Controllable Unlearning for Image-to-Image Generative Models via ε-Constrained Optimization

Xiaohua Feng[1,*], Yuyuan Li[2,*], Chaochao Chen[1,†], Li Zhang[1], Longfei Li[3], Jun Zhou[3], Xiaolin Zheng[1]

Zhejiang University[1], Hangzhou Dianzi University[2], Ant Group[3]

2025.4.26

Table of Contents

Privacy Concerns in Recommender Systems:

Existing problem:

- Generative models absorb biases and expose private information from large datasets.
- Generative models recall training instances, raising bias and privacy concerns.
- Personal information is entitled to the right to be forgotten.

Naive solution: Single-objective optimization that combines performance on both forget and retain sets.

I2I Generative Models

Image-to-Image (I2I) generative models, including AEs, GANs, and diffusion models, are used for tasks like style transfer, each with varying strengths and challenges.

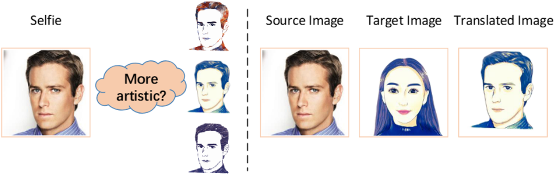


Figure: An example of style transfer in I2I translation [1].

I2I models use encoder-decoder structures, with E_γ mapping images to latent space and D_ϕ reconstructing them. For model I_θ with input x, the output is:

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

I2I Generative Model Unlearning

Unlearning objective: To obtain a model I_θ that fails on D_f while maintaining performance on D_R with KL divergence used to measure the distributional distance, formulated as:

$$\max_\theta Div(\mathbb{P}_{X_f}||\mathbb{P}_{\hat{X}_f}), \text{ and } \min_\theta Div(\mathbb{P}_{X_r}||\mathbb{P}_{\hat{X}r}), \qquad (2)$$

Definition: I2I generative model's inability to reconstruct a full image from a partial one [2].

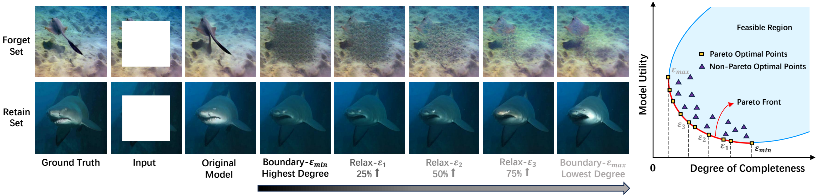
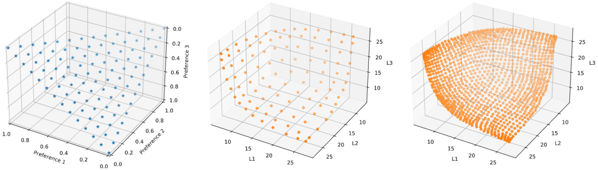


Figure: An overview of generative model unlearning.

Evaluation Metrics

- ***Inception Score (IS).*** Assesses the quality of generated images independently.
- ***Frechét Inception Distance (FID).*** Measures similarity between generated and ground truth images.
- ***Cosine Similarity of CLIP Embeddings.*** Assesses whether the generated images capture similar semantics to the ground truth images.

Pareto Optimality



(a) 105 Preferences (b) 105 Solutions (c) 1, 035 Solutions

Figure: Pareto Set Approximation in Multi-Objective Optimization [3].

In a multi-objective optimization problem:

1. **Pareto dominance:** θ^a dominates θ^b if $f_i(\theta^a) \leq f_i\left(\theta^b\right)$ for all i, and for some j, $f_j(\theta^a) < f_j\left(\theta^b\right)$.
2. **Pareto optimal:** A point θ^* is Pareto optimal if no other point $\hat{\theta}$ dominates it.

The collection of Pareto optimal points forms the Pareto set, and its projection in the objective space is the Pareto front.

A Controllable Unlearning Framework

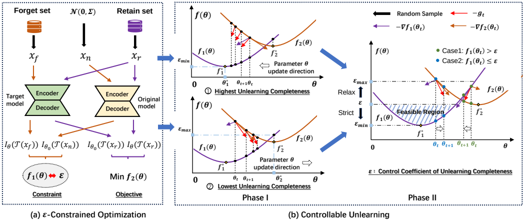


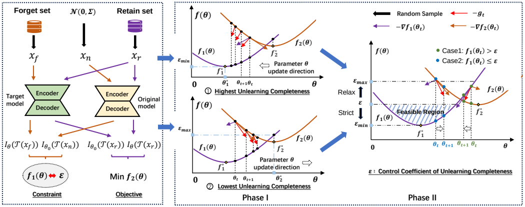
Figure: Pipeline of the controllable unlearning framework.

Phase I: Boundaries of Unlearning: We solve for two boundary solutions: the highest and the lowest unlearning completeness. The highest completeness is formulated as:

$$\min_{\theta \in \mathbb{R}^d} f_2(\theta) \quad \text{s.t.} \quad f_1(\theta) \leq f_1^* \tag{3}$$

f_1^* is the infimum of $f_1(\theta)$.

