NOT EVERY TIME AND FREQUENCY NEED TO BE FORGOTTEN IN DIFFUSION UNLEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Data unlearning aims to remove the influence of specific training samples from a trained model. In fine-tuning methods, data unlearning relies primarily on loss maximization over forget samples, which often leads to quality degradation or incomplete forgetting. Existing methods perform unlearning uniformly across diffusion stages, ignoring diffusion dynamics from noise to data. Our systematic study of diffusion phases shows that forgetting in diffusion models is *uneven* across time and frequency, with theoretical justification of distributive distortion and forgetting-utility trade-off. By selectively forgetting time and frequency in diffusion models, we achieve both higher unlearning success rates and improved generation quality across diverse settings, including both conditional and unconditional scenarios. We also introduce an improved SSCD metric that measures dissimilarity using a normalized perturbation distance. Together, we provide practical insights for understanding and improving data unlearning in diffusion models.

1 INTRODUCTION

The ability to remove the influence of some 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 diffusion models (Ho et al., 2020), unlearning unsafe concepts or memorization has been actively explored through fine-tuning (Wang et al., 2025) and training-free sampling (Kim et al., 2025; 2026).

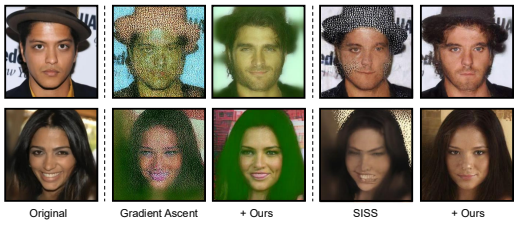


Figure 1: Illustration of quality degradation of unlearned images after data unlearning. **Left:** two images to unlearn. **Middle:** Gradient Ascent (GA, first column) and GA with our approach (second column). **Right:** SISS (Alberti et al., 2025) (first) and SISS with our approach (second). Our approach improves the quality of unlearning.

In machine unlearning with generative models, two major objectives exist: (a) concept or class unlearning, which refers to preventing the generation of a particular concept or class in samples (Fan et al., 2024), and (b) data unlearning, which focuses on removing individual samples rather than concepts. As noted by Alberti et al. (2025), data unlearning is less studied than concept unlearning. Furthermore, removing a single individual while preserving the generation of similar outputs is more challenging than deleting the entire mode of similar examples. Finetuning-based unlearning methods, such as gradient ascent (Alberti et al., 2025) or negative preference optimization (Wang et al., 2025), rely on loss maximization objectives. The generated images often become noisier and aesthetically degraded, while forgetting might not be complete nor precise, as shown in Figure 1.

To mitigate the drawbacks of the data unlearning in diffusion models, we analyze the dynamics of data unlearning. We argue that forgetting does not occur uniformly, but rather unevenly across time and spectral domains. Diffusion models learn phase-dependent behaviour during training (Choi et al., 2022): later diffusion time steps close to Gaussian noise capture coarse semantics, while earlier time steps close to the data refine fine-grained details. However, previous unlearning methods

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

treat all diffusion time steps equally, potentially leading to the inadvertent forgetting of unintended details, as illustrated in Figure 2. Furthermore, we investigate the spectral domain to selectively remove unwanted high-frequency components in the forget set. Overall, our framework effectively implements forgetting and preserves sample quality by focusing on selective components in both the time and frequency domains. We summarize our contributions below:

- We provide a *systematic* study of data unlearning in diffusion models in time and spectral space, considering both conditional and unconditional scenarios.
- Our time- and frequency-selective unlearning approach is *simple, compatible* with existing unlearning objectives, and performs effective unlearning across various settings.
- We devise an *improved* metric, $\text{SSCD}^{\text{norm}}$. Unlike the original SSCD (Pizzi et al., 2022), which only measures dissimilarity, $\text{SSCD}^{\text{norm}}$ standardizes perturbation distance.

2 Background

2.1 Diffusion Models

Diffusion models (Ho et al., 2020) aim to generate samples from an unknown data distribution by considering forward and reverse processes. The forward process gradually injects noise into samples $\mathbf{x}_0$ drawn from the data distribution p_0, forwarding them into a fixed standard Gaussian distribution $p_T = \mathcal{N}(\mathbf{0}, \boldsymbol{I})$. The corresponding forward stochastic differential equation (SDE) is:

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

where $f : \mathbb{R}^d \times [0, T] \to \mathbb{R}^d$ represents the drift function, $g : [0, T] \to \mathbb{R}$ 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:

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

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

$$\mathbb{E}_{\mathbf{x}_0 \sim p_0} \mathbb{E}_{t \in [0,T]} \Big[w_t \big\| s_\theta(\mathbf{x}_t, t) - \nabla_{\mathbf{x}_t} \log q(\mathbf{x}_t|\mathbf{x}_0) \big\|_2^2 \Big]. \tag{3}$$

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 as concept unlearning and data unlearning. Concept unlearning (Gandikota et al., 2023) 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 (Gandikota et al., 2023; Srivatsan et al., 2025; Park et al., 2024).

Data unlearning, on the other hand, is closely related to individual data memorization and aims to remove specific data examples in accordance with the "Right to be Forgotten" by the aforementioned GDPR. 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, diffusion unlearning $\min_\theta \mathcal{L}_{\text{UL}}(\theta)$ is

$$\min_\theta \big(-\mathcal{L}_{\mathcal{D}_\mathcal{F}}(\theta) + \mathcal{L}_{\mathcal{D}_\mathcal{R}}(\theta) \big). \tag{4}$$

Gradient ascent (GA) or negative gradient on forget data samples $\mathcal{D}_F$ is a key component for data unlearning (Golatkar et al., 2020). For GA, the loss function on the forget set $\mathcal{L}_{\text{GA}}(\theta)$ is

$$-\mathbb{E}_{\mathbf{x}_0 \sim \mathcal{D}_F} \mathbb{E}_{t \in [0,T]} \Big[w_t \big\| s_\theta(\mathbf{x}_t, t) - \nabla_{\mathbf{x}_t} \log q(\mathbf{x}_t|\mathbf{x}_0) \big\|_2^2 \Big]. \tag{5}$$

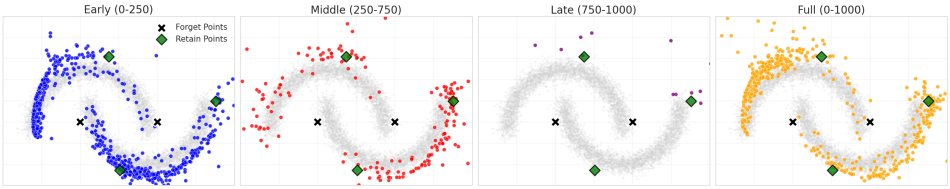


Figure 2: Illustration of unlearning a specific stage of time steps. See Section 3.1 for details.

EraseDiff (Wu et al., 2025) replaces the true score of the forget samples with a randomly sampled noisy image. SISS (Alberti et al., 2025) investigated the unlearning as a mixture distribution of forget and retain data with importance sampling. Previous works investigate how to formulate Equation (4).

2.3 Evaluation of Data Unlearning in Diffusion Models

We follow the evaluation of Alberti et al. (2025) based on unconditional or conditional scenarios.

Memorization on Unconditional Reconstruction. In unconditional generation, memorization is assessed by the diffusion model's ability to reconstruct the training data. In particular, we noise a clean image $\mathbf{x}_0$ via the DDPM forward process to an intermediate time step $\mathbf{x}_t$ (e.g., $t = 250$) where the model can easily reconstruct, and measure the similarity between the original $\mathbf{x}_0$ and the reconstructed $\hat{\mathbf{x}}_0$ using model θ from $\mathbf{x}_t$.

Memorization on Conditional Guidance. In conditional generation, conditioning on text or class embeddings can lead to stronger overfitting to the conditional distribution over the images given those conditions (Zhai et al., 2024). Therefore, measuring the effects of condition $\epsilon_\theta(\mathbf{x}_t, e_p)$ or its guidance norm $\|\epsilon_\theta(\mathbf{x}_t, e_p) - \epsilon_\theta(\mathbf{x}_t, e_\emptyset)\|_2^2$ is crucial to track conditional memorization (Wen et al., 2024; Jain et al., 2025; Alberti et al., 2025), with conditional embedding e_p and null embedding $e_\emptyset$. Unlike unconditional memorization, its evaluation starts by sampling from $t = T = 1000$ with conditions.

3 Time and Frequency Selection for Data Unlearning in Diffusion

In this section, we first empirically investigate the effects of time steps and frequency on diffusion unlearning and propose a selective framework for data unlearning, followed by a theoretical analysis.

3.1 Not Every Time Step Needs to Be Forgotten

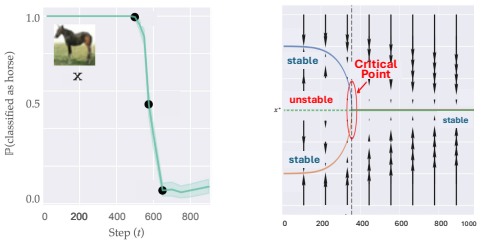
Several theoretical papers analyze the dynamics of diffusion models (Li & Chen, 2024; Biroli et al., 2024; Sclocchi et al., 2025; Li et al., 2025). Starting from the Gaussian noise, diffusion models learn distinct attributes across different phases. **Phase I** involves learning coarse features at late time steps near time T. **Phase II** involves generating content at intermediate steps, and **Phase III** involves refining details toward convergence at early time steps near $t = 0$ (Choi et al., 2022). Biroli et al. (2024) describes the diffusion dynamics in terms of these three distinct phases divided by two important transition points in time: (a) **speciation** time t_S; and (b) **collapse** time t_C.

The denoising trajectories begin to separate by semantic class at the **speciation time** t_S (phase I-II), for example via embedding similarity or consistency scores. For $t > t_S$, trajectories are nearly class-agnostic noise, whereas for $t < t_S$, they encode semantic identity. Afterwards, trajectories converge toward individual training samples at the **collapse time** t_C (phase II-III), which quantifies instance-level attraction, such as nearest-neighbor consistency. For $t < t_C$, diffusion paths are drawn to a single point and lose their generalization capability, indicating memorization.

