

DATA UNLEARNING BEYOND UNIFORM FORGETTING VIA DIFFUSION TIME AND FREQUENCY SELECTION

Jinseong Park

Korea Institute for Advanced Study, Seoul
jinseong@kias.re.kr

Mijung Park*

Department of Computer Science
University of British Columbia, Vancouver
mijungp@cs.ubc.ca

ABSTRACT

Data unlearning aims to remove the influence of specific training samples from a trained model without requiring full retraining. Unlike concept unlearning, data unlearning in diffusion models remains underexplored and often suffers from quality degradation or incomplete forgetting. To address this, we first observe that most existing methods attempt to unlearn the samples at all diffusion time steps equally, leading to poor-quality generation. We argue that forgetting occurs disproportionately across time and frequency, depending on the model and scenarios. By selectively focusing on specific time–frequency ranges during training, we achieve samples with higher aesthetic quality and lower noise. We validate this improvement by applying our time–frequency selective approach to diverse settings, including gradient-based and preference optimization objectives, as well as both image-level and text-to-image tasks. Finally, to evaluate both deletion and quality of unlearned data samples, we propose a simple normalized version of SSCD. Together, our analysis and methods establish a clearer understanding of the unique challenges in data unlearning for diffusion models, providing practical strategies to improve both evaluation and unlearning performance.

1 INTRODUCTION

The ability to remove the influence of training samples from a learned model, often referred to as *machine unlearning* (Bourtoule et al., 2021), has become increasingly important. Regulatory frameworks such as the “right to be forgotten” in the General Data Protection Regulation (GDPR) by the European Union and growing concerns about sensitive or proprietary data have created demand for methods that allow models to forget without costly retraining from scratch. Recently, with the development of generative models such as diffusion models (Ho et al., 2020), unlearning the unsafe concept or memorization has been actively explored through training-free sampling (Kim et al., 2025), output filtering (Yoon et al., 2025), and fine-tuning (Wang et al., 2025a).

In machine unlearning, we can consider two major scenarios: (a) Concept or class unlearning, which refers to preventing the generation of certain types of samples (categorized in a particular concept or class) (Fan et al., 2024); and (b) Data unlearning, which focuses on removing individual samples. While extensive research has been conducted in the context of classification tasks, its extension to generative modelling, in particular when using diffusion models, remains underexplored (Alberti et al., 2025). Recent work on fine-tuning has begun to address this challenge with unlearning objectives such as gradient ascent, gradient importance sampling (Alberti et al., 2025), or preference optimization (Park et al., 2024). However, these approaches typically unlearn the samples in the forget set at all diffusion time steps equally. We demonstrate that this uniform unlearning in all time steps results in artifacts: generated images become noisier and aesthetically degraded, while forgetting is neither complete nor precise, as shown in Figure 1.

In this paper, we analyze the dynamics of data unlearning and argue that forgetting does not occur uniformly, but rather disproportionately across time and frequency domains. Diffusion models learn stage-dependent behaviour during training (Choi et al., 2022): later training steps close to

*Corresponding author

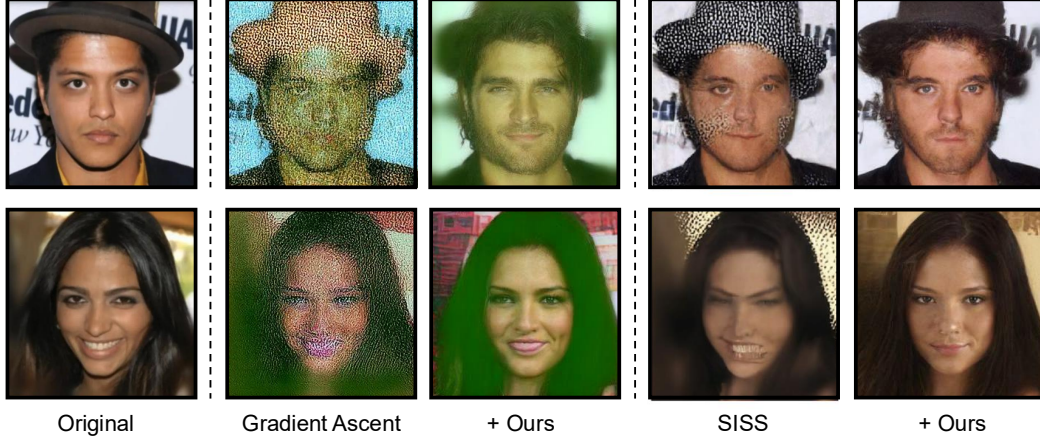


Figure 1: Illustration of quality degradation of unlearned images after data unlearning. **Left:** two images to unlearn. **Middle:** the results of Gradient Ascent (GA, first column) and those of GA together with our approach (second column). **Right:** the results of SISS (first column) and those of SISS together with our approach (second column). Our approach, combined with existing unlearning objective functions, improves the quality of generated unlearned samples.

Gaussian noise capture coarse semantics, while earlier steps close to data refine fine-grained details. Consequently, enforcing deletion across all steps may inadvertently erase either global features or only local details. To push this further, we investigate the application of a low-pass filter in the frequency domain, which encourages the model to unlearn only semantic contexts without deleting the high-frequency components. Overall, our framework offers practical strategies for achieving effective forgetting on selective components in both time and frequency domains, while preserving the quality of unlearned samples. Our contributions are summarized as follows:

- We first observe that most existing methods attempt to unlearn the samples in the forget set at all diffusion time steps equally, leading to poor quality. We argue that forgetting occurs disproportionately across time and frequency, depending on the model and scenarios.
- We therefore introduce a novel time–frequency selective unlearning approach, which is compatible with existing unlearning objectives and effective in various experimental setups to enhance the quality of generated unlearned samples.
- We argue that current unlearning metrics, such as FID or SSCD, fail to capture the quality of unlearned samples. To address this, we introduce a simple, modified metric of SSCD.

2 PRELIMINARIES

2.1 DIFFUSION MODELS

Diffusion models (Ho et al., 2020) aim to learn a data distribution. within the diffusion process. Diffusion gradually injects noise into samples $\mathbf{x}_0$ drawn from the data distribution p_0 , forwarding them into a fixed standard Gaussian distribution $p_T = \mathcal{N}(\mathbf{0}, \mathbf{I})$. The corresponding forward SDE is:

$$d\mathbf{x}_t = f(\mathbf{x}_t, t)dt + g(t)d\mathbf{w}_t, \quad (1)$$

where $f : \mathbb{R}^d \times [0, T] \rightarrow \mathbb{R}^d$ represents the drift function, $g : [0, T] \rightarrow \mathbb{R}$ denotes the diffusion coefficient, and $\mathbf{w}_t$ is the Wiener process. As diffusion models learn to reverse the above process, the reverse SDE (Anderson, 1982) is formulated as follows:

$$d\mathbf{x}_t = [f(\mathbf{x}_t, t) - g(t)^2 \nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t)]dt + g(t)d\bar{\mathbf{w}}_t, \quad (2)$$

where $p(\mathbf{x}_t)$ refers to the probability density of $\mathbf{x}_t$. Diffusion models learn to match the score function $\nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t)$ for denoising (Song et al., 2021). With standard DDPM forward noising process $q(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1 - \bar{\alpha}_t)\mathbf{I})$, the conditional score can be written as $\nabla_{\mathbf{x}_t} \log q(\mathbf{x}_t|\mathbf{x}_0)$. With a weighting factor w_t , the training objective is then

$$\mathcal{L}_{\mathcal{D}}(\theta) = \mathbb{E}_{\mathbf{x}_0 \sim p_0} \mathbb{E}_{t \in [0, T]} \left[w_t \left\| s_{\theta}(\mathbf{x}_t, t) - \nabla_{\mathbf{x}_t} \log q(\mathbf{x}_t|\mathbf{x}_0) \right\|_2^2 \right]. \quad (3)$$

2.2 DATA UNLEARNING IN DIFFUSION MODELS

Machine Unlearning. Suppose a model θ^* is already trained on a dataset $\mathcal{D}$. Then, our goal is to delete the influence of the forget set $\mathcal{D}_F$ from θ^* , while maintaining the model utility on the retain set $\mathcal{D}_R = \mathcal{D} \setminus \mathcal{D}_F$. Specifically, our goal is to fine-tune the trained model θ^* using the forget dataset $\mathcal{D}_F$, often using the retain dataset $\mathcal{D}_R$ to mitigate quality degradation (Bourtoule et al., 2021).

