Forget All That Should Be Forgotten: Separable, Recoverable, and Sustainable Multi-Concept Erasure from Diffusion Models

Abstract—**Text-to-image diffusion models raise concerns regarding their social impact, such as the imitation of copyrighted styles. While recent methods have successfully erased inappropriate concepts from these models, they overlook critical issues caused by multi-concept erasure, including interference from concurrent concept unlearning, irrecoverability of erased concepts, degradation of model performance and watermark, and significant memory overhead.**

In this work, we propose a novel Separable, Recoverable, and Sustainable Multi-concept Eraser (SRS-ME), enabling diffusion models to forget all concepts that they should forget without necessitating retraining from scratch. Specifically, through theoretical analysis, we propose a novel weight decoupling paradigm for constructing separable weight shifts. This can effectively decouple interactions among weight shifts targeting diverse concepts, providing flexibility in both erasing and recovering arbitrary concepts. Meanwhile, it preserves model watermarks, such as predefined images triggered by specific text prompts. To effectively erase inappropriate concepts while preserving model performance on regular concepts, we design an innovative concept-irrelevant unlearning process. This process defines concept representations and introduces a concept correlation loss along with a momentum statistic-based stopping condition. Besides, to reduce memory usage, we demonstrate the feasibility of optimization decoupling for separated weight shifts. Benchmarked against prior work, extensive experiments show that our SRS-ME framework excels in concept manipulation, effectively preserves model performance, and significantly reduces memory consumption.

1. Introduction

The field of text-to-image generation has witnessed remarkable development [1], [2], [3], [4], especially the occurrence of diffusion models (DMs) like DALL-E2 [5] and Stable Diffusion [6]. As the integration of DMs into practical applications [7], [8], [9] proves advantageous, addressing challenges related to their societal impact increasingly attracts the attention of researchers [10], [11], [12], [13]. One crucial challenge arises from diverse training data sources, potentially leading to unsafe image generation [14], [15], such as violent content or mimicking specific artistic styles. To resolve this concern, the machine unlearning (MU) technique has been proposed [16], [17], [18], [19], which involves erasing the impact of specific data points or

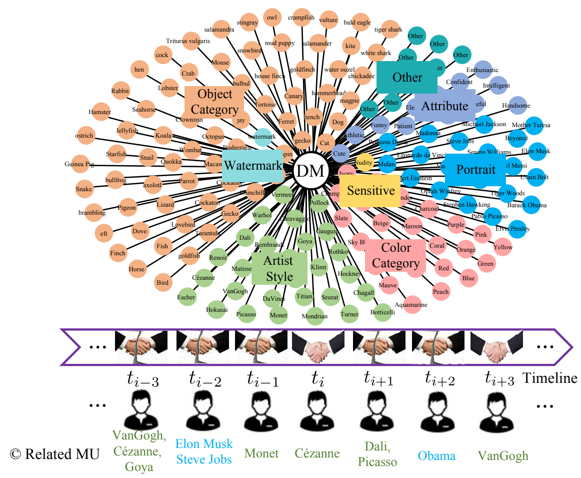


Figure 1. Concept recovery application.

concepts to enhance model security, without necessitating complete retraining from scratch.

Recent MU work such as Erased Stable Diffusion (ESD) [15], Forget-me-not (FMN) [14], Safe self-distillation diffusion (SDD) [20], and AbConcept [17], can be broadly categorized into untargeted (*e.g.*, FMN) and targeted erasures (*e.g.*, ESD, AbConcept and SDD). Specifically, FMN minimizes the values of the attention maps associated with forgotten concepts. In contrast, ESD, AbConcept, and SDD align the denoising distributions of forgotten concepts with predefined distributions that are unrelated to these concepts.

Research Gaps. Despite recent progresses in MU [21], [22], there exist several unresolved gaps, as listed below.

(1) *Concept restoration:* As illustrated in Figure 1, the agreement breakdown between concept owners and DM owners may be temporary, and DM owners need to recover these forgotten concepts after regaining their copyrights. However, prior work has not considered the scenario of concept restoration.

(2) *Multi-concept erasure:* Current erasure procedures are confined to single-concept elimination and pose challenges when extending them to multi-concept erasure. As described in Figure 1, multi-concept erasure can take two forms: simultaneous erasure of multiple concepts (*e.g.*, unlearning 'Van Gogh', 'Cezanne', and 'Goya' at t_{i-3}) and iterative concept erasure (*e.g.*, unlearning 'Van Gogh' at t_{i-3} and then unlearning 'Elon Musk' at t_{i-2}). The former encounters memory overload, while both involve interactions between fine-tuned weights for erasing various concepts.

(3) *Model performance preservation:* Although prior methods successfully erase concepts, they lead to significant performance degradation in the overall generative capabilities of DMs. In particular, these methods can destroy model watermarks, *e.g.*, watermarks such as specific logos or images triggered by predefined prompts for text-guided DMs [23], [24]. As illustrated in Figure 2, existing MU approaches consistently affect the generation performance of other concepts.

Contributions. To fulfill these gaps, we propose an innovative framework called SRS-ME for Separable, Recoverable, and Sustainable Multi-Concept Erasure. Specifically, it features three new components: weight decoupling to construct independent weight shifts, concept-irrelevant unlearning to effectively optimize these weight shifts, and optimization decoupling to reduce memory consumption.

Weight decoupling. Through theoretical analysis, we establish the paradigm of weight decoupling for multi-concept erasure. Specifically, we decompose the weight shift for erasing multiple concepts into *independent* weight shifts. Each of them aims to erase a specific forgotten concept (or multiple inappropriate concepts at a specific timestamp) without compromising the generation performance of DMs regarding other forgotten concepts. These independent weights shifts are formulated as linear combinations of constant particular solutions calculated based on other known undesirable concepts. This enables concept restoration, alleviates mutual interactions among various fine-tuned weights, and preserves model watermarks.

Concept-irrelevant unlearning. To effectively optimize independent weight shifts, we introduce a concept-irrelevant unlearning approach, which can effectively preserve model performance on regular concepts and erase undesirable concepts. Within each schedule of DMs, we measure concept representation by observing feature changes upon concept introduction. Furthermore, we define the unlearning loss as the correlation degree between the concept representations of unlearned and original DMs, with the latter concept representation viewed as pseudo-ground truth. Considering the instability of this loss among noisy inputs, we additionally propose a momentum statistic-based stopping condition.

Optimization decoupling. We theoretically prove the feasibility of separately optimizing independent weight shifts, significantly reducing memory consumption at the cost of training time. Furthermore, optimization decoupling effectively circumvents the need for researchers to balance erasure performance across multiple concepts.

Our main contributions are summarized as follows:

- To the best of our knowledge, the scenarios of concept restoration and watermark preservation remain unexplored in prior unlearning work. The proposed weight decoupling fills these crucial gaps by innovatively constructing independent weight shifts. This enables combinations of diverse weight shifts for flexible erasure and restoration of erased concepts.
- To effectively unlearn undesirable concepts and preserve overall model performance, we propose a novel concept-irrelevant unlearning approach.
- We indicate the feasibility of optimization decoupling, which mitigates memory overload and obviates the necessity to balance multi-concept erasures.
- We conduct an extensive array of experiments to demonstrate that our SRS-ME can flexibly manipulate arbitrary concepts, preserve model generation capabilities, and address memory overhead.

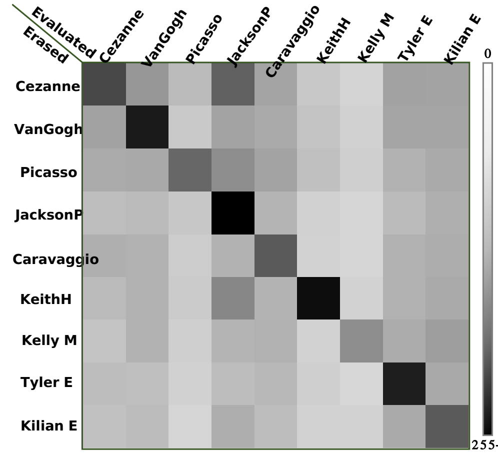


Figure 2. Interference between various concepts during unlearning with AbConcept. A larger Fréchet Inception Distance [25] indicates a greater impact on the DM generation performance of the evaluation concept.

2. BACKGROUND AND RELATED WORK

The image generation field has experienced rapid development in recent years, evolving from autoencoder [26], [27], [28], generative adversarial networks [29], [30], [31], unconditional diffusion models (DMs) [32], [33] to DMs enhanced with large-scale pre-trained image-text models [34], [35], [36] like CLIP [37]. These text-guided DMs, exemplified by DALL-E 2 [5] and Stable Diffusion [6], exhibit excellent generative abilities across various prompts c. The constraint for training DMs is formulated as

$$\mathcal{L}_{dm} = \mathbb{E}_{\boldsymbol{x}_0 \in \mathcal{D}, c, t, \boldsymbol{\epsilon}_{gt} \in \mathcal{N}(0,\mathbf{I})} \left[\|\boldsymbol{\epsilon}_{gt} - \boldsymbol{\epsilon}(\boldsymbol{x}_t, c; \boldsymbol{\theta}_{dm})\|_2^2 \right],$$

where $\boldsymbol{x}_t$ represents the noised data or the noised latent representation [38], $\boldsymbol{x}_t = \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}$, $\bar{\alpha}_t$ is the noise variance schedule, $\boldsymbol{x}_0$ denotes the original reference image, and $\boldsymbol{\epsilon} \in \mathcal{N}(0, \mathbf{I})$. $\mathcal{D}$ is the training dataset. $\boldsymbol{\epsilon}_{gt}$ means the ground truth noise. $\boldsymbol{\epsilon}(\boldsymbol{x}_t, c; \boldsymbol{\theta}_{dm})$ denotes the t-th step noise predicted by DMs with parameters $\boldsymbol{\theta}_{dm}$. $\| \cdot \|_2^2$ is the squared ℓ_2-norm function.

However, DMs also induce potential risks associated with privacy violations and copyright infringement, such as the training data leakage [39], [40], [41], the imitation of various artistic styles [42], [43], and the generation of sensitive content [44]. Hence, there is a growing focus on erasing specific outputs from pre-trained DMs [10], [17].

Existing research for DM unlearning primarily falls into three distinct directions: removal of unsafe data and model retraining [45], integration of additional plug-ins to guide

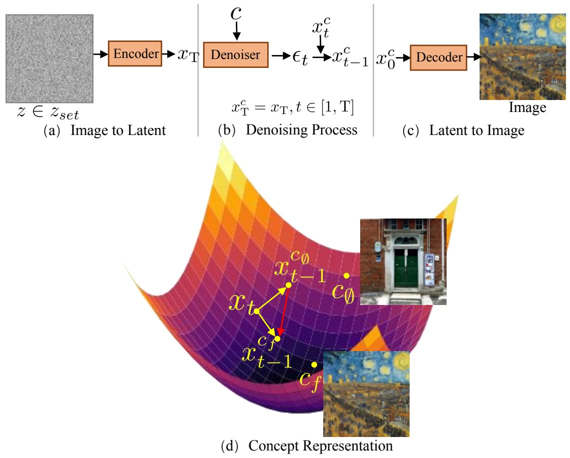


Figure 3. Denoising process of DMs (a∼c) and concept representation generation (d). $c_\emptyset$ and c_f are blank and forgotten concepts, respectively.

model outputs [46], [47], and fine-tuning of trained model weights [14], [15], [17]. Considering a practical scenario where, after unlearning, DM owners update their unlearned model weights online. The drawback of the first direction is that large-scale model retraining demands considerable computational resources and time. The risk of the second direction is that, with the public availability of model structures and weights, malicious users may easily remove plugins. Therefore, this work focuses on the third direction.

Recent finetuning work can be summarized as:

$$\min_{\boldsymbol{\theta}_{op}} \mathcal{L}_{unlearn} = \begin{cases} \|\boldsymbol{\epsilon}(\boldsymbol{x}_t, c_f; \boldsymbol{\theta}_{op}) - \boldsymbol{\epsilon}_{target}\|_2 & \text{if } \boldsymbol{x}_0 \in \mathcal{D}_f \\ \|\boldsymbol{\epsilon}(\boldsymbol{x}_t, c_{gt}; \boldsymbol{\theta}_{op}) - \boldsymbol{\epsilon}_{gt}\|_2 & \text{otherwise}, \end{cases} \tag{1}$$

where $\boldsymbol{\theta}_{op}$ represents optimizable model weights, *e.g.,* the parameters of cross-attention modules in DMs. $\boldsymbol{x}_t$ can be obtained through either the diffusion process or the sampling process. $\boldsymbol{\epsilon}(\boldsymbol{x}_t, c_f; \boldsymbol{\theta}_{op})$ denotes the noise predicted by unlearned DMs at the t-th step. $\mathcal{D}_f$ refers to the dataset containing the forgotten concepts c_f. $\boldsymbol{\epsilon}_{target}$ and $\boldsymbol{\epsilon}_{gt}$ represent the noise of predefined target concepts and the ground-truth noise added in the diffusion process, respectively.

For instance, ESD [15] leverages the predicted noise for both concept-free $c_\emptyset$ and c_f to construct $\boldsymbol{\epsilon}_{target}$,

$$\boldsymbol{\epsilon}_{target} = (1+\eta)\boldsymbol{\epsilon}(\boldsymbol{x}_t, c_\emptyset; \boldsymbol{\theta}_{dm}) - \eta\boldsymbol{\epsilon}(\boldsymbol{x}_t, c_f; \boldsymbol{\theta}_{dm}),$$

where $\boldsymbol{\theta}_{dm}$ represents parameters of the frozen DMs. η is the hyperparameter. SDD [20] directly maps the prediction distribution of erased concepts c_f to the prediction distribution of $c_\emptyset$, $\boldsymbol{\epsilon}_{target} = \boldsymbol{\epsilon}(\boldsymbol{x}_t, c_\emptyset; \boldsymbol{\theta}_{dm})$. AbConcept [17] assigns anchor concepts c^* for each erased concept c_f, *e.g.,* c_f is "VanGogh's painting" and c^* is "painting" when erasing "VanGogh", or c_f is "a photo of Grumpy cat" and c^* is "a photo of cat" when erasing 'Grumpy cat', $\boldsymbol{\epsilon}_{target} = \boldsymbol{\epsilon}(\boldsymbol{x}_t, c^*; \boldsymbol{\theta}_{dm})$. Besides, FMN [14] is an untargeted erasure method, which minimizes the values of attention maps corresponding to the forgotten concepts c_f.

In contrast, this work highlights the challenges of concept restoration, model preservation, and memory overload.

Figure 4. Expectation illustration. c_f denotes unlearned concepts, including 'Picasso', 'VanGogh', and 'Cezanne'. Here, $\mathcal{M}_{c_f^*}$ represent the model unlearning c_f^*, 'VanGogh' $\notin c_f^*$. 'KellyMcKernan' is a regular concept and ' ' is concept-free. The yellow font in the red box indicates text prompts.

Our SRS-ME offers a solution for flexible erasure or restoration of concepts while preserving model performance with limited memory consumption.

3. Proposed SRS-ME

3.1. Problem Definition

We first describe the denoising process of text-guided DMs, as illustrated in Figure 3 (a∼c). The encoder converts the noisy image into latent representations, the denoiser iteratively removes the predicted noise from these representations, and the decoder reconstructs the image from the denoised representations.

Recent finetuning approaches for DM unlearning primarily focus on the single concept erasure. However, they overlook several issues mentioned in Section 1. This work aims to flexibly *unlearn* or *recover* concepts while *preserving the model performance on both regular concepts and watermark prompts* with *limited memory consumption*.

① **Unlearning.** The MU techniques for DMs should effectively erase all undesirable concepts;

② **Concept Restoration.** We denote the forgotten and other concepts as c_f and $c_{\notin f}$ respectively. For copyright-related unlearning, DM owners should have the right to restore erased concepts.

On one hand, concept restoration should not compromise the unlearning performance of previously erased concepts $c_{sub,f}$, where $c_{sub,f}$ is an arbitrary subset of c_f, $c_{sub,f} \in c_f$.

$$\mathcal{M}_{c^*_{sub,f}}(\boldsymbol{z}, c_{i,f}) = \mathcal{M}_{c_{sub,f}}(\boldsymbol{z}, c_{i,f}),\ s.t. \forall c_{i,f} \in c^*_{sub,f}, \tag{2}$$

where $c_{i,f} \in c_{sub,f}$. $c_{j,f}$ signifies the recovered concept, and $c^*_{sub,f}$ means $c_{sub,f}$ that removes $c_{j,f}$, *i.e.,* $c_{j,f} \in c_{sub,f}$ and $c_{j,f} \notin c^*_{sub,f}$. $\mathcal{M}_{c_{sub,f}}(\cdot)$ denotes DMs with $c_{sub,f}$ erased. $\boldsymbol{z}$ is randomly initialized Gaussian noise. On the other hand, unlearned DMs should be able to flawlessly reconstruct the generation performance of the recovered concept.

$$\mathcal{M}_{c^*_{sub,f}}(\boldsymbol{z}, c_{j,f}) = \mathcal{M}(\boldsymbol{z}, c_{j,f}), \tag{3}$$

where $\mathcal{M}(\cdot)$ represents the original DMs.

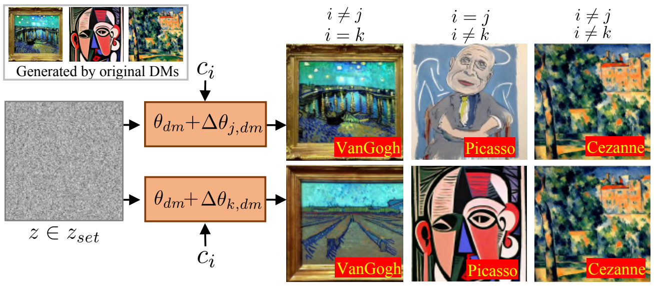


Figure 5. Decoupling weights for erasing different concepts. $\boldsymbol{\Delta\theta}_{j,dm}$ and $\boldsymbol{\Delta\theta}_{k,dm}$ are weight shifts of DMs designed to erase 'Picasso' and 'VanGogh' respectively.