Table 1: Summary of diffusion transition points t_S and t_C ($T = 1000$) across existing studies. † denotes intervals.

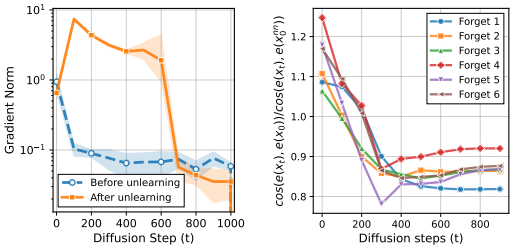
Reference	Basis of Analysis	t_S (phase I-II)	t_C (phase II-III)
Biroli et al. (2024)†	Spectral / Entropy	$[500, 800]$	$[100, 250]$
Choi et al. (2022)	SNR $(10^{-2}, 10^0)$	675	259
Yang et al. (2024)	Structure Pruning	750	250
Kim et al. (2025)	Conditional Sampling	780	–
Jain et al. (2025)†	Conditional Sampling	$[600, 800]$	–

Optimal Transition Points. Finding the optimal t_S, t_C is an open question in diffusion models. We summarize previous re-



(a) Speciation time t_S (Phase I-II) (Georgiev et al., 2023) (b) Collapse time t_C (Phase II-III) (Raya & Ambrogioni, 2023)

Figure 3: Empirical evidence on transition points.



(a) Gradient norm of forget set during unlearning (b) Embedding similarity of $\mathbf{x}_t$ w.r.t. $\mathbf{x}_0$ and $\mathbf{x}_0^{\text{nn}}$

Figure 4: Analysis of diffusion time steps.

search in diverse views of diffusion models in Table 1. In Figure 3[1], Georgiev et al. (2023) observed the likelihood that the noisy data $\mathbf{x}_t$ is classified into a certain class at speciation time t_S. Also, Raya & Ambrogioni (2023) observed that data converges to a stable point, indicating collapse time t_C.

To investigate data unlearning in the aforementioned view, we will exploit the use of the transition points in data unlearning. We conduct a *toy experiment* with a shallow diffusion model on a half-moon dataset, as visualized in Figure 2. We only change the unlearning time steps while using the same data samples, with the loss function of Equation (5). Deleting only the early time steps (phase III) does not help forget data samples; instead, it leads to quality degradation in real samples. Unlearning the middle time steps (phase II) is most aligned with the instance-level attraction before reaching the collapse time t_C. For stronger forgetting of semantic conditioning, targeting the late time steps (phase I) preceding speciation at t_S is effective. t_S and t_C are aligned with $(t_C, t_S) = (250, 750)$ in Table 1.

3.1.1 Time-selective Diffusion Forgetting

Based on the previous analysis, we suggest a simple framework of selective time steps as

$$\mathbb{E}_{\mathbf{x}_0 \sim p_0} \underbrace{\mathbb{E}_{t \sim P(t)}}_{\text{Time selection}} \left[w_t \| s_\theta(\mathbf{x}_t, t) - \nabla_{\mathbf{x}_t} \log q(\mathbf{x}_t | \mathbf{x}_0) \|^2 \right]. \tag{6}$$

Here, $P(t)$ is a time step weighting. 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:

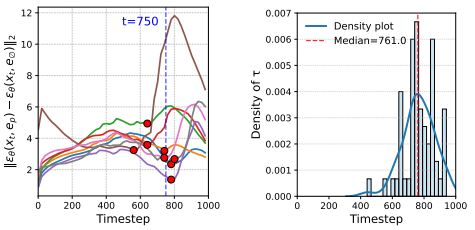
$$P(t) = \begin{cases} (1-k)/(t_2 - t_1), & t_1 \leq t \leq t_2, \\ k/(T - (t_2 - t_1)), & \text{otherwise}, \end{cases} \tag{7}$$

where $0 \leq k \leq 1$ is the suppression intensity and $t_1, t_2 \in (0, T]$ with $t_1 < t_2$ and we use $k = 0$ for simplicity. As explained in Section 2.3, the evaluation scenarios for unconditional and conditional settings differ significantly. Therefore, we investigate appropriate time steps $P(t)$ in Equation (7).

Unconditional Unlearning. The gradient norm $\|\nabla_\theta \mathcal{L}(\mathbf{x}; \theta)\|_2^2$ 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. In Figure 4a, to compare the actual changes after standard unlearning with uniform time steps in the image-level unlearning with CelebA-HQ, we unlearn a pre-trained model with SISS. The gradient norm increases significantly after unlearning, with the gap being most pronounced in the middle time steps (Phase II). The gradient gap indicates that unlearning induces a substantial shift in the model's parameter space during Phase II.

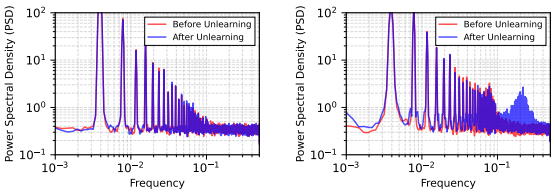
Furthermore, as shown in Figure 4b, when we calculate the similarity on DINOv3 (Siméoni et al., 2025) of the noisy data $\mathbf{x}_t$ between other clean data samples, the similarity towards the original data $\mathbf{x}_0$ is drastically increased in the refinement stage, compared to other nearest samples. Thus, to prevent the diffusion trajectory from heading toward a memorization point, we need to control the time steps preceding the memorization at t_C. Note that the evaluation scheme also compares the difference between the original and reconstructed images from intermediate steps $t = 250, 500$.

[1]Figure 3 is a slight modification of figures from (Raya & Ambrogioni, 2023) and (Georgiev et al., 2023).



(a) Guidance norm (b) Distribution of τ

Figure 5: Adaptive conditional memorization.



(a) Retain samples (b) Over-forgotten samples

Figure 6: Power spectral density before and after unlearning. Over-forgotten samples differ in high frequencies, while retain samples remain unchanged.

Conditional Unlearning. In conditional memorization, the effect of conditional guidance, such as class or text embeddings, acts more strongly than its unconditional counterpart (Zhai et al., 2024). Therefore, calculating the guidance magnitude $\|\epsilon_\theta(\mathbf{x}_t, e_p) - \epsilon_\theta(\mathbf{x}_t, e_\emptyset)\|_2^2$ is key for conditional memorization (Wen et al., 2024; Jeon et al., 2025). In Figure 5a, we measured the trajectory of conditional guidance with memorized prompts in Stable Diffusion v1.4 (Rombach et al., 2022). We can observe a peak in guidance norms in phase I before speciation begins (Jain et al., 2025).

Based on the experiments, we can hypothesize that controlling the diffusion trajectory for $t > t_S$ before speciation begins is important for the conditional case. Similarly, training-free methods observed that guiding only late time steps $t > t_S$ is sufficient to steer the generation away from harmful concepts. For example, Kim et al. (2025) applied guidance within $t > 780$, while Kirchhof et al. (2025) noted that their repellency terms are significant in late time steps.

Adaptive Time Selection. Since the guideline in Table 1 is not sample-wise, we further consider an propmt-adaptive method on conditional unlearning based on Jain et al. (2025), analogue to the speciation time t_S: (i) the guidance magnitude $\|\epsilon_\theta(\mathbf{x}_t, e_p) - \epsilon_\theta(\mathbf{x}_t, e_\emptyset)\|_2^2$ is computed, and (ii) set the transition point τ as the first local minimum observed in sampling, detailed in Algorithm 1 (Appendix B). As the τ searching method is prompt-adaptive, Figure 5b shows the optimal τ varies across prompts, the median of τ is 761.1, similar to $t_S = 750$.

In summary, $t_C < t < t_S$ (phase II) is critical for controlling unconditional behavior, while $t > t_S$ (phase I) matters for conditional generation, where conditional guidance dominates the diffusion dynamics after speciation, and the trajectory should be regulated before each critical time step begins.

3.2 Not Every Frequency Mode Needs to Be Forgotten

From a time-selective perspective, Phase I ($t > t_S$) and Phase II ($t_C < t < t_S$) represent semantic- and individual-level memorization are established, respectively. Maintaining time steps $t < t_C$ (Phase III) is essential for quality preservation because this phase only refines the quality of the generated sample determined in earlier stages. This motivates the use of low-pass filters in diffusion models to avoid forgetting fine-grained and high-frequency details (Lian et al., 2025; Choi et al., 2025).

We evaluate the impact of unlearning in the frequency domain via the power spectral density in Figure 6. For the retain set, the high-frequency distribution is preserved post-unlearning. Conversely, the forget set exhibits *unwanted high-frequency components* that risk degrading sample quality.

3.2.1 Frequency-selective Diffusion Forgetting

Thus, to preserve fine-grained image quality and make less perturbation in the manifold during unlearning, we add a low-pass filter to remove the high-frequency components of the forget set as

$$\mathbb{E}_{\mathbf{x}_0 \sim p_0} \mathbb{E}_t \Big[w_t \big\| \underbrace{\mathcal{F}\big(s_\theta(\mathbf{x}_t, t) - \nabla_{\mathbf{x}_t} \log q(\mathbf{x}_t \mid \mathbf{x}_0)\big)}_{\text{Low-pass filter}} \big\|^2 \Big] \tag{8}$$

Notice that the denoiser input remains the noisy natural sample $\mathbf{x}_t$ during unlearning; **only the loss residual is projected to low frequencies**. Hence, the input distribution during unlearning and inference is identical, and only the gradient signal used for unlearning is projected.