Diffusion Unlearning. Unlearning in diffusion models can be categorized into two: concept unlearning (Gandikota et al., 2023) and data unlearning. **Concept unlearning** refers to prohibiting a diffusion model from producing images categorized in a particular type of high-level concept, including not safe for work (NSFW) material (nude and violent images, for example), or copyrighted content. Most existing work addresses concept unlearning by fine-tuning the model to suppress the desired concepts, with a focus on text-to-image tasks (Gandikota et al., 2023; Srivatsan et al., 2025), adversarial training (Zhang et al., 2024b), or preference optimization frameworks (Park et al., 2024).

Due to the high flexibility of diffusion models in controlling the generation process during sampling, unlike one-step generative models such as GANs (Goodfellow et al., 2020), training-free steering methods have been actively investigated (Singhal et al., 2025; Kim et al., 2025; Koulischer et al., 2025). These approaches leverage guidance techniques to repel the generation from specific points or embeddings, thereby achieving concept erasure without the need for fine-tuning.

On the other hand, **data unlearning** is closely related to individual data memorization and aims to remove specific data examples in accordance with the ‘‘Right to be Forgotten.’’ For example, if a user requests the deletion of a particular face image used to train a diffusion model, the goal is to eliminate the influence of that image from the trained model. In contrast to concept erasing, for which various practical solutions have been developed, data unlearning remains relatively underexplored (Alberti et al., 2025). In general, the objective of diffusion unlearning $\mathcal{L}_{UL}(\theta)$ can be formulated as follows:

$$\min_{\theta} \mathcal{L}_{UL}(\theta) = \begin{cases} \max_{\theta} \mathcal{L}_{\mathcal{D}_F}(\theta) = \min_{\theta} -\mathcal{L}_{\mathcal{D}_F}(\theta), & \mathcal{D}_F \text{ only} \\ \min_{\theta} (-\mathcal{L}_{\mathcal{D}_F}(\theta) + \mathcal{L}_{\mathcal{D}_R}(\theta)), & \mathcal{D}_F \text{ and } \mathcal{D}_R. \end{cases} \quad (4)$$

Gradient ascent (GA) or negative gradient on forget data samples $\mathcal{D}_F$ is a key component for data unlearning (Golatkhar et al., 2020). For GA, we can rewrite the loss function on the forget set as

$$\mathcal{L}_{GA}(\theta) = - \underbrace{\mathbb{E}_{\mathbf{x}_0 \sim \mathcal{D}_F} \mathbb{E}_{t \in [0, T]} \left[w_t \| s_{\theta}(\mathbf{x}_t, t) - \nabla_{\mathbf{x}_t} \log q(\mathbf{x}_t | \mathbf{x}_0) \|_2^2 \right]}_{\text{Diffusion loss in Equation (3) on forget set } \mathcal{D}_F}. \quad (5)$$

EraseDiff (Wu et al., 2025) replaces the true score of the forget samples with a randomly sampled noisy image, while training normally on the retain set with scaler β :

$$\mathcal{L}_{\text{EraseDiff}}(\theta) = \mathbb{E}_{\mathbf{x}_0 \sim \mathcal{D}_F} \mathbb{E}_{\epsilon \sim \mathcal{N}(0, \mathbf{I})} \mathbb{E}_{t \in [0, T]} \left[w_t \| s_{\theta}(\mathbf{x}_t, t) - (-\frac{1}{\sigma_t} \epsilon) \|_2^2 \right] + \beta \mathcal{L}_{\mathcal{D}_R}(\theta), \quad (6)$$

where ϵ is data-independent noise. Recently, SISS (Alberti et al., 2025) investigated the unlearning as a mixture distribution of forget and retain data with importance sampling as

$$\mathcal{L}_{\text{SISS}}(\theta) = \mathbb{E}_{\mathbf{x} \sim \mathcal{D}_F, \mathbf{x}' \sim \mathcal{D}_R} \mathbb{E}_{\mathbf{m}_t \sim q_{\lambda}(\cdot | \mathbf{x}, \mathbf{x}')} \mathbb{E}_{t \in [0, T]} \left[w_{\text{keep}} \| s_{\theta}(\mathbf{m}_t, t) - \nabla_{\mathbf{m}_t} \log q(\mathbf{m}_t | \mathbf{x}') \|_2^2 \right. \\ \left. - (1 + \beta) w_{\text{forget}} \| s_{\theta}(\mathbf{m}_t, t) - \nabla_{\mathbf{m}_t} \log q(\mathbf{m}_t | \mathbf{x}) \|_2^2 \right], \quad (7)$$

where $q_{\lambda}(\mathbf{m}_t | \mathbf{x}, \mathbf{x}') = (1 - \lambda)q(\mathbf{m}_t | \mathbf{x}') + \lambda q(\mathbf{m}_t | \mathbf{x})$ and $w_{\text{keep}}, w_{\text{forget}}$ are importance sampling weights determined by forget and retain data, and mixture proportion λ (e.g., $\lambda = 0.5$). In conclusion, previous methods investigate how to well-formulate Equation (4) for diffusion models.

Quality Degradation in Unlearning. As the goal of data unlearning is to erase information for generating a specific example, it inevitably forces the model parameters apart from a well-converged minimum. Consequently, the model also forgets previously learned features and produces lower-quality outputs. Zhang et al. (2024a) refers to this quick drop in model utility when deleting the forget set as ‘‘catastrophic collapse’’. To address the quality degradation on the retain data, additional regularization in terms of the retain set is often employed to preserve stability, i.e., $\mathcal{D}_F$ and $\mathcal{D}_R$ in Equation (4) (Alberti et al., 2025; Wu et al., 2025). Moreover, to mitigate the negative effects of forgetting on the forget data samples $\mathcal{D}_F$, negative preference optimization (Zhang et al., 2024a; Wang et al., 2025a) or positive anchoring in near point (Zhao et al., 2024). However, we argue that none of the previous methods investigate the quality degradation of unlearned samples.

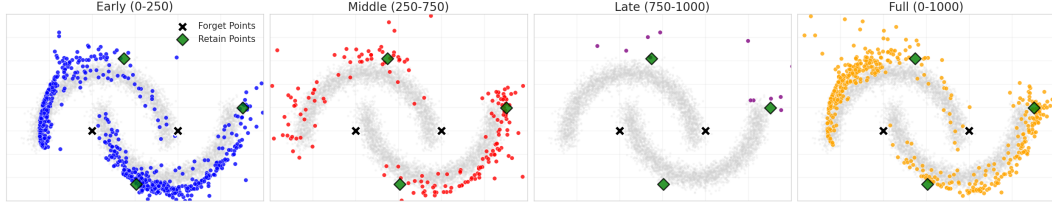


Figure 2: Illustration of unlearning the specific stage of time steps only in diffusion models.

