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

Anonymous Author(s)*

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 introduce the paradigm of weight decoupling for constructing separable weight shifts, which can decouple interactions among weight shifts targeting diverse concepts. This approach also provides flexibility in both erasing and recovering arbitrary concepts while preserving model watermarks. To effectively erase inappropriate concepts and preserve model performance on regular concepts, we design an innovative concept-irrelevant unlearning optimization process. By defining concept representations, this process introduces the concept correlation loss and the 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 demonstrate the flexibility of our SRS-ME in concept manipulation, as well as its efficacy in preserving model performance and reducing memory consumption.

CCS CONCEPTS

• **Computing methodologies** → Machine learning; • **Security and privacy** → Software and application security.

KEYWORDS

unlearning, separable multi-concept erasure, weight decoupling, concept-irrelevant unlearning, optimization decoupling

ACM Reference Format:

Anonymous Author(s). 2018. Forget All That Should Be Forgotten: Separable, Recoverable, and Sustainable Multi-Concept Erasure from Diffusion Models. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (Conference acronym 'XX)*. ACM, New York, NY, USA, 19 pages. https://doi.org/XXXXXXX.XXXXXXX

Conference acronym 'XX, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/XXXXXXX.XXXXXXX

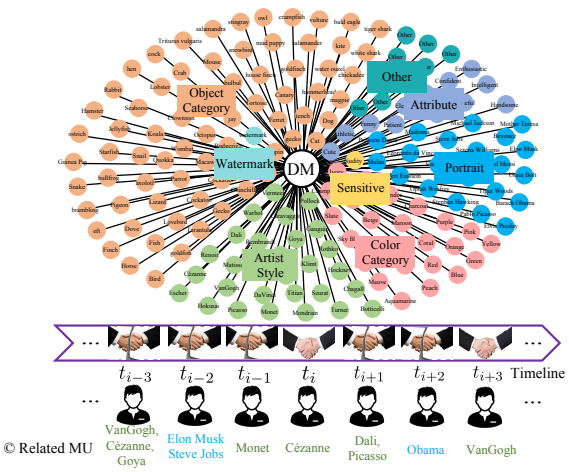

Figure 1: Concept recovery application.

1 INTRODUCTION

The field of text-to-image generation has witnessed remarkable development [12, 26, 39, 56], especially the occurrence of diffusion models (DMs) like DALL-E2 [38] and Stable Diffusion [40]. As the integration of DMs into practical applications [22, 23, 50] proves advantageous, addressing challenges related to their societal impact increasingly attracts the attention of researchers [4, 13, 21, 30]. One crucial challenge arises from diverse training data sources, potentially leading to unsafe image generation [11, 52], such as violent content or mimicking specific artistic styles. To resolve this concern, the machine unlearning (MU) technique has been proposed [28, 42, 47, 54], which involves erasing the impact of specific data points or concepts to enhance model security, without necessitating complete retraining from scratch. Recent MU work such as Erased Stable Diffusion (ESD) [11], Forget-me-not (FMN) [52], Safe self-distillation diffusion (SDD) [25], and Ablation Concept [28], can be broadly categorized into untargeted (*e.g.*,, FMN) and targeted concept erasures (*e.g.*,, ESD, Ablation and SDD). Specifically, FMN minimizes the values of the attention maps associated with forgotten concepts. In contrast, ESD, Ablation, 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 [18, 35], there exist several unresolved gaps, as listed below.

- G1 *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.
- G2 *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 forms involve interactions between fine-tuned weights for erasing various concepts.

G3 *Model performance preservation:* Prior efforts focus on concept erasure, leading to a considerable performance degradation in the overall generative capability of DMs. Particularly, they may destroy model watermarks, *i.e.* watermarks triggered by pre-defined prompts for text-guided DMs [32, 55]. For instance, in Figure 2, existing MU methods always affect the generation performance of other concepts.