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

$$\mathcal{T}(u,v) = \text{FFT}(\mathbf{x}) = \sum_{x=1}^{H}\sum_{y=1}^{W} \mathbf{x}(x,y) e^{-j2\pi\left(\frac{u}{H}x + \frac{v}{W}y\right)}, \tag{9}$$

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. 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 with radius r as $\mathcal{F}(\mathbf{x}_{i,t}) = \text{IFFT}\left(\text{FFT}(\mathbf{x}_{i,t}) \odot \beta_{i,t}(r)\right)$, where $\beta_{i,t}(r) = s$ if the radius $r > r_t, 1$ otherwise. r_t is a radius cutoff frequency threshold. $0 \leq s \leq 1$ is a high frequency weight, where we set $s = 0$ to exclude high frequency during unlearning.

3.3 Theoretical Justification

Data unlearning must remove the influence of specific samples while preserving the overall data manifold. To this end, we analyze the enhanced forgetting within the selected region and the minimized distortion in the non-selected region, showing the strength in the forgetting–preservation trade-off.

Setup. Let $\mathbb{P}_{\text{orig}}$ and $\mathbb{P}_m$ ($m \in \{\text{full}, \text{sel}\}$) be the path measures of the original and unlearned diffusion models, with score perturbation $\Delta s_m(x_t,t) := s_{\theta_m}(x_t,t) - s_{\theta_{\text{orig}}}(x_t,t)$. By Girsanov's theorem (Liptser & Shiriaev, 1977), the KL divergence between the original and unlearned path is

$$D_{\text{KL}}(\mathbb{P}_{\text{orig}} \,\|\, \mathbb{P}_m) = \frac{1}{2}\int_0^T g(t)^2 \mathbb{E}_{x_t}\left[\|\Delta s_m(x_t,t)\|_2^2\right] dt. \tag{10}$$

To evaluate unlearning efficacy, we decompose this total unlearning-induced distortion into a selected target window $\mathcal{S}$ (e.g., selected time-steps or specific frequencies) and its complement $\mathcal{S}^c$. Let $D_{\text{KL}}^{\mathcal{W}}(\mathbb{P}_{\text{orig}} \,\|\, \mathbb{P}_m) := \frac{1}{2}\int_{\mathcal{W}} g(t)^2 \mathbb{E}_{x_t}[\|\Delta s_m\|_2^2] dt$ denote the regional divergence over a window $\mathcal{W}$. Thus, the total divergence can be formulated as $D_{\text{KL}}(\mathbb{P}_{\text{orig}} \,\|\, \mathbb{P}_m) = D_{\text{KL}}^{\mathcal{S}}(\mathbb{P}_{\text{orig}} \,\|\, \mathbb{P}_m) + D_{\text{KL}}^{\mathcal{S}^c}(\mathbb{P}_{\text{orig}} \,\|\, \mathbb{P}_m)$. For $\beta > 0$, we define the forgetting–preservation utility as:

$$\mathcal{J}_\beta(m) := D_{\text{KL}}^{\mathcal{S}}(\mathbb{P}_{\text{orig}} \,\|\, \mathbb{P}_m) - \beta D_{\text{KL}}^{\mathcal{S}^c}(\mathbb{P}_{\text{orig}} \,\|\, \mathbb{P}_m). \tag{11}$$

Note that $D_{\text{KL}}^{\mathcal{S}}(\mathbb{P}_{\text{orig}} \,\|\, \mathbb{P}_m)$ is a proxy for the *intended forgetting* and, $D_{\text{KL}}^{\mathcal{S}^c}(\mathbb{P}_{\text{orig}} \,\|\, \mathbb{P}_m)$ is a proxy for *collateral distortion* without intend to unlearn, detailed in Appendix A.

Assumptions. (i) *Finite divergence*: $D_{\text{KL}}(\mathbb{P}_{\text{orig}} \,\|\, \mathbb{P}_m) < \infty$. (ii) *Focused updates in $\mathcal{S}$*: Due to the frequent update on the selected region under the same training budget, selective unlearning amplifies changes in the target region ($\|\Delta s_{\text{sel}}\|_2 \geq \mu \|\Delta s_{\text{full}}\|_2$ for $\mu \geq 1$ in $\mathcal{S}$) and weaker changes in $\mathcal{S}^c$ ($\|\Delta s_{\text{sel}}\|_2 \leq \eta \|\Delta s_{\text{full}}\|_2$ for $\eta \in [0,1]$ in $\mathcal{S}^c$). See Appendix A for empirical validation and proofs.

Proposition 3.1. *Under the assumptions above, for every $\beta > 0$, and $\lambda := \frac{D_{\text{KL}}^{\mathcal{S}^c}(\mathbb{P}_{\text{orig}} \,\|\, \mathbb{P}_{\text{full}})}{D_{\text{KL}}(\mathbb{P}_{\text{orig}} \,\|\, \mathbb{P}_{\text{full}})} \in [0,1]$, the trade-off utility gap of $\mathcal{J}_\beta(\text{sel})$ and $\mathcal{J}_\beta(\text{full})$ satisfies:*

$$\mathcal{J}_\beta(\text{sel}) - \mathcal{J}_\beta(\text{full}) \geq \left[(\mu^2 - 1)(1-\lambda) + \beta(1-\eta^2)\lambda\right] D_{\text{KL}}(\mathbb{P}_{\text{orig}} \,\|\, \mathbb{P}_{\text{full}}), \tag{12}$$

where λ indicates the proportion of collateral distortion compared to the full unlearning.

Remark 1. *The bracketed term $(\mu^2 - 1)(1-\lambda) + \beta(1-\eta^2)\lambda$ is always non-negative under our assumptions ($\mu \geq 1, \eta \leq 1, \beta > 0, \lambda \in [0,1]$). Then, the improvement in utility is strictly positive.*

Remark 2. *The utility gap $\mathcal{J}_\beta(\text{sel}) - \mathcal{J}_\beta(\text{full}) \geq 0$ indicates our selective mechanism focuses score updates on the selected forget-region while mitigating unwanted distortion in non-selected parts.*

4 Experiments

4.1 Unconditional Data Unlearning

Experimental Setups. We follow the settings of Alberti et al. (2025). For CelebA-HQ (Liu et al., 2015), we evaluate the deletion of individual samples at the image level in unconditional image generation. The objective is to delete six randomly sampled celebrity faces selected by Alberti et al.

Table 2: Comparison of unconditional unlearning. 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 results. Positive gains are in blue, and degradations in red.

Method			Denoising from $t = 250$			Denoising from $t = 500$		
		FID-10K ↓	SSCD ↓	$\text{SSCD}^{\text{norm}}$ ↓	Aesthetic ↑	SSCD ↓	$\text{SSCD}^{\text{norm}}$ ↓	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	$-3{\cdot}10^{3}$	+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

(2025) from pre-trained unconditional diffusion models. Unlearning is applied image by image in a continual manner. For GA, we initialize the model as a pretrained point after each individual deletion.

Based on Table 1, we mainly use $(t_1, t_2) = (250, 750)$ for the unconditional case. Following Figure 6, as our goal is to filter out the high-frequency, we tune $r_t \in [0.1, 0.15]$, and unless otherwise specified, $r_t = 0.15$. We fine-tune the model with Adam using a batch size of 64 and a learning rate of $5 \cdot 10^{-6}$.

Evaluation. As explained in Section 2.3, we first inject diffusion noise at certain time steps $t = 250, 500$ (only $t = 250$ in (Alberti et al., 2025)) and compare the reversed image using the unlearned model and the original image. Evaluation is twofold: the quality of retained data samples, measured by 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 the generated 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 the normalized distance between the generated image $\hat{\mathbf{x}}_0(\mathbf{x}_t, \theta)$ and the original image $\mathbf{x}_0$. Motivated by the perturbation in adversarial attacks (Madry et al., 2018), we project the difference onto ℓ_2 ball with radius ρ as

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

Here, ε is a small constant to avoid division by zero, and we use $\rho = 100$ as a default with a $[-1, 1]$ scale normalized image with 256x256 resolution. As a quality metric, we compute sample-wise aesthetic scores using the LAION-Aesthetic V2 predictor[2]. We tested with **five runs** and report their mean value since variance is comparably small, shown in Table 14.

Main Results. We apply our selective unlearning framework to various optimization methods (GA, SISS, DPO (Rafailov et al., 2023; Wallace et al., 2024), and KTO (Ethayarajh et al., 2024; Li et al., 2024)) in Table 2. 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. At $t = 250$, it maintains the SSCD while improving $\text{SSCD}^{\text{norm}}$, and at $t = 500$, it improves both metrics compared to base models. Furthermore, we observe a minor increase in FID for retained data.

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

[2]https://github.com/christophschuhmann/improved-aesthetic-predictor

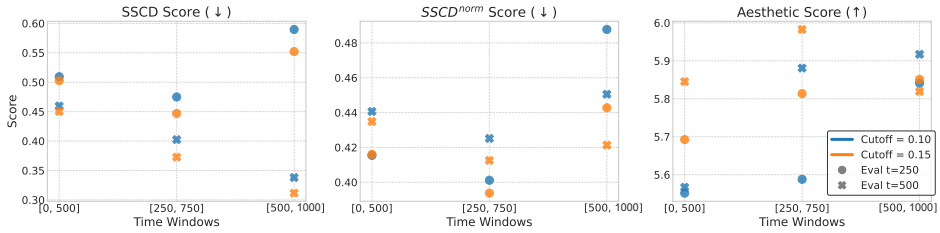


Figure 7: Ablation study with different diffusion time steps and cutoff for low-pass filter.

Table 3: Ablation studies on filter band (left) and unlearning time window (right), both applied on SISS. Lower is better for SSCD-based metrics; higher is better for Aesthetic.