3 SELECTIVE DATA UNLEARNING IN DIFFUSION MODELS

3.1 IMPORTANT FEATURES IN DIFFUSION UNLEARNING

The quality degradation issue in diffusion unlearning is illustrated in Figure 1. To generate the images, we follow the memorization experiment in (Alberti et al., 2025), where the objective is to unlearn an original data sample in diffusion models. After unlearning, the process begins by taking the training face to be removed and injecting noise (e.g., $t = 250$) as part of the DDPM forward noising process. Then, we compare the reversed image using the unlearned model θ and the original image. We observe that when gradient ascent is used for unlearning, the generated images after unlearning tend to lose not only the targeted information but also useful features. When combined with a retain loss, such as SISS in Equation (7), the quality is better preserved compared to using only the forget loss, but important details are still removed simultaneously.

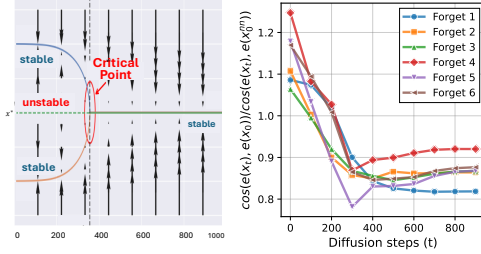
Allouah et al. (2025) pointed out that the KL divergence between forget and retain information plays a key role in machine unlearning. In contrast to concept unlearning, where the distributions of forget and retain data are clearly separated, data unlearning often involves considerable overlap between the distributions. Thus, we need to design a selective deletion strategy, preventing the erasure of useful information. In this section, we investigate two practical solutions: (i) analysis of diffusion time stages and (ii) low-pass filtering to mitigate the quality degradation after unlearning.

Hypothesis 1: Unlearning the specific diffusion stage is enough. The diffusion model has its own intrinsic characteristics to use diffusion time steps $[0, T]$. Starting from Gaussian noise, diffusion models are known to learn different attributes in different time steps (Choi et al., 2022), i.e., (i) initially learn coarse features in late time steps close to time T , (ii) generate content during intermediate steps, and (iii) refine the details for convergence in initial time steps near time 0. However, a practical analysis of which time steps are most crucial for unlearning is missing based on fine-tuning methods. Only Zhang et al. (2025) investigated a time-aware unlearning framework, uncovering that in T2I concept unlearning, a wider range of diffusion time steps is required to fine-tune as the target concept becomes more abstract (i.e., from instance to style, class, and NSFW).

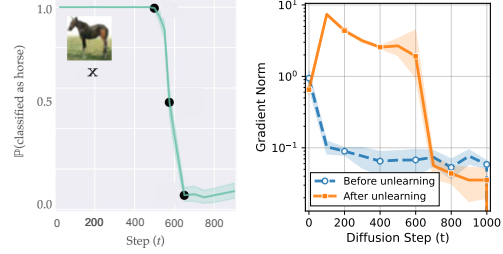
To clarify the effect of unlearning each stage, we conduct a toy experiment with a shallow diffusion model on a two half-moon dataset, visualized in Figure 2. We only split diffusion time steps for unlearning while using the same forget and retain data samples with Equation (5). Our observations are as follows: deleting only the early phase cannot prevent the model from generating forget points. On the other hand, unlearning the middle phase is most aligned with the data level, balancing the repelling of the forget dataset and the maintenance of the data manifold. Targeting the late phase is a powerful way to delete the forget samples, but also delete the distributional properties, only generating some samples near specific retain data used in unlearning fine-tuning.

Near the convergence in early time steps, Raya & Ambrogioni (2023) argued that data converges into a stable point or not, after a certain critical point in Figure 3. When we calculate the embedding similarity on DINOv3 (Siméoni et al., 2025) of the noisy data $\mathbf{x}_t$ between other clean data samples, the similarity towards the original data $\mathbf{x}_0$ is drastically increased in the refinement stage, compared to other nearest samples. Practically, Wang et al. (2025b) gave multiple times of training on the early refinement stage to achieve the fast convergence of diffusion training. Therefore, as the early diffusion stage mainly performs refinement, unlearning early time steps results in forgetting the details rather than deleting the memorization of a data point.

In Figure 4, Georgiev et al. (2023) observed that the likelihood that the noisy data $\mathbf{x}_t$ is classified to a certain instance around time step $t = 650$. This suggests that the core identity of a sample



(a) Early stage (Raya & Ambrogioni, 2023) (b) Embedding similarity of $\mathbf{x}_t$ w.r.t. $\mathbf{x}_0$ and $\mathbf{x}_0^{nn}$



(a) Middle-late stage (Georgiev et al., 2023) (b) Gradient norm of forget set before/after unlearning

Figure 3: Analysis on early refinement stage.

Figure 4: Analysis on middle-late stage.

is primarily formed in these middle-to-late stages. To compare the actual changes after standard unlearning with uniform time steps in the image-level unlearning with CelebA-HQ, we measure the gradient norm $\|\nabla_{\theta} \mathcal{L}(\mathbf{x}; \theta)\|_2^2$, which reflects the magnitude of the update required to forget a sample $\mathbf{x}$ (Paul et al., 2021; Pal et al., 2025), allowing us to identify which stages are most affected by the unlearning. Interestingly, the gap in the gradient norm between the model before and after unlearning is largest in the middle phase, justifying a selective unlearning with a specific window.

Similarly, in training-free sampling, including text-to-image models, the role of different time steps has been investigated in controllable generation (Kim et al., 2025; Kirchhof et al., 2025; Koulischer et al., 2025). It has been empirically shown that applying steering methods is sufficient in the late stage to maintain image quality while avoiding globally harmful concepts. For example, Kim et al. (2025) explicitly applied their guidance in the range of $t \in [780, \dots, 1000]$, and similarly, Kirchhof et al. (2025) observed that their repency term is most active in the late time steps.

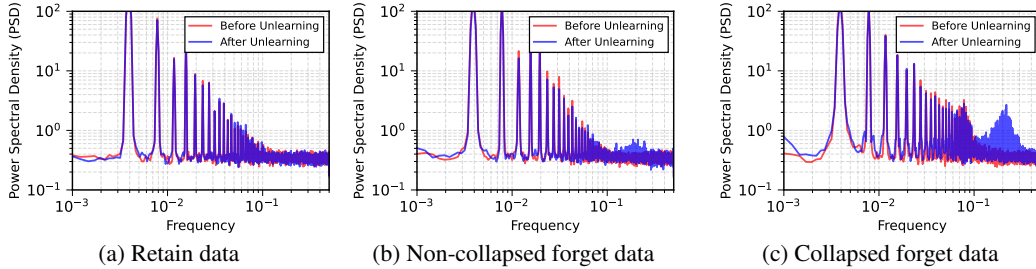


Figure 5: Power Spectral Density (PSD) before and after unlearning. For collapsed forget data, the difference in PSD is significant in high-frequency regions.

Hypothesis 2: High-frequency components are not necessarily unlearned. Based on the previous section, we aim to prevent the model from removing fine-grained components. We further test whether applying a low-pass filter provides selective target features for unlearning.

To clarify the differences observed during unlearning, we visualize the power spectral density, which quantifies the distribution of signal energy across frequencies, in Figure 5. For the retain dataset, the frequency domain shows no noticeable difference before and after unlearning, consistent with the results in the image space. In contrast, the collapsed forget data, which loses fine-grained details as illustrated in Figure 1, exhibits similar behavior in the low-frequency domain compared to the non-collapsed forget data. However, in the high-frequency domain, the collapsed data show a much larger difference after unlearning, providing evidence for why these data yield lower-quality images.