Contribution. To fulfill these gaps, we propose an innovative framework SRS-ME for separable, recoverable, and sustainable multi-concept erasure. SRS-ME relies on weight decoupling to construct independent weight shifts, concept-irrelevant unlearning to effectively optimize 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 independent weight shift 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 is expressed as a linear combination of constant particular solutions calculated based on other known undesirable concepts. This decoupling mechanism enables concept restoration, alleviates mutual interactions among various fine-tuned weights, and preserves model watermarks.

Concept-irrelevant unlearning. To 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 layer 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 the 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, thereby significantly reducing memory consumption at the cost of training time. Furthermore, when simultaneously unlearning multiple concepts, researchers are required to introduce additional hyperparameters to balance the erasure performance across these concepts. Optimization decoupling effectively circumvents this issue.

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. Our weight decoupling fills these crucial gaps by 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 demonstrate the feasibility of optimization decoupling, which alleviates memory overload.
- We conduct an extensive array of experiments to demonstrate that our method can flexibly manipulate arbitrary concepts, preserve model watermarks and generation capabilities, and address memory overhead.

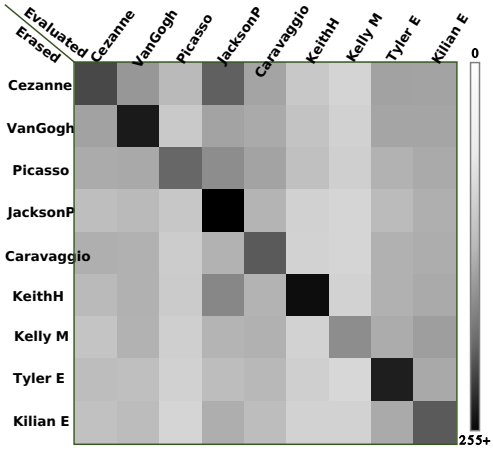

Figure 2: Interference between various concepts during unlearning with Ablation50 [28]. A larger Fréchet Inception Distance [16] indicates a greater impact on the DM generation performance of the evaluation concept.

2 BACKGROUND AND RELATED WORK

The image generation field has experienced rapid development in recent years, evolving from autoencoder [33, 36, 44], generative adversarial networks [7, 29, 31], unconditional diffusion models (DMs) [6, 17] to DMs enhanced with large-scale pre-trained image-text models [14, 24, 48] like CLIP [37]. These text-guided DMs, exemplified by DALL-E 2 [38] and Stable Diffusion [40], exhibit excellent generative abilities across various prompts c. The constraint for training DMs is formulated as

$$\mathcal{L}_{DM} = \mathbb{E}_{\mathbf{x}_t \in \mathcal{D}, c, t, \epsilon_{GT} \in \mathcal{N}(0,1)} \left[\| \epsilon_{GT} - \epsilon(\mathbf{x}_t, c, \theta_{dm}) \|_2^2 \right],$$

where $\mathbf{x}_t$ represents the noised data or the noised latent representation [27], $\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon$, $\bar{\alpha}_t$ is the noise variance schedule, $\mathbf{x}_0$ denotes the original reference image, and $\epsilon \in \mathcal{N}(0, \mathbf{I})$. $\mathcal{D}$ is the training dataset. ϵ_{GT} means the ground truth noise. $\epsilon(\mathbf{x}_t, c, \theta_{dm})$ denotes the t-th step noise predicted by DMs with parameters θ_{dm}. $\| \cdot \|_2^2$ is the squared ℓ_2-norm function. Additionally, researchers indicate the unknown concept generation capability of DMs by fine-tuning partial model weights on small reference sets [10, 20, 41, 49].

However, DMs also induce potential risks associated with privacy violations and copyright infringement, such as the training data leakage [2, 8, 9], the imitation of various artistic styles [43, 46], and the generation of sensitive content [51]. Hence, there is a growing focus on erasing specific outputs from pre-trained DMs [4, 28].

Figure 3: Expectation illustration. c_f denotes unlearned concepts, including 'Picasso', 'VanGogh', and 'Cezanne'. Here, $\mathcal{DM}_{c_f^*}$ 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.