(a) Filter band					**(b) Time window**				
	$t = 250$		$t = 500$			$t = 250$		$t = 500$	
Band	SSCD↓	Aes.↑	SSCD↓	Aes.↑	Window	$\text{SSCD}^{\text{norm}}$ ↓	Aes.↑	$\text{SSCD}^{\text{norm}}$ ↓	Aes.↑
Base (No filter)	0.336	4.094	0.299	4.845	Base (Full)	0.430	4.094	0.501	4.845
Low-pass (ours)	**0.331**	**5.553**	**0.264**	**5.904**	$[250, 750]$	**0.349**	5.520	**0.399**	**6.095**
Mid-pass	0.364	4.374	0.329	5.043	$[250, 500]$	0.415	**5.790**	0.439	5.951
High-pass	0.379	4.143	0.227	4.670	$[333, 666]$	0.399	5.706	0.420	5.696

scores, which is merely an artifact of severe image quality degradation. In contrast, our selective framework demonstrates a superior unlearning direction by achieving significant $\text{SSCD}_{\text{norm}}$ gains.

Ablation Study. Figure 7 and Table 3 show the ablation studies on time and frequency selections. Across both timesteps, the low-pass filter achieves a better trade-off of low SSCD and high Aesthetic score compared to mid-pass and high-pass alternatives, confirming that preserving high-frequency detail is critical for quality-preserving forgetting. Similarly, the $[250, 750]$ time window yields the best, while naively chosen $[250, 500]$ and $[333, 666]$ outperform the full ($[0, 1000]$) baseline.

Membership inference attack (MIA). We further evaluate forgetting with membership inference attacks using SecMI (Duan et al., 2023) and PIAN (Kong et al., 2024). We measure $\text{AUC}_{\text{sep}} = \max(\text{AUC}, 1 - \text{AUC})$, where 0.5 indicates chance-level separation and 1 indicates perfect separation and the absolute standardized mean distance $\Delta_d = \frac{|\mu_{\text{unlearned}} - \mu_{\text{hold}}|}{\sqrt{(\sigma^2_{\text{unlearned}} + \sigma^2_{\text{hold}})/2}}$ to calculate distance from hold-out data. For the pretrained reference, the comparison is between member and hold-out records; for each unlearning method, it is between unlearned and hold-out records (Table 4). We conclude that selective unlearning mitigates over-forgetting relative to the corresponding baseline.

Table 4: Separability after unlearning using MIA for diffusion. Lower is better.

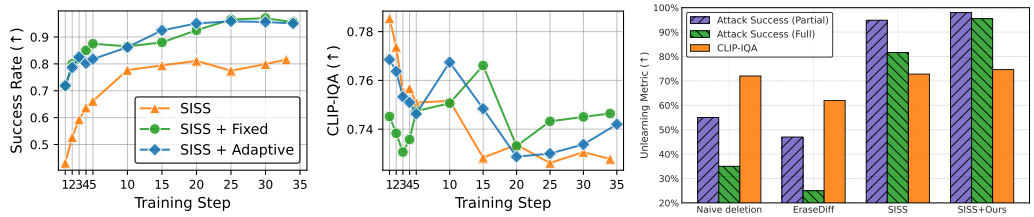
		AUC_{sep} ↓		Δ_d ↓	
		Base	+Ours	Base	+Ours
SecMI	Pre.	0.58		0.03	
	SISS	1.00	**0.79**	4.04	**1.07**
	KTO	0.76	**0.68**	1.60	**0.34**
PIAN	Pre.	0.68		0.54	
	SISS	1.00	1.00	8.59	**4.12**
	KTO	0.99	**0.94**	0.87	**0.86**

4.2 Conditional Text-to-Image Data Unlearning.

Experimental Setups. For Stable Diffusion (Rombach et al., 2022), we evaluate the memorization of the LAION dataset corresponding to each prompt. We train with a time window $(t_1, t_2) = (750, 1000)$ and the adaptive method in Algorithm 1 with $r_t = 0.15$. We fine-tune Stable Diffusion from Hugging Face Diffusers with Adam using a batch size of 16 and a learning rate of 10^{-5}.

For evaluation, we measure 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. Fully-memorized prompts consistently generate exact replicas of the training images, whereas partially-memorized prompts contain weaker memorization, thus easier to unlearn. Further experimental details for both settings are in the Appendix B.

Main Results. The results of Stable Diffusion 1.4 are shown in Figure 8, where our method achieves a higher unlearning success rate and faster convergence than the baseline SISS throughout the training



(a) Unlearning Ratio (Full ASR) (b) CLIP-IQA (Quality) (c) Overall Comparison

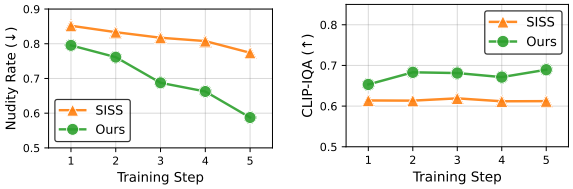
Figure 8: Unlearning attack success rate (ASR) and CLIP-IQA during unlearning. As shown in (a) and (b), selective unlearning shows a faster and higher unlearning rate. As shown in (c), our method shows strong forgetting while preserving quality compared to existing unlearning methods.

process. Figure 8c further confirms that our approach yields superior unlearning (lower attack success) and better image quality (higher CLIP-IQA). These observations suggest that applying selective unlearning is also efficient for conditional cases. As illustrated in Figure 8 (a-b), both fixed and adaptive methods improve unlearning, while the adaptive one shows stable performance in the earlier training steps. The frequency scheduling also improves quality, as shown in Table 12 in the Appendix.

Furthermore, we extend to Stable Diffusion v3.0. built upon the Diffusion Transformer (DiT) architecture (Peebles & Xie, 2023). Since Stable Diffusion v3.0 has practical guardrail methods to prevent memorization, we use the same prompts as in v1.4. Table 5 shows our framework works in unlearning models using flow matching (Lipman et al., 2022).

Table 5: Results on Stable Diffusion v3.0, having Diffusion Transformer backbone.

	Partial		Full	
Measure	SISS	+Ours	SISS	+Ours
CLIP-IQA ↑	**0.863**	0.856	0.886	**0.890**
SSCD ↓	0.349	**0.302**	0.421	**0.362**
$\text{SSCD}^{\text{norm}}$ ↓	0.704	**0.698**	0.724	**0.707**



(d) Nudity Detection rate (e) CLIP-IQA (Quality)

Figure 9: Unlearning concept "nudity" on Ring-a-bell.

Extension to Concept Unlearning. We examine whether the choice of timesteps also affects concept generation in terms of "memorization of certain concepts." We utilize the nudity prompts used in Ring-a-bell (Tsai et al., 2024) in SAFREE (Yoon et al., 2025). As shown in Figure 9, the proposed method shows faster convergence and superior forgetting behavior.

Computation. Table 6 shows computational costs in seconds using Stable Diffusion v1.4. Our method requires an additional computational cost for FFT of 5.99% per epoch. Additionally, the adaptive method requires a sampling step before training begins. However, since our method converges significantly faster, the actual total training time can be reduced. Importantly, inference time remains unchanged for our methods.

Table 6: Computational costs.

Metric	1 Epoch (s)	Total (s)
SISS	8.51	297.68
Ours	9.02	315.61
Adaptive	–	370.66
Sampling	–	55.05

Refer to Appendix C for supplementary experiments and Appendix D for visualizations.

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 spectral domains. We focus on memorization and data unlearning, whose individual influence is measurable and poses the individual privacy risk. As a limitation, this work provides limited application to concept unlearning or automatically discovering the optimal unlearning regions for different tasks. We believe that this work assists the social impact of machine unlearning within generative models.

AI USE STATEMENT

In this work, we used generative AI tools for checking grammatical errors and fixing the code implementation errors. We have not used generative AI tools for research ideation or for producing or analyzing any experimental result. Additionally, we used generative AI tools to check the manuscript for internal consistency. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work aims at data unlearning, which is closely related to societal aspects of generative models. All prompts and pre-trained models come from published benchmarks. Every generation was made for this study alone, and we do not release a new attack.

REPRODUCIBILITY STATEMENT

Our method relies on a simple implementation on existing data unlearning code. Appendix B explains the details of the backbone, prompts, and optimization we used for implementation. In the supplementary material, we append the code for our method. We will publish the full code on GitHub.

REFERENCES

Silas Alberti, Kenan Hasanaliyev, Manav Shah, and Stefano Ermon. Data unlearning in diffusion models. In *The Thirteenth International Conference on Learning Representations*, 2025.

Brian DO Anderson. Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, 12(3):313–326, 1982.

Martin Bertran, Shuai Tang, Michael Kearns, Jamie H Morgenstern, Aaron Roth, and Steven Z Wu. Reconstruction attacks on machine unlearning: Simple models are vulnerable. *Advances in Neural Information Processing Systems*, 37:104995–105016, 2024.

Giulio Biroli, Tony Bonnaire, Valentin de Bortoli, and Marc Mézard. Dynamical regimes of diffusion models. *Nature Communications*, 15(1):9957, November 2024.

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.

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.

June Suk Choi, Kyungmin Lee, Sihyun Yu, Yisol Choi, Jinwoo Shin, and Kimin Lee. Enhancing motion dynamics of image-to-video models via adaptive low-pass guidance. *arXiv preprint arXiv:2506.08456*, 2025.

Jinhao Duan, Fei Kong, Shiqi Wang, Xiaoshuang Shi, and Kaidi Xu. Are diffusion models vulnerable to membership inference attacks? In *International Conference on Machine Learning*, pp. 8717–8730. PMLR, 2023.

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.

Hongcheng Gao, Tianyu Pang, Chao Du, Taihang Hu, Zhijie Deng, and Min Lin. Meta-unlearning on diffusion models: Preventing relearning unlearned concepts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2131–2141, 2025.

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.

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.

Anubhav Jain, Yuya Kobayashi, Takashi Shibuya, Yuhta Takida, Nasir Memon, Julian Togelius, and Yuki Mitsufuji. Classifier-free guidance inside the attraction basin may cause memorization. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 12871–12879, 2025.

Dongjae Jeon, Dueun Kim, and Albert No. Understanding and mitigating memorization in generative models via sharpness of probability landscapes. In *Forty-second International Conference on Machine Learning*, 2025.