We also evaluate the gradient norm of each high and low frequency with a threshold of 10^{-1} in Figure 6. The change in the high-frequency region is significant in forget images, highlighting the utility of a low-pass filter for unlearning to filter out the fine-grained details.

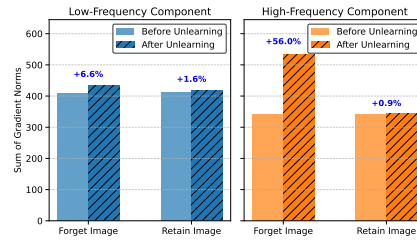


Figure 6: Gradient norm of low and high frequency parts in unlearning.

3.2 SELECTIVE UNLEARNING WITH TIME STEP SELECTION AND FREQUENCY FILTERING

Based on the two observations, we propose a simple framework for selective forgetting as follows:

$$\mathcal{L}(\theta) = \underbrace{\mathbb{E}_{\tilde{\mathbf{x}}_t = \mathcal{F}(\mathbf{x}_t), \mathbf{x}_0 \sim p_0}}_{\text{Low-pass filter}} \underbrace{\mathbb{E}_{t \sim P(t)}}_{\text{Time selection}} \left[w_t \|s_\theta(\tilde{\mathbf{x}}_t, t) - \nabla_{\tilde{\mathbf{x}}_t} \log q(\tilde{\mathbf{x}}_t | \mathbf{x}_0)\|^2 \right]. \quad (8)$$

Here, $P(t)$ assigns a non-uniform time step weighting and $\tilde{\mathbf{x}}_t$ denotes the data after applying the low-pass filter $\mathcal{F}(\cdot)$. In detail, we utilize the time scales that give a higher probability to a specific interval $t_1 \leq t \leq t_2$ for the forget data samples as follows:

$$P(t) = \begin{cases} 1 - k/(t_2 - t_1), & t_1 \leq t \leq t_2, \\ k/(T - (t_2 - t_1)), & \text{otherwise,} \end{cases} \quad (9)$$

where $0 \leq k \leq 1$ is the suppression intensity and $t_1, t_2 \in (0, T]$ with $t_1 < t_2$.

For the low-pass filter, we employ the discrete Fast Fourier Transform (FFT) as follows:

$$\mathcal{T}(u, v) = \text{FFT}(\mathbf{x}) = \sum_{x=1}^H \sum_{y=1}^W \mathbf{x}(x, y) e^{-j2\pi(\frac{u}{H}x + \frac{v}{W}y)}, \quad (10)$$

where $\mathbf{x}(x, y)$ is the pixel intensity at position (x, y) , $\mathcal{T}(u, v)$ is the complex coefficient at frequency (u, v) , and e and j represent Euler's constant and the imaginary unit, respectively. Its inverse is

$$\mathbf{x}(x, y) = \text{IFFT}(\mathbf{x}) = \frac{1}{HW} \sum_{u=1}^H \sum_{v=1}^W \mathcal{T}(u, v) e^{j2\pi(\frac{u}{H}x + \frac{v}{W}y)}. \quad (11)$$

To remove only the high-frequency components, we apply a masking function to the FFT results and then reconstruct the image using the inverse FFT as follows:

$$\mathcal{F}(\mathbf{x}_{i,t}) = \text{IFFT}(\text{FFT}(\mathbf{x}_{i,t}) \odot \beta_{i,t}(r)), \quad (12)$$

where $\beta_{i,t}(r) = s$ if the radius $r > r_t$, 1 otherwise. $0 \leq s \leq 1$ is a weight of high frequency and r_t is a radius cutoff frequency threshold.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUPS

We follow the recent experimental setups concerning data unlearning in (Alberti et al., 2025). Our primary goal is to maintain the quality of the unlearned samples and accelerate the unlearning.

Datasets and Baselines. For CelebA-HQ (Liu et al., 2015), we test the deletion of individual samples at the image level in unconditional image generation. For Stable Diffusion v1.4 (Rombach et al., 2022), we evaluated the memorization of the LAION dataset corresponding to each prompt in text-to-image generation. We compare EraseDiff (Wu et al., 2025), Gradient Ascent (GA), and SISS, under the settings of (Alberti et al., 2025). We use SISS (No IS) to reproduce the results of SISS among its variants. We also evaluate preference optimization-based approaches, including Direct Preference Optimization (DPO) (Rafailov et al., 2023) and its diffusion variant (Wallace et al., 2024), as well as Kahneman-Tversky Optimization (KTO) (Ethayarajh et al., 2024) and its diffusion variant (Li et al., 2024). Further details of experimental setups are provided in the Appendix A.

Hyperparameter. Following (Alberti et al., 2025), we use a pretrained model from the Hugging Face diffusers package. For CelebA-HQ and Stable Diffusion, we then fine-tune the model with the Adam optimizer with a batch size of 64 and 16 with a learning rate of $5 \cdot 10^{-6}$ and 10^{-5} , respectively. To unify the training, we use epsilon-matching (Ho et al., 2020) rather than a score-based matching.

We set the time suppression intensity $k = 0$ in Equation (9) and the high-frequency suppression weight $s = 0$ in Equation (12) to simplify experimental designs. Thus, we only tune two parameter sets: the forgetting time steps $[t_1, t_2]$ and the cutoff frequency radius r_t in the FFT low-pass filter.

Table 1: Comparison of unlearning methods with a relative Gain. For each method, we show the baseline scores, the scores with our method applied (+ Ours), and the resulting relative gain (%) in a separate row. For the 'Gain (%)' row, a higher value always indicates better performance. Positive gains are colored in blue, and degradations are in red.

Method		FID-10K ↓	Denoising from $t = 250$			Denoising from $t = 500$		
			SSCD ↓	SSCD ^{norm} ↓	Aesthetic ↑	SSCD ↓	SSCD ^{norm} ↓	Aesthetic ↑
Pre-trained		30.3	1.257	-	-	-	-	-
Naive deletion		19.61	0.874	0.607	6.118	0.726	0.641	6.077
EraseDiff		117.81	0.133	0.551	3.702	0.096	0.517	4.359
GA	Base	359.79	0.131	0.548	3.079	0.002	0.783	3.470
	+ Ours	375.18	0.113	0.499	3.699	0.062	0.600	4.577
	Gain (%)	-4.28	+13.74	+8.94	+20.14	-3.10 ³	+23.37	+31.90
SISS	Base	23.42	0.336	0.430	4.094	0.299	0.501	4.845
	+ Ours	23.65	0.345	0.349	5.520	0.282	0.399	6.095
	Gain (%)	-0.98	-2.68	+18.84	+34.83	+5.69	+20.36	+25.80
DPO	Base	20.58	0.369	0.404	5.128	0.313	0.459	5.058
	+ Ours	21.15	0.344	0.332	5.614	0.292	0.393	6.138
	Gain (%)	-2.77	+6.78	+17.82	+9.48	+6.71	+14.38	+21.35
KTO	Base	23.48	0.363	0.366	5.355	0.289	0.442	5.879
	+ Ours	23.58	0.373	0.340	5.470	0.280	0.377	6.274
	Gain (%)	-0.43	-2.75	+7.10	+2.15	+3.11	+14.71	+6.72

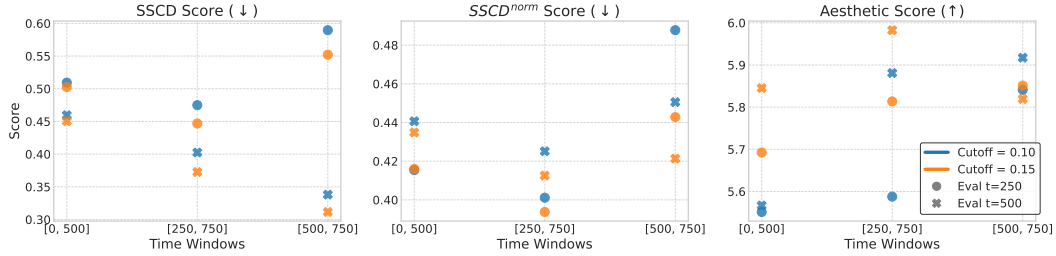


Figure 7: Image-level unlearning with different diffusion time steps and cutoff for low-pass filter.