③ **Regular Concept Preservation.** MU techniques should preserve the generative capability of DMs for $c_{\notin f}$,

$$\mathcal{M}_{c_f}(\boldsymbol{z}, c_{\notin f}) \approx \mathcal{M}(\boldsymbol{z}, c_{\notin f}). \tag{4}$$

④ **Watermark Preservation.** MU operations, including both concept erasure and restoration, should not compromise the generative capabilities of DMs for watermarks, such as specific images triggered by watermark prompts c_{wm}.

$$\mathcal{M}_{c^*_{sub,f}}(\boldsymbol{z}, c_{wm}) = \mathcal{M}(\boldsymbol{z}, c_{wm}). \tag{5}$$

⑤ **Memory Consumption.** DM unlearning can be implemented with limited memory consumption.

These objectives are illustrated in in Figure 4.

3.2. Fundamentals of SRS-ME

The follow questions cover the basic design aspects of weight decoupling (Q1∼Q3), concept-irrelevant unlearning (Q4∼Q6), and optimization decoupling (Q7). A1∼A3 provide the reason, feasibility, and solution for weight decoupling respectively. A4∼A6 define the concept representation, the unlearning loss and the stopping condition during the unlearning process, respectively. A7 theoretically validates the feasibility of optimization decoupling.

Q1 *Why is the proposal for weight decoupling made?*

A1: As formulated in Eqs. (2) and (3), weight shifts aimed at erasing various concepts should not interfere with each other. Therefore, we propose to decouple $\boldsymbol{\Delta\theta}_{dm}$ into $\boldsymbol{\Delta\theta}_{1\sim\mathrm{N},dm}$, where N is the number of erased concepts. $\boldsymbol{\Delta\theta}_{k,dm}$ is utilized to manipulate the specific forgotten concept $c_{k,f}$. Figure 5 shows the expected results.

NOTE. If readers prefer to avoid digging into the mathematical details of weight decoupling, they may skip answers for Q2 and Q3. Lines 1∼7 of Alg. 1 show the implementation details of weight decoupling.

Q2 *Can weights for erasing concepts be decoupled?*

A2: According to *A1*, we derive the paradigm of independent weight shifts that satisfy Eqs. (2), (3) and (5). We summarize these equations as follows

$$\mathcal{M}(\boldsymbol{z}, c_{j,f}; \boldsymbol{\theta}_1) = \mathcal{M}(\boldsymbol{z}, c_{j,f}; \boldsymbol{\theta}_2) \Rightarrow \boldsymbol{x}_0^{\Diamond} = \boldsymbol{x}_0^{\Box}, \tag{6}$$

where $\boldsymbol{x}_0^{\Diamond}$ and $\boldsymbol{x}_0^{\Box}$ represent latent representations produced by the DMs with parameters $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$, respectively. The index '0' denotes the final schedule time. Taking DDIM [18] as an example, given $\boldsymbol{x}_\mathrm{T} \sim \mathcal{N}(0, \mathbf{I})$, $\boldsymbol{z} \sim \mathcal{N}(0, \mathbf{I})$, the denoising process is expressed as

$$\boldsymbol{x}_{t-1} = \sqrt{\overline{\alpha}_{t-1}}\frac{\boldsymbol{x}_t - \sqrt{1-\overline{\alpha}_t}\boldsymbol{\epsilon}(\boldsymbol{x}_t)}{\sqrt{\overline{\alpha}_t}} + \sqrt{1-\overline{\alpha}_{t-1}-\delta_t^2}\boldsymbol{\epsilon}(\boldsymbol{x}_t) + \delta_t\boldsymbol{z}. \tag{7}$$

where $\boldsymbol{\epsilon}(\boldsymbol{x}_t)$ is predicted noise at the t-th timestamp. For clarity, we simply Eq. (7) as

$$\boldsymbol{x}_{t-1} = \lambda_1\boldsymbol{x}_t - \lambda_2\boldsymbol{\epsilon}(\boldsymbol{x}_t) + \lambda_3\boldsymbol{z}, \tag{8}$$

where $\lambda_1 = \frac{\sqrt{\overline{\alpha}_{t-1}}}{\sqrt{\overline{\alpha}_t}}$, $\lambda_2 = \sqrt{1-\overline{\alpha}_{t-1}-\delta_t^2} - \frac{\sqrt{\overline{\alpha}_{t-1}}\sqrt{1-\overline{\alpha}_t}}{\sqrt{\overline{\alpha}_t}}$, and $\lambda_3 = \delta_t$. Combined with Eqs. (6) and (8), we have

$$\begin{aligned} &\boldsymbol{x}_0^{\Diamond} = \boldsymbol{x}_0^{\Box} \\ &\Rightarrow \lambda_1\boldsymbol{x}_1^{\Diamond} - \lambda_2\boldsymbol{\epsilon}(\boldsymbol{x}_1^{\Diamond})^{\Diamond} = \lambda_1\boldsymbol{x}_1^{\Box} - \lambda_2\boldsymbol{\epsilon}(\boldsymbol{x}_1^{\Box})^{\Box}. \end{aligned} \tag{9}$$

Hence, $\boldsymbol{x}_0^{\Diamond} = \boldsymbol{x}_0^{\Box}$ could be satisfied when

$$\begin{aligned} &\boldsymbol{x}_1^{\Diamond} = \boldsymbol{x}_1^{\Box}, \\ &\boldsymbol{\epsilon}(\boldsymbol{x}_1^{\Diamond})^{\Diamond} = \boldsymbol{\epsilon}(\boldsymbol{x}_1^{\Box})^{\Box}. \end{aligned} \tag{10}$$

Similarly, $\boldsymbol{x}_t^{\Diamond} = \boldsymbol{x}_t^{\Box}$ can be satisfied when

$$\begin{aligned} &\boldsymbol{x}_{t+1}^{\Diamond} = \boldsymbol{x}_{t+1}^{\Box}, \\ &\boldsymbol{\epsilon}(\boldsymbol{x}_{t+1}^{\Diamond})^{\Diamond} = \boldsymbol{\epsilon}(\boldsymbol{x}_{t+1}^{\Box})^{\Box}. \end{aligned} \tag{11}$$

Notably, $\boldsymbol{x}_\mathrm{T}^{\Diamond} = \boldsymbol{x}_\mathrm{T}^{\Box} = \boldsymbol{x}_\mathrm{T}$. Therefore, the sufficient condition for $\boldsymbol{x}_0^{\Diamond} = \boldsymbol{x}_0^{\Box}$ can be formulated as

$$\forall_{t\in[0,\mathrm{T}]}\boldsymbol{\epsilon}(\boldsymbol{x}_t^{\Diamond})^{\Diamond} = \boldsymbol{\epsilon}(\boldsymbol{x}_t^{\Box})^{\Box}. \tag{12}$$

For text-guided DMs, we represent $\boldsymbol{\epsilon}(\boldsymbol{x})$ simply as:

$$\boldsymbol{\epsilon}(\boldsymbol{x}) = \boldsymbol{\epsilon}(\boldsymbol{x}, c_\emptyset) + \lambda_4(\boldsymbol{\epsilon}(\boldsymbol{x}, c_{j,f}) - \boldsymbol{\epsilon}(\boldsymbol{x}, c_\emptyset)) \tag{13}$$

Combined with Eq. (13), Eq. (12) could be satisfied when

$$\begin{aligned} &\forall_{t\in[0,\mathrm{T}]}\boldsymbol{\epsilon}(\boldsymbol{x}_t^{\Diamond}, c_\emptyset)^{\Diamond} = \boldsymbol{\epsilon}(\boldsymbol{x}_t^{\Box}, c_\emptyset)^{\Box}, \\ &\forall_{t\in[0,\mathrm{T}]}\boldsymbol{\epsilon}(\boldsymbol{x}_t^{\Diamond}, c_{j,f})^{\Diamond} = \boldsymbol{\epsilon}(\boldsymbol{x}_t^{\Box}, c_{j,f})^{\Box}. \end{aligned} \tag{14}$$

Based on Eqs. (6) and (14), Eq. (2) can be resolved when

$$\begin{aligned} &\forall_{t\in[0,\mathrm{T}]}\boldsymbol{\epsilon}(\boldsymbol{x}_t, c_\emptyset; \boldsymbol{\theta_1}) = \boldsymbol{\epsilon}(\boldsymbol{x}_t, c_\emptyset; \boldsymbol{\theta_2}), \\ &\forall_{t\in[0,\mathrm{T}]}\boldsymbol{\epsilon}(\boldsymbol{x}_t, c_{i,f}; \boldsymbol{\theta_1}) = \boldsymbol{\epsilon}(\boldsymbol{x}_t, c_{i,f}; \boldsymbol{\theta_2}). \end{aligned} \tag{15}$$

Similarly, Eqs. (3) and (5) will be solved when

$$\begin{aligned} &\forall_{t\in[0,\mathrm{T}]}\boldsymbol{\epsilon}(\boldsymbol{x}_t, c_\emptyset; \boldsymbol{\theta_1}) = \boldsymbol{\epsilon}(\boldsymbol{x}_t, c_\emptyset; \boldsymbol{\theta}_{dm}), \\ &\forall_{t\in[0,\mathrm{T}]}\boldsymbol{\epsilon}(\boldsymbol{x}_t, c_{j,f}; \boldsymbol{\theta_1}) = \boldsymbol{\epsilon}(\boldsymbol{x}_t, c_{j,f}; \boldsymbol{\theta}_{dm}), \\ &\forall_{t\in[0,\mathrm{T}]}\boldsymbol{\epsilon}(\boldsymbol{x}_t, c_{wm}; \boldsymbol{\theta_1}) = \boldsymbol{\epsilon}(\boldsymbol{x}_t, c_{wm}; \boldsymbol{\theta}_{dm}). \end{aligned} \tag{16}$$

Here, $\boldsymbol{x}_t$ and t can be any images (or latent representations of images) and timestamps. $\boldsymbol{\theta_1} = \boldsymbol{\theta}_{dm} + \sum_{k\in I^*_{sub}}\boldsymbol{\Delta\theta}_{k,dm}$ and $\boldsymbol{\theta_2} = \boldsymbol{\theta}_{dm} + \sum_{k\in I_{sub}}\boldsymbol{\Delta\theta}_{k,dm}$. I_{sub} denotes the arbitrary subset of $[1,\mathrm{N}]$, $c_{j,f}$ denotes the recovered concept, I^*_{sub} represents I_{sub} that removes the index j, $j \notin I^*_{sub}$. Conditions (**c1**∼**2**) enable Eqs. (15∼16) to have solutions.

To satisfy Eqs. (15∼16) for any image, the nonzero positions of $\boldsymbol{\Delta\theta}_{k,dm}$ should be image-independent (II),

$$\boldsymbol{\Delta\theta}_{k,dm} = \begin{cases} \boldsymbol{\Delta w}_k & II \\ \mathbf{0} & else. \end{cases} \tag{17}$$

c1: There are image-independent embedding update modules, *e.g.*, the 'to_k' layer of cross-attention modules is solely used for updating text embeddings,

$$\boldsymbol{e}_{to\_k}(c) = \boldsymbol{e}(c) \otimes \boldsymbol{w}_{to\_k}, \boldsymbol{w}_{to\_k} \in \mathbb{R}^{\mathrm{d}_{in} \times \mathrm{d}_{out}}, \quad (18)$$

where $\boldsymbol{e}(\cdot)$ represents fixed models such as CLIP, $\boldsymbol{e}(c) \in \mathbb{R}^{\mathrm{d}_{emb} \times \mathrm{d}_{in}}$ signifies embeddings of concepts c. d_{emb}, d_{in} and d_{out} indicate feature dimensions. $\otimes$ is matrix multiplication.

Based on Eqs. (17) and (18), we have

$$\begin{aligned} &\forall_{t \in [0,\mathrm{T}]} \boldsymbol{\epsilon}(\boldsymbol{x}_t, c; \boldsymbol{\theta}_{dm} + \boldsymbol{\Delta\theta}_{k,dm}) = \boldsymbol{\epsilon}(\boldsymbol{x}_t, c; \boldsymbol{\theta}_{dm}) \\ &\Rightarrow \boldsymbol{e}(c) \otimes \boldsymbol{\Delta w}_{k,II} = \boldsymbol{0}, \end{aligned} \quad (19)$$

where $\boldsymbol{w}_{II}$ indicates image-independent weights within $\boldsymbol{\theta}_{dm}$, $\boldsymbol{w}_{II} \in \boldsymbol{\theta}_{dm}$ and $\boldsymbol{\Delta w}_{k,II} \in \boldsymbol{\Delta\theta}_{k,dm}$. Hence, we further expand Eqs. (15$\sim$16) as

$$\begin{aligned} &\boldsymbol{e}(c_\emptyset) \otimes \Delta \boldsymbol{w}_{j,II} = \boldsymbol{0}, \\ &\boldsymbol{e}(c_{i,f}) \otimes \Delta \boldsymbol{w}_{j,II} = \boldsymbol{0}, \\ &\boldsymbol{e}(c_\emptyset) \otimes \sum_{k \in I^*_{sub}} \Delta \boldsymbol{w}_{k,II} = \boldsymbol{0}, \\ &\boldsymbol{e}(c_{j,f}) \otimes \sum_{k \in I^*_{sub}} \Delta \boldsymbol{w}_{k,II} = \boldsymbol{0}, \\ &\boldsymbol{e}(c_{wm}) \otimes \sum_{k \in I^*_{sub}} \Delta \boldsymbol{w}_{k,II} = \boldsymbol{0}, \end{aligned} \quad (20)$$

Eq. (20) could be satisfied when

$$\begin{aligned} &\forall_{k \in [1,\mathrm{N}]} \boldsymbol{e}(c_\emptyset) \otimes \Delta \boldsymbol{w}_{k,II} = \boldsymbol{0}, \\ &\forall_{k \in [1,\mathrm{N}]} \boldsymbol{e}(c_{wm}) \otimes \Delta \boldsymbol{w}_{k,II} = \boldsymbol{0}, \\ &\forall_{j,k \in [1,\mathrm{N}], j \neq k} \boldsymbol{e}(c_{j,f}) \otimes \Delta \boldsymbol{w}_{k,II} = \boldsymbol{0}. \end{aligned} \quad (21)$$

Namely, for each $\Delta \boldsymbol{w}_{k,II}$, it should satisfy the condition

$$\begin{aligned} &\boldsymbol{e}(c_\emptyset) \otimes \Delta \boldsymbol{w}_{k,II} = \boldsymbol{0}, \\ &\boldsymbol{e}(c_{wm}) \otimes \Delta \boldsymbol{w}_{k,II} = \boldsymbol{0}, \\ &\forall_{j \in [1,\mathrm{N}], j \neq k} \boldsymbol{e}(c_{j,f}) \otimes \Delta \boldsymbol{w}_{k,II} = \boldsymbol{0}. \end{aligned} \quad (22)$$

For clarity, Eq. (22) is expressed as

$$\boldsymbol{e}_m \otimes \Delta \boldsymbol{w}_{k,II} = \boldsymbol{0}. \quad (23)$$

The matrix $\boldsymbol{e}_m \in \mathbb{R}^{((\mathrm{N}+1)\cdot \mathrm{d}_{emb}) \times \mathrm{d}_{in}}$ represents

$$[\boldsymbol{e}(c_\emptyset)^\top; \boldsymbol{e}(c_{wm})^\top; \cdots; \boldsymbol{e}(c_{k-1,f})^\top; \boldsymbol{e}(c_{k+1,f})^\top; \cdots; \boldsymbol{e}(c_{\mathrm{N},f})^\top]^\top, \quad (24)$$

where $\top$ means the transpose operation.

Eq. (23) has solutions when $\mathrm{d}_{in} > r$, where r is the rank of $\boldsymbol{e}_m$ and $r \leq \min((\mathrm{N}+1) \cdot \mathrm{d}_{emb}, \mathrm{d}_{in})$. The answer *A3* will provide a detailed explanation for this.

c2: $\boldsymbol{e}(c) \in \mathbb{R}^{\mathrm{d}_{emb} \times \mathrm{d}_{in}}$, where $\mathrm{d}_{in} \gg \mathrm{d}_{emb}$ in DMs.

According to c2, it is evident that weight decoupling is feasible when erasing a limited number of concepts.