Existing research primarily falls into three distinct directions: removal of unsafe data and model retraining [5], integration of additional plug-ins to guide model outputs [3, 34], and fine-tuning of trained model weights [11, 28, 52]. The drawback of the first direction is that large-scale model retraining demands considerable computational resource and time. The risk of the second direction is that, with the public availability of model structures and weights, malicious users can easily remove plug-ins. Therefore, this work focuses on the third direction.

Most finetuning work for DM unlearning can be summarized as:

$$\min_{\theta_{op}} \mathcal{L}_{unlearn} = \begin{cases} \|\epsilon(\mathbf{x}_t, c_f, \theta_{op}) - \epsilon_{target}\|_2 & \text{if } \mathbf{x}_0 \in \mathcal{D}_f \\ \|\epsilon(\mathbf{x}_t, c_f, \theta_{op}) - \epsilon_{GT}\|_2 & \text{otherwise,} \end{cases} \tag{1}$$

where θ_{op} represents optimizable model weights, *e.g.,* the parameters of cross-attention modules in DMs. $\mathbf{x}_t$ can be obtained through either the diffusion process or the sampling process. $\epsilon(\mathbf{x}_t, c_f, \theta_{op})$ 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. ϵ_{target} and ϵ_{GT} represent the noise of predefined target concepts and the ground-truth noise added in the diffusion process, respectively.

For instance, ESD [11] leverages the predicted noise for both concept-free $c_\emptyset$ and forgotten concepts c_f to construct ϵ_{target},

$$\epsilon_{target} = (1+\eta)\epsilon(\mathbf{x}_t, c_\emptyset, \theta_{dm}) - \eta\epsilon(\mathbf{x}_t, c_f, \theta_{dm}),$$

where θ_{dm} represents parameters of the frozen DMs. η is the hyperparameter. SDD [25] directly maps the prediction distribution of erased concepts c_f to the prediction distribution of concept-free $c_\emptyset$, $\epsilon_{target} = \epsilon_{\theta_{dm}}(\mathbf{x}_t, c_\emptyset, t)$. Ablation [28] 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', $\epsilon_{target} = \epsilon_{\theta_{dm}}(\mathbf{x}_t, c^*, t)$. Additionally, FMN [52] is an untargeted concept erasure method, which minimizes the values of attention maps corresponding to the forgotten concepts.

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

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, such as artist styles, 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. On the other hand, unlearned DMs should be able to flawlessly reconstruct the generation performance of the recovered concept.

$$\begin{aligned} &\mathcal{DM}_{c^*_{sub,f}}(\mathbf{z}, c_{i,f}) = \mathcal{DM}_{c_{sub,f}}(\mathbf{z}, c_{i,f}),\ s.t.\ \forall c_{i,f} \in c^*_{sub,f}, \\ &\mathcal{DM}_{c^*_{sub,f}}(\mathbf{z}, c_{j,f}) = \mathcal{DM}(\mathbf{z}, c_{j,f}), \end{aligned} \tag{2}$$

where $c_{sub,f}$ is an arbitrary subset of c_f, $c_{sub,f} \in c_f$, $c_{j,f}$ 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 $c_{j,f} \notin c^*_{sub,f}$. $\mathcal{DM}(\cdot)$ represents the original DMs. $\mathcal{DM}_{c_{sub,f}}(\cdot)$ denotes DMs with $c_{sub,f}$ erased. $\mathbf{z}$ is randomly initialized Gaussian noise.

③ **Regular Concept Preservation.** The MU techniques should preserve the generative capability of DMs for regular concepts,

$$\mathcal{DM}_{c_f}(\mathbf{z}, c_{\notin f}) \approx \mathcal{DM}(\mathbf{z}, c_{\notin f}). \tag{3}$$

④ **Watermark Preservation.** The MU techniques should not affect the generative capability of DMs for watermark prompts,

$$\mathcal{DM}_{c_f}(\mathbf{z}, c_{watermark}) = \mathcal{DM}(\mathbf{z}, c_{watermark}). \tag{4}$$

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

These objectives are illustrated in in Figure 3.

