix H.3 demonstrate that our framework is robust to different crop ratios. Furthermore, we find that the visual effects of unlearning control are more prominent with larger crop ratios.

Summary: These results validate the effectiveness of our proposed method, which is universally applicable to mainstream I2I generative models as well as a variety of image generation tasks, consistently achieving favorable outcomes across all these tasks.

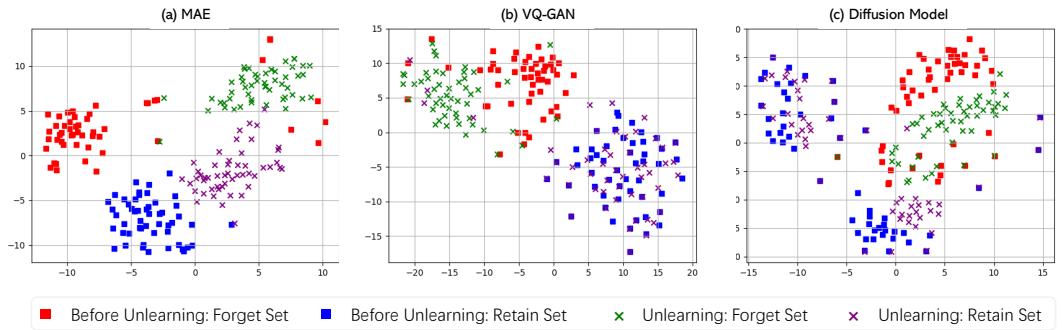

Figure 4: T-SNE analysis between images generated by our method and ground truth images.

Table 2: Results of center cropping 50% of the images under different unlearning completeness. "Highest" and "Lowest" respectively represent the two boundary points of unlearning identified in Phase I. ε is a coefficient used to control the unlearning completeness in Phase II.

	MAE						VQ-GAN						Diffusion Models					
	IS		FID		CLIP		IS		FID		CLIP		IS		FID		CLIP	
	F↓	R↑	F↑	R↓	F↓	R↑	F↓	R↑	F↑	R↓	F↓	R↑	F↓	R↑	F↑	R↓	F↓	R↑
Original	21.59	21.83	16.28	14.87	0.88	0.88	23.74	24.06	21.80	18.17	0.78	0.85	16.90	19.65	82.12	81.51	0.89	0.91
Highest	12.33	17.47	154.60	68.453	0.69	0.75	13.23	22.55	139.21	26.39	0.46	0.82	11.84	18.47	165.05	95.42	0.55	0.81
ε-25%	17.93	20.55	85.36	59.09	0.74	0.77	14.14	22.65	130.71	24.57	0.46	0.82	15.12	19.27	137.95	84.21	0.60	0.81
ε-50%	19.47	21.42	57.81	50.99	0.77	0.79	14.60	22.25	123.32	22.65	0.47	0.83	15.92	18.70	118.76	71.43	0.66	0.83
ε-75%	20.68	22.87	42.51	31.80	0.80	0.82	15.20	22.53	116.59	20.63	0.47	0.84	16.33	19.53	104.21	63.62	0.73	0.83
Lowest	21.23	22.92	31.28	25.83	0.82	0.84	15.77	22.75	109.28	20.26	0.48	0.84	16.36	20.78	90.03	52.96	0.77	0.84

5.3 Controllable Unlearning

We also evaluate the controllability of our method which provides a set of solutions for varied user expectations. First, we obtain two boundary points of unlearning, thereby establishing the valid range of values for ε. We linearly increase the value of ε within this range, adding 25% of the range interval each time, to obtain optimum solutions corresponding to different ε values. We provide some generated images corresponding to these solutions in Figure 1. Due to the space limit, please refer to Appendix G for more examples. For results of more fine-grained control (i.e., smaller increments of the linear increase of ε), please refer to Appendix H.2.

We verify the unlearned models at different ε values, and report results in Table 2. As ε increases, we observe a trade-off: the unlearning completeness decreases, while the generated images' performance on the forget set progressively improves, and, simultaneously, the performance on the retain set also improves. This observation clearly demonstrates the controllability of our proposed method, which can cater to varied user expectations. Please refer to Appendix I for additional results of the generated images and T-SNE analysis, which corroborates the above numerical results.