Q3 *How to resolve decoupled weight shifts?*

A3: The preceding discussion has clarified that decoupled weight shifts $\boldsymbol{\Delta w}_{k,II}$ should satisfy Eq. (23). To resolve this, we first employ the Gaussian Elimination approach to compute a set of constant particular solutions $\mathcal{S}$, which adheres to the condition

$$\boldsymbol{e}_m \otimes \mathcal{S}_i = \boldsymbol{0}, \quad (25)$$

where $\boldsymbol{e}_m \in \mathbb{R}^{((\mathrm{N}+1)\cdot \mathrm{d}_{emb}) \times \mathrm{d}_{in}}$, $\mathcal{S}_i \in \mathbb{R}^{\mathrm{d}_{in}}$ and $\mathcal{S} \in \mathbb{R}^{(\mathrm{d}_{in} - r) \times \mathrm{d}_{in}}$. $\mathrm{d}_{in} - r$ quantifies the number of particular solutions, and should be a constant greater than 0. r is the rank of $\boldsymbol{e}_m$, $r \leq \min((\mathrm{N}+1) \cdot \mathrm{d}_{emb}, \mathrm{d}_{in})$. Then, each column of layer weights within $\boldsymbol{\Delta w}_{k,II}$ can be formulated as a linear combination of solutions $\mathcal{S}$. For $\forall_{k \in [1,\mathrm{N}]} \boldsymbol{\Delta\theta}_{k,dm}$,

$$\boldsymbol{\Delta\theta}_{k,dm} = \begin{cases} (\boldsymbol{w}_{k,l} \otimes \mathcal{S})^\top & II \\ \boldsymbol{0} & else, \end{cases} \quad (26)$$

where l denotes the l-th layer of *II*, and $\boldsymbol{w}_{k,l}$ represents the learnable coefficients for linear combinations, $\boldsymbol{w}_{k,l} \in \mathbb{R}^{\mathrm{d}_{out} \times (\mathrm{d}_{in} - r)}$. Notably, to eliminate original biases in $\mathcal{S}$, we normalize each $\mathcal{S}_i$ to a unit vector.

Algorithm 1: SRS-ME†.

Input: The diffuser $\boldsymbol{\epsilon}(\cdot; \boldsymbol{\theta})$, the weights of original DMs $\boldsymbol{\theta}_{dm}$, N forgotten concepts $c_{i,f} \in c_f$, the inference dataset $\boldsymbol{x}_0 \in D$, the noise schedule $\bar{\alpha}_t$, the hyperparameters λ and β, image-independent layers II within DMs.

Output: The fine-tuned weight shifts $\boldsymbol{\Delta\theta}_{i \in [1,\mathrm{N}], dm}$.

```
/*Weight decoupling for constructing Δθ_{j,dm}.*/
for c_{j,f} ∈ c_f do
    e_m = [e(c_∅)^⊤, e(c_cm)^⊤, e(c_{1,f})^⊤, ···, e(c_{j−1,f})^⊤, e(c_{j+1,f})^⊤, ···, e(c_{N,f})^⊤]^⊤;
    Obtain solutions S for e_m ⊗ S = 0;
    Initialize learnable variables w_{j,l} with zero.
    Δθ_{j,dm} = { (w_{j,l} ⊗ (βS))^⊤   II
                { 0                    else,
end
/*Concept-irrelevant unlearning.*/
for n, x_0 ∈ D do
    Randomly select a sampling step t;
    x_t = sqrt(ᾱ_t) x_0 + sqrt(1 − ᾱ_t) ε, ε ∈ N(0, I);
    ε_{c_f} = ε(x_t, c_f; θ_dm); ε_{c_∅} = ε(x_t, c_∅; θ_dm);
    for c_{j,f} ∈ c_f do
        ε'_{c_{j,f}} = ε(x_t, c_{j,f}; θ_dm + Σ_{i∈[1,N]} Δθ_{i,dm});
        ε'_{c_∅} = ε(x_t, c_∅; θ_dm + Σ_{i∈[1,N]} Δθ_{i,dm}));
        Calculate L_cor(c_{j,f}, Σ_{i∈[1,N]} Δθ_{i,dm})) with Eq. (28);
        Calculate η_j using Eq. (35);
    end
    Calculate L^n_mom using Eq. (37);
    if L^n_mom ≤ τ then
        break;
    end
    min_W L^† = λ|| Σ_{i∈[1,N]} Δθ_{i,dm})||_p + Σ_{j=1}^{N} η_j L_cor(c_{j,f}, Σ_{i∈[1,N]} Δθ_{i,dm})
end
```

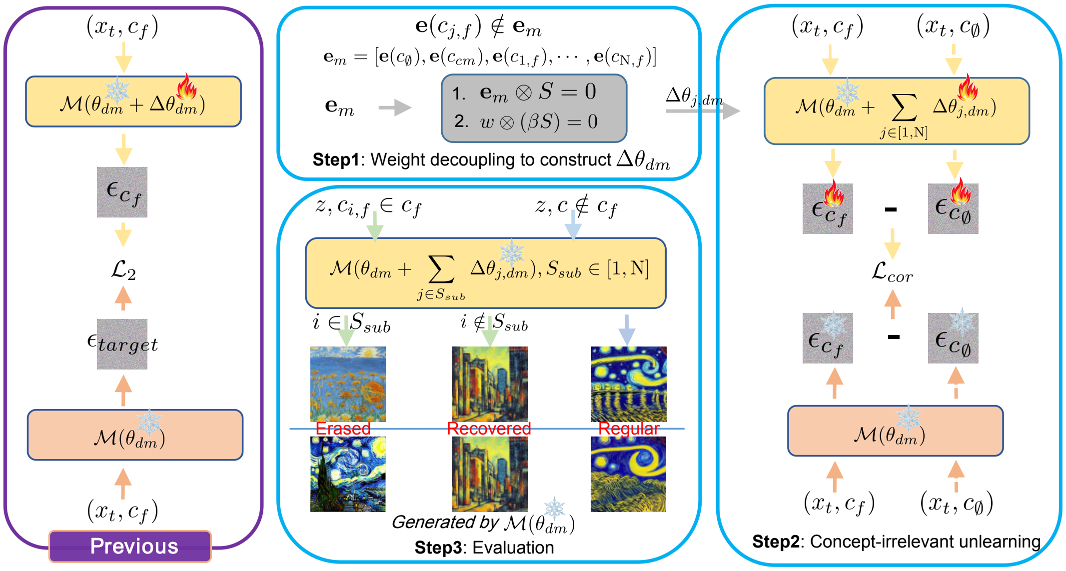


Figure 6. Comparison between our SRS-ME† and previous DM unlearning methods. 'Ice flowers' and 'flames' represent frozen and optimizable model weights respectively. $\boldsymbol{\epsilon}_{target}$ is the predefined noise distribution unrelated to c_f. $\mathcal{L}_2$ is ℓ_2-norm. $\boldsymbol{x}_t$ means latent image representations. $\boldsymbol{z} \in \mathcal{N}(0, \mathbf{I})$. c_f and $c_\emptyset$ signify N forgotten prompts and one blank prompt, respectively. $\mathcal{L}_{cor}$ is the correlation loss. SRS-ME† separates optimizable weights as $\boldsymbol{\Delta\theta}_{1\sim \mathrm{N},dm}$. $\boldsymbol{\Delta\theta}_{j,dm}$ aims to erase a specific concept $c_{j,f} \in c_f$.

After constructing optimizable variables $\boldsymbol{w}_{k,l}$, we proceed to illustrate their supervision function.

Q4 *How to define the representations $\boldsymbol{r}_{c_f}$ of c_f?*

A4: $\boldsymbol{r}_{c_f}$ represents the memory within DMs for concepts c_f. In practical terms, $\boldsymbol{r}_{c_f}$ can be quantified by observing how generation changes when c_f is introduced. Specifically, during the iterative denoising process, we calculate the degree of generation change $\boldsymbol{r}_{c_f,t-1}(\boldsymbol{\theta}_{dm}) \in \mathbb{R}^{\mathrm{h}\times\mathrm{w}\times\mathrm{d}}$ at each denoising timestamp t, as depicted in Figure 3 (d).

$$\boldsymbol{r}_{c_f,t-1}(\boldsymbol{\theta}_{dm}) = \boldsymbol{\epsilon}(\boldsymbol{x}_t, c_f; \boldsymbol{\theta}_{dm}) - \boldsymbol{\epsilon}(\boldsymbol{x}_t, c_\emptyset; \boldsymbol{\theta}_{dm}) \propto \boldsymbol{x}^{c_f}_{t-1} - \boldsymbol{x}^{c_\emptyset}_{t-1}, \tag{27}$$

where $\boldsymbol{\theta}_{dm}$ is weights of original DMs. $\boldsymbol{x}^{c_f}$ and $\boldsymbol{x}^{c_\emptyset}$ are denoised representations supervised by c_f and $c_\emptyset$, respectively. $c_\emptyset$ means the blank prompt. c_f and $c_\emptyset$ can be replaced with phrases, *e.g.*, using "a picture of a VanGogh's painting" to replace c_f and "a picture of a painting" to replace $c_\emptyset$.

Q5 *How to estimate the unlearning degree?*

A5: According to *A4*, we regard the concept representations $\boldsymbol{r}_{c_f,t}(\boldsymbol{\theta}_{dm})$ calculated by original DMs as pseudo-ground truth. The unlearning objective is to ensure that $\boldsymbol{r}_{c_f,t}(\boldsymbol{\theta}_{dm} + \boldsymbol{\Delta\theta}_{k,dm})$ calculated by unlearned DMs deviates from its corresponding pseudo-ground truth, where $\boldsymbol{\Delta\theta}_{k,dm}$ denotes learnable weight shifts. Hence, we calculate the concept correlation between unlearned and original DMs to quantify the unlearning degree,

$$\mathcal{L}^t_{cor}(c_f, \boldsymbol{\Delta\theta}_{k,dm}) = \frac{\sum \boldsymbol{r}_{c_f,t}(\boldsymbol{\theta}_{dm} + \boldsymbol{\Delta\theta}_{k,dm}) \odot \boldsymbol{r}_{c_f,t}(\boldsymbol{\theta}_{dm})}{\mathrm{h}\cdot\mathrm{w}\cdot\mathrm{d}}, \tag{28}$$

where $\odot$ is the element-wise product.

Q6 *Excessive forgetting significantly affects the model generation performance for regular concepts. Can we define a condition to monitor the unlearning degree for c_f and timely cease the unlearning process?*

A6: In Eq. (28), $\boldsymbol{\Delta\theta}_{dm}$ is initially set to zero, leading to a high initial value for $\mathcal{L}^t_{cor}(c_f, \boldsymbol{\Delta\theta}_{k,dm})$. As the unlearning process progresses, this value is expected to decrease, aiming to decorrelate $\boldsymbol{r}_{c_f,t}(\boldsymbol{\theta}_{dm} + \boldsymbol{\Delta\theta}_{k,dm})$ from $\boldsymbol{r}_{c_f,t}(\boldsymbol{\theta}_{dm})$. Inspired by vector orthogonality, we employ $\mathcal{L}^t_{cor}(c_f, \boldsymbol{\Delta\theta}_{k,dm}) = \mathbf{0}$ as the stopping condition, where the concept representations $\boldsymbol{r}_{c_f}$ of the unlearned and original DMs are considered to be uncorrelated.

Q7 *Can the training for $\forall_k \boldsymbol{\Delta\theta}_{k,dm}$ be separated?*

A7: The training process for decoupled weight shifts can be separated when the following condition is satisfied,

$$\mathcal{L}^t_{cor}(c_{i,f}, \sum_{k\in I_{sub}, k\neq i} \boldsymbol{\Delta\theta}_{k,dm}) = \mathcal{L}^t_{cor}(c_{i,f}, \mathbf{0}). \tag{29}$$

This occurs because the unlearning loss of $c_{i,f}$ is unrelated to $\forall_{k\in I_{sub}, i\neq k}$ $\boldsymbol{\Delta\theta}_{k,dm}$, resulting in no gradient backward propagation. $\mathbf{0}$ indicates the zero matrices with the same shape as $\boldsymbol{\theta}_{dm}$. According to Eqs. (17) and (28), we have

$$\begin{aligned}
&\mathcal{L}^t_{cor}(c_{i,f}, \boldsymbol{\Delta}_w) - \mathcal{L}^t_{cor}(c_{i,f}, \mathbf{0}) \\
&\propto \sum \left(\boldsymbol{r}_{c_{i,f},t}(\boldsymbol{w}_{II} + \boldsymbol{\Delta}_w) - \boldsymbol{r}_{c_{i,f},t}(\boldsymbol{w}_{II})\right) \odot \boldsymbol{r}_{c_{i,f},t}(\boldsymbol{w}_{II}) \\
&\propto \sum \left(\boldsymbol{r}_{c_{i,f},t}(\boldsymbol{w}_{II} + \boldsymbol{\Delta}_w) - \boldsymbol{r}_{c_{i,f},t}(\boldsymbol{w}_{II})\right)
\end{aligned} \tag{30}$$

where $\boldsymbol{\Delta}_w = \sum_{k\in I_{sub}, k\neq i} \Delta\boldsymbol{w}_{k,II}$. According to Eq. (27), we express Eq. (30) as

$$\begin{aligned}
&\boldsymbol{r}_{c_{i,f},t}(\boldsymbol{w}_{II} + \boldsymbol{\Delta}_w) - \boldsymbol{r}_{c_{i,f},t}(\boldsymbol{w}_{II}) \\
&= \left(\boldsymbol{\epsilon}(\boldsymbol{x}_t, c_{i,f}; \boldsymbol{w}_{II} + \boldsymbol{\Delta}_w) - \boldsymbol{\epsilon}(\boldsymbol{x}_t, c_\emptyset; \boldsymbol{w}_{II} + \boldsymbol{\Delta}_w)\right) - \\
&\qquad \left(\boldsymbol{\epsilon}(\boldsymbol{x}_t, c_{i,f}; \boldsymbol{w}_{II}) - \boldsymbol{\epsilon}(\boldsymbol{x}_t, c_\emptyset; \boldsymbol{w}_{II})\right), \\
&= \left(\boldsymbol{\epsilon}(\boldsymbol{x}_t, c_{i,f}; \boldsymbol{w}_{II} + \boldsymbol{\Delta}_w) - \boldsymbol{\epsilon}(\boldsymbol{x}_t, c_{i,f}; \boldsymbol{w}_{II})\right) - \\
&\qquad \left(\boldsymbol{\epsilon}(\boldsymbol{x}_t, c_\emptyset; \boldsymbol{w}_{II} + \boldsymbol{\Delta}_w) - \boldsymbol{\epsilon}(\boldsymbol{x}_t, c_\emptyset; \boldsymbol{w}_{II})\right),
\end{aligned} \tag{31}$$

Since $\epsilon(\boldsymbol{x}_t, c; \boldsymbol{\theta}_{dm} + \Delta\boldsymbol{\theta}_{k,dm}) = \epsilon(\boldsymbol{x}_t, c; \boldsymbol{\theta}_{dm})$ when $\boldsymbol{e}(c) \otimes \Delta\boldsymbol{w}_{k,II} = \mathbf{0}$, and the weight decoupling makes $\boldsymbol{e}(c_\emptyset) \otimes \boldsymbol{\Delta}_w = \mathbf{0}$, $\boldsymbol{e}(c_{i,f}) \otimes \boldsymbol{\Delta}_w = \mathbf{0}$, we have

$$\begin{aligned} &\epsilon(\boldsymbol{x}_t, c_{i,f}; \boldsymbol{w}_{II} + \boldsymbol{\Delta}_w) - \epsilon(\boldsymbol{x}_t, c_{i,f}; \boldsymbol{w}_{II}) = \mathbf{0}, \\ &\epsilon(\boldsymbol{x}_t, c_\emptyset; \boldsymbol{w}_{II} + \boldsymbol{\Delta}_w) - \epsilon(\boldsymbol{x}_t, c_\emptyset; \boldsymbol{w}_{II}) = \mathbf{0}. \end{aligned} \tag{32}$$

Namely, for each $c_{i,f} \in c_f$,

$$\begin{aligned} &\mathcal{L}^t_{cor}(c_{i,f}, \boldsymbol{\Delta}_w) - \mathcal{L}^t_{cor}(c_{i,f}, \mathbf{0}) \\ &\propto \sum \left(\boldsymbol{r}_{c_{i,f},t}(\boldsymbol{w}_{II} + \boldsymbol{\Delta}_w) - \boldsymbol{r}_{c_{i,f},t}(\boldsymbol{w}_{II})\right) = \mathbf{0}. \end{aligned} \tag{33}$$

This makes the training process of $\forall_k \boldsymbol{\Delta\theta}_{k,dm}$ separable.

3.3. Variants of SRS-ME

The preceding part has established the unlearning loss and stopping condition while demonstrating the feasibility of weight and optimization decoupling. Next, we introduce three variants, SRS-ME, SRS-ME†, and SRS-ME‡, each tailored to address distinct scenarios.

SRS-ME†. SRS-ME† optimizes the weight shifts $\forall_{j\in[1,\mathrm{N}]} \boldsymbol{\Delta\theta}_{j,dm}$ simultaneously, which incurs higher memory consumption but accelerates the unlearning process,

$$\begin{aligned} &\min_{\mathcal{W}} \mathcal{L}^\dagger = \sum_{i=1}^{\mathrm{N}} \eta_i \mathcal{L}_{cor}(c_{i,f}, \sum_{j=1}^{\mathrm{N}} \boldsymbol{\Delta\theta}_{j,dm}) + \lambda \| \sum_{j=1}^{\mathrm{N}} \boldsymbol{\Delta\theta}_{j,dm} \|_p, \\ &s.t.\ \forall_{i\in[1,\mathrm{N}]} \mathcal{L}_{cor}(c_{i,f}, \sum_{j=1}^{\mathrm{N}} \boldsymbol{\Delta\theta}_{j,dm}) = \mathbf{0}, \end{aligned} \tag{34}$$

where λ denotes the hyperparameter, η_i is used to balance the losses of multiple concepts,

$$\eta_i = \frac{\| \max_{k\in[1,\mathrm{N}]} \mathcal{L}_{cor}(c_{k,f}, \sum_{j=1}^{\mathrm{N}} \boldsymbol{\Delta\theta}_{j,dm}) \|_2}{\| \mathcal{L}_{cor}(c_{i,f}, \sum_{j=1}^{\mathrm{N}} \boldsymbol{\Delta\theta}_{j,dm}) \|_2}. \tag{35}$$

$\boldsymbol{\Delta\theta}_{j,dm}$ is formulated as

$$\boldsymbol{\Delta\theta}_{j,dm} = \begin{cases} (\boldsymbol{w}_{j,l} \otimes (\beta\mathcal{S}))^\top & II \\ \mathbf{0} & else, \end{cases} \tag{36}$$

where β represents a scaling factor. II means the image-independent layers, such as the 'to_k' and 'to_v' layers of cross-attention modules. $\mathcal{W}$ is the set of optimizable variables $\boldsymbol{w}_{j,l}$ utilized to replace image-independent layers. $\|\cdot\|_p$ denotes the p-norm function.