Mingyu Kim, Dongjun Kim, Amman Yusuf, Stefano Ermon, and Mijung Park. Training-free safe denoisers for safe use of diffusion models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Mingyu Kim, Young-Heon Kim, and Mijung Park. SAFETY-GUIDED FLOW (SGF): A UNIFIED FRAMEWORK FOR NEGATIVE GUIDANCE IN SAFE GENERATION. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=EA80Zib9UI.

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.

Fei Kong, Jinhao Duan, RuiPeng Ma, Heng Tao Shen, Xiaoshuang Shi, Xiaofeng Zhu, and Kaidi Xu. An efficient membership inference attack for the diffusion model by proximal initialization. In *The Twelfth International Conference on Learning Representations*, 2024.

Marvin Li and Sitan Chen. Critical windows: non-asymptotic theory for feature emergence in diffusion models, 2024. URL https://arxiv.org/abs/2403.01633.

Marvin Li, Aayush Karan, and Sitan Chen. Blink of an eye: a simple theory for feature localization in generative models. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 35047–35080. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/li25at.html.

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.

Puwei Lian, Yujun Cai, Songze Li, and Bingkun Bao. Unveiling impact of frequency components on membership inference attacks for diffusion models. *arXiv preprint arXiv:2505.20955*, 2025.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Robert Shevilevich Liptser and Albert Nikolaevich Shiriaev. *Statistics of random processes: General theory*, volume 394. Springer, 1977.

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=UdXE5V2d0O.

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.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4195–4205, 2023.

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.

Antonio Sclocchi, Alessandro Favero, and Matthieu Wyart. A phase transition in diffusion models reveals the hierarchical nature of data. *Proceedings of the National Academy of Sciences*, 122(1): e2408799121, 2025. doi: 10.1073/pnas.2408799121. URL https://www.pnas.org/doi/abs/10.1073/pnas.2408799121.

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.

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.

Yu-Lin Tsai, Chia-Yi Hsu, Chulin Xie, Chih-Hsun Lin, Jia You Chen, Bo Li, Pin-Yu Chen, Chia-Mu Yu, and Chun-Ying Huang. Ring-a-bell! how reliable are concept removal methods for diffusion models? In *The Twelfth International Conference on Learning Representations*, 2024.

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, Jingtan 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*, 2025.

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.

Yuxin Wen, Yuchen Liu, Chen Chen, and Lingjuan Lyu. Detecting, explaining, and mitigating memorization in diffusion models. In *The Twelfth International Conference on Learning Representations*, 2024.

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.

Shuai Yang, Yukang Chen, Luozhou Wang, Shu Liu, and Ying-Cong Chen. Denoising diffusion step-aware models. In *The Twelfth International Conference on Learning Representations*, 2024.

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=hgTFotBRKl.

Kôsaku Yosida. *Functional analysis*. Springer Science & Business Media, 2012.

Shengfang Zhai, Huanran Chen, Yinpeng Dong, Jiajun Li, Qingni Shen, Yansong Gao, Hang Su, and Yang Liu. Membership inference on text-to-image diffusion models via conditional likelihood discrepancy. *Advances in Neural Information Processing Systems*, 37:74122–74146, 2024.

Ruiqi Zhang, Licong Lin, Yu Bai, and Song Mei. Negative preference optimization: From catastrophic collapse to effective unlearning. In *First Conference on Language Modeling*, 2024.

A JUSTIFICATION OF PROPOSITION 3.1

Setup. Let $\mathbb{P}_{\text{orig}}$ and $\mathbb{P}_m$ ($m \in \{\text{full}, \text{sel}\}$) be the path measures of the original and unlearned diffusion models. We denote the score perturbation as

$$\Delta s_m(x_t, t) := s_{\theta_m}(x_t, t) - s_{\theta_{\text{orig}}}(x_t, t). \tag{14}$$

By Girsanov's theorem (Liptser & Shiriaev, 1977), the KL divergence between the original and unlearned path measures is

$$D_{\text{KL}}(\mathbb{P}_{\text{orig}} \,\|\, \mathbb{P}_m) = \frac{1}{2} \int_0^T g(t)^2 \mathbb{E}_{x_t} \left[\|\Delta s_m(x_t, t)\|_2^2 \right] dt. \tag{15}$$

To evaluate unlearning efficacy, we decompose this total unlearning-induced distortion into a selected target window $\mathcal{S}$ (e.g., selected time-steps or specific frequencies) and its complement $\mathcal{S}^c$. Let

$$D_{\text{KL}}^{\mathcal{W}}(\mathbb{P}_{\text{orig}} \,\|\, \mathbb{P}_m) := \frac{1}{2} \int_{\mathcal{W}} g(t)^2 \mathbb{E}_{x_t}[\|\Delta s_m\|_2^2] dt \tag{16}$$

denote the regional divergence over a window $\mathcal{W}$. Thus, the total divergence can be formulated as

$$D_{\text{KL}}(\mathbb{P}_{\text{orig}} \,\|\, \mathbb{P}_m) = D_{\text{KL}}^{\mathcal{S}}(\mathbb{P}_{\text{orig}} \,\|\, \mathbb{P}_m) + D_{\text{KL}}^{\mathcal{S}^c}(\mathbb{P}_{\text{orig}} \,\|\, \mathbb{P}_m). \tag{17}$$

For $\beta > 0$, we define the forgetting–preservation utility as:

$$\mathcal{J}_\beta(m) := D_{\text{KL}}^{\mathcal{S}}(\mathbb{P}_{\text{orig}} \,\|\, \mathbb{P}_m) - \beta D_{\text{KL}}^{\mathcal{S}^c}(\mathbb{P}_{\text{orig}} \,\|\, \mathbb{P}_m). \tag{18}$$

Note that $D_{\text{KL}}^{\mathcal{S}}(\mathbb{P}_{\text{orig}} \,\|\, \mathbb{P}_m)$ is a proxy for the *forgetting action*. If $\mathcal{S}$ is chosen to contain the components that encode the forget-set information, a larger divergence on $\mathcal{S}$ means the method is actively modifying the targeted information. Conversely, $D_{\text{KL}}^{\mathcal{S}^c}(\mathbb{P}_{\text{orig}} \,\|\, \mathbb{P}_m)$ is a proxy for *collateral distortion*. Since $\mathcal{S}^c$ is not the region we intend to unlearn, score changes there are undesirable side effects that can degrade retained content and sample quality.

Assumptions.

1. *Finite divergence*: $D_{\text{KL}}(\mathbb{P}_{\text{orig}} \,\|\, \mathbb{P}_m) < \infty$.
2. *Focused updates in $\mathcal{S}$*: Due to the frequent update on the selected region under the same training budget, selective unlearning amplifies changes in the target region ($\|\Delta s_{\text{sel}}\|_2 \geq \mu \|\Delta s_{\text{full}}\|_2$ for $\mu \geq 1$ in $\mathcal{S}$) and weaker changes in $\mathcal{S}^c$ ($\|\Delta s_{\text{sel}}\|_2 \leq \eta \|\Delta s_{\text{full}}\|_2$ for $\eta \in [0, 1]$ in $\mathcal{S}^c$). Empirical validations are presented in Appendix A.1.

Proposition A.1. *Under the assumptions above, the trade-off utility of $\mathcal{J}_\beta(\text{sel})$ is larger than $\mathcal{J}_\beta(\text{full})$. For every $\beta > 0$ and $\lambda := \frac{D_{\text{KL}}^{\mathcal{S}^c}(\mathbb{P}_{\text{orig}} \,\|\, \mathbb{P}_{\text{full}})}{D_{\text{KL}}(\mathbb{P}_{\text{orig}} \,\|\, \mathbb{P}_{\text{full}})} \in [0, 1]$, their gap satisfies:*

$$\mathcal{J}_\beta(\text{sel}) - \mathcal{J}_\beta(\text{full}) \geq \left[(\mu^2 - 1)(1 - \lambda) + \beta(1 - \eta^2)\lambda \right] D_{\text{KL}}(\mathbb{P}_{\text{orig}} \,\|\, \mathbb{P}_{\text{full}}), \tag{19}$$

where λ indicates the proportion of collateral distortion compared to the full-unlearning.

Proof. From the definition of the utility function given in Eq. 11, the difference is:

$$\begin{aligned} \mathcal{J}_\beta(\text{sel}) - \mathcal{J}_\beta(\text{full}) = &\left(D_{\text{KL}}^{\mathcal{S}}(\mathbb{P}_{\text{orig}} \,\|\, \mathbb{P}_{\text{sel}}) - D_{\text{KL}}^{\mathcal{S}}(\mathbb{P}_{\text{orig}} \,\|\, \mathbb{P}_{\text{full}}) \right) \\ &- \beta \left(D_{\text{KL}}^{\mathcal{S}^c}(\mathbb{P}_{\text{orig}} \,\|\, \mathbb{P}_{\text{sel}}) - D_{\text{KL}}^{\mathcal{S}^c}(\mathbb{P}_{\text{orig}} \,\|\, \mathbb{P}_{\text{full}}) \right). \end{aligned}$$

By Assumption 2, the condition $\|\Delta s_{\text{sel}}\|_2 \geq \mu \|\Delta s_{\text{full}}\|_2$ in $\mathcal{S}$ ensures the divergence in the forget-window satisfies:

$$D_{\text{KL}}^{\mathcal{S}}(\mathbb{P}_{\text{orig}} \,\|\, \mathbb{P}_{\text{sel}}) \geq \mu^2 D_{\text{KL}}^{\mathcal{S}}(\mathbb{P}_{\text{orig}} \,\|\, \mathbb{P}_{\text{full}}).$$

Moreover, the condition $\|\Delta s_{\text{sel}}\|_2 \leq \eta \|\Delta s_{\text{full}}\|_2$ in $\mathcal{S}^c$ (Assumption 2) ensures the divergence in the non-target window satisfies:

$$D_{\text{KL}}^{\mathcal{S}^c}(\mathbb{P}_{\text{orig}} \,\|\, \mathbb{P}_{\text{sel}}) \leq \eta^2 D_{\text{KL}}^{\mathcal{S}^c}(\mathbb{P}_{\text{orig}} \,\|\, \mathbb{P}_{\text{full}}).$$