Measures. The traditional evaluation of unlearning methods is two-fold: the quality of retained data samples, measured by Frechet Inception Distance (FID) (Heusel et al., 2017), and the dissimilarity of forgotten samples from the originals, measured by SSCD (Pizzi et al., 2022). However, SSCD ignores the quality of unlearned samples, e.g., a blurry image achieves a very low SSCD, though it is not a meaningful outcome. To address this, we introduce a normalized SSCD score that considers directionality by projecting the difference between the generated image $\hat{\mathbf{x}}_0(\mathbf{x}_t, \theta)$ from the unlearned model θ and the original image $\mathbf{x}_0$, onto an ℓ_2 -bounded perturbation similar to the constraints in adversarial attacks (Madry et al., 2018):

$$\text{SSCD}^{\text{norm}} = \text{SSCD}(\mathbf{x}_0, \mathbf{x}_0 + \rho \frac{\hat{\mathbf{x}}_0(\mathbf{x}_t, \theta) - \mathbf{x}_0}{\|\hat{\mathbf{x}}_0(\mathbf{x}_t, \theta) - \mathbf{x}_0\|_2^2 + \varepsilon}). \quad (13)$$

Here, ε is a small constant to avoid division by zero, and ρ is the radius for the perturbation, where we use $\rho = 100$ with a normalized image with 256x256 resolution. For generated samples, we also compute sample-wise aesthetic scores using the LAION-Aesthetic V2 predictor¹ as a quality metric.

For Stable Diffusion, we evaluate generation quality using CLIP-IQA (Wang et al., 2023) and measure unlearning performance through the unlearning success rate, defined as the proportion of cases where all 16 memorized samples are successfully removed (Alberti et al., 2025).

¹<https://github.com/christophschuhmann/improved-aesthetic-predictor>

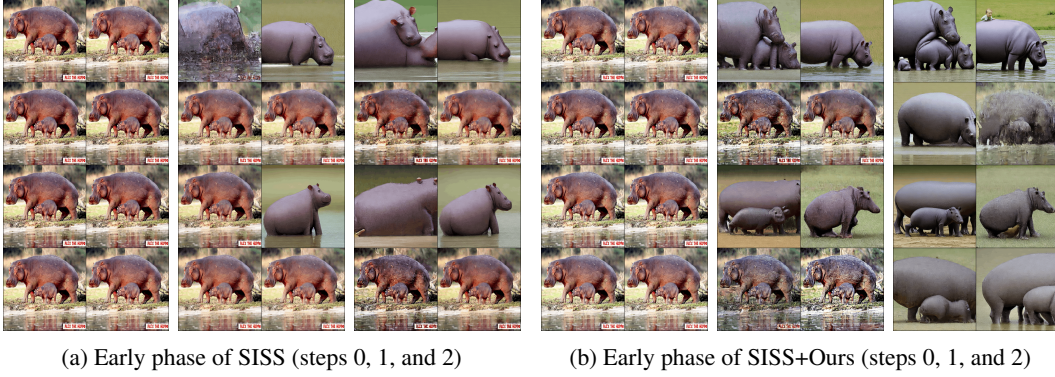


Figure 8: Visualization of generated images from the memorization with **fully-memorized prompt** “Mothers influence on her young hippo” as unlearning progresses. The proposed method shows faster forgetting performance while maintaining quality.

4.2 IMAGE-LEVEL DATA UNLEARNING

For CelebA-HQ, our objective is to delete six randomly sampled celebrity faces selected by (Alberti et al., 2025) from pre-trained unconditional DDPM models (Ho et al., 2020). Unlearning is applied image-by-image, and performance is evaluated after training. For SISS, DPO, and KTO, we use a single model in a continual setting, while GA erases all details after one deletion. Accordingly, we assess GA by measuring the unlearned model after each individual deletion. As explained in Section 3.1, we first inject diffusion noise to certain time steps $t = 250, 500$ and compare the reversed image using the unlearned model and the original image with SSCD score. While Alberti et al. (2025) only measured the exact memorization at $t = 250$, we also include the experiment of partial memorization at $t = 500$.

For selecting important time steps and frequency components, we first analyze which intervals are most critical for image-level data unlearning. In Figure 7, we examine three ranges of unlearning time steps: $[0, 500]$, $[250, 750]$, and $[500, 750]$. We select $[500, 750]$ because the late time steps $[750, 1000]$ often result in deleting nothing, as our evaluation begins from earlier steps $t = 250$ (or $t = 500$). Similar to the toy results in Figure 2, unlearning only early steps is not effective for removing individual information; instead, it reduces the aesthetic score, resulting in noisy but still memorized samples. Therefore, we focus on the middle and middle-late intervals, where the middle steps show relatively stable forgetting and better quality preservation in image-level data unlearning. Although unlearning middle-late steps can be advantageous at $t = 500$, we ultimately choose the middle range $[250, 750]$ for image-level data unlearning. For a low-pass filter, a lower cutoff threshold effectively reduces noisy artifacts in unlearned images. The removal of high-frequency components eliminates fragile details that are most closely tied to memorized data.

We apply our selective unlearning framework to various optimization methods (GA, SISS, DPO, and KTO), with results shown in Table 1. To ensure that a positive value consistently indicates an improvement over metric V , the relative gain (%) is calculated as $(V_{\text{ours}} - V_{\text{base}})/V_{\text{base}}$ for higher is better metric and $(V_{\text{base}} - V_{\text{ours}})/V_{\text{base}}$ for lower is better metric. The results indicate that both time and frequency selection achieve forgetting and preserve the quality of unlearned data examples. Overall, adding our method results in only a minor increase in FID for retained data samples, while significantly improving image quality, as indicated by the substantial gains in Aesthetic scores. Our framework demonstrates effective unlearning; at $t = 250$, it generally maintains the raw SSCD while improving $\text{SSCD}^{\text{norm}}$, and at $t = 500$, it improves both metrics compared to base models.

The comparison between SSCD and $\text{SSCD}_{\text{norm}}$ highlights the limitations of using raw similarity to evaluate unlearning. For instance, deletion-focused methods like EraseDiff and GA achieve low SSCD scores. While the results are well-aligned with target forgetting, it is merely an artifact of severe image quality degradation. Their high $\text{SSCD}_{\text{norm}}$ scores correctly reveal that their perturbation direction is ineffective for true unlearning. In contrast, our selective framework demonstrates a superior unlearning direction by achieving significant $\text{SSCD}_{\text{norm}}$ gains without this quality collapse.