To mitigate the impact of unlearning on regular concepts, as described in Eq. (4), we propose the following settings to restrict modifications to model weights:

- **s1** The scaling factor β is set to a small value;
- **s2** $\boldsymbol{w}_{j,l}$ is initialized with zero matrices;
- **s3** $\| \sum_{j=1}^{\mathrm{N}} \boldsymbol{\Delta\theta}_{j,dm} \|_p$ restricts the weight deviation of the unlearned DMs from the original ones;
- **s4** The stopping condition avoids unlearning substantial information associated with regular concepts.

To realize the condition of zero relevance in Eq. (34), we utilize the momentum statistic method since $\mathcal{L}_{cor}(c_{i,f}, \boldsymbol{\Delta\theta}_{j,dm})$ is affected by noisy inputs $\boldsymbol{x}_t$. Early stopping is activated once $\mathcal{L}^n_{mom} \leq \tau$, where τ denotes a threshold with a small value.

$$\mathcal{L}^n_{mom} = \alpha \mathcal{L}^{n-1}_{mom} + (1-\alpha) \sum_{i=1}^{\mathrm{N}} \eta_i \mathcal{L}_{cor}(c_{i,f}, \sum_{j=1}^{\mathrm{N}} \boldsymbol{\Delta\theta}_{j,dm}). \tag{37}$$

where n represents the number of iterations.

Taking SRS-ME† as an example, we show figure descriptions in Figure 6 and implementation details in Alg. 1.

SRS-ME. Section 3.2 demonstrates the feasibility of optimization decoupling, namely, separately optimizing the decoupled weight shifts $\forall_{j\in[1,\mathrm{N}]} \boldsymbol{\Delta\theta}_{j,dm}$. Fine-tunning each $\boldsymbol{\Delta\theta}_{j,dm}$ can be realized by setting N in SRS-ME† to 1. While this setting is more time-consuming, it significantly reduces memory consumption.

Evaluation for SRS-ME and SRS-ME†. Benefiting from weight decoupling, one can randomly combine various $\boldsymbol{\Delta\theta}_{i,dm}$ to erase associated concepts, *e.g.*, DMs with $\boldsymbol{\theta}_{dm} + \sum_{i\in\{j,k\}} \boldsymbol{\Delta\theta}_{i,dm}$ eliminate concepts $c_{j,f}$ and $c_{k,f}$. Furthermore, concept restoration is achieved by directly removing the corresponding weight shifts. Additionally, watermark preservation is accomplished by incorporating watermark prompts as constant terms in $\boldsymbol{e}_m$ of Eq. (24).

SRS-ME‡. SRS-ME and SRS-ME† are specifically designed to flexibly manipulate concepts and preserve model watermarks, allowing only the image-independent layers to be fine-tuned. However, certain concepts, such as 'Nudity,' should not be recovered. Furthermore, in cases where researchers completely abandon watermark considerations, they can manipulate arbitrary model weights, as done in prior work [14], [15]. To erase concepts under such scenarios, we introduce SRS-ME‡ and formulate it as follows:

$$\min_{\boldsymbol{\Delta\theta}_{dm}} \mathcal{L}^\ddagger = \sum_{i=1}^{\mathrm{N}} \eta_i \mathcal{L}_{cor}(c_{i,f}, \boldsymbol{\Delta\theta}_{dm}) + \lambda \|\boldsymbol{\Delta\theta}_{dm}\|_p, s.t.\ \mathcal{L}_{mom} \leq \tau, \tag{38}$$

where η_i and $\mathcal{L}_{mom}$ can be calculated by replacing $\sum_{j=1}^{\mathrm{N}} \boldsymbol{\Delta\theta}_{j,dm}$ with $\boldsymbol{\Delta\theta}_{dm}$ in Eqs. (35) and (37), repestively. The comparison between variants is depicted in Table 1.

TABLE 1. COMPARISON BETWEEN SRS-ME VARIANTS. 'W-P', 'U-F' AND 'W-F' DENOTE THE WATERMARK PRESERVATION, UNLEARNING FLEXIBILITY AND WEIGHT FLEXIBILITY, RESPECTIVELY.

Methods	Time-efficient	Memory-efficient	W-P	U-F	W-F
SRS-ME†	✓	-	✓	✓	-
SRS-ME	-	✓	✓	✓	-
SRS-ME‡	✓	-	-	-	✓

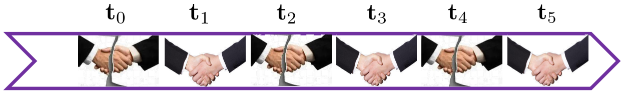
4. Experiments

4.1. Experimental Settings

Implementation Details. We follow prior works [15], [20] to unlearn concepts from Stable Diffusion [6]. For SRS-ME and SRS-ME†, the optimization process utilizes the Adam optimizer with a learning rate of 0.1. Only image-independent layers are fine-tuned. For SRS-ME‡, we set

TABLE 2. Quantitative results (FID/ACC/LPIPS) of SRS-ME on style unlearning. ORI denotes the results of original DMs. ▬, ▬, and ▬ denote the evaluation performance for erased, recovered and regular concepts, respectively.

Scene	t	c_0:Cezanne	c_1:VanGogh	c_2:Picasso	c_3:JacksonPo...	c_4:Caravaggio	c_5:KeithHaring	c_6:KellyMcK...	c_7:TylerEdlin	c_8:KilianEng
ORI	-	0.00/98.0/.00	0.00/90.4/.00	0.00/98.8/.00	0.00/96.0/.00	0.00/99.6/.00	0.00/98.4/.00	0.00/99.6/.00	0.00/100/.00	0.00/100/.00
Scene$_1$	t_0	209/12.4/.338	220/36.8/.388	200/6.40/.336	275/26.4/.329	203/38.0/.312	217/15.2/.596	56.0/95.2/.153	135/98.2/.125	110/99.2/.126
	t_1	0.75/98.0/.00	2.12/90.8/.00	199/6.00/.336	277/26.8/.328	204/38.3/.313	209/25.2/.526	47.1/98.8/.096	119/100/.102	101/99.2/.107
	t_3	0.78/98.4/.00	3.28/90.4/.00	1.25/98.8/.00	277/26.8/.329	203/38.0/.312	209/25.6/.526	44.7/99.6/.097	99.9/100/.080	85.8/100/.066
Scene$_2$	t_0	209/12.4/.338	220/36.8/.388	200/6.40/.336	275/26.4/.329	203/38.0/.312	217/15.2/.596	56.0/95.2/.153	135/98.2/.125	110/99.2/.126
	t_1	209/12.4/.339	220/38.0/.388	200/6.40/.336	0.18/96.0/.00	2.83/99.6/.00	219/16.0/.596	75.1/86.4/.192	180/98.8/.141	117/98.8/.141
	t_3	209/12.4/.338	220/38.0/.388	200/6.40/.336	0.14/96.0/.00	2.08/100/.00	2.11/98.4/.00	61.0/92.4/.168	94.7/100/.076	115/98.4/.151
Scene$_3$	t_0	255/15.6/.349	95.9/86.2/.153	218/2.00/.351	170/64.5/.151	182/49.2/.308	148/94.4/.435	57.6/95.6/.134	97.2/100/.090	146/84.0/.239
	t_1	1.04/98.0/.00	67.4/87.8/.125	1.01/98.8/.00	39.5/96.0/.026	182/48.0/.307	42.2/99.2/.168	30.4/98.4/.035	60.7/100/.024	76.3/.053/100.
	t_2	1.00/98.0/.00	206/66.8/.393	1.15/98.8/.00	280/29.2/.325	182/48.0/.308	59.8/98.0/.264	46.6/97.6/.096	103/100./.086	99.5/99.2/.123
	t_3	1.02/98.0/.00	67.4/90.4/.125	0.98/98.8/.00	280/29.2/.326	182/48.0/.308	62.4/96.8/.262	47.5/98.4/.097	60.5/100/.023	77.0/100/.053
	t_4	1.04/98.0/.00	67.5/89.6/.125	1.14/98.8/.00	280/28.8/.326	182/48.0/.307	210/.527/25.2	52.2/97.2/.116	88.5/100/.063	83.0/99.6/.061
Scene$_4$	t_0	188/36.0/.256	219/37.2/.395	214/5.20/.365	184/60.4/.245	96.7/98.8/.182	52.5/97.2/.189	65.1/95.6/.160	50.8/100./.016	108/99.2/.147
	t_1	0.77/98.0/.00	219/37.2/.395	215/5.20/.364	176/64.0/.211	96.0/99.6/.166	45.4/96.8/.147	49.7/98.4/.096	49.9/100/.016	113/100/.156
	t_2	0.81/98.0/.00	219/37.6/.395	215/5.20/.365	365/5.60/.463	246/19.6/.379	135/59.6/.402	77.4/86.0/.176	61.0/100/.024	113/99.6/.151
	t_3	0.76/98.0/.00	3.82/90.4/.00	215/5.20/.364	355/2.80/.458	219/28.0/.340	132/58.2/.395	73.3/90.0/.166	52.0/100/.016	101/98.4/.123
	t_4	0.87/98.0/.00	2.82/90.4/.00	214/5.20/.365	354/2.80/.458	219/28.4/.340	288/11.6/.623	82.1/85.6/.185	77.0/100/.045	103/98.4/.123
	t_5	0.92/98.0/.00	4.11/90.4/.00	1.06/98.8/.00	274/26.4/.343	231/28.0/.368	278/17.6/.615	68.4/90.8/.154	78.3/100/.046	91.8/99.6/.098



	$\mathbf{t}_0$	$\mathbf{t}_1$	$\mathbf{t}_2$	$\mathbf{t}_3$	$\mathbf{t}_4$	$\mathbf{t}_5$
Scene$_1$	$\mathbf{c}_{0\sim5}$	$\mathbf{c}_{0\sim1}$	-	$\mathbf{c}_2$	-	-
Scene$_2$	$\mathbf{c}_{0\sim5}$	$\mathbf{c}_{3\sim4}$	-	$\mathbf{c}_5$	-	-
Scene$_3$	$\mathbf{c}_{0,2,4}$	$\mathbf{c}_{0,2}$	$\mathbf{c}_{1,3}$	$\mathbf{c}_1$	$\mathbf{c}_5$	-
Scene$_4$	$\mathbf{c}_{0\sim2}$	$\mathbf{c}_0$	$\mathbf{c}_{3\sim4}$	$\mathbf{c}_1$	$\mathbf{c}_5$	$\mathbf{c}_2$

Figure 7. Experimental scenario clarification.

the learning rate to 1e-5 and fine-tune all weights of cross-attention modules. For all variants, the maximum iteration is set to 1000, and an early stopping strategy is employed. Without specific statement, hyperparameters α in Eq. (37), β in Eq. (36), λ in Eq. (34) and τ are set to 0.9, 1e-4, 1e-6, 1e-4 respectively. $\|\boldsymbol{\theta}\|_p = \frac{\|\boldsymbol{\theta}\|_1}{\mathrm{M}}$, where M represents the number of layers. All experiments are conducted on 2 RTX 3090 GPUs. The code is accessible at https://anonymous.4open.science/r/SRS-ME-140B/README.md.

Evaluation Metrics. ① Fréchet Inception Distance (FID) [25] between images generated by unlearned and original DMs; ② Classification accuracy (ACC) of pre-trained classification models for images generated by DMs; ③ Perceptual Image Patch Similarity (LPIPS) [48] between images generated by unlearned and original DMs.

For the *style classification model*, we consider a blank concept and nine artist styles: 'Cezanne', 'Van Gogh', 'Picasso', 'Jackson Pollock', 'Caravaggio', 'Keith Haring', 'Kelly McKernan', 'Tyler Edlin', and 'Kilian Eng'. For each category, we generate 1000 images using original DMs with artist styles (or '') as prompts. 70% data is allocated for training purposes, while the remaining 30% is reserved for testing. Only fully connected (FC) layers of the pre-trained ResNet18 model [49] is optimized with 20 epochs. The cyclical learning rate [50] is employed with the maximum learning rate of 1e-2.

For the *object classification network*, we directly utilize the pre-trained ResNet50 [49]. The object unlearning evaluation utilizes categories in Imagenette [51] following [15], [20], including 'chain saw', 'church', 'gas pump', 'tench', 'garbage truck', 'english springer', 'golf ball', 'parachute', and 'french horn'. We omit 'cassette player' because the pretrained ResNet50 exhibits classification accuracy lower than 40% on data generated by original DMs with 'cassette player' as the prompt, *e.g.*, ResNet50 confidently misclassifies the 'cassette player' guided data as 'tape' or 'radio'.

Baselines. We use advanced unlearning methods as baselines, *i.e.*, FMN [14], ESD [15] and AbConcept [20].

Evaluation Data. We yield 250 samples using unlearned (or original) DMs for each evaluation prompt, *i.e.*, 50 seeds per prompt and 5 samples per seed.

4.2. Experimental Scenarios

We divide evaluated concepts into *forgotten, recovered, and regular* concepts. Then, we assume four different scenarios and illustrate them in Figure 7. For example, in t_3 of *Scene*$_1$, $c_{3\sim5}$, $c_{0\sim2}$, and $c_{>5}$ are designated as forgotten concepts, recovered concepts, and regular concepts, respectively. These scenarios involve two cases:

- Simultaneous erasure of multiple concepts, *e.g.*, unlearning $c_{0\sim5}$ simultaneously at t_0 in *Scene*$_1$;
- Iterative concept erasures, such as unlearning $c_{0\sim2}$ at t_0 and then $c_{3\sim4}$ at t_2 in *Scene*$_4$.

Our proposed methods can handle these cases with the same setting. At each timestamp, $\boldsymbol{e}_m$ contains all other previously erased (or recovered) concepts. For instance, $\boldsymbol{\Delta\theta}_1$ at t_0 in *Scene*$_1$ is calculated based on $\boldsymbol{e}_m = [\boldsymbol{e}(c_\emptyset), \boldsymbol{e}(c_{cm}), \boldsymbol{e}(c_0), \boldsymbol{e}(c_{2\sim5})]$, and $\boldsymbol{\Delta\theta}_1$ at t_2 in *Scene*$_3$ is calculated based on $\boldsymbol{e}_m = [\boldsymbol{e}(c_\emptyset), \boldsymbol{e}(c_{cm}), \boldsymbol{e}(c_0), \boldsymbol{e}(c_{2\sim4})]$.

Notably, in iterative concept erasure, we do not re-calculate the weight shifts for previously erased concepts. Only weight shifts for erasing new concepts are added. For example, in *Scene*$_4$ at t_4, we only finetune the weight shift $\boldsymbol{\Delta\theta}_{5,dm}$ for erasing c_5, and directly utilize $\boldsymbol{\theta}_{dm} + \sum_{i\in[2,3,4,5]} \boldsymbol{\Delta\theta}_{i,dm}$ as the unlearned model weights.

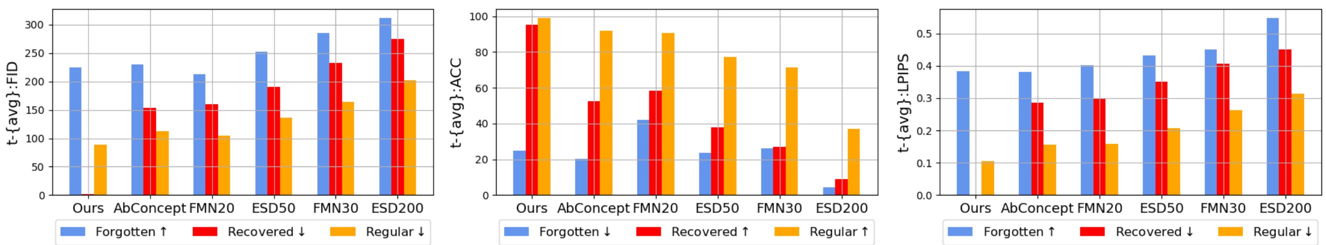


Figure 8. Performance comparison on $Scene_1$ of style unlearning.

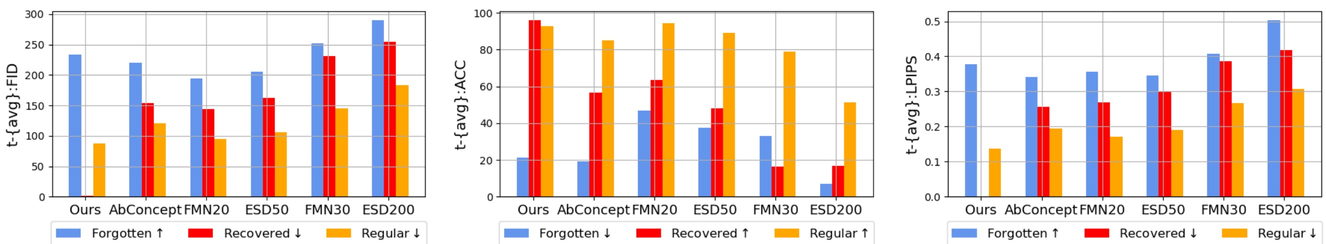


Figure 9. Performance comparison on $Scene_4$ of style unlearning.