Substituting these bounds into the utility difference yields:

$$\mathcal{J}_\beta(\text{sel}) - \mathcal{J}_\beta(\text{full}) \geq (\mu^2 - 1) D_{\text{KL}}^{\mathcal{S}}(\mathbb{P}_{\text{orig}} \,\|\, \mathbb{P}_{\text{full}}) + \beta(1 - \eta^2) D_{\text{KL}}^{\mathcal{S}^c}(\mathbb{P}_{\text{orig}} \,\|\, \mathbb{P}_{\text{full}}).$$

By the definitions of λ and total D_{KL}, we can express the regional divergences as:

$$D_{\mathrm{KL}}^{\mathcal{S}^c}(\mathbb{P}_{\mathrm{orig}} \,\|\, \mathbb{P}_{\mathrm{full}}) = \lambda D_{\mathrm{KL}}(\mathbb{P}_{\mathrm{orig}} \,\|\, \mathbb{P}_{\mathrm{full}}), \quad D_{\mathrm{KL}}^{\mathcal{S}}(\mathbb{P}_{\mathrm{orig}} \,\|\, \mathbb{P}_{\mathrm{full}}) = (1-\lambda) D_{\mathrm{KL}}(\mathbb{P}_{\mathrm{orig}} \,\|\, \mathbb{P}_{\mathrm{full}}).$$

Substituting these exact expressions into the lower bound directly yields the stated inequality:

$$\begin{aligned} \mathcal{J}_\beta(\mathrm{sel}) - \mathcal{J}_\beta(\mathrm{full}) &\geq (\mu^2 - 1)(1-\lambda) D_{\mathrm{KL}}(\mathbb{P}_{\mathrm{orig}} \,\|\, \mathbb{P}_{\mathrm{full}}) + \beta(1-\eta^2)\lambda D_{\mathrm{KL}}(\mathbb{P}_{\mathrm{orig}} \,\|\, \mathbb{P}_{\mathrm{full}}) \\ &= \left[(\mu^2 - 1)(1-\lambda) + \beta(1-\eta^2)\lambda\right] D_{\mathrm{KL}}(\mathbb{P}_{\mathrm{orig}} \,\|\, \mathbb{P}_{\mathrm{full}}). \end{aligned}$$

□

Remark 1. *The bracketed term $(\mu^2 - 1)(1-\lambda) + \beta(1-\eta^2)\lambda$ is always non-negative under our assumptions ($\mu \geq 1, \eta \leq 1, \beta > 0, \lambda \in [0,1]$). Then, the improvement in utility is strictly positive.*

Remark 2. *The utility gap $\mathcal{J}_\beta(\mathrm{sel}) - \mathcal{J}_\beta(\mathrm{full}) \geq 0$ indicates our selective mechanism focuses score updates on the selected forget-region while mitigating unwanted distortion in non-selected parts.*

A.1 Empirical Validation of Assumptions

Our theoretical analysis relies on two key assumptions: enhanced forgetting in the target region (Assumption 2, $\mu \geq 1$) and suppressed distortion in the non-target region (Assumption 2, $\eta \leq 1$). We empirically validate both assumptions across the time and frequency domains.

Enhanced Forgetting ($\mu \geq 1$). To verify that targeted training induces stronger score updates within the specified region compared to full unlearning, we evaluate the μ-ratio ($\|\Delta s_{\mathrm{sel}}\|_2^2 / \|\Delta s_{\mathrm{full}}\|_2^2$) strictly within the targeted window $\mathcal{S}$. As shown in Table 7, the relative ratios are strictly greater than 1 across all targeted time windows (Table 7a) and frequency bands (Table 7b). This confirms that selective unlearning strengthens its update energy on the selected region.

Table 7: Empirical validation of $\mu^2 \geq 1$. We report the ratio $\|\Delta s_{\mathrm{sel}}\|_2^2 / \|\Delta s_{\mathrm{full}}\|_2^2$ strictly within the target region $\mathcal{S}$. Values > 1 indicate that targeted training induces stronger score updates compared to full unlearning. W_1, W_2, and W_3 denote the early $[0, 250)$, middle $[250, 750)$, and late $[750, 1000]$ diffusion time windows. Low, Mid, and High correspond to the spatial frequency bands.

(a) Time-selective Unlearning

Target Region ($\mathcal{S}$)	**Ratio** ($\|\Delta s_{\mathrm{sel}}\|_2^2 / \|\Delta s_{\mathrm{full}}\|_2^2$)
W_1 (Early)	1.2515
W_2 (Middle)	1.4950
W_3 (Late)	1.3095

(b) Frequency-selective Unlearning

Target Region ($\mathcal{S}$)	**Ratio** ($\|\Delta s_{\mathrm{sel}}\|_2^2 / \|\Delta s_{\mathrm{full}}\|_2^2$)
Low-pass	1.2403
Mid-pass	13.3800
High-pass	1.3850

Suppressed Distortion ($\eta \leq 1$). To quantify the bounded out-of-region effect, we measure the normalized perturbation across different evaluation domains when the model is trained on a specific target region. The results in Table 8 indicate that selective unlearning predominantly affects the intended phase or band compared to non-target counterparts. The off-diagonal elements in both the time domain (Table 8a) and the frequency domain (Table 8b) experimentally support $\eta < 1$.

Table 8: Empirical validation of $\eta^2 \leq 1$. We report the normalized perturbation $\|\Delta s_{\mathrm{sel}}\|_2^2$ across evaluation regions (rows) when trained on a specific target $\mathcal{S}$ (columns). Diagonal elements are scaled to 1, while off-diagonal elements show the bounded effects of collateral distortion.

(a) Time-selective (Cross-Region)

	Target Region ($\mathcal{S}$)		
Eval Region	W_1	W_2	W_3
W_1	1.0000	0.2511	0.0009
W_2	0.7532	1.0000	0.0010
W_3	0.4939	0.7037	1.0000

(b) Frequency-selective (Cross-Region)

	Target Region ($\mathcal{S}$)		
Eval Region	Low	Mid	High
Low	1.0000	0.2535	0.1681
Mid	0.4116	1.0000	0.0511
High	0.0502	0.0814	1.0000

A.2 Time and Frequency Understanding for λ

Time-Selective Unlearning. Defining the target region as a specific time window $\mathcal{S} = \mathcal{T}_s \subset [0, T]$ formulates time-selective unlearning. In this case, the ratio λ_t represents the fraction of the full-unlearning distortion that occurs outside the targeted time window $\mathcal{T}_s$:

$$\lambda_t \;:=\; \frac{\int_{[0,T]\setminus\mathcal{T}_s} g(t)^2\, \mathbb{E}_{x_t}\left[\|\Delta s_{\text{full}}(x_t,t)\|_2^2\right] dt}{\int_0^T g(t)^2\, \mathbb{E}_{x_t}[\|\Delta s_{\text{full}}(x_t,t)\|_2^2]\, dt} \;\in\; [0,1]. \tag{20}$$

Frequency-Selective Unlearning. Plancherel's theorem (Yosida, 2012) establishes an equivalence between spatial and spectral energy:

$$\|\Delta s_m(x_t,t)\|_2^2 \;=\; \sum_{\boldsymbol{\omega}} \|\widehat{\Delta s}_m(\boldsymbol{\omega},t)\|_2^2, \tag{21}$$

where $\widehat{\Delta s}_m(\boldsymbol{\omega},t)$ is the discrete Fourier transform of the score perturbation across spatial dimensions at frequency $\boldsymbol{\omega}$. By applying a low-pass filter with cutoff r, we define our targeted forgetting region in the frequency domain as $\mathcal{S} = \Omega_r := \{|\boldsymbol{\omega}| \le r\}$. Consequently, the collateral distortion ratio λ_ω, corresponding to the unselected high-frequency band, is defined as:

$$\lambda_\omega \;:=\; \frac{\int_0^T g(t)^2 \sum_{|\boldsymbol{\omega}|>r} \mathbb{E}_{x_t}\left[\|\widehat{\Delta s}_{\text{full}}(\boldsymbol{\omega},t)\|_2^2\right] dt}{\int_0^T g(t)^2 \sum_{\boldsymbol{\omega}} \mathbb{E}_{x_t}\left[\|\widehat{\Delta s}_{\text{full}}(\boldsymbol{\omega},t)\|_2^2\right] dt} \;\in\; [0,1]. \tag{22}$$

B 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 (7) and the high-frequency suppression weight at $s = 0$ in Equation (**??**) 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 v1.4. 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., 2024) has been proposed as an alternative to gradient descent. Unlike gradient ascent (Zhang et al., 2024), 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. (2025) 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).

To formulate this, let the per-sample diffusion loss be $\ell(\theta, \mathbf{x}) = \mathbb{E}_t[w_t \| s_\theta(\mathbf{x}_t, t) - \nabla_{\mathbf{x}_t} \log q(\mathbf{x}_t|\mathbf{x})\|_2^2]$. We define the implicit reward as the loss reduction relative to the reference model: $r(\theta, \mathbf{x}) = \ell(\theta_{\text{ref}}, \mathbf{x}) - \ell(\theta, \mathbf{x})$. By maximizing the implicit reward gap defined by the difference in diffusion loss relative to the reference model θ_{ref}, the framework effectively minimizes the likelihood of generating forget concepts while preserving general capabilities. The objectives for DPO and KTO are then given by for forget data $\mathbf{x}_f$ and retain data $\mathbf{x}_r$:

$$\mathcal{L}_{\text{DPO}}(\theta) = -\mathbb{E}_{\mathbf{x}_r \sim \mathcal{D}_R, \mathbf{x}_f \sim \mathcal{D}_F}\Big[\log \sigma\Big(\beta\big(r(\theta, \mathbf{x}_r) - r(\theta, \mathbf{x}_f)\big)\Big)\Big]. \tag{23}$$

$$\mathcal{L}_{\text{KTO}}(\theta) = \mathbb{E}_{\mathbf{x}_r \sim \mathcal{D}_R}\Big[1 - \sigma\big(\beta r(\theta, \mathbf{x}_r) - z_{\text{ref}}\big)\Big] + \lambda \mathbb{E}_{\mathbf{x}_f \sim \mathcal{D}_F}\Big[1 - \sigma\big(z_{\text{ref}} - \beta r(\theta, \mathbf{x}_f)\big)\Big]. \tag{24}$$

σ is the sigmoid function, β is a temperature hyperparameter, and z_{ref} is the reference point for the reward.