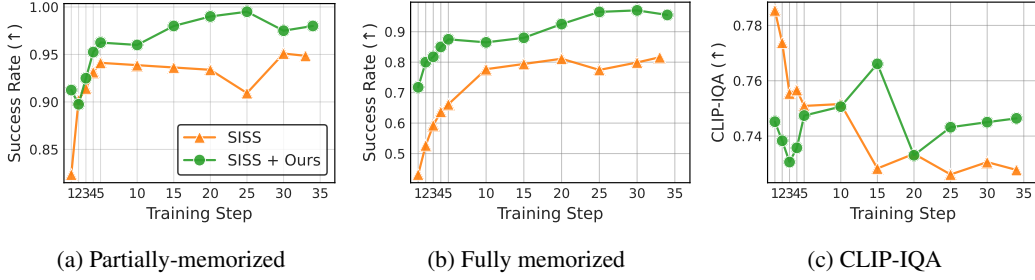


Figure 9: Unlearning success rate of 45 memorized prompts in Stable Diffusion. Our method shows faster unlearning convergence and a high unlearning success rate.

However, we also observe the failure cases in quality preservation within our framework, which leaves room for additional selective design.

4.3 TEXT-TO-IMAGE DATA UNLEARNING

For Stable Diffusion, we tested our strength on 45 memorized prompts within a specific target prompts selected by Alberti et al. (2025) from the LAION datasets. We performed unlearning experiments on specific time steps to identify the most effective range for Stable Diffusion. Interestingly, the results differ significantly from those in image-level unlearning cases. In the $[0, 750]$ range, we observed no significant changes to the forgotten images. However, we found that the later time steps, $[750, 1000]$, are more effective for data unlearning in the text-to-image setting. While we do not claim these specific time steps are universally optimal, we offer several insights to explain the difference. The generated images of SISS and SISS+Ours, where we use time steps $[750, 1000]$ and $r_t = 0.15$, are visualized in Figure 8. Refer to Appendix C for the visualization of other results.

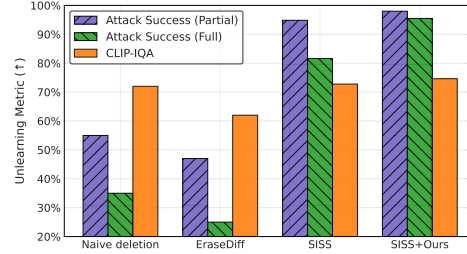


Figure 10: Comparison of unlearning methods in Stable Diffusion with CLIP-IQA for quality evaluation and attack success rate at the final step.

First, the sampling procedures for evaluation differ between image-level and text-level tasks. As explained previously, image-level reconstruction starts from intermediate steps like $t = 250$ or $t = 500$, where later steps are relatively impractical. In contrast, even when the objective is the same to delete a memorized example in Stable Diffusion, the sampling process begins from random noise ($t = 1000$), utilizing the condition on a text embedding of “memorized prompt” rather than a “memorized image” itself. Furthermore, because Stable Diffusion is trained on a vast number of image-text pairs, its generalization ability remains strong even if it loses information about a specific sample. Therefore, the effective time steps or low-level filter may vary based on the scenario.

Our experimental results, which assess both unlearning efficacy and image quality, are presented in Figure 9 and Figure 10. As shown in Figure 9, our method achieves a higher unlearning success rate and converges faster than the baseline SISS throughout the training process. Although the CLIP-IQA score temporarily decreases during the initial phase of rapid convergence, it remains at a high level overall. The final comparison in Figure 10 further confirms that our approach yields superior unlearning (lower attack success) and significantly better image quality (higher CLIP-IQA). These observations suggest that applying unlearning uniformly across all time steps and frequency components is indeed ineffective for unlearning.

5 CONCLUSION

In this paper, we address the critical issues of quality degradation and slow convergence in data unlearning for diffusion models. Our analysis reveals that forgetting is not a uniform process, but

an effective region exists across time steps and frequency domains. Based on this insight, we introduce a novel time-frequency selective framework that targets specific diffusion steps and frequency components. Our work provides a practical path toward high-fidelity machine unlearning that is compatible with existing methods. As a limitation, this work does not provide a method for automatically discovering the optimal unlearning regions. A promising direction for future research is to develop adaptive strategies that can dynamically identify the most effective components to unlearn.

ETHICS STATEMENT

The primary motivation for this research is to enhance data privacy and empower individuals with control over their personal data, in alignment with regulatory principles such as the “right to be forgotten.” However, we acknowledge that the unlearning process itself is not safe. The process of removing data could potentially be attacked to infer information about the unlearned information, like re-learning attacks. We recognize that robust safeguards and further research are necessary to mitigate such privacy leakages during the unlearning procedure.

REPRODUCIBILITY

Please refer to Appendix A for experimental settings. Our code is built upon the official PyTorch implementation of SISS (Alberti et al., 2025), and our source code will be released publicly.

THE USE OF LARGE LANGUAGE MODELS (LLMS)

The authors only use LLMs for the purpose of grammar correction and code modification.

REFERENCES

- Silas Alberti, Kenan Hasanaliyev, Manav Shah, and Stefano Ermon. Data unlearning in diffusion models. In *The Thirteenth International Conference on Learning Representations*, 2025.
- Youssef Allouah, Rachid Guerraoui, and Sanmi Koyejo. Distributional unlearning: Forgetting distributions, not just samples. *arXiv preprint arXiv:2507.15112*, 2025.
- Brian DO Anderson. Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, 12(3):313–326, 1982.
- Lucas Bourtole, 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.
- Jooyoung Choi, Jungbeom Lee, Chaehun Shin, Sungwon Kim, Hyunwoo Kim, and Sungroh Yoon. Perception prioritized training of diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11472–11481, 2022.
- Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. Model alignment as prospect theoretic optimization. In *Forty-first International Conference on Machine Learning*, 2024.
- Chongyu Fan, Jiancheng Liu, Yihua Zhang, Eric Wong, Dennis Wei, and Sijia Liu. Salun: Empowering machine unlearning via gradient-based weight saliency in both image classification and generation. In *The Twelfth International Conference on Learning Representations*, 2024.
- 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.
- Kristian Georgiev, Joshua Vendrow, Hadi Salman, Sung Min Park, and Aleksander Madry. The journey, not the destination: How data guides diffusion models. *arXiv preprint arXiv:2312.06205*, 2023.

- 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.
- Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.
- 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.
- Mingyu Kim, Dongjun Kim, Amman Yusuf, Stefano Ermon, and Mijung Park. Training-free safe denoisers for safe use of diffusion models. *arXiv preprint arXiv:2502.08011*, 2025.
- Michael Kirchhof, James Thornton, Louis Béthune, Pierre Ablin, Eugene Ndiaye, and marco cuturi. Shielded diffusion: Generating novel and diverse images using sparse repellency. In *Forty-second International Conference on Machine Learning*, 2025.
- Felix Koulischer, Johannes Deleu, Gabriel Raya, Thomas Demeester, and Luca Ambrogioni. Dynamic negative guidance of diffusion models. In *The Thirteenth International Conference on Learning Representations*, 2025.
- Shufan Li, Konstantinos Kallidromitis, Akash Gokul, Yusuke Kato, and Kazuki Kozuka. Aligning diffusion models by optimizing human utility. *Advances in Neural Information Processing Systems*, 37:24897–24925, 2024.
- Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.
- Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.
- Soumyadeep Pal, Changsheng Wang, James Diffenderfer, Bhavya Kailkhura, and Sijia Liu. LLM unlearning reveals a stronger-than-expected coreset effect in current benchmarks. In *Second Conference on Language Modeling*, 2025.
- Yong-Hyun Park, Sangdoo Yun, Jin-Hwa Kim, Junho Kim, Geonhui Jang, Yonghyun Jeong, Junghyo Jo, and Gayoung Lee. Direct unlearning optimization for robust and safe text-to-image models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL <https://openreview.net/forum?id=UdXE5V2d00>.
- Mansheej Paul, Surya Ganguli, and Gintare Karolina Dziugaite. Deep learning on a data diet: Finding important examples early in training. *Advances in neural information processing systems*, 34: 20596–20607, 2021.
- Ed Pizzi, Sreya Dutta Roy, Sugosh Nagavara Ravindra, Priya Goyal, and Matthijs Douze. A self-supervised descriptor for image copy detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14532–14542, 2022.
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.
- Gabriel Raya and Luca Ambrogioni. Spontaneous symmetry breaking in generative diffusion models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.
- 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.

- Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3, 2025. URL <https://arxiv.org/abs/2508.10104>.
- Raghav Singhal, Zachary Horvitz, Ryan Teehan, Mengye Ren, Zhou Yu, Kathleen McKeown, and Rajesh Ranganath. A general framework for inference-time scaling and steering of diffusion models. In *Forty-second International Conference on Machine Learning*, 2025.
- Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2020.
- Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.
- Koushik Srivatsan, Fahad Shamshad, Muzammal Naseer, Vishal M Patel, and Karthik Nandakumar. Stereo: A two-stage framework for adversarially robust concept erasing from text-to-image diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 23765–23774, 2025.
- Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8228–8238, 2024.
- Fu-Yun Wang, Yunhao Shui, Jintan Piao, Keqiang Sun, and Hongsheng Li. Diffusion-NPO: Negative preference optimization for better preference aligned generation of diffusion models. In *The Thirteenth International Conference on Learning Representations*, 2025a.
- Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 2555–2563, 2023.
- Kai Wang, Mingjia Shi, Yukun Zhou, Zekai Li, Zhihang Yuan, Yuzhang Shang, Xiaojiang Peng, Hanwang Zhang, and Yang You. A closer look at time steps is worthy of triple speed-up for diffusion model training. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 12934–12944, 2025b.
- Jing Wu, Trung Le, Munawar Hayat, and Mehrtash Harandi. Erasing undesirable influence in diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 28263–28273, 2025.
- Jaehong Yoon, Shoubin Yu, Vaidehi Patil, Huaxiu Yao, and Mohit Bansal. SAFREE: Training-free and adaptive guard for safe text-to-image and video generation. In *The Thirteenth International Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=hgTFotBRK1>.
- Chaoshuo Zhang, Chenhao Lin, Zhengyu Zhao, Le Yang, Qian Wang, and Chao Shen. Concept unlearning by modeling key steps of diffusion process. *arXiv preprint arXiv:2507.06526*, 2025.
- Ruiqi Zhang, Licong Lin, Yu Bai, and Song Mei. Negative preference optimization: From catastrophic collapse to effective unlearning. In *First Conference on Language Modeling*, 2024a.
- 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. *Advances in neural information processing systems*, 37:36748–36776, 2024b.
- Mengnan Zhao, Lihe Zhang, Xingyi Yang, Tianhang Zheng, and Baocai Yin. Advanchor: Enhancing diffusion model unlearning with adversarial anchors. *arXiv preprint arXiv:2501.00054*, 2024.

A EXPERIMENTAL DETAILS

Experimental Setups. We basically follow the experimental setups in SISS (Alberti et al., 2025). We first make note of an important detail in their paper.

All the diffusion models were provided in the Huggingface `diffusers` library with a U-Net backbone. For the CelebA-HQ dataset, we used a pretrained checkpoint released by (Ho et al., 2020) at <https://huggingface.co/google/ddpm-celebahq-256>. For the Stable Diffusion experiments, we used version 1.4 at <https://huggingface.co/CompVis/stable-diffusion-v1-4> as the pretrained checkpoint and used a 50-step DDIM sampler (Song et al., 2020).

Regarding hyperparameters, we follow (Alberti et al., 2025) without additional tuning. For the proposed method, we fix the time suppression intensity at $k = 0$ in Equation (9) and the high-frequency suppression weight at $s = 0$ in Equation (12) to simplify the design. Consequently, we only adjust the key time steps $[t_1, t_2]$ with $t_1 < t_2$ by selecting two values in $[0, 250, 500, 750, 1000]$ and the cutoff frequency radius of the FFT low-pass filter, $r_t \in [0.05, 0.1, 0.15, 0.2]$. Unless otherwise specified, we set $r_t = 0.15$ as it consistently shows good performance.

For the experimental setups, we mainly use NVIDIA A40 GPUs with the PyTorch library and utilize NVIDIA A100 GPUs for parallel runs.

Memorized Prompts of Stable Diffusion. Alberti et al. (2025) constructed 45 prompts in Stable Diffusion 1.4. Since only one LAION image corresponds to each prompt, synthetic datasets are generated by sampling 128 images and applying k-means clustering for classification. A “fully-memorized” prompt refers to a case where Stable Diffusion repeatedly reproduces the same outcome, whereas a “partially-memorized” prompt is obtained by manually adding or deleting tokens, producing more diverse outputs that are easier to unlearn.

Preference Optimization. Direct Preference Optimization (DPO) (Rafailov et al., 2023) is widely used to evaluate preference alignment in language models. Originally developed for Reinforcement Learning from Human Feedback (RLHF), it has recently also been applied to diffusion fine-tuning (Wallace et al., 2024). For unlearning, Negative Preference Optimization (NPO) (Zhang et al., 2024a) has been proposed as an alternative to gradient descent. Unlike gradient ascent (Zhang et al., 2024a), NPO leverages the initial model as a reference point, which helps mitigate overfitting by keeping the optimization close to the initialization. In the diffusion setting, Wang et al. (2025a) applied NPO for alignment. Another line of work, Kahneman–Tversky Optimization (KTO) (Ethayarajh et al., 2024), has its own strength since the method does not require positive–negative pairs. Li et al. (2024) extended KTO to diffusion for pair-free feedback alignment. In this paper, we follow the formulation of forget loss with DPO (Wallace et al., 2024) and KTO (Li et al., 2024).

B ADDITIONAL EXPERIMENTS

We illustrate some details through a toy example. For Figure 2, we construct a multi-layer perceptron diffusion model trained with the DDPM objective (Ho et al., 2020). The model is trained until convergence. After training, we delete the forget samples ($\times$) using gradient ascent, while applying gradient descent to certain retain data points ($\blacklozenge$).

Except for Figure 2, all experiments in Section 3 are conducted on the CelebHQ dataset used in the main table 1. We now provide additional results. In Figure 3b, we utilize the DINOv3 embedding (Siméoni et al., 2025) to calculate the similarity between data samples. We also measured in pixel-wise similarity in Figure 12, where we failed to observe similar patterns at the embedding level. The nearest images of individual data samples are in Figure 11. Individual gradient norms of Figure 4b are in Figure 13. Finally, we report the remaining results of Figure 5 in Figures 14 and 15. Furthermore, we clarify that Figures 3a and 4a are slight modifications of figures from (Raya & Ambrogioni, 2023) and (Georgiev et al., 2023), respectively, which we adjusted to align their time steps with our analysis.