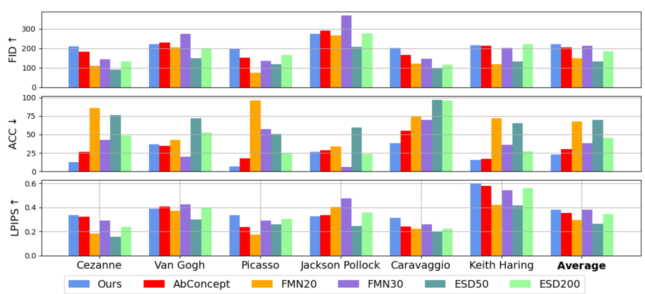


Figure 10. Performance comparison of single style unlearning for models combined to erase multiple styles. The x-axis indicates the erased styles.

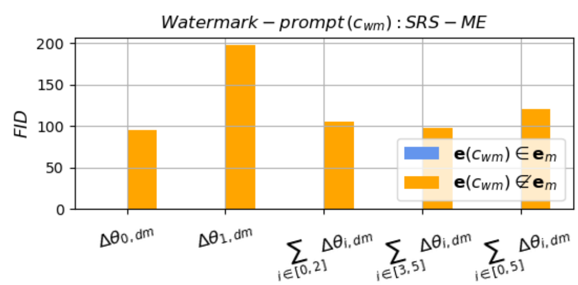


Figure 11. Watermark preservation.

4.3. Evaluation for SRS-ME

SRS-ME optimizes decoupled weight shifts separately.

4.3.1. Style unlearning. Table 2 presents the quantitative results of SRS-ME on style unlearning. Our findings indicate that: (1) SRS-ME effectively unlearns forgotten styles. For forgotten styles, SRS-ME significantly increases their FID and LPIPS metric values, and decreases their ACC metric values; (2) Concept restoration of SRS-ME does not compromise the unlearning performance of other forgotten styles. After concept restoration, the metrics FID/ACC/LPIPS of other previously erased styles remain unchanged. (3) Additionally, the DM unlearning at all timestamps does not affect the generation performance on the blank and watermark prompts. This is because we utilize these prompts as a constant vectors in $\boldsymbol{e}_m$ of Eq. (24). This demonstrates the feasibility of our SRS-ME in preserving model watermarks, as shown in Figure 11; (3) While SRS-ME affects the generation performance of DMs regarding regular styles, this effect remains within acceptable bounds. For instance, the classification model achieves high classification accuracy for data generated by unlearned DMs with regular styles as prompts.

To further explore the effectiveness of our approach in multi-style erasure, we compare it with state-of-the-art DM unlearning methods. (1) For each scene, we calculate the average metric values for forgotten, recovered, and regular styles separately. These values respectively measure the unlearning, restoration, and preservation performance. (2) For all methods, we optimize model weights separately for each forgotten style instead of sequentially fine-tuning them, as restoring early weights inevitably impacts the erasure performance of later styles in sequentially fine-tuning. (3) Notably, *since all DM unlearning methods are capable of removing undesirable styles if model preservation performance is not considered, our comparative experiments are conducted under comparable unlearning performance and focus on comparing preservation and restoration performance.* To ensure fairness in comparison, we carefully adjust the attack iterations for each unlearning method. Figure 10 displays a comparison of single style unlearning performance for models combined to erase multiple styles.

Figures 8 and 9 illustrate the performance comparison on $Scene_1$ and $Scene_4$ respectively. The findings reveal that existing methods show significant interactions

TABLE 3. QUANTITATIVE RESULTS (FID/ACC/LPIPS/) OF SRS-ME ON OBJECT UNLEARNING. ORI DENOTES THE RESULTS OF ORIGINAL DMS. (blue), (red), AND (gray) DENOTE THE EVALUATION PERFORMANCE FOR ERASED, RECOVERED AND REGULAR CONCEPTS, RESPECTIVELY.

Scene	t	c_0:ChainSaw	c_1:Church	c_2:GasPump	c_3:Tench	c_4:GarbageT...	c_5:E.Springer	c_6:GolfBall	c_7:Parachute	c_8:FrenchHorn
ORI	-	0.00/91.6/0.00	0.00/80.4/0.00	0.00/60.0/0.00	0.00/81.6/0.00	0.00/84.8/0.00	0.00/95.6/0.00	0.00/97.6/0.00	0.00/93.2/0.00	0.00/100/0.00
Scene$_1$	t_0	331/1.20/.331	216/48.8/.386	261/2.00/.443	167/12.4/.358	317/2.80/.479	326/0.40/.393	18.2/98.0/.195	75.4/73.6/.411	57.3/85.2/.347
	t_1	0.79/91.2/.000	0.19/79.6/.000	262/1.60/.443	167/12.8/.358	317/2.80/.479	326/0.40/.393	19.7/98.8/.262	32.3/89.2/.273	16.0/96.8/.264
	t_3	1.06/91.2/.000	0.20/80.4/.000	0.58/60.0/.000	166/12.4/.358	315/2.80/.479	325/0.40/.393	20.0/98.8/.255	27.3/92.8/.228	14.6/97.6/.219
Scene$_2$	t_0	331/1.20/.331	216/48.8/.386	261/2.00/.443	167/12.4/.358	317/2.80/.479	326/0.40/.393	18.2/98.0/.195	75.4/73.6/.411	57.3/85.2/.347
	t_1	333/1.20/.331	216/48.4/.386	262/2.80/.443	0.24/82.0/.000	0.36/85.2/.000	325/0.40/.392	19.7/96.0/.184	55.3/84.4/.343	24.5/95.2/.293
	t_3	332/1.20/.331	216/49.2/.386	262/2.00/.443	0.19/82.0/.000	0.23/84.8/.000	0.24/95.2/.000	18.2/96.0/.157	60.0/82.0/.361	13.9/99.6/.253
Scene$_3$	t_0	286/10.4/.319	34.1/90.0/.206	271/3.20/.397	35.1/68.8/.183	189/30.8/.340	26.8/82.8/.147	8.51/96.8/.038	45.1/86.8/.241	9.80/100/.150
	t_1	0.73/90.4/.000	15.2/81.2/.088	0.47/60.4/.177	26.7/70.8/.339	188/30.8/.061	11.6/92.8/.061	4.68/97.2/.018	13.8/92.4/.049	18.4/97.2/.182
	t_2	0.79/90.4/.000	268/28.8/.390	0.44/60.0/.000	228/9.20/.412	189/30.4/.340	17.2/91.2/.113	18.0/93.2/.134	23.3/92.0/.131	10.3/99.6/.145
	t_3	0.60/91.2/.000	15.2/81.2/.088	0.50/59.6/.000	228/.096/.412	189/29.6/.340	15.4/93.6/.112	7.53/96.4/.039	23.3/90.4/.140	13.8/98.4/.165
	t_4	0.63/91.2/.000	15.2/81.2/.088	.57/60.0/.000	228/9.20/.412	189/30.4/.340	312/0.00/.371	10.8/96.8/.058	19.6/92.0/.101	13.3/99.2/.179
Scene$_4$	t_0	244/22.4/.276	169/48.0/.362	90.0/40.0/.332	57.4/53.6/.184	26.0/76.8/.165	60.4/54.8/.185	19.4/98.0/.184	22.5/90.8/.205	12.6/99.6/.222
	t_1	1.12/91.2/.000	169/48.0/.361	90.0/40.0/.292	30.3/65.6/.277	37.0/58.8/.263	16.9/94.8/.075	13.8/95.6/.126	21.1/95.6/.154	9.73/99.6/.179
	t_2	0.59/91.2/.000	169/47.2/.362	89.9/40.0/.291	217/7.20/.502	237/8.00/.431	50.6/83.2/.357	35.1/89.2/.267	30.6/89.2/.213	27.4/99.6/.444
	t_3	0.66/92.4/.000	0.23/79.6/.000	90.0/40.4/.291	209/5.60/.414	206/11.2/.371	77.0/69.2/.428	35.5/88.4/.242	44.9/83.2/.245	19.2/100/.333
	t_4	0.62/90.8/.000	0.22/79.6/.000	90.1/40.0/.291	209/6.40/.414	206/11.2/.371	304/2.80/.431	33.2/91.6/.265	31.0/88.4/.210	14.7/100/.290
	t_5	0.63/91.2/.000	15.2/81.2/.088	0.57/60.0/.000	228/9.20/.412	189/30.4/.340	312/0.00/.371	10.8/96.8/.058	19.6/92.0/.101	13.3/99.2/.179

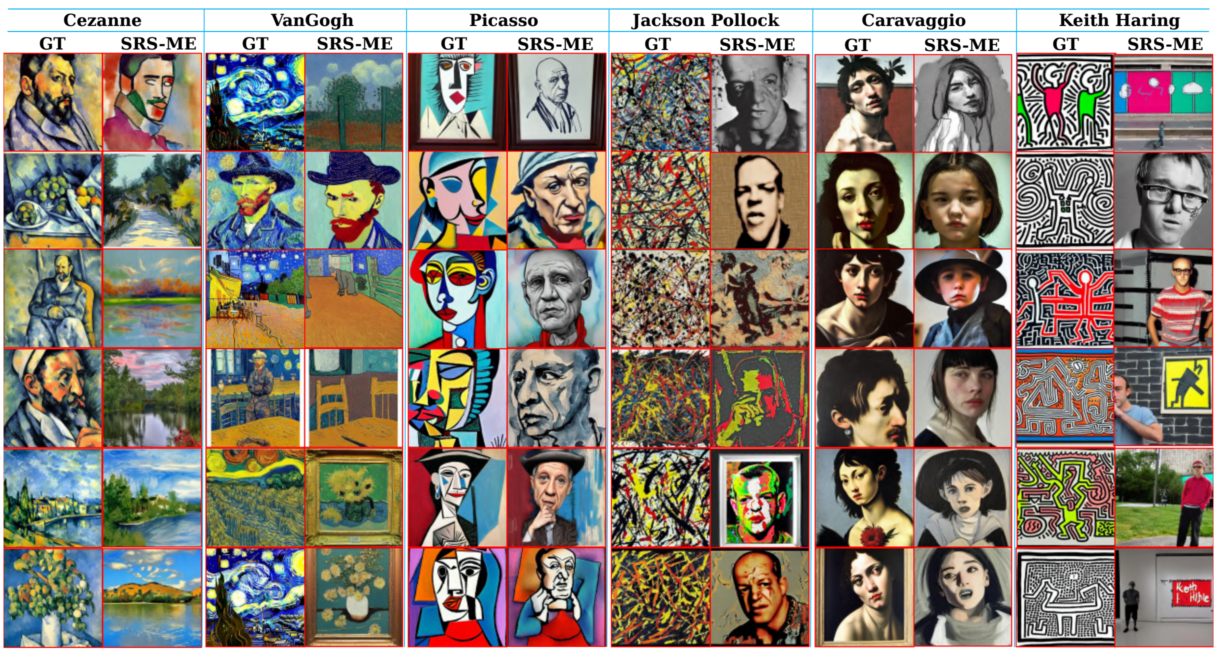


Figure 12. Visual examples of SRS-ME on style unlearning.

among various finetuned weight shifts during multi-style erasure. For instance, the average FID/ACC/LPIPS values of FMN30 and ESD200 in Figure 10 are 211.5/38.5/0.381 and 184.6/45.5/0.346, respectively. However, in Figure 8, these values are increased to 284.8/25.9/0.449 and 311.4/4.53/0.547, respectively. This interaction among various fine-tuned weights limits the applicability of existing methods in iterative style erasure scenarios: (1) Existing methods exhibit limited style restoration capability, as evidenced by the recovery metric in both Figures 8 and 9; (2) They may fail to erase previously forgotten styles after style restoration; (3) Determining the unlearning degree for new forgotten styles becomes challenging. An insufficient erasure level may not effectively remove styles, while too aggressive erasure could significantly degrade the model generation capability for regular styles. For instance, despite ESD200 shows inferior single-style erasure performance compared to our SRS-ME, it particularly struggles to produce effective images in iterative style erasure scenarios, *i.e.*, the average FID value of ESD200 in Figures 8 and 9 for regular styles exceeds 200; (4) Weight interactions can easily affect the generation performance of DMs on regular styles. In Figure 9, when compared with AbConcept, FMN20 and ESD50, our SRS-ME shows superior restoration and preservation performance, even with a higher degree of unlearning. *These results indicate the effectiveness of our SRS-ME in achieving style unlearning, style restoration, and model preservation during multi-style erasure scenarios.*

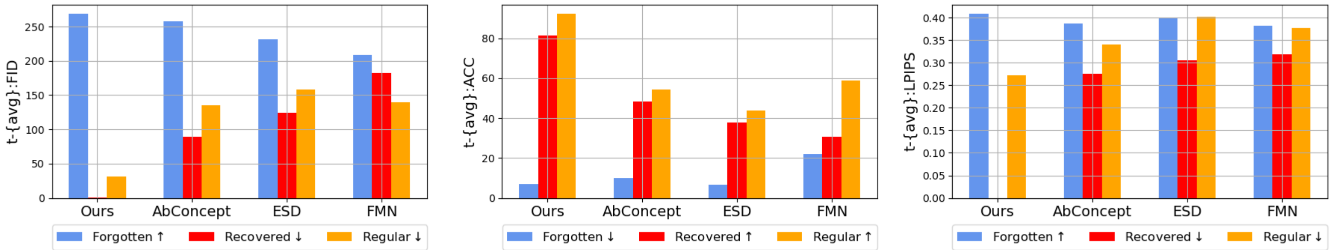


Figure 13. Performance comparison on $Scene_1$ of object unlearning.

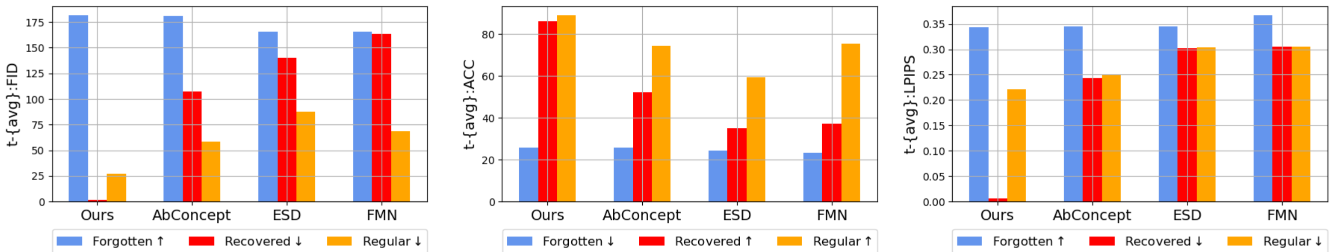


Figure 14. Performance comparison on $Scene_4$ of object unlearning.

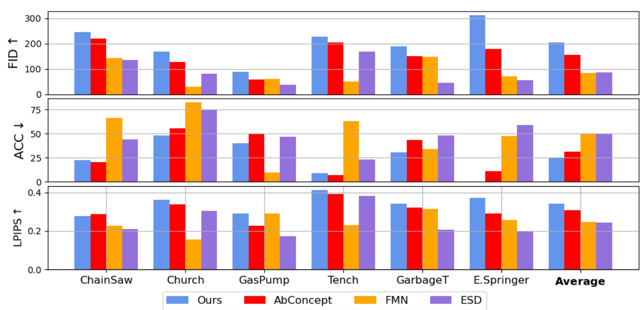


Figure 15. Performance comparison of single object unlearning for models combined to erase multiple objects. The x-axis indicates the erased objects.

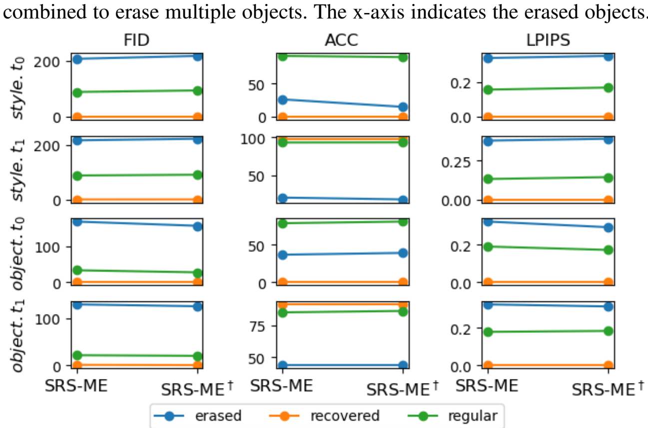


Figure 16. Performance comparison at $Scene_4$.

Additionally, we present visual examples of SRS-ME in Figure 12. As observed, the proposed SRS-ME effectively removes undesirable concepts while preserving the overall layout for most images.

4.3.2. Object Unlearning. Similar to experiments on style unlearning, we compare our SRS-ME with state-of-the-art DM unlearning methods in multi-object erasure. Table 3 presents the quantitative results of SRS-ME. Figure 15 displays a comparison of single object unlearning performance for models combined to erase multiple objects. Additionally, Figures 13 and 14 provide performance comparisons with existing advanced methods for $Scene_1$ and $Scene_4$, respectively. Similar to its behavior in style unlearning, the proposed *SRS-ME demonstrates effective erasing, restoration, and preservation capabilities in object unlearning*.

We offer visual examples of SRS-ME on object unlearning in Figure 17. As observed, even when prompts include previously forgotten objects, our SRS-ME successfully prevents these objects from appearing in the generated images.

4.4. Evaluation for SRS-ME†

SRS-ME† optimizes decoupled weight shifts simultaneously. However, due to GPU resource limitations, the maximum batch size is set to 3. We selected two specific timestamps, t_0 and t_1 at $Scene_4$, to conduct a comparative analysis between our SRS-ME and SRS-ME†. The experimental results are shown in Figure 16. It can be observed that these two variants achieve comparable performance in terms of unlearning, restoration, and preservation metrics. *This demonstrates the feasibility of optimization decoupling*.