Algorithm of Adaptive Search. Based on Jain et al. (2025), we present pseudocode of the adaptive transition point $\tau = t_S$ search method in Algorithm 1.

Algorithm 1 Adaptive Transition Time Search and Selective Unlearning

Require: Pretrained model θ^*; prompt embedding e_p; unconditional embedding $e_\emptyset$;
Require: Forget set $\mathcal{D}_F$; retain set $\mathcal{D}_R$; total time steps T
Phase 1: Adaptive search for transition time τ
Sample $x_T \sim \mathcal{N}(0, \mathbf{I})$
$\tau \leftarrow$ None
$d_{\text{prev}} \leftarrow +\infty$
for t following the denoising schedule from T to 1 **do**
  $d_t \leftarrow \|\epsilon_\theta(x_t, e_p) - \epsilon_\theta(x_t, e_\emptyset)\|_2$
  if $\tau =$ None **and** $d_t > d_{\text{prev}}$ **then**
    $\tau \leftarrow t_{\text{prev}}$ ▷ first local minimum along the trajectory
  end if
  $d_{\text{prev}} \leftarrow d_t$
  $t_{\text{prev}} \leftarrow t$
  $x_{t-1} \leftarrow \text{DenoisingStep}(x_t, \epsilon_\theta)$
end for
Phase 2: Selective Unlearning
$t_S \leftarrow \tau$
Perform unlearning for timesteps $t \in [t_S, T]$

C Supplementary Experiments

C.1 Different Backbones

To evaluate the adaptability of latent models for image-level generation, we conduct experiments using the Latent Diffusion Model (LDM) (Rombach et al., 2022) pretrained on the CelebHQ dataset[3]. The results are presented in Table 9. Tables 9, in addition to5, demonstrate the adaptability and effectiveness of our proposed method across diverse architectures. We did not perform additional hyperparameter tuning for these experiments.

[3] https://huggingface.co/CompVis/ldm-celebahq-256

Table 9: Comparison of unlearning methods on latent diffusion models 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			Denoising from $t = 250$			Denoising from $t = 500$		
		FID-10K ↓	SSCD ↓	$\text{SSCD}^{\text{norm}}$ ↓	Aesthetic ↑	SSCD ↓	$\text{SSCD}^{\text{norm}}$ ↓	Aesthetic ↑
SISS	Base	17.800	0.431	0.547	4.601	0.281	0.588	5.092
	+ Ours	16.960	0.437	0.520	5.076	0.250	0.536	4.955
	Gain (%)	+4.72	-1.54	+4.95	+10.32	+10.90	+8.84	-2.69
DPO	Base	28.630	0.428	0.573	4.505	0.372	0.630	4.880
	+ Ours	28.630	0.372	0.630	4.880	0.295	0.546	4.996
	Gain (%)	0.00	+13.02	-9.89	+8.34	+20.63	+13.34	+2.38
KTO	Base	28.630	0.372	0.630	4.880	0.458	0.472	5.812
	+ Ours	31.360	0.498	0.576	5.072	0.457	0.493	6.446
	Gain (%)	-9.54	-33.65	+8.58	+3.93	+0.19	-4.48	+10.90

C.2 Importance of Unlearned Quality

We argue that significant degradation or distinct artifacts in these outputs can create a privacy vulnerability. If the unlearned model generates evident anomalies for specific queries, it provides a clue to adversaries, allowing them to identify which data points were deleted (i.e., membership inference via outlier detection).

To verify this hypothesis, we conducted an experiment using advanced Large Language Models (LLMs) as anomaly detectors. We provided the LLMs with generated images and instructed them to evaluate whether the images are realistic (in-distribution) or abnormal (out-of-distribution) based on the following prompt:

> *Prompt: "You are an image quality and anomaly detection expert. You are given two image sets: The first set contains reference (in-distribution) images. Assign a score to each query image: 1.0 = Realistic / In-distribution (no visible generation problem), 0.0 = Abnormal / Out-of-distribution (clear generation failure or artifacts)."*

As shown in Table 10, baseline methods like SISS exhibited low scores, indicating that their outputs are easily distinguishable as anomalies. In contrast, our method maintains higher generation quality, effectively masking the trace of unlearning. Furthermore, recent studies (Bertran et al., 2024; Gao et al., 2025) warn that unlearning mechanisms are vulnerable against relearning attacks, which can recover deleted data using residual information such as embedding similarity. Therefore, ensuring the quality of unlearned samples is crucial not only for aesthetics but also for preventing privacy attacks.

Table 10: Anomaly detection scores evaluated by various LLMs on unlearned outputs. A higher score indicates the generated image is perceived as realistic (in-distribution), making it harder to detect that unlearning has occurred.

LLM Evaluator	SISS	SISS + Ours
ChatGPT-5	0.22	0.67
Claude Sonnet 4.5	0.13	0.83
Gemini 2.5 Flash	0.00	0.50

To investigate the impact of unlearning on similar samples, we choose the five nearest neighbors for each forgotten image, and report the mean SSCD between their original and reconstructed images as well as the aesthetic score of the reconstructed samples in Table 11. As data unlearning focuses on individual deletion, the SSCD on these nearest images does not decrease. However, we do observe a slight drop in aesthetic quality after unlearning. By implementing our selective unlearning, this aesthetic degradation is mitigated. Consequently, selective unlearning helps preserve both the semantics and the overall quality of nearby examples.

Table 11: Reconstruction on the five nearest neighbors of each forgotten image. Selective unlearning preserves both semantics (SSCD) and visual quality (aesthetic score) of nearby samples.

	$t = 250$		$t = 500$	
Model	SSCD ↓	Aesthetic ↑	SSCD ↓	Aesthetic ↑
Orig.	0.8817	6.3894	0.7405	6.7483
SISS	0.8812	6.2441	0.7394	6.5261
+ Ours	0.8797	6.4467	0.7439	6.5873

C.3 Ablation on Frequency Filtering

Frequency cutoff scheduling. Our frequency selection can be further improved with different filtering schedules. Therefore, we test four additional variants during training, including Linear Increase ($0.05 \rightarrow 0.2$), Linear Decay ($0.2 \rightarrow 0.05$), and Stepwise strategies, to assess the impact of filtering high-frequency content over the unlearning fine-tuning. As shown in Table 12, the Linear Increase strategy, which retains a greater amount of high-frequency information during the later training stages, results in the highest image quality. This indicates that frequency filtering should be dynamically adjusted to preserve fine-grained details during the later stages of unlearning.

Table 12: Different schedules of the low-pass cutoff r_t during unlearning fine-tuning. Linear Increase ($0.05 \rightarrow 0.2$) gives the best quality at both evaluation timesteps.

	$t = 250$			$t = 500$		
Schedule	SSCD ↓	$\text{SSCD}^{\text{norm}}$ ↓	Aesth. ↑	SSCD ↓	$\text{SSCD}^{\text{norm}}$ ↓	Aesth. ↑
SISS	0.336	0.430	4.094	0.299	0.501	4.845
Fixed	0.345	0.349	5.520	0.282	0.399	**6.095**
Linear Increase	**0.323**	**0.322**	**5.709**	0.260	**0.382**	5.832
Linear Decay	0.429	0.422	5.296	0.317	0.448	5.722
Stepwise Increase	0.328	0.331	5.631	**0.255**	0.392	5.918
Stepwise Decay	0.431	0.419	5.186	0.317	0.438	5.535

C.4 Ablation on the Perturbation Radius ρ

The perturbation radius ρ in Equation (13) controls the ℓ_2 projection scale of $\text{SSCD}^{\text{norm}}$. We select $\rho = 100$ as the empirical sweet spot of the privacy-utility trade-off: for $\rho > 200$ the projected perturbation degrades quality beyond human perception, while for $\rho < 20$ the projected image remains too similar to the original. We further ablate $\rho \in \{50, 100, 150, 200\}$ on SISS and KTO at both evaluation timesteps in Table 13. Our method achieves superior performance in **18 out of 20 evaluations (90.0% win rate)**, indicating that the improvement is robust to the choice of ρ rather than an artifact of the default value.

Table 13: Sensitivity of $\text{SSCD}^{\text{norm}}$ and Aesthetic score to the perturbation radius ρ on CelebA-HQ. Lower is better for $\text{SSCD}^{\text{norm}}$, higher is better for Aesthetic.

	$\rho = 50$		$\rho = 100$		$\rho = 150$		$\rho = 200$		Aesthetic ↑	
Method	t=250	t=500	t=250	t=500	t=250	t=500	t=250	t=500	t=250	t=500
SISS	0.786	**0.778**	0.430	0.501	0.395	0.346	0.295	0.259	4.094	4.845
+ Ours	**0.742**	0.810	**0.349**	**0.399**	**0.251**	**0.223**	**0.203**	**0.179**	**5.520**	**6.095**
KTO	0.727	0.808	0.366	0.442	0.283	0.277	0.218	**0.193**	5.355	5.879
+ Ours	**0.702**	**0.793**	**0.340**	**0.377**	**0.275**	**0.239**	**0.214**	0.201	**5.470**	**6.274**

C.5 STATISTICAL RELIABILITY

To verify that the gap is statistically significant, we further evaluate the SISS and KTO experiments in Table 14 across five independent runs, which yield small standard deviations.