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 Eq. (2), weight shifts aimed at erasing various concepts should not interfere with each other. Therefore, we propose to decouple $\Delta\theta_{dm}$ into $\Delta\theta_{1\sim N,dm}$, where N is the number of erased concepts. $\Delta\theta_{k,dm}$ is utilized to manipulate the specific forgotten concept $c_{k,f}$. Figure 4 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 Algorithm 1 show the implementation details of it.

Q2 Can weights for erasing various concepts be decoupled?

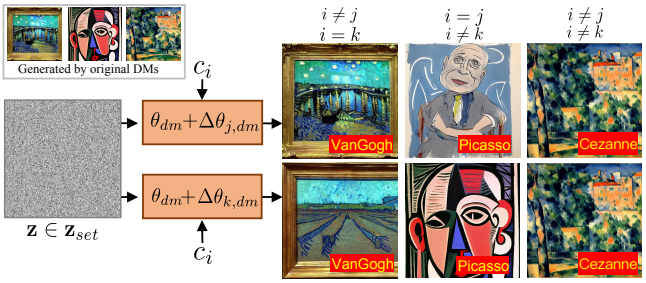

Figure 4: Decoupling weights for erasing different concepts. $\Delta\theta_{j,dm}$ and $\Delta\theta_{k,dm}$ are weight shifts of DMs designed to erase 'Picasso' and 'VanGogh' respectively.

A2: The paradigm of the independent weight shift $\Delta\theta_{k,dm}$ is derived from Eq. (2). Specifically, Eq. (2) will be satisfied when [1]

$$\begin{aligned}
&\epsilon(\mathbf{x}_t, c_\emptyset; \theta_{dm} + \sum_{k\in S_{sub}} \Delta\theta_{k,dm}) = \epsilon(\mathbf{x}_t, c_\emptyset; \theta_{dm}),\\
&\epsilon(\mathbf{x}_t, c_{i,f}; \theta_{dm} + \sum_{k\in S^*_{sub}} \Delta\theta_{k,dm}) = \epsilon(\mathbf{x}_t, c_{i,f}; \sum_{k\in S_{sub}} \Delta\theta_{k,dm}),\\
&\epsilon(\mathbf{x}_t, c_{j,f}; \theta_{dm} + \sum_{k\in S^*_{sub}} \Delta\theta_{k,dm}) = \epsilon(\mathbf{x}_t, c_{j,f}; \theta_{dm}),
\end{aligned} \tag{5}$$

where $\mathbf{x}_t$ and t can be any images and timestamps. S_{sub} denotes an arbitrary subset of $[1, \mathrm{N}]$, $i, j \in S_{sub}$, $i \neq j$. $c_{j,f}$ is the recovered concept. S^*_{sub} represents S_{sub} that removes the index j.

Conditions (**c1**~**2**) enable Eq. (5) to have solutions.

To satisfy Eq. (5) for any image, the nonzero positions of $\Delta\theta_{k,dm}$ should be image-independent (II),

$$\Delta\theta_{k,dm} = \begin{cases} \Delta\mathbf{w}_k & II \\ 0 & else. \end{cases} \tag{6}$$

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

$$\mathbf{e}_{to\_k}(c) = \mathbf{e}(c) \otimes \mathbf{w}_{to\_k}, \mathbf{w}_{to\_k} \in \mathbb{R}^{d_{in}\times d_{out}}, \tag{7}$$

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

Based on Eqs. (6) and (7), $\epsilon(\mathbf{x}_t, c; \theta_{dm} + \Delta\theta_{k,dm}) = \epsilon(\mathbf{x}_t, c; \theta_{dm})$ can be simplified as $\mathbf{e}(c) \otimes \Delta\mathbf{w}_{k,II} = 0$, where $\mathbf{w}_{II}$ indicates image-independent weights within θ_{dm}, $\mathbf{w}_{II} \in \theta_{dm}$ and $\Delta\mathbf{w}_{k,II} \in \Delta\theta_{k,dm}$. Accordingly, Eq. (5) will be realized when [2]

$$\mathbf{e}_m \otimes \Delta\mathbf{w}_{k,II} = 0,\ s.t.\ k \in [1, \mathrm{N}], \tag{8}$$

where the matrix $\mathbf{e}_m \in \mathbb{R}^{(\mathrm{N}\cdot d_{emb})\times d_{in}}$ represents

$$[\mathbf{e}(c_\emptyset)^\top; \mathbf{e}(c_{1,f})^\top; \cdots; \mathbf{e}(c_{k-1,f})^\top; \mathbf{e}(c_{k+1,f})^\top; \cdots; \mathbf{e}(c_{\mathrm{N},f})^\top]^\top, \tag{9}$$

where $\top$ means the transpose operation.