4.5. Evaluation for SRS-ME‡

'nudity' unlearning. To assess the efficacy of our approach in erasing 'nudity' and preserving model performance, we conduct a comparative analysis of various DM unlearning methods by fine-tuning all layers within cross-attention modules. For assessing erasure performance, we utilize I2P prompts from [19] and categorize images exposing body parts into different nudity classes with Nudenet [52]. For evaluating model preservation performance, we

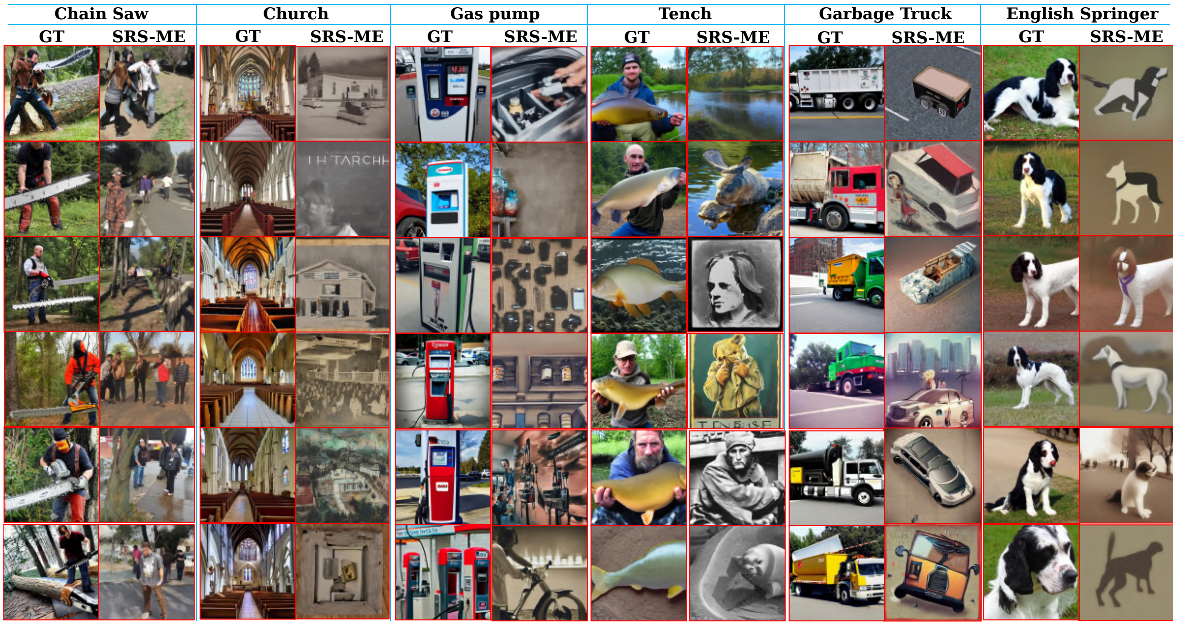


Figure 17. Visual examples of SRS-ME on object unlearning.

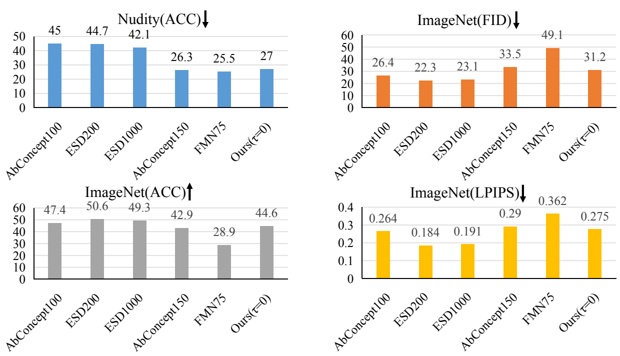


Figure 18. Comparative experiments on 'nudity' unlearning.

adopt 1859 images generated using prompts from 1000 different categories listed in the ImageNet dataset [1]. Results in Figure 18 show that *our approach exhibits superior model preservation performance compared to previous methods under comparable erasure performance*, *e.g.*, the results of AbConcept150, FMN75, and our SRS-ME[‡].

TABLE 4. Impact of weight regularization on SRS-ME (FID/ACC/LPIPS). $Scene_1$ and $Scene_2$ share the same results at t_0. Bold font indicates the best results.

Methods	Scene	t	Forgotten	Recovered	Regular
wo.reg	$Scene_1$	t_0	**220.9/24.9/**.379	0.00/0.00/.00	146.8/92.0/.167
		t_1	**227.9/23.7/.390**	2.46/94.2/.00	100.2/97.6/.118
		t_2	**241.8/28.5/.416**	1.98/95.7/.342	78.1/99.6/.096
	$Scene_2$	t_1	207.2/20.9/.393	**1.44**/97.8/.00	171.1/83.6/.199
		t_2	199.5/21.1/.342	**1.29**/98.1/.00	112.4/93.2/.168
w.reg	$Scene_1$	t_0	220.4/20.0/**.383**	0.00/0.00/.00	**100.2/97.5/.135**
		t_1	222.2/24.1/.376	**1.44/94.4/.00**	**89.2/99.3/.102**
		t_2	229.4/30.1/.389	**1.77/95.9/.00**	**76.8/99.9/.081**
	$Scene_2$	t_1	**211.7/18.2/.415**	1.51/97.8/.00	**123.8/94.7/.158**
		t_2	**209.2/18.9/.354**	1.44/98.1/.00	**90.2/96.9/.132**

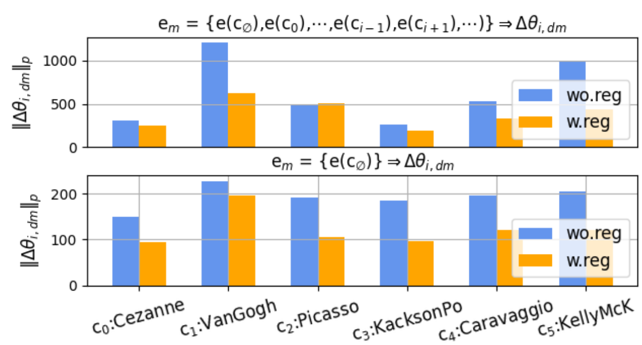


Figure 19. Influence of the weight regularization on SRS-ME.

4.6. Ablation Studies

4.6.1. Impact of the weight regularization. Ablation studies on weight regularization are conducted on $Scene_1$ and $Scene_2$ of style unlearning. For each timestamp, we calculate the average metric values for forgotten, recovered, and regular styles separately, as summarized in Table 4. As observed, the regularization term significantly improves the generation performance of SRS-ME for regular concepts, with only a slight impact on its erasure performance. Additionally, Figure 19 illustrates the influence of weight regularization on the degree of weight modifications. It is evident that weight regularization greatly reduces the extent of weight modification, which also explains why SRS-ME with weight regularization exhibits superior performance in regular concept-related generation.

4.6.2. Impact of the momentum statistic. Figure 20 shows that the unlearning loss exhibits instability, and the integra-

1. https://github.com/rohitgandikota/erasing/blob/main/data

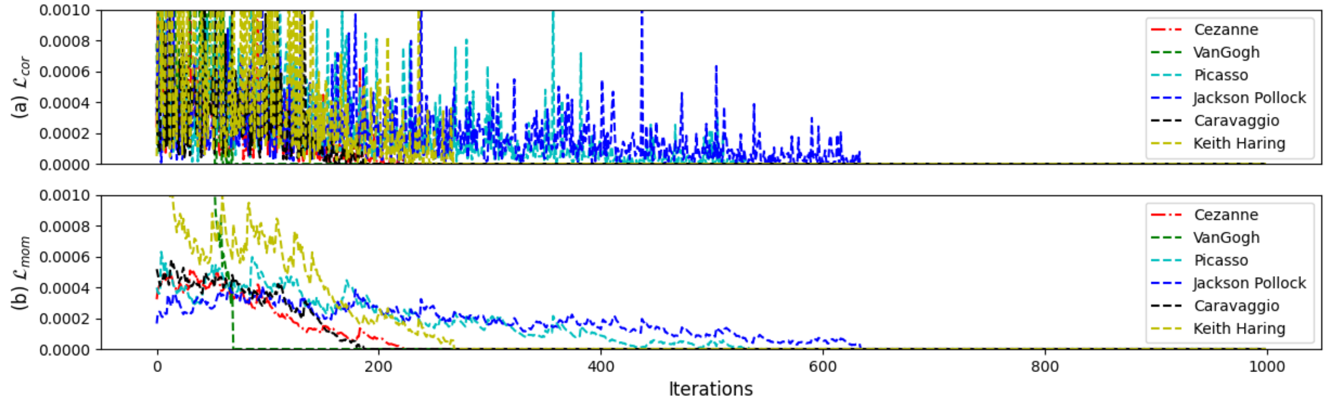


Figure 20. The impact of momentum statistics.

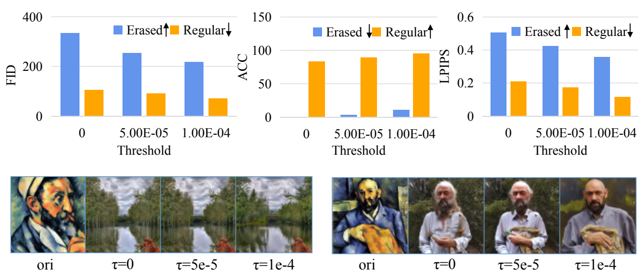


Figure 21. Impact of the threshold τ on SRS-ME.

tion of momentum statistics can mitigate this effect.

4.6.3. Impact of the threshold τ. We evaluate the impact of τ on SRS-ME in erasing the concept "Cezanne", setting τ to 0, 5e-5, and 1e-4, respectively. Experimental results in Figure 21 show that a smaller threshold enhances concept erasure performance, but also negatively affects the generation performance for regular concepts. Furthermore, visual examples indicate that when the threshold is set to 1e-4, SRS-ME can effectively eliminate "Cezanne" from DMs.

4.6.4. Impact of the hyperparameter β. Figure 22 illustrates that increasing β accelerates the unlearning optimization but also causes more model weight modifications.

4.6.5. Impact of the number of decoupled concepts. Figure 19 demonstrates that as the number of decoupled concepts increases, the required weight modification for concept erasure also increases. This phenomenon could be attributed to the correlations among different concepts, namely, decoupling concepts that are similar to the forgotten concept increases the unlearning difficulty.

4.6.6. Impact of similar concepts on weight decoupling. We erase the concept 'Cezanne' by integrating various concepts into e_m. Experimental results are depicted in Figure 23. As observed, the closer the decoupled concept aligns with the forgotten concept, the more challenging the erasure becomes. Therefore, we can conclude that even very similar concepts, such as "plane" and "aircraft", "man" and

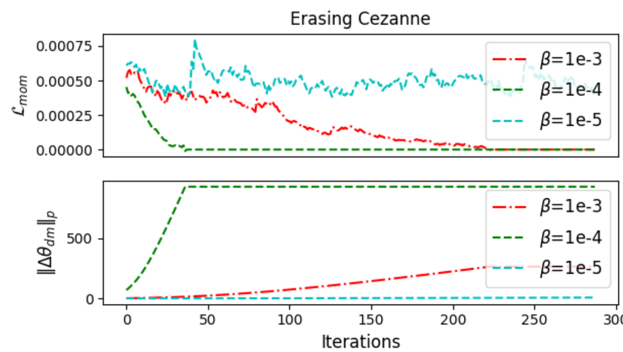


Figure 22. Impact of β on SRS-ME. τ and λ are set to 0.

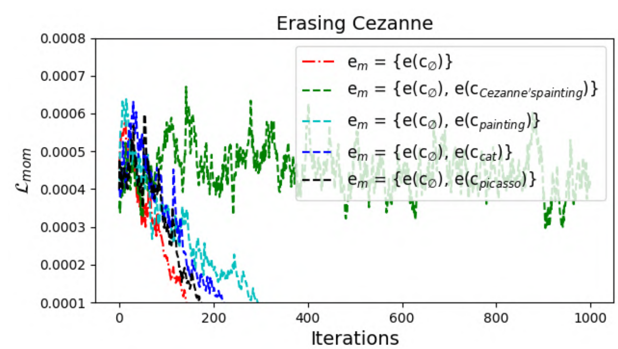


Figure 23. Impact of similar concepts on weight decoupling.

"bear", can still be decoupled. However, decoupling similar concepts makes erasure more difficult or even impossible.

5. Conclusion

In this study, we introduce an innovative machine unlearning technique for diffusion models termed Separable, Recoverable, and Sustainable Multi-Concept Eraser (SRS-ME). It enables flexible manipulation of forgotten concepts without requiring retraining from scratch. SRS-ME tackles concerns related to unlearning performance, concept restoration, model preservation performance, watermark preservation, and memory overload. It expands the horizon of diffusion model unlearning beyond mere concept erasure.

References

[1] Y. Zhou, R. Zhang, C. Chen, C. Li, C. Tensmeyer, T. Yu, J. Gu, J. Xu, and T. Sun, "Towards language-free training for text-to-image generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 17 907–17 917.

[2] S. Ge, T. Park, J.-Y. Zhu, and J.-B. Huang, "Expressive text-to-image generation with rich text," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 7545–7556.

[3] Y. Kim, J. Lee, J.-H. Kim, J.-W. Ha, and J.-Y. Zhu, "Dense text-to-image generation with attention modulation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 7701–7711.

[4] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, "Zero-shot text-to-image generation," in *International Conference on Machine Learning*. PMLR, 2021, pp. 8821–8831.

[5] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with clip latents," *arXiv preprint arXiv:2204.06125*, vol. 1, no. 2, p. 3, 2022.

[6] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695.

[7] L. Yang, Z. Zhang, Y. Song, S. Hong, R. Xu, Y. Zhao, W. Zhang, B. Cui, and M.-H. Yang, "Diffusion models: A comprehensive survey of methods and applications," *ACM Computing Surveys*, vol. 56, no. 4, pp. 1–39, 2023.

[8] B. Kawar, S. Zada, O. Lang, O. Tov, H. Chang, T. Dekel, I. Mosseri, and M. Irani, "Imagic: Text-based real image editing with diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 6007–6017.

[9] A. Kazerouni, E. K. Aghdam, M. Heidari, R. Azad, M. Fayyaz, I. Hacihaliloglu, and D. Merhof, "Diffusion models for medical image analysis: A comprehensive survey," *arXiv preprint arXiv:2211.07804*, 2022.

[10] R. Chourasia and N. Shah, "Forget unlearning: Towards true data-deletion in machine learning," in *International Conference on Machine Learning*. PMLR, 2023, pp. 6028–6073.

[11] C.-P. Huang, K.-P. Chang, C.-T. Tsai, Y.-H. Lai, and Y.-C. F. Wang, "Receler: Reliable concept erasing of text-to-image diffusion models via lightweight erasers," *arXiv preprint arXiv:2311.17717*, 2023.

[12] A. Golatkar, A. Achille, A. Swaminathan, and S. Soatto, "Training data protection with compositional diffusion models," *arXiv preprint arXiv:2308.01937*, 2023.

[13] H. Li, C. Shen, P. Torr, V. Tresp, and J. Gu, "Self-discovering interpretable diffusion latent directions for responsible text-to-image generation," *arXiv preprint arXiv:2311.17216*, 2023.

[14] E. Zhang, K. Wang, X. Xu, Z. Wang, and H. Shi, "Forget-me-not: Learning to forget in text-to-image diffusion models," *arXiv preprint arXiv:2303.17591*, 2023.

[15] R. Gandikota, J. Materzynska, J. Fiotto-Kaufman, and D. Bau, "Erasing concepts from diffusion models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

[16] Y. Zhang, J. Jia, X. Chen, A. Chen, Y. Zhang, J. Liu, K. Ding, and S. Liu, "To generate or not? safety-driven unlearned diffusion models are still easy to generate unsafe images... for now," *arXiv preprint arXiv:2310.11868*, 2023.

[17] N. Kumari, B. Zhang, S.-Y. Wang, E. Shechtman, R. Zhang, and J.-Y. Zhu, "Ablating concepts in text-to-image diffusion models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

[18] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," *arXiv preprint arXiv:2010.02502*, 2020.

[19] P. Schramowski, M. Brack, B. Deiseroth, and K. Kersting, "Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 22 522–22 531.

[20] S. Kim, S. Jung, B. Kim, M. Choi, J. Shin, and J. Lee, "Towards safe self-distillation of internet-scale text-to-image diffusion models," *arXiv preprint arXiv:2307.05977*, 2023.

[21] S. Hong, J. Lee, and S. S. Woo, "All but one: Surgical concept erasing with model preservation in text-to-image diffusion models," *arXiv preprint arXiv:2312.12807*, 2023.

[22] Z. Ni, L. Wei, J. Li, S. Tang, Y. Zhuang, and Q. Tian, "Degeneration-tuning: Using scrambled grid shield unwanted concepts from stable diffusion," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 8900–8909.

[23] Y. Liu, Z. Li, M. Backes, Y. Shen, and Y. Zhang, "Watermarking diffusion model," *arXiv preprint arXiv:2305.12502*, 2023.

[24] Y. Zhao, T. Pang, C. Du, X. Yang, N.-M. Cheung, and M. Lin, "A recipe for watermarking diffusion models," *arXiv preprint arXiv:2303.10137*, 2023.

[25] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," *Advances in neural information processing systems*, vol. 30, 2017.

[26] T. Park, J.-Y. Zhu, O. Wang, J. Lu, E. Shechtman, A. Efros, and R. Zhang, "Swapping autoencoder for deep image manipulation," *Advances in Neural Information Processing Systems*, vol. 33, pp. 7198–7211, 2020.