C.6 STATISTICAL RELIABILITY

To verify that the gap is statistically significant, we further evaluate the SISS and KTO experiments in Table 14 across five independent runs with the fixed target records, which yield small standard deviations. Since each selected face defines a separate single-record unlearning task in a continual manner, we additionally test the sensitivity to the choice of target records by sampling six random faces in each of five independent runs (30 deletions in total). As shown in Table 15, the gain of our selective scheme is not confined to the fixed images, showing its effectiveness on various seeds.

Table 14: Mean $\pm$ standard deviation over five independent runs on CelebA-HQ while fixing the six selected images. Adding our selective scheme yields consistent gains on the perturbation-normalized metric $\text{SSCD}^{\text{norm}}$ and on the aesthetic score, with negligible variance.

		$t = 250$			$t = 500$		
Method	FID ↓	SSCD ↓	$\text{SSCD}^{\text{norm}}$ ↓	Aesth. ↑	SSCD ↓	$\text{SSCD}^{\text{norm}}$ ↓	Aesth. ↑
SISS	$23.42_{\pm 0.31}$	$0.336_{\pm 0.002}$	$0.430_{\pm 0.003}$	$4.094_{\pm 0.054}$	$0.299_{\pm 0.006}$	$0.501_{\pm 0.004}$	$4.845_{\pm 0.051}$
+ Ours	$23.65_{\pm 0.35}$	$0.345_{\pm 0.003}$	$0.349_{\pm 0.002}$	$5.520_{\pm 0.063}$	$0.282_{\pm 0.004}$	$0.399_{\pm 0.003}$	$6.095_{\pm 0.049}$
KTO	$23.48_{\pm 0.15}$	$0.363_{\pm 0.002}$	$0.366_{\pm 0.001}$	$5.355_{\pm 0.062}$	$0.289_{\pm 0.007}$	$0.442_{\pm 0.005}$	$5.879_{\pm 0.089}$
+ Ours	$23.58_{\pm 0.12}$	$0.373_{\pm 0.003}$	$0.340_{\pm 0.003}$	$5.470_{\pm 0.049}$	$0.280_{\pm 0.010}$	$0.377_{\pm 0.006}$	$6.274_{\pm 0.024}$

Table 15: Mean $\pm$ standard deviation over five independently sampled runs, each deleting six randomly selected faces for every run.

	$t = 250$		$t = 500$	
Method	$\text{SSCD}^{\text{norm}}$ ↓	Aesthetic ↑	$\text{SSCD}^{\text{norm}}$ ↓	Aesthetic ↑
SISS	$0.427_{\pm 0.099}$	$4.729_{\pm 0.660}$	$0.399_{\pm 0.044}$	$5.255_{\pm 0.667}$
SISS + Ours	$\mathbf{0.380}_{\pm 0.095}$	$\mathbf{5.373}_{\pm 0.336}$	$\mathbf{0.377}_{\pm 0.073}$	$\mathbf{5.680}_{\pm 0.310}$

C.7 BASE 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 2. We now provide additional results. In Figure 4, 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 11, where we failed to observe similar patterns. The nearest images of individual data samples are in Figure 10. Individual gradient norms are in Figure 12. Finally, we report the remaining results of Figure 6 in Figures 13 and 14.

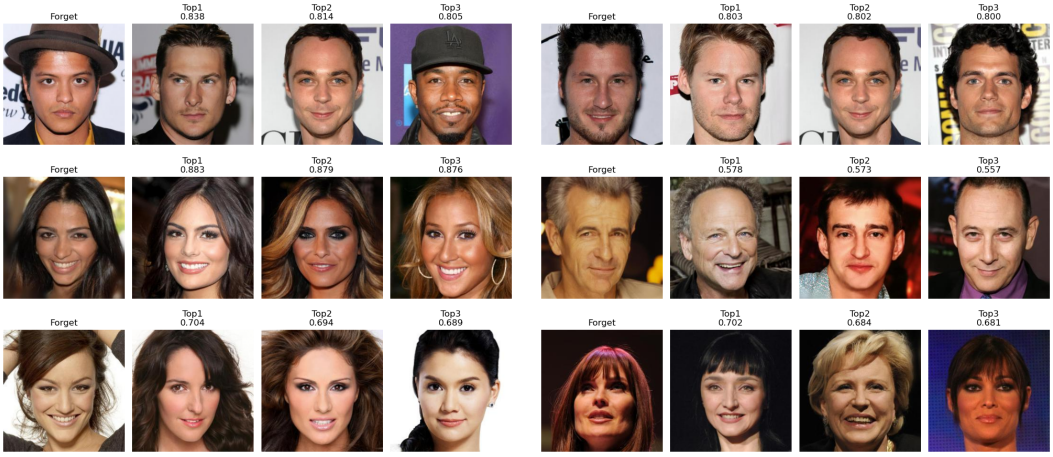


Figure 10: Forget samples and their nearest neighborhood on DINOv3 embedding (Siméoni et al., 2025).

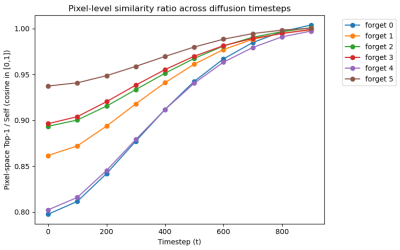


Figure 11: Similarity in pixel-level.

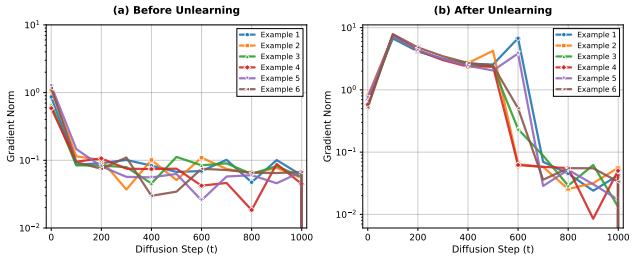


Figure 12: Gradient norms for each data sample before and after unlearning.

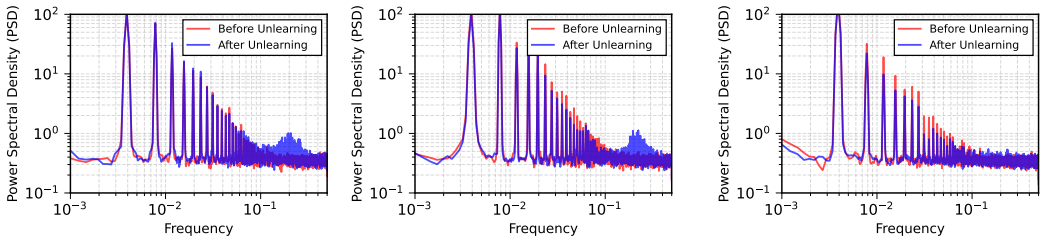


Figure 13: Additional results of Power Spectral Density (PSD) before and after unlearning on the forget dataset. As shown in the right figure, when the generated data from the unlearned model preserves aesthetic quality, the difference in high-frequency components is diminished.

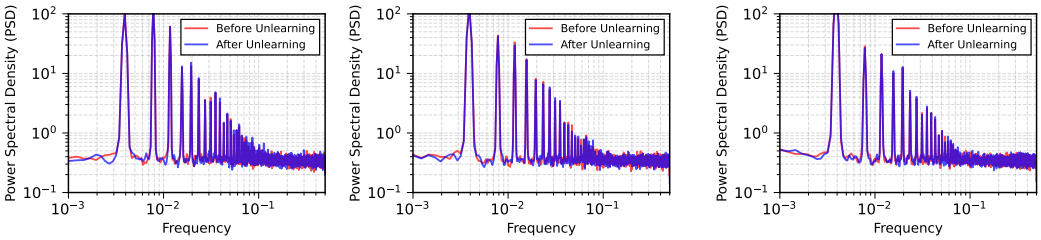


Figure 14: Additional results of Power Spectral Density (PSD) before and after unlearning on the retain dataset.

D VISUALIZATION

CelebA-HQ Results. We first visualize the results for each method: SISS, SISS+Ours, and variational results on different time steps in Figures 15 to 19.



Figure 15: Unlearning with SISS.



Figure 16: Unlearing with SISS+Ours.



Figure 17: Unlearning on time steps [0, 500].

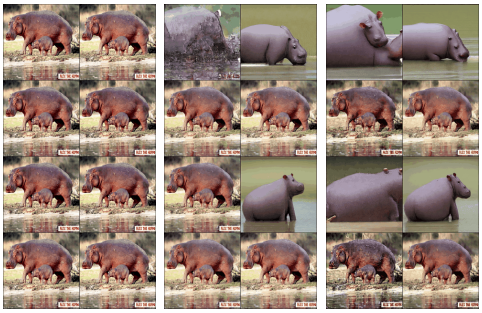


Figure 18: Unlearning on time steps [500, 1000].

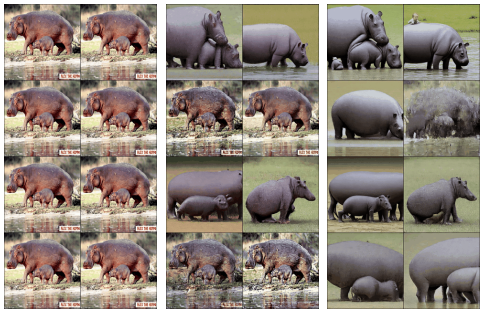


Figure 19: Unlearning on time steps [750, 1000].

Stable Diffusion Results. We visualize the text-to-image data unlearning in Figures 20 to 24.



(a) Generated samples of SISS at early steps



(b) Generated samples of SISS+Ours at early steps

Figure 20: Visualization of generated images from the memorization with **fully-memorized prompt** "Mothers influence on her young hippo" at early unlearning steps $[0, 1, 2]$. The proposed method shows faster forgetting performance while maintaining quality.



(a) Memorized



(b) SISS



(c) SISS + Ours

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".



(a) Time steps on [0,500] (b) Time steps on [250,750] (c) Time steps [500,1000]

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.



(a) Memorized (b) SISS (c) SISS + Ours

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".



(a) Memorized (b) SISS (c) SISS + Ours

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