Eq. (8) has solutions when $d_{in} > r$, where r is the rank of $\mathbf{e}_m$ and $r \leq \min(\mathrm{N}\cdot d_{emb}, d_{in})$. The answer A3 will provide a detailed explanation for this.

c2: $\mathbf{e}(c) \in \mathbb{R}^{d_{emb}\times d_{in}}$, where $d_{in} \gg d_{emb}$ in DMs.

[1]The detailed reasoning process is provided in the Appendix A.1.
[2]The detailed reasoning process is provided in the Appendix A.2.

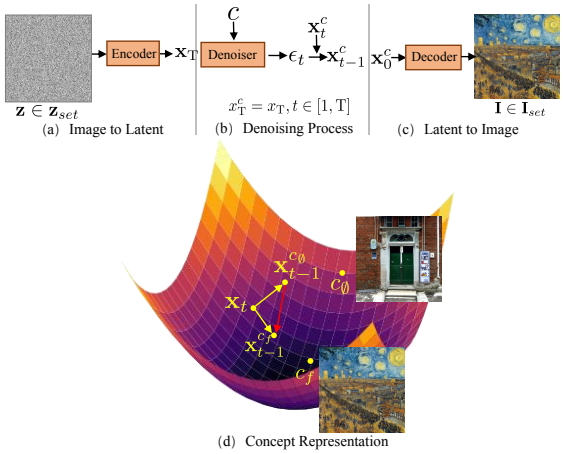

Figure 5: The denoising process of DMs (a~c) and the concept representation generation (d). $c_\emptyset$ and c_f represent the blank and forgotten concepts, respectively.

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 $\Delta\mathbf{w}_{k,II}$ should satisfy Eq. (8). 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

$$\mathbf{e}_m \otimes \mathcal{S}_i = 0, \tag{10}$$

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

$$\Delta\theta_{k,dm} = \begin{cases} (\mathbf{w}_l \otimes \mathcal{S})^\top & II \\ 0 & else, \end{cases} \tag{11}$$

where $\mathbf{w}_l$ represents the coefficients for linear combinations, $\mathbf{w}_l \in \mathbb{R}^{d_{out}\times(d_{in}-r)}$. Notably, to eliminate original biases in $\mathcal{S}$, we normalize each $\mathcal{S}_i$ to a unit vector.

After constructing optimizable variables, we proceed to illustrate the unlearning supervision function.

Q4 How to define the representations $\mathcal{R}_{c_f}$ of c_f?

A4: We first describe the denoising process of text-guided DMs, as illustrated in Figure 5 (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.

$\mathcal{R}c_f$ represents the memory within DMs for concepts c_f. In practical terms, $\mathcal{R}c_f$ 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 at each denoising time point t, as depicted in Figure 5 (d).

$$\mathcal{R}_{c_f,t-1}(\theta_{dm}) = \epsilon(\mathbf{x}_t, c_f; \theta_{dm}) - \epsilon(\mathbf{x}_t, c_\emptyset; \theta_{dm}) \propto \mathbf{x}_{t-1}^{c_f} - \mathbf{x}_{t-1}^{c_\emptyset}, \tag{12}$$

where $\epsilon(\cdot)$ means the denoiser in DMs. θ_{dm} is weights of original DMs. $\mathbf{x}^{c_f}$ and $\mathbf{x}^{c_\emptyset}$ are denoised representations supervised by c_f