5.4 Unlearning Efficiency

To enhance the efficiency of our controllable unlearning framework, we modify the selections of control function $\psi(\theta)$ during various phases. Specifically, we empirically examine the convergence under these conditions to assess the framework's unlearning performance of efficiency. *In Phase I*, with the control function satisfying $\psi(\theta) = \alpha\|\nabla f_1(\theta)\|^\delta$, we manipulate the value of the exponent δ to change the control function. Additionally, we verify the changes in the convergence rates of $f_1(\theta)$ and $f_2(\theta)$ under four different δ values across three models, with results shown in Appendix J. It is evident that $f_1(\theta)$ and $f_2(\theta)$ achieve an optimal balance in convergence rates when $\delta = 2$, and the overall rate of convergence is fastest. *In Phase II*, where the control function satisfies $\psi(\theta) = \beta(f_1(\theta) - \varepsilon)^\delta$, we test the changes in the convergence rates of $f_1(\theta)$ and $f_2(\theta)$ for two different δ values on three models. To stabilize the optimization process, we scale the form of the control function (i.e., $\psi(\theta) = \beta(f_1(\theta) - \varepsilon)^\delta\|\nabla f_1(\theta)\|^2$), selecting two different δ values, with results presented in Appendix J. It can be observed that at $\delta = 1$ the overall rate of convergence was optimized.

6 Conclusion

In this paper, we propose a controllable unlearning framework for I2I generative models to overcome the limitation of the existing method's incapability to fulfill varied user expectations. Our approach allows for a controllable trade-off between unlearning completeness and model utility by introducing a control coefficient ε to control the degrees of unlearning completeness. We reformulate unlearning as a ε-constrained optimization problem and solve it with a gradient-based method to find two boundary points that guide the valid range for ε. Within this range, every chosen value of ε will lead to a Pareto optimal solution, addressing the existing method's issue of lacking theoretical guarantee. Extensive experiments on two large datasets (i.e., ImageNet-1K and Places-365) across three mainstream I2I models (i.e., MAE, VQ-GAN, diffusion model) demonstrate significant advantages of our method over the SOTA methods with higher unlearning efficiency, and a controllable balance between the unlearning completeness and model utility.

REFERENCES

Guillaume Alain and Yoshua Bengio. What regularized auto-encoders learn from the data-generating distribution. *The Journal of Machine Learning Research*, 15(1):3563–3593, 2014.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pp. 214–223. PMLR, 2017.

Samyadeep Basu, Philip Pope, and Soheil Feizi. Influence functions in deep learning are fragile. In *International Conference on Learning Representations (ICLR)*, 2021.

Joseph-Frédéric Bonnans, Jean Charles Gilbert, Claude Lemaréchal, and Claudia A Sagastizábal. *Numerical optimization: theoretical and practical aspects*. Springer Science & Business Media, 2006.

Lucas Bourtoule, Varun Chandrasekaran, Christopher A Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. Machine unlearning. In *2021 IEEE Symposium on Security and Privacy (SP)*, pp. 141–159. IEEE, 2021.

Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. In *International Conference on Learning Representations (ICLR)*, 2019.

Nicolas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramer, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models. In *32nd USENIX Security Symposium (USENIX Security 23)*, pp. 5253–5270, 2023.

Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman. Maskgit: Masked generative image transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11315–11325, 2022.

Vira Chankong and Yacov Y Haimes. On the characterization of noninferior solutions of the vector optimization problem. *Automatica*, 18(6):697–707, 1982.

Chong Chen, Fei Sun, Min Zhang, and Bolin Ding. Recommendation unlearning. In *Proceedings of the ACM Web Conference 2022*, pp. 2768–2777, 2022.

Lisha Chen, Heshan Fernando, Yiming Ying, and Tianyi Chen. Three-way trade-off in multi-objective learning: Optimization, generalization and conflict-avoidance. *Advances in Neural Information Processing Systems*, 36, 2024.

Dasol Choi and Dongbin Na. Towards machine unlearning benchmarks: Forgetting the personal identities in facial recognition systems. *arXiv preprint arXiv:2311.02240*, 2023.

T.M. Cover and J.A. Thomas. *Elements of Information Theory*. Number 12. Wiley, 2012. ISBN 9781118585771.

Stephen Dempe, Nguyen Dinh, and Joydeep Dutta. Optimality conditions for a simple convex bilevel programming problem. *Variational Analysis and Generalized Differentiation in Optimization and Control: In Honor of Boris S. Mordukhovich*, pp. 149–161, 2010.

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 (CVPR)*, pp. 248–255. Ieee, 2009.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021.

Mohamed Elasri, Omar Elharrouss, Somaya Al-Maadeed, and Hamid Tairi. Image generation: A review. *Neural Processing Letters*, 54(5):4609–4646, 2022.

Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 12873–12883, 2021.

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*, pp. 2426–2436, 2023.

Philip E Gill and Elizabeth Wong. Sequential quadratic programming methods. In *Mixed integer nonlinear programming*, pp. 147–224. Springer, 2011.

Aditya Golatkar, Alessandro Achille, and Stefano Soatto. 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.

Chengyue Gong, Xingchao Liu, and Qiang Liu. Automatic and harmless regularization with constrained and lexicographic optimization: A dynamic barrier approach. *Advances in Neural Information Processing Systems*, 34:29630–29642, 2021.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

Laura Graves, Vineel Nagisetty, and Vijay Ganesh. Amnesiac machine learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 11516–11524, 2021.

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. *Advances in neural information processing systems*, 30, 2017.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 16000–16009, 2022.

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*, 30, 2017.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Robert Janin. *Directional derivative of the marginal function in nonlinear programming*. Springer, 1984.

Il Yong Kim and Oliver L De Weck. Adaptive weighted-sum method for bi-objective optimization: Pareto front generation. *Structural and multidisciplinary optimization*, 29:149–158, 2005.

Diederik P Kingma, Max Welling, et al. An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning*, 12(4):307–392, 2019.

Zhifeng Kong and Kamalika Chaudhuri. Data redaction from conditional generative models. *arXiv preprint arXiv:2305.11351*, 2023.

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*, pp. 22691–22702, 2023.

Aditya Kuppa, Lamine Aouad, and Nhien-An Le-Khac. Towards improving privacy of synthetic datasets. In *Annual Privacy Forum*, pp. 106–119. Springer, 2021.

Meghdad Kurmanji, Peter Triantafillou, Jamie Hayes, and Eleni Triantafillou. Towards unbounded machine unlearning. *Advances in Neural Information Processing Systems*, 36, 2024.

Guihong Li, Hsiang Hsu, Radu Marculescu, et al. Machine unlearning for image-to-image generative models. In *International Conference on Learning Representations (ICLR)*, 2024a.

Tianhong Li, Huiwen Chang, Shlok Mishra, Han Zhang, Dina Katabi, and Dilip Krishnan. Mage: Masked generative encoder to unify representation learning and image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2142–2152, 2023a.

Yuyuan Li, Chaochao Chen, Xiaolin Zheng, Yizhao Zhang, Biao Gong, Jun Wang, and Linxun Chen. Selective and collaborative influence function for efficient recommendation unlearning. *Expert Systems with Applications*, 234:121025, 2023b.

Yuyuan Li, Chaochao Chen, Yizhao Zhang, Weiming Liu, Lingjuan Lyu, Xiaolin Zheng, Dan Meng, and Jun Wang. Ultrare: Enhancing receraser for recommendation unlearning via error decomposition. *Advances in Neural Information Processing Systems*, 36, 2024b.

Yuyuan Li, Chaochao Chen, Xiaolin Zheng, Junlin Liu, and Jun Wang. Making recommender systems forget: Learning and unlearning for erasable recommendation. *Knowledge-Based Systems*, 283:111124, 2024c.

Xi Lin, Hui-Ling Zhen, Zhenhua Li, Qing-Fu Zhang, and Sam Kwong. Pareto multi-task learning. *Advances in neural information processing systems*, 32, 2019.

Shilin Lu, Zilan Wang, Leyang Li, Yanzhu Liu, and Adams Wai-Kin Kong. Mace: Mass concept erasure in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6430–6440, 2024.

James Martens. *Second-order optimization for neural networks*. University of Toronto (Canada), 2016.

Kaisa Miettinen. *Nonlinear multiobjective optimization*, volume 12. Springer Science & Business Media, 1999.

Thanh Tam Nguyen, Thanh Trung Huynh, Phi Le Nguyen, Alan Wee-Chung Liew, Hongzhi Yin, and Quoc Viet Hung Nguyen. A survey of machine unlearning. *arXiv preprint arXiv:2209.02299*, 2022.

Jorge Nocedal and Stephen J Wright. *Numerical optimization*. Springer.

OpenAI. GPT-4 technical report. *CoRR*, abs/2303.08774, 2023. doi: 10.48550/arXiv.2303.08774. URL `https://doi.org/10.48550/arXiv.2303.08774`.

Panos M Pardalos, Antanas Žilinskas, Julius Žilinskas, et al. *Non-convex multi-objective optimization*. Springer, 2017.