A Controllable Unlearning Framework



(a) ε-Constrained Optimization (b) Controllable Unlearning

Figure: Pipeline of the controllable unlearning framework.

Phase II: Controllable Unlearning: The unlearning constraint is relaxed by adjusting ε between f_1^* and f_2^*, controlling unlearning completeness. The problem is reformulated as:

$$\min_{\theta \in \mathbb{R}^d} f_2(\theta) \quad \text{s.t.} \quad f_1(\theta) \leq \varepsilon \tag{4}$$

Solution to ε-Constrained Optimization Problem

A gradient-based optimization method is used to solve the ε-constrained optimization problem. The update rule is:

$$\theta_{t+1} \leftarrow \theta_t - \mu_t g_t \tag{5}$$

where g_t is determined by solving the following convex quadratic programming problem:

$$g_t = \min_{g \in \mathbb{R}^d} \left\{ \|\nabla f_2(\theta_t) - g\|^2 \quad \text{s.t.} \quad \nabla f_1(\theta_t)^\top g \geq f_1(\theta_t) - \varepsilon \right\}. \tag{6}$$

Unlearning Performance

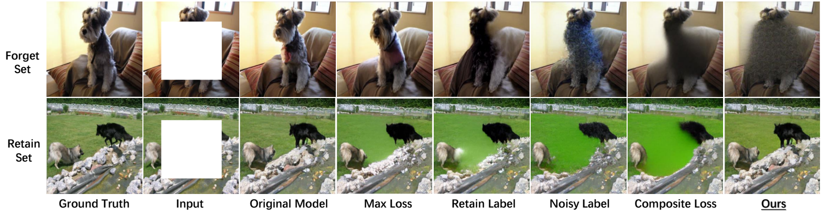


Figure: Generated images of cropping 50% at the center of the image on VQ-GAN.

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.

Controllability of Unlearning

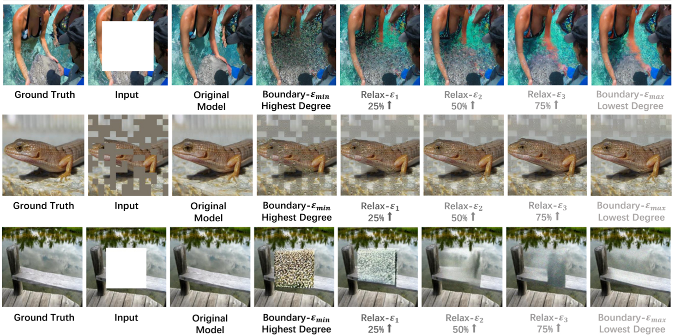


Figure: Controllability performance of our unlearning framework using VQ-GAN (above), MAE (middle), and the diffusion model (below).

The results in Figure 6 indicate that our method can effectively control the completeness of unlearning in image inpainting tasks as well as image reconstruction tasks.

Unlearning Efficiency

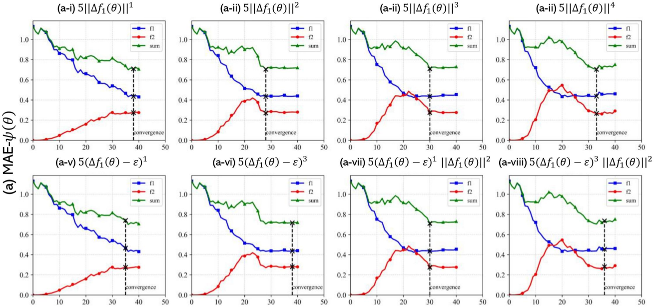


Figure: The convergence rates under different control functions $\psi(\theta)$ using VQ-GAN. Each section contains two rows, corresponding to Phase I and Phase II, respectively. The titles on each subplot indicate the forms of the control function $\psi(\theta)$.

In Phase I, the optimal parameter is $\delta = 2$, while in Phase II, the optimal parameter is $\delta = 1$ for the fastest convergence rate.

Takeaways

- **Controllable Unlearning.** We reformulate machine unlearning as a ε-constrained optimization, with unlearning the forget set as a constraint, ensuring optimal theoretical solutions.
- **Pareto Optimal Solutions.** By progressively relaxing the unlearning constraint, we obtain a Pareto set and plot the corresponding Pareto front, using gradient-based methods to solve the optimization problem.
- **Experimental Validation.** Experiments on large I2I generative models show our method outperforms baselines, offering controllable unlearning that balances user expectations and model utility.

References

- Pang, Y., Lin, J., Qin, T. and Chen, Z. (2022). Image-to-Image Translation: Methods and Applications. IEEE Transactions on Multimedia, 24, pp.3859–3881.
- Guihong Li, Hsiang Hsu, Radu Marculescu, et al. Machine unlearning for image-to-image generative models. In International Conference on Learning Representations (ICLR), 2024.
- Lin, X., Yang, Z. and Zhang, Q. (2022). Pareto Set Learning for Neural Multi-objective Combinatorial Optimization. [online] arXiv.org.