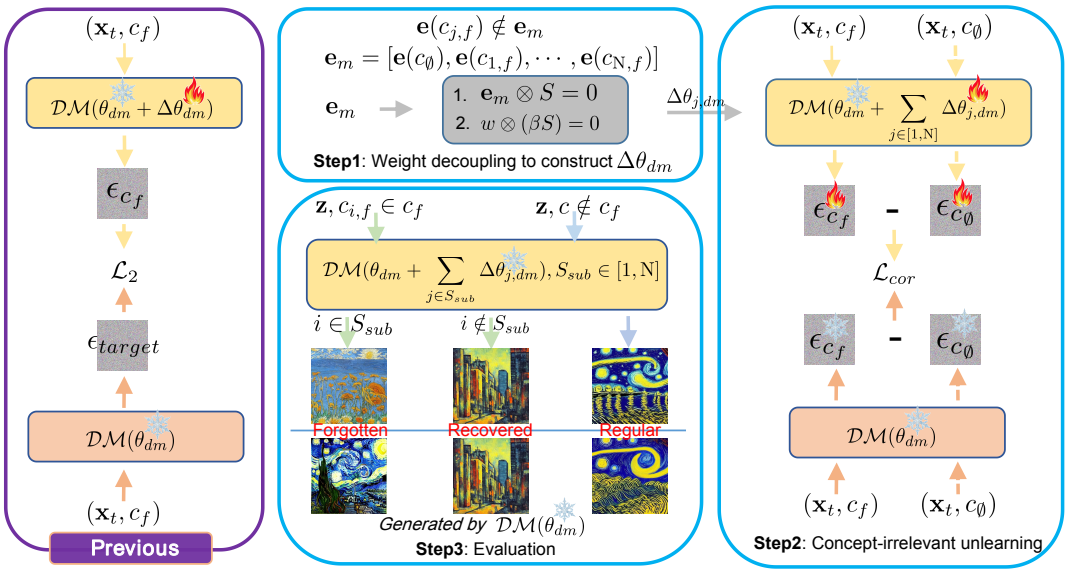

Figure 6: Comparison between our SRS-ME† and previous DM unlearning methods. 'Ice flowers' and 'flames' represent frozen and optimizable model weights respectively. ϵ_{target} is the predefined noise distribution unrelated to c_f. $\mathcal{L}_2$ is ℓ_2-norm. $\mathbf{x}_t$ means latent image representations. $\mathbf{z} \in \mathcal{N}(0, \mathbf{I})$. 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 $\Delta\theta_{1\sim\mathrm{N},dm}$. $\Delta\theta_{j,dm}$ aims to erase a specific concept $c_{j,f} \in c_f$.

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" as c_f and "a picture of a painting" as $c_\emptyset$. $\mathbf{x} \in \mathbb{R}^{h\times w\times d}$. h, w, and d are the height, width, and channel dimensions of $\mathbf{x}$, respectively.

Q5 How to estimate the unlearning degree?

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

$$\mathcal{L}^t_{cor}(c_f, \Delta\theta_{dm}) = \frac{1}{h\cdot w\cdot d}\sum \mathcal{R}_{c_f,t}(\theta_{dm}+\Delta\theta_{dm}) \odot \mathcal{R}_{c_f,t}(\theta_{dm}), \tag{13}$$

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

Q7 Can the training process for weight shifts be separated?

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

$$\mathcal{L}^t_{cor}(c_{i,f}, \sum_{i,k\in S_{sub}, i\neq k} \Delta\theta_{k,dm}) = \mathcal{L}^t_{cor}(c_{i,f}, \mathbf{0}). \tag{14}$$

This occurs because the unlearning loss of $c_{i,f}$ is unrelated to $\forall_{i,k\in S_{sub},i\neq k}\ \Delta\theta_{k,dm}$, resulting in no gradient backward propagation. $\mathbf{0}$ indicates the zero matrices with the same shape as θ_{dm}.