Vitali Petsiuk and Kate Saenko. Concept arithmetics for circumventing concept inhibition in diffusion models. In *European Conference on Computer Vision*, pp. 309–325. Springer, 2025.

Minh Pham, Kelly O Marshall, Niv Cohen, Govind Mittal, and Chinmay Hegde. Circumventing concept erasure methods for text-to-image generative models. In *The Twelfth International Conference on Learning Representations*, 2023.

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*, pp. 8748–8763. PMLR, 2021.

James Renegar. *A mathematical view of interior-point methods in convex optimization*. SIAM, 2001.

Stephen M Robinson. Perturbed kuhn-tucker points and rates of convergence for a class of nonlinear-programming algorithms. *Mathematical programming*, 7:1–16, 1974.

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*, pp. 10684–10695, 2022.

Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 conference proceedings*, pp. 1–10, 2022a.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022b.

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *International Conference on Learning Representations (ICLR)*, 2022.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016.

Katja Schwarz, Yiyi Liao, and Andreas Geiger. On the frequency bias of generative models. *Advances in Neural Information Processing Systems*, 34:18126–18136, 2021.

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*, 34:18075–18086, 2021.

Yang Song and Stefano Ermon. Improved techniques for training score-based generative models. *Advances in neural information processing systems*, 33:12438–12448, 2020.

Gilbert W Stewart. On the perturbation of pseudo-inverses, projections and linear least squares problems. *SIAM review*, 19(4):634–662, 1977.

Piotr Teterwak, Aaron Sarna, Dilip Krishnan, Aaron Maschinot, David Belanger, Ce Liu, and William T Freeman. Boundless: Generative adversarial networks for image extension. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10521–10530, 2019.

Kushal Tirumala, Aram Markosyan, Luke Zettlemoyer, and Armen Aghajanyan. Memorization without overfitting: Analyzing the training dynamics of large language models. *Advances in Neural Information Processing Systems*, 35:38274–38290, 2022.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

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.

Bram Wallace, Akash Gokul, and Nikhil Naik. Edict: Exact diffusion inversion via coupled transformations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22532–22541, 2023.

Alexander Warnecke, Lukas Pirch, Christian Wressnegger, and Konrad Rieck. Machine unlearning of features and labels. In *30th Annual Network and Distributed System Security Symposium NDSS*, 2023.

Max Welling and Yee W Teh. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pp. 681–688. Citeseer, 2011.

Weihao Xia, Yulun Zhang, Yujiu Yang, Jing-Hao Xue, Bolei Zhou, and Ming-Hsuan Yang. Gan inversion: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 45(3): 3121–3138, 2022.

Heng Xu, Tianqing Zhu, Lefeng Zhang, Wanlei Zhou, and Philip S. Yu. Machine unlearning: A survey. *Association for Computing Machinery*, 56:36, 2023.

Haonan Yan, Xiaoguang Li, Ziyao Guo, Hui Li, Fenghua Li, and Xiaodong Lin. Arcane: An efficient architecture for exact machine unlearning. In *IJCAI*, volume 6, pp. 19, 2022.

Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. Diffusion models: A comprehensive survey of methods and applications. *ACM Computing Surveys*, 56(4):1–39, 2023.

Özgür Yeniay. Penalty function methods for constrained optimization with genetic algorithms. *Mathematical and computational Applications*, 10(1):45–56, 2005.

Lu Yu, Joost van de Weijer, et al. Deepi2i: Enabling deep hierarchical image-to-image translation by transferring from gans. *Advances in Neural Information Processing Systems*, 33:11803–11815, 2020.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021.

Xulong Zhang, Jianzong Wang, Ning Cheng, Yifu Sun, Chuanyao Zhang, and Jing Xiao. Machine unlearning methodology based on stochastic teacher network. In *International Conference on Advanced Data Mining and Applications*, pp. 250–261. Springer, 2023.

Yihua Zhang, Yuguang Yao, Parikshit Ram, Pu Zhao, Tianlong Chen, Mingyi Hong, Yanzhi Wang, and Sijia Liu. Advancing model pruning via bi-level optimization. *Advances in Neural Information Processing Systems*, 35:18309–18326, 2022.

Yihua Zhang, Prashant Khanduri, Ioannis Tsaknakis, Yuguang Yao, Mingyi Hong, and Sijia Liu. An introduction to bilevel optimization: Foundations and applications in signal processing and machine learning. *IEEE Signal Processing Magazine*, 41(1):38–59, 2024.

Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pp. 2223–2232, 2017.