[27] Z.-S. Liu, W.-C. Siu, and L.-W. Wang, "Variational autoencoder for reference based image super-resolution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 516–525.

[28] H. Shao, S. Yao, D. Sun, A. Zhang, S. Liu, D. Liu, J. Wang, and T. Abdelzaher, "Controlvae: Controllable variational autoencoder," in *International Conference on Machine Learning*. PMLR, 2020, pp. 8655–8664.

[29] X. Ding, Y. Wang, Z. Xu, W. J. Welch, and Z. J. Wang, "Ccgan: Continuous conditional generative adversarial networks for image generation," in *International conference on learning representations*, 2020.

[30] R. Liu, Y. Ge, C. L. Choi, X. Wang, and H. Li, "Divco: Diverse conditional image synthesis via contrastive generative adversarial network," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 16 377–16 386.

[31] B. Li, X. Qi, P. Torr, and T. Lukasiewicz, "Lightweight generative adversarial networks for text-guided image manipulation," *Advances in Neural Information Processing Systems*, vol. 33, pp. 22 020–22 031, 2020.

[32] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Advances in neural information processing systems*, 2020, pp. 6840–6851.

[33] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," *Advances in neural information processing systems*, vol. 34, pp. 8780–8794, 2021.

[34] S. Gu, D. Chen, J. Bao, F. Wen, B. Zhang, D. Chen, L. Yuan, and B. Guo, "Vector quantized diffusion model for text-to-image synthesis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10 696–10 706.

[35] X. Xu, Z. Wang, G. Zhang, K. Wang, and H. Shi, "Versatile diffusion: Text, images and variations all in one diffusion model," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 7754–7765.

[36] G. Kim, T. Kwon, and J. C. Ye, "Diffusionclip: Text-guided diffusion models for robust image manipulation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 2426–2435.

[37] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[38] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[39] N. Carlini, J. Hayes, M. Nasr, M. Jagielski, V. Sehwag, F. Tramer, B. Balle, D. Ippolito, and E. Wallace, "Extracting training data from diffusion models," in *32nd USENIX Security Symposium (USENIX Security 23)*, 2023, pp. 5253–5270.

[40] J. Duan, F. Kong, S. Wang, X. Shi, and K. Xu, "Are diffusion models vulnerable to membership inference attacks?" *arXiv preprint arXiv:2302.01316*, 2023.

[41] J. Dubiński, A. Kowalczuk, S. Pawlak, P. Rokita, T. Trzciński, and P. Morawiecki, "Towards more realistic membership inference attacks on large diffusion models," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 4860–4869.

[42] G. Somepalli, V. Singla, M. Goldblum, J. Geiping, and T. Goldstein, "Diffusion art or digital forgery? investigating data replication in diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 6048–6058.

[43] S. Shan, J. Cryan, E. Wenger, H. Zheng, R. Hanocka, and B. Y. Zhao, "Glaze: Protecting artists from style mimicry by text-to-image models," *arXiv preprint arXiv:2302.04222*, 2023.

[44] Y. Yang, R. Gao, X. Wang, N. Xu, and Q. Xu, "Mma-diffusion: Multimodal attack on diffusion models," *arXiv preprint arXiv:2311.17516*, 2023.

[45] F.-A. Croitoru, V. Hondru, R. T. Ionescu, and M. Shah, "Diffusion models in vision: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[46] Z.-Y. Chin, C.-M. Jiang, C.-C. Huang, P.-Y. Chen, and W.-C. Chiu, "Prompting4debugging: Red-teaming text-to-image diffusion models by finding problematic prompts," *arXiv preprint arXiv:2309.06135*, 2023.

[47] N. Mehrabi, P. Goyal, C. Dupuy, Q. Hu, S. Ghosh, R. Zemel, K.-W. Chang, A. Galstyan, and R. Gupta, "Flirt: Feedback loop in-context red teaming," *arXiv preprint arXiv:2308.04265*, 2023.

[48] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 586–595.

[49] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[50] L. N. Smith, "Cyclical learning rates for training neural networks," in *2017 IEEE winter conference on applications of computer vision (WACV)*. IEEE, 2017, pp. 464–472.

[51] J. Howard and S. Gugger, "Fastai: A layered api for deep learning," *Information*, vol. 11, no. 2, p. 108, 2020.

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

Appendix

1. Proof for Eq. (27)

In each denoising timestamp, the generation changes brought by introducing forgotten concepts c_f can be formulated as

$$\mathcal{M}(\boldsymbol{z}, c_f, \boldsymbol{\theta}_{dm}) - \mathcal{M}(\boldsymbol{z}, c_\emptyset, \boldsymbol{\theta}_{dm}). \tag{39}$$

For convenience, we utilize $\boldsymbol{x}_0^\diamondsuit$ and $\boldsymbol{x}_0^\square$ to represent latent representations of DMs for prompts c_f and $c_\emptyset$ respectively. The index '0' denotes the final schedule time. According to Eqs. (8) and (13), we have

$$\begin{aligned}\boldsymbol{x}_0^\diamondsuit &= \lambda_1\boldsymbol{x}_1 - \lambda_2(\boldsymbol{\epsilon}(\boldsymbol{x}_1, c_\emptyset) + \lambda_4(\boldsymbol{\epsilon}(\boldsymbol{x}_1, c_f) - \boldsymbol{\epsilon}(\boldsymbol{x}_1, c_\emptyset))) + \lambda_3\boldsymbol{z},\\ \boldsymbol{x}_0^\square &= \lambda_1\boldsymbol{x}_1 - \lambda_2(\boldsymbol{\epsilon}(\boldsymbol{x}_1, c_\emptyset) + \lambda_4(\boldsymbol{\epsilon}(\boldsymbol{x}_1, c_\emptyset) - \boldsymbol{\epsilon}(\boldsymbol{x}_1, c_\emptyset))) + \lambda_3\boldsymbol{z}.\end{aligned} \tag{40}$$

Eq. (39) is further expressed as

$$\boldsymbol{x}_0^\diamondsuit - \boldsymbol{x}_0^\square = \lambda_2\lambda_4(\boldsymbol{\epsilon}(\boldsymbol{x}_1, c_f) - \boldsymbol{\epsilon}(\boldsymbol{x}_1, c_\emptyset)). \tag{41}$$

Similarly, Eq. (41) can be generalized to arbitrary timestamp,

$$\begin{aligned}\boldsymbol{x}_{t-1}^\diamondsuit - \boldsymbol{x}_{t-1}^\square &= \lambda_2\lambda_4(\boldsymbol{\epsilon}(\boldsymbol{x}_t, c_f) - \boldsymbol{\epsilon}(\boldsymbol{x}_t, c_\emptyset)),\\ &\propto \boldsymbol{\epsilon}(\boldsymbol{x}_t, c_f) - \boldsymbol{\epsilon}(\boldsymbol{x}_t, c_\emptyset).\end{aligned} \tag{42}$$

2. Proofs for the answer A7

The detailed quantitative results of various methods on style and object unlearning are given in Tables 5 and 6, respectively.

Additionally, the quantitative results of the ablation study on weight regulation are provided in Table 7.

Illustration for the unlearning loss: Norm functions are unsuitable for supervising the unlearning process since we initialize the weight shifts to zero matrices, resulting in initial norm function values of 0. Additionally, determining the stopping condition when employing norm functions as the unlearning loss poses a challenge.

TABLE 5. Comparative results (ACC/LPIPS/FID) on style unlearning. ■, ■, and ■ denote the evaluation performance for erased, recovered and regular concepts, respectively.

Scene	t	c_0:**Cezanne**	c_1:**VanGogh**	c_2:**Picasso**	c_3:**JacksonPo...**	c_4:**Caravaggio**	c_5:**KeithHaring**	c_6:**KellyMcK...**	c_7:**TylerEdlin**	c_8:**KilianEng**
ORI	-	0.00/98.0/0.00	0.00/90.4/0.00	0.00/98.8/0.00	0.00/96.0/0.00	0.00/99.6/0.00	0.00/98.4/0.00	0.00/99.6/0.00	0.00/100/0.00	0.00/100/0.00
						SRS-ME				
	t_0	209/12.4/.338	220/36.8/.388	200/6.40/.336	275/26.4/.329	203/38.0/.312	217/15.2/.596	56.0/95.2/.153	135/98.2/.125	110/99.2/.126
Scene$_1$	t_1	0.75/98.0/.00	2.12/90.8/.00	199/6.00/.336	277/26.8/.328	204/38.3/.313	209/25.2/.526	47.1/98.8/.096	119/100/.102	101/99.2/.107
	t_3	0.78/98.4/.00	3.28/90.4/.00	1.25/98.8/.00	277/26.8/.329	203/38.0/.312	209/25.6/.526	44.7/99.6/.097	99.9/100/.080	85.8/100/.066
	t_0	188/36.0/.256	219/37.2/.395	214/5.20/.365	184/60.4/.245	96.7/98.8/.182	52.5/97.2/.189	65.1/95.6/.160	50.8/100./.016	108/99.2/.147
	t_1	0.77/98.0/.00	219/37.2/.395	215/5.20/.364	176/64.0/.211	96.0/99.6/.166	45.4/96.8/.147	49.7/98.4/.096	49.9/100/.016	113/100/.156
Scene$_4$	t_2	0.81/98.0/.00	219/37.6/.395	215/5.20/.365	365/5.60/.463	246/19.6/.379	135/59.6/.402	77.4/86.0/.176	61.0/100/.024	113/99.6/.151
	t_3	0.76/98.0/.00	3.82/90.4/.00	215/5.20/.364	355/2.80/.458	219/28.0/.340	132/58.2/.395	73.3/90.0/.166	52.0/100/.016	101/98.4/.123
	t_4	0.87/98.0/.00	2.82/90.4/.00	214/5.20/.365	354/2.80/.458	219/28.4/.340	288/11.6/.623	82.1/85.6/.185	77.0/100/.045	103/98.4/.123
	t_5	0.92/98.0/.00	4.11/90.4/.00	1.06/98.8/.00	274/26.4/.343	231/28.0/.368	278/17.6/.615	68.4/90.8/.154	78.3/100/.046	91.8/99.6/.098
						AbConcept[75,75,50,50,50,50]				
	t_0	223/19.2/.382	246/9.60/.422	238/0.80/.337	300/14.8/.343	236/10.4/.319	230/12.4/.589	82.7/76.8/.194	176/75.6/.214	157/90.0/.214
Scene$_1$	t_1	145/40.8/.255	219/33.2/.401	165/6.40/.252	298/12.8/.366	197/26.0/.270	227/19.2/.576	70.7/92.0/.158	124/98.8/.118	127/98.4/.166
	t_3	106/76.4/.186	183/50.8/.363	83.6/77.2/.188	291/18.0/.345	179/51.2/.269	200/29.6/.510	57.2/95.6/.123	109/100/.093	109/99.6/.128
	t_0	210/20.0/.371	235/20.8/.413	199/0.80/.288	236/56.0/.266	158/67.2/.235	129/64.4/.427	64.3/96.0/.158	148/93.6/.151	130/98.4/.168
	t_1	141/55.2/.202	239/20.8/.415	174/6.40/.254	165/87.6/.178	112/96.0/.165	97.1/80.0/.368	57.8/96.4/.130	109/100/.089	103/98.8/.120
Scene$_4$	t_2	175/39.2/.272	240/16.8/.415	208/1.60/.287	279/21.2/.331	213/23.2/.287	124/60.8/.409	61.7/93.6/.154	130/98.4/.132	119/98.0/.146
	t_3	130/82.0/.170	163/83.2/.296	182/5.60/.268	287/18.8/.338	195/28.0/.259	87.6/81.6/.313	56.7/96.0/.124	99.7/99.6/0.09	101/99.2/.111
	t_4	145/40.8/.255	219/33.2/.401	165/6.40/.252	298/12.8/.366	197/26.0/.270	227/19.2/.576	70.7/92.0/.158	124/98.8/.118	127/98.4/.166
	t_5	106/76.4/.186	183/50.8/.363	83.6/77.2/.188	291/18.0/.345	179/51.2/.269	200/29.6/.510	57.2/95.6/.123	109/100/.093	109/99.6/.128
						FMN30				
	t_0	344/0.40/.416	315/29.6/.517	272/16.4/.428	404/13.6/.520	341/15.6/.453	253/35.2/.558	278/20.2/.439	247/23.6/.434	209/76.8/.302
Scene$_1$	t_1	248/10.4/.410	282/31.6/.469	187/30.0/.357	414/20.0/.417	242/35.2/.396	244/24.4/.564	102/63.2/.239	166/88.4/.211	151/93.2/.228
	t_3	201/19.6/.357	265/20.8/.447	136/58.8/.318	368/18.4/.430	190/54.0/.333	225/23.2/.544	77.7/82.4/.187	122/98.0/.141	125/97.6/.189
	t_0	277/2.00/.369	281/32.4/.415	159/44.4/.358	323/27.6/.476	187/80.0/.338	175/58.5/.534	112/64.4/.268	196/56.0/.265	174/88.4/.268
	t_1	168/31.6/.323	290/19.2/.440	155/56.0/.357	281/32.0/.398	129/92.4/.280	95.6/88.0/.382	68.1/89.6/.177	106/99.2/.120	137/97.2/.198
Scene$_4$	t_2	251/4.40/.391	279/44.0/.444	167/43.2/.347	392/17.2/.499	229/41.2/.385	174/68.4/.521	143/57.2/.310	207/.464/.281	175/87.6/.260
	t_3	204/11.6/.372	264/33.2/.424	168/42.4/.329	379/18.4/.457	194/53.6/.349	123/74.8/.435	75/82.4/.186	122/98.0/.139	138/98.0/.197
	t_4	248/10.4/.410	282/31.6/.469	187/30.0/.357	414/20.0/.417	242/35.2/.396	244/24.4/.564	102/63.2/.239	166/88.4/.211	151/93.2/.228
	t_5	201/19.6/.357	265/20.8/.447	136/58.8/.318	368/18.4/.430	190/54.0/.333	225/23.2/.544	77.7/82.4/.187	122/98.0/.141	125/97.6/.189
						FMN20				
	t_0	213/0.80/.382	274/34.0/.440	156/46.0/.339	372/25.6/.506	208/53.2/.349	221/27.2/.575	103/67.2/.257	196/67.6/.256	163/92.8/.246
Scene$_1$	t_1	142/52.8/.265	248/32.0/.411	153/48.8/.315	328/26.8/.450	157/63.2/.292	173/39.2/.498	54.5/91.6/.141	96.4/100/.102	104/99.6/.139
	t_3	104/89.2/.188	174/53.6/.332	93.7/80.4/.249	306/26.4/.434	136/71.2/.255	142/54.0/.462	48.4/98.0/.111	85.4/100/.071	91.7/99.2/.097
	t_0	145/30.8/.310	263/18.8/.430	113/71.2/.279	268/45.6/.395	113/97.6/.236	81.1/86.0/.332	55.5/94.8/.147	98.8/100/.108	124/98.8/.171
	t_1	95.2/98.8/.177	264/20.8/.423	89.6/78.4/.240	205/71.6/.294	84.0/99.2/.173	64.5/96.8/.268	45.7/98.4/.111	82.3/100./.065	97.1/100/.115
Scene$_4$	t_2	155/32.8/.306	278/17.6/.428	152/44.4/.316	341/28.8/.449	164/69.2/.294	115/86.4/.418	74.6/87.6/.194	122/98.0/.152	135/97.6/.207
	t_3	117/83.6/.202	187/54.4/.342	110/69.2/.258	305/28.8/.429	142/70.0/.268	80.4/86.8/.338	47.1/98.4/.112	87.7/100/.077	98.0/99.6/.109
	t_4	142/52.8/.265	248/32.0/.411	153/48.8/.315	328/26.8/.450	157/63.2/.292	173/39.2/.498	54.5/91.6/.141	96.4/100/.102	104/99.6/.139
	t_5	104/89.2/.188	174/53.6/.332	93.7/80.4/.249	306/26.4/.434	136/71.2/.255	142/54.0/.462	48.4/98.0/.111	85.4/100/.071	91.7/99.2/.097
						ESD200				
	t_0	306/1.60/.465	303/6.80/.526	298/2.40/.498	398/0.00/.573	280/8.00/.482	313/2.00/.689	234/24.4/.347	249/8.00/.351	212/36.4/.366
Scene$_1$	t_1	305/2.00/.442	278/10.0/.469	283/2.40/.487	386/0.80/.559	281/8.80/.474	301/2.40/.666	184/38.4/.310	256/8.80/.321	198/43.2/.343
	t_3	278/4.40/.430	243/27.2/.443	254/4.80/.462	366/1.20/.557	258/14.0/.454	290/4.40/.659	122/62.8/.260	211/34.8/.268	168/76.4/.265
	t_0	320/0.00/.467	283/7.20/.496	262/4.40/.475	290/17.2/.386	234/30.8/.425	247/22.0/.603	108/68.0/.253	179/58.4/.233	167/84.4/.251
	t_1	160/50.4/.294	251/22.8/.441	215/9.20/.408	212/57.6/.265	126/90.8/.218	174/51.6/.503	65.9/89.6/.168	116/98.0/.123	134/96.4/.174
Scene$_4$	t_2	299/0.00/.459	297/5.60/.516	277/3.20/.505	378/0.00/.573	273/12.4/.491	279/6.00/.649	165/46.8/.297	244/12.4/.315	188/58.4/.323
	t_3	285/4.00/.431	240/24.8/.442	265/4.40/.485	362/0.00/.538	262/14.4/.464	258/10.0/.620	97.7/71.2/.238	193/45.2/.249	174/84.8/.249
	t_4	305/2.00/.442	278/10.0/.469	283/2.40/.487	386/0.80/.559	281/8.80/.474	301/2.40/.666	184/38.4/.310	256/8.80/.321	198/43.2/.343
	t_5	278/4.40/.430	243/27.2/.443	254/4.80/.462	366/1.20/.557	258/14.0/.454	290/4.40/.659	122/62.8/.260	211/34.8/.268	168/76.4/.265
						ESD50				
	t_0	301/1.20/.462	300/6.40/.492	256/5.20/.468	350/6.40/.480	254/28.8/.448	284/10.4/.628	142/49.6/.277	225/23.2/.284	179/78.0/.266
Scene$_1$	t_1	227/10.8/.384	240/36.4/.418	218/8.00/.416	307/15.2/.416	194/50.4/.343	262/13.6/.610	79.3/84.4/.205	144/83.2/.176	148/94.0/.195
	t_3	125/53.6/.233	165/69.6/.345	153/39.6/.321	282/26.0/.378	148/63.6/.283	236/27.2/.559	63.2/89.6/.168	120/96.8/.122	124/97.2/.160
	t_0	169/24.0/.308	234/36.8/.420	165/28.0/.339	162/75.2/.207	121/96.0/.210	113/79.6/.395	56.6/94.4/.146	110/98.8/.106	115/98.0/.141
	t_1	82.4/88.4/.137	177/66.0/.357	135/45.2/.288	107/84.4/.125	89/99.6/.126	78.0/87.6/.295	46.6/98.8/.097	86.5/100/.057	95.7/98.0/.094
Scene$_4$	t_2	207/14.0/.367	252/25.2/.434	194/17.6/.395	286/22.0/.379	183/58.8/.342	187/47.6/.500	68.5/86.8/.185	131/89.6/.150	139/94.4/.180
	t_3	138/53.2/.268	148/70.8/.323	176/21.2/.366	264/27.2/.351	155/63.2/.289	144/62.4/.455	59.4/94.0/.152	113/98.8/.110	119/97.2/.144
	t_4	227/10.8/.384	240/36.4/.418	218/8.00/.416	307/15.2/.416	194/50.4/.343	262/13.6/.610	79.3/84.4/.205	144/83.2/.176	148/94.0/.195
	t_5	125/53.6/.233	165/69.6/.345	153/39.6/.321	282/26.0/.378	148/63.6/.283	236/27.2/.559	63.2/89.6/.168	120/96.8/.122	124/97.2/.160

TABLE 6. COMPARATIVE RESULTS (FID/ACC/LPIPS/) ON OBJECT UNLEARNING. (blue), (red), AND (grey) DENOTE THE EVALUATION PERFORMANCE FOR ERASED, RECOVERED AND REGULAR CONCEPTS, RESPECTIVELY.