We omit positions where $\Delta\theta_{k,dm} = 0$ and simplify Eq. (14) as [3]

$$\begin{aligned}&\mathcal{L}^t_{cor}(c_{i,f}, \Delta_w) - \mathcal{L}^t_{cor}(c_{i,f}, \mathbf{0})\\ &\propto \mathcal{R}_{c_{i,f},t}(\mathbf{w}_{II}+\Delta_w) - \mathcal{R}_{c_{i,f},t}(\mathbf{w}_{II}) = 0,\end{aligned} \tag{15}$$

where $\Delta_w = \sum_{i,k\in S_{sub},i\neq k} \Delta\mathbf{w}_{k,II}$. Namely, for each $c_{i,f} \in c_f$, SRS-ME fulfills Eq. (14), making its training process 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 detail three variants, SRS-ME, SRS-ME†, and SRS-ME‡, each tailored to address distinct scenarios.

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

$$\begin{aligned}&\min_{\mathcal{W}} \mathcal{L}_{SRS-ME^\dagger} = \sum_{i=1}^{\mathrm{N}} \eta_i \mathcal{L}_{cor}(c_{i,f}, \sum_{j=1}^{\mathrm{N}} \Delta\theta_{j,dm}) + \lambda \| \sum_{j=1}^{\mathrm{N}} \Delta\theta_{j,dm}\|_p,\\ &s.t.\ \forall_{i\in[1,\mathrm{N}]} \mathcal{L}_{cor}(c_{i,f}, \sum_{j=1}^{\mathrm{N}} \Delta\theta_{j,dm}) = 0,\end{aligned} \tag{16}$$

[3] The detailed reasoning process is provided in the Appendix A.4.

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‡	✓	-	-	-	✓

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

$$\eta_i = \frac{\|\mathcal{L}_{cor}(c_{1,f}, \sum_{j=1}^{\mathrm{N}} \Delta\theta_{j,dm})\|_2}{\|\mathcal{L}_{cor}(c_{i,f}, \sum_{j=1}^{\mathrm{N}} \Delta\theta_{j,dm})\|_2}. \tag{17}$$

$\Delta\theta_{j,dm}$ is formulated as

$$\Delta\theta_{j,dm} = \begin{cases} (\mathbf{w}_l \otimes (\beta\mathcal{S}))^{\top} & II \\ 0 & else, \end{cases} \tag{18}$$

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 $\mathbf{w}_l$ 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. (3), we employ the following settings to restrict modifications to model weights:

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

To realize the condition of zero relevance in Eq. (16), we utilize the momentum statistic method since $\mathcal{L}_{cor}(c_{i,f}, \Delta\theta_{j,dm})$ is affected by noisy inputs $\mathbf{x}_t$. Early stopping is activated once $\mathcal{L}_{mom}^n \leq \tau$, where τ denotes a threshold with a small value.

$$\mathcal{L}_{mom}^n = \alpha\mathcal{L}_{mom}^{n-1} + (1-\alpha)\sum_{i=1}^{\mathrm{N}} \eta_i \mathcal{L}_{cor}(c_{i,f}, \sum_{j=1}^{\mathrm{N}} \Delta\theta_{j,dm}). \tag{19}$$

where n represents the number of iterations.

Taking SRS-ME† as an example, we provide the figure description in Figure 6 and the implementation details in Algorithm 1.

SRS-ME. Section 3.2 demonstrates the feasibility of optimization decoupling, namely, separately optimizing the decoupled weight shifts $\forall_{j\in[1,\mathrm{N}]}\Delta\theta_{j,dm}$. Fine-tunning each $\Delta\theta_{j,dm}$ 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 $\Delta\theta_{i,dm}$ to erase associated concepts, *e.g.*, DMs with $\theta_{dm} + \sum_{i\in\{j,k\}} \Delta\theta_{i,dm}$ 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 $\mathbf{e}_m$ of Eq. (9).

Algorithm 1: SRS-ME†.

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

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

/**Weight decoupling for constructing $\Delta\theta_{j,dm}$.*/
for $c_{j,f} \in c_f$ **do**
  $\mathbf{e}_m = [\mathbf{e}(c_\emptyset)^{\top}, \mathbf{e}(c_{1,f})^{\top}, \cdots, \mathbf{e}(c_{j-1,f})^{\top}, \mathbf{e}(c_{j+1,f})^{\top}, \cdots, \mathbf{e}(c_{\mathrm{N},f})^{\top}]^{\top}$;
  Obtain solutions $\mathcal{S}$