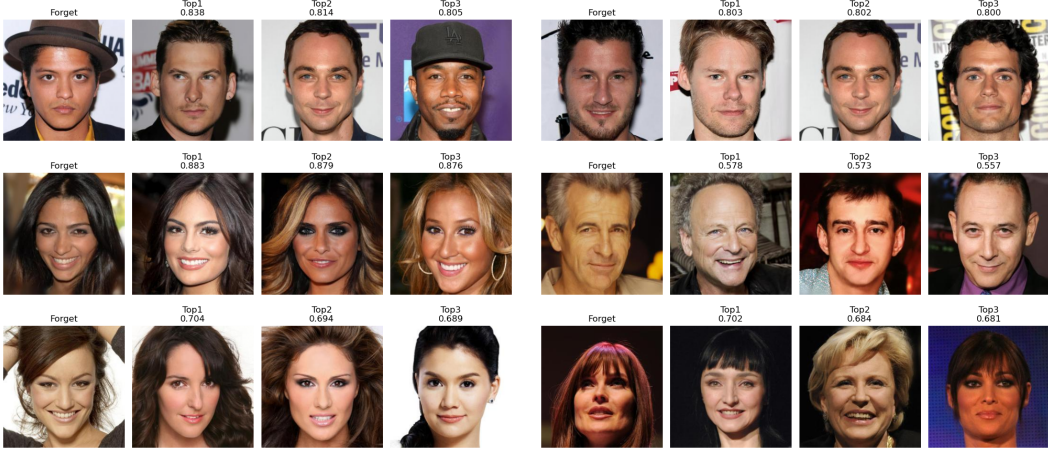


Figure 11: Forget samples and their nearest neighborhood on DINOv3 embedding (Siméoni et al., 2025).

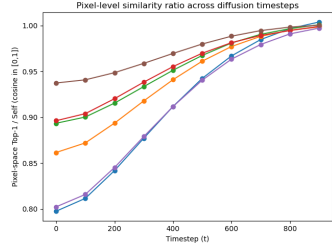


Figure 12: Similarity in pixel-level.

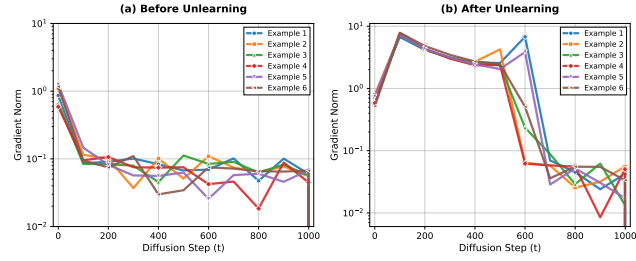


Figure 13: Gradient norms for each data sample before and after unlearning.

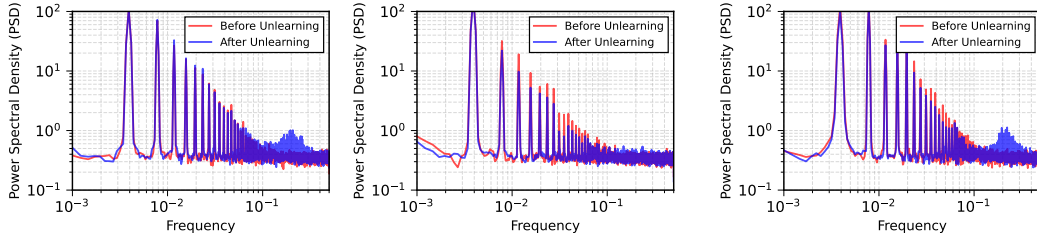


Figure 14: Additional results of Power Spectral Density (PSD) before and after unlearning on the forget dataset.

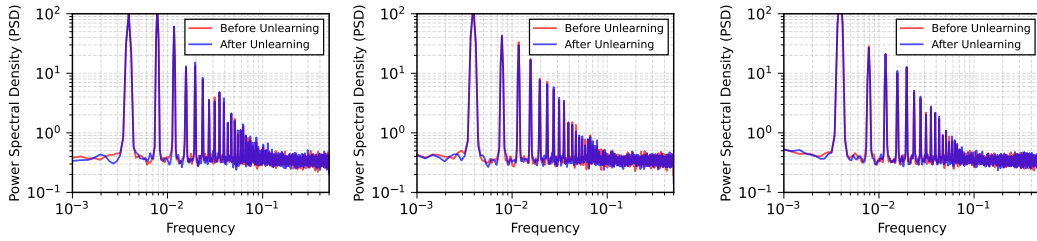


Figure 15: Additional results of Power Spectral Density (PSD) before and after unlearning on the retain dataset.

C VISUALIZATION

CelebA-HQ Results. We first visualize the results for each method: SISS, SISS+Ours, and variational results on different time steps in Figures 16 to 20.



Figure 16: Unlearning with SISS.



Figure 17: Unlearning with SISS+Ours.

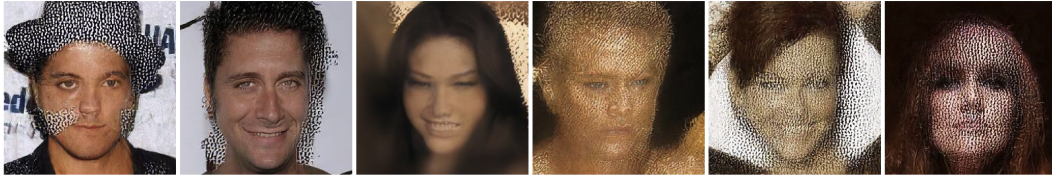


Figure 18: Unlearning on time steps [0, 500].



Figure 19: Unlearning on time steps [500, 1000].



Figure 20: Unlearning on time steps [750, 1000].

Stable Diffusion Results. We visualize the results of text-to-image data unlearning in Figures 21 to 24.



Figure 21: Visualization of (left) Partially-memorized and (right) Fully-memorized results after unlearning of the prompt "Living in the Light with Ann Graham Lotz".



Figure 22: Visualization of (left) Partially-memorized and (right) Fully-memorized results after unlearning of the prompt "Living in the Light with Ann Graham Lotz" using different time steps for unlearning.



Figure 23: Visualization of (left) Partially-memorized and (right) Fully-memorized results after unlearning of the prompt "Rambo 5 und Rocky Spin-Off - Sylvester Stallone gibt Updates".

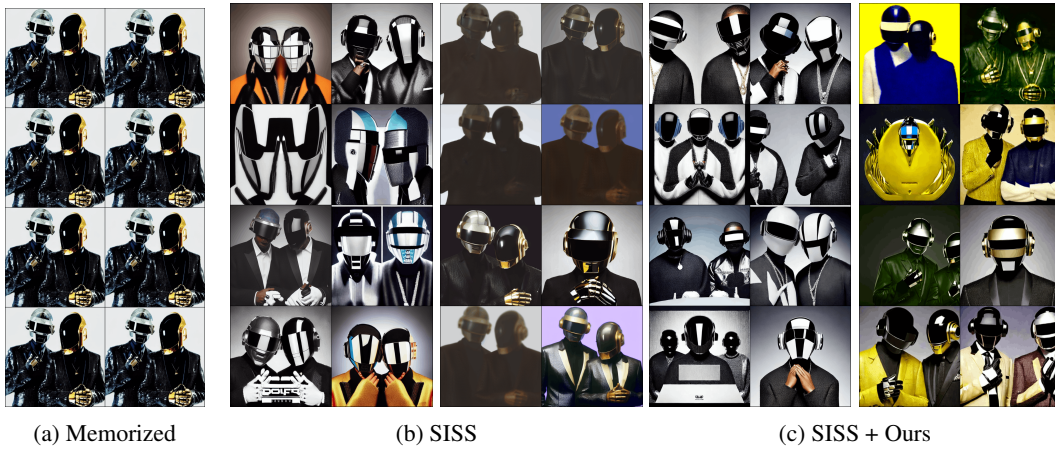


Figure 24: Visualization of (left) Partially-memorized and (right) Fully-memorized results after unlearning of the prompt "Daft Punk, Jay Z Collaborate on { "Computerized" }".