Scene	t	c_0:ChainSaw	c_1:Church	c_2:GasPump	c_3:Tench	c_4:GarbageT...	c_5:E.Springer	c_6:GolfBall	c_7:Parachute	c_8:FrenchHorn
ORI	-	0.00/91.6/0.00	0.00/80.4/0.00	0.00/60.0/0.00	0.00/81.6/0.00	0.00/84.8/0.00	0.00/95.6/0.00	0.00/97.6/0.00	0.00/93.2/0.00	0.00/100/0.00
SRS-ME										
	t_0	331/1.2/.331	216/48.8/.386	261/2.0/.443	167/12.4/.358	317/2.8/.479	326/0.4/.393	18.2/98.0/.195	75.4/73.6/.411	57.3/85.2/.347
Scene$_1$	t_1	0.79/91.2/.000	0.19/79.6/.000	262/1.6/.443	167/12.8/.358	317/2.8/.479	326/0.4/.393	19.7/98.8/.262	32.3/89.2/.273	16.0/96.8/.264
	t_3	1.06/91.2/.000	0.20/80.4/.000	0.58/60.0/.000	166/12.4/.358	315/2.8/.479	325/0.4/.393	20.0/98.8/.255	27.3/92.8/.228	14.6/97.6/.219
	t_0	244/22.4/.276	169/48.0/.362	90.0/40.0/.332	57.4/53.6/.184	26.0/76.8/.165	60.4/54.8/.185	19.4/98.0/.184	22.5/90.8/.205	12.6/99.6/.222
	t_1	1.12/91.2/.000	169/48.0/.361	90.0/40.0/.292	30.3/65.6/.277	37.0/58.8/.263	16.9/94.8/.075	13.8/95.6/.126	21.1/95.6/.154	9.73/99.6/.179
Scene$_4$	t_2	0.59/91.2/.000	169/47.2/.362	89.9/40.0/.291	217/7.20/.502	237/8.00/.431	50.6/83.2/.357	35.1/89.2/.267	30.6/89.2/.213	27.4/99.6/.444
	t_3	0.66/92.4/.000	0.23/79.6/.000	90.0/40.4/.291	209/5.60/.414	206/11.2/.371	77.0/69.2/.428	35.5/88.4/.242	44.9/83.2/.245	19.2/100/.333
	t_4	0.62/90.8/.000	0.22/79.6/.000	90.1/40.0/.291	209/6.40/.414	206/11.2/.371	304/2.80/.431	33.2/91.6/.265	31.0/88.4/.210	14.7/100/.290
	t_5	0.63/91.2/.000	15.2/81.2/.088	0.57/60.0/.000	228/9.20/.412	189/30.4/.340	312/0.00/.371	10.8/96.8/.058	19.6/92.0/.101	13.3/99.2/.179
AbConcept[75,75,50,50,75,75]										
	t_0	364/2.00/.371	266/2.80/.409	280/3.20/.391	329/0.40/.476	298/9.20/.418	367/0.80/.418	223/28.4/.436	280/7.20/.467	302/14.4/.415
Scene$_1$	t_1	190/29.2/.281	72.4/63.2/.300	120/30.8/.315	293/0.80/.462	187/30.4/.341	291/2.40/.357	72.2/76.4/.306	115/48.4/.348	66.6/80.4/.263
	t_3	159/41.2/.262	63.4/67.6/.282	38.8/42.8/.234	275/1.60/.457	167/29.2/.334	260/2.80/.348	59.1/80.0/.282	79.8/61.2/.323	20.6/93.2/.223
	t_0	234/13.6/.300	180/25.6/.366	86.3/30.0/.261	103/39.2/.310	30.6/72.8/.161	35.0/78.0/.176	39.5/88.0/.274	78.6/56.0/.293	71.5/81.6/.253
	t_1	98.1/65.2/.181	150/44.0/.356	57.2/40.8/.230	30.8/74.0/.180	22.2/77.2/.120	22.1/86.4/.111	26.6/91.6/.210	35.2/79.2/.218	21.7/96.8/.177
Scene$_4$	t_2	177/35.6/.265	203/18.4/.377	118/22.0/.296	264/2.00/.454	189/22.4/.354	50.8/66.4/.213	61.5/77.6/.310	135/40.0/.359	96.5/71.6/.280
	t_3	124/53.6/.212	45.2/74.0/.234	87.8/37.6/.285	238/5.20/.422	164/37.6/.332	33.2/80.0/.157	34.5/89.6/.217	51.4/73.2/.243	40.3/90.8/.216
	t_4	190/29.2/.281	72.4/63.2/.300	120/30.8/.315	293/0.80/.462	187/30.4/.341	291/2.40/.357	72.2/76.4/.306	115/48.4/.348	66.6/80.4/.263
	t_5	159/41.2/.262	63.4/67.6/.282	38.8/42.8/.234	275/1.60/.457	167/29.2/.334	260/2.80/.348	59.1/80.0/.282	79.8/61.2/.323	20.6/93.2/.223
FMN[50,50,50,30,50,50]										
	t_0	350/7.60/.349	375/0.00/.601	254/0.00/.433	341/0.00/.483	329/2.40/.475	350/0.40/.476	259/38.0/.526	318/1.60/.470	429/0.40/.470
Scene$_1$	t_1	195/22.8/.335	345/2.00/.473	175/5.60/.436	310/0.40/.462	283/6.80/.414	288/1.20/.358	92.7/76.0.393	176/31.6/.437	144/57.2/.309
	t_3	141/41.2/.293	187/42.0/.393	48.0/35.6/.292	266/2.40/.466	187/26.0/.357	206/12.8/.338	58.5/84.4/.386	88.1/65.2/.369	30.8/97.6/.299
	t_0	180/39.6/.273	336/4.00/.488	99.0/3.60/.363	127/45.6/.393	78.9/41.2/.294	54.0/64.0/.258	36.6/93.6/.324	75.3/64.4/.372	53.0/88.8/.285
	t_1	98.4/62.4/.234	234/24.4/.438	74.9/9.60/.322	42.0/76.8/.283	38.7/60.8/.238	35.2/76.0/.219	24.0/97.6/.276	45.9/77.6/.311	15.3/98.8/.237
Scene$_4$	t_2	207/18.8/.331	377/1.20/.486	179/4.40/.444	309/0.40/.450	266/4.80/.403	136/33.2/.312	91.0/76.0/.371	219/16.4/.474	189/36.4/.321
	t_3	155/38.0/.319	298/11.2/.442	178/1.60/.444	304/1.60/.463	226/8.00/.383	75.3/48.8/.276	75.6/77.6/.415	157/36.4/.429	94.8/74.8/.300
	t_4	195/22.8/.335	345/2.00/.473	175/5.60/.436	310/0.40/.462	283/6.80/.414	288/1.20/.358	92.7/76.0.393	176/31.6/.437	144/57.2/.309
	t_5	141/41.2/.293	187/42.0/.393	48.0/35.6/.292	266/2.40/.466	187/26.0/.357	206/12.8/.338	58.5/84.4/.386	88.1/65.2/.369	30.8/97.6/.299
FMN[50,30,50,20,50,50]										
	t_0	304/14.4/.338	368/0.80/.538	205/2.00/.426	328/0.00/.477	316/2.00/.458	336/0.40/.435	164/63.6/.426	271/9.20/.487	358/4.40/.415
Scene$_1$	t_1	178/28.8/.323	330/2.80/.274	166/4.00/.432	290/0.80/.464	262/8.00/.412	268/4.40/.350	86.5/74.4/.412	149/40.4/.421	106/72.8/.308
	t_3	122/50.4/.282	159/48.4/.386	48.5/38.4/.284	202/20.8/.433	173/29.6/.358	178/19.6/.332	54.8/86.4/.349	68.8/69.6/.349	21.0/100/.298
	t_0	163/46.8/.255	200/33.2/.409	79.2/7.60/.345	110/53.6/.361	62.6/48.0/.267	39.1/72.8/.231	32.7/95.6/.314	56.6/72.4/.338	31.2/96.0/.274
	t_1	84.3/70.0/.200	90.9/63.6/.316	63.8/7.20/.303	29.9/83.2/.239	32.4/71.6/.212	25.2/.87.2/.176	20.2/98.0/.238	38.1/85.2/.273	13.4/99.2/.228
Scene$_4$	t_2	162/28.4/.315	323/4.40/.467	168/2.80/.436	243/12.0/.443	224/8.80/.389	80.3/46.8/.284	67.1/81.2/.366	150/38.0/.426	87.6/76.0/.285
	t_3	150/36.0/.311	262/18.0/.437	162/3.20/.431	255/6.40/.451	213/9.60/.381	59.3/54.8/.263	73.0/78.0/.424	125/50.0/.403	65.7/86.4/.300
	t_4	178/28.8/.323	330/2.80/.274	166/4.00/.432	290/0.80/.464	262/8.00/.412	268/4.40/.350	86.5/74.4/.412	149/40.4/.421	106/72.8/.308
	t_5	122/50.4/.282	159/48.4/.386	48.5/38.4/.284	202/20.8/.433	173/29.6/.358	178/19.6/.332	54.8/86.4/.349	68.8/69.6/.349	21.0/100/.298
ESD[100,100,50,75,100,100]										
	t_0	330/0.00/.429	314/0.40/.440	276/0.40/.444	293/0.00/.490	291/0.00/.411	296/0.00/.463	203/31.6/.475	288/1.60/.501	333/2.40/.501
Scene$_1$	t_1	190/24.0/.320	143/32.8/.361	130/11.6/.336	276/0.00/.500	188/10.0/.346	235/8.40/.368	72.0/73.6/.368	198/18.0/.432	114/58.4/.341
	t_3	125/45.2/.257	84.9/64.8/.317	36.9/30.8/.235	274/1.20/.494	107/22.8/.285	179/13.2/.326	45.1/85.2/.319	126/40.8/.358	42.7/83.6/.255
	t_0	212/19.6/.305	203/22.8/.396	99.6/18.8/.299	145/32.0/.374	38.4/52.0/.213	37.6/72.4/.152	34.1/91.2/.282	134/33.6/.368	50.9/81.6/.244
	t_1	81.5/71.6/.153	118/55.2/.341	49.0/39.6/.213	33.4/78.0/.217	26.0/69.2/.145	20.7/84.8/.096	18.8/95.6/.182	53.4/70.8/.251	15.7/96.8/.155
Scene$_4$	t_2	210/16.4/.342	287/3.20/.421	170/6.00/.362	280/0.00/.498	208/6.00/.361	102/47.2/.273	86.8/65.2/.383	234/9.20/.473	188/39.2/.389
	t_3	133/50.0/.260	86.7/59.2/.318	91.8/20.0/.296	266/0.00/.493	125/18.4/.305	49.2/71.6/.194	48.7/82.8/.313	142/31.6/.377	56.6/82.8/.268
	t_4	190/24.0/.320	143/32.8/.361	130/11.6/.336	276/0.00/.500	188/10.0/.346	235/8.40/.368	72.0/73.6/.368	198/18.0/.432	114/58.4/.341
	t_5	125/45.2/.257	84.9/64.8/.317	36.9/30.8/.235	274/1.20/.494	107/22.8/.285	179/13.2/.326	45.1/85.2/.319	126/40.8/.358	42.7/83.6/.255

TABLE 7. ABLATION STUDY ON THE WEIGHT REGULATION. (blue), (red), AND (gray) DENOTE THE EVALUATION PERFORMANCE FOR ERASED, RECOVERED AND REGULAR CONCEPTS, RESPECTIVELY.

Scene	t	0:Cezanne	1:VanGogh	2:Picasso	3:JacksonPo...	4:Caravaggio	5:KeithHaring	6:KellyMcK...	7:TylerEdlin	8:KilianEng
wo.reg: $\|\boldsymbol{\Delta\theta}_{0,dm}\|_p$=309.5; $\|\boldsymbol{\Delta\theta}_{1,dm}\|_p$=1204.4; $\|\boldsymbol{\Delta\theta}_{2,dm}\|_p$=485.5; $\|\boldsymbol{\Delta\theta}_{3,dm}\|_p$=256.9; $\|\boldsymbol{\Delta\theta}_{4,dm}\|_p$=531.0; $\|\boldsymbol{\Delta\theta}_{5,dm}\|_p$=1004.3										
Scene$_1$	t_0	199/14.0/.331	215/41.6/.382	186/8.00/.314	318/15.6/.403	179/50.4/.302	230/20.0/.544	77.7/86.4/.193	235/93.6/.152	128/96.0/.156
	t_1	1.40/98.0/.00	3.52/90.4/.00	186/8.80/.313	318/16.0/.402	178/49.6/.302	230/20.4/.544	68.9/95.2/.140	120/99.6/.089	112/98.0/.126
	t_3	0.61/98.0/.00	3.83/90.4/.00	1.49/98.8/.00	318/16.0/.402	178/50.0/.302	230/19.6/.544	61.5/99.6/.159	86.5/100./.057	86.3/99.2/.072
Scene$_2$	t_1	199/14.0/.331	215/41.2/.382	185/8.80/.314	0.31/96.0/.00	2.56/99.6/.00	230/19.6/.544	119/65.2/.238	254/88.8/.166	141/96.8/.192
	t_3	199/13.6/.331	215/41.2/.382	185/8.40/.314	0.30/96.0/.00	1.61/100/.00	1.96/98.4/.00	85.2/82.8/.207	115/100/.095	137/96.8/.201
w.reg: $\|\boldsymbol{\Delta\theta}_{0,dm}\|_p$=253.2; $\|\boldsymbol{\Delta\theta}_{1,dm}\|_p$=623.2; $\|\boldsymbol{\Delta\theta}_{2,dm}\|_p$=510.4; $\|\boldsymbol{\Delta\Delta\theta}_{3,dm}\|_p$=190.6; $\|\boldsymbol{\Delta\theta}_{4,dm}\|_p$=328.4; $\|\boldsymbol{\Delta\theta}_{5,dm}\|_p$=442.1										
Scene$_1$	t_0	209/12.4/.338	220/36.8/.388	200/6.40/.336	275/26.4/.329	203/38.0/.312	217/15.2/.596	56.0/95.2/.153	135/98.2/.125	110/99.2/.126
	t_1	0.75/98.0/.00	2.12/90.8/.00	199/6.00/.336	277/26.8/.328	204/38.3/.313	209/25.2/.526	47.1/98.8/.096	119/100/.102	101/99.2/.107
	t_3	0.78/98.4/.00	3.28/90.4/.00	1.25/98.8/.00	277/26.8/.329	203/38.0/.312	209/25.6/.526	44.7/99.6/.097	99.9/100/.080	85.8/100/.066
Scene$_2$	t_1	209/12.4/.339	220/38.0/.388	200/6.40/.336	0.18/96.0/.00	2.83/99.6/.00	219/16.0/.596	75.1/86.4/.192	180/98.8/.141	117/98.8/.141
	t_3	209/12.4/.338	220/38.0/.388	200/6.40/.336	0.14/96.0/.00	2.08/100/.00	2.11/98.4/.00	61.0/92.4/.168	94.7/100/.076	115/98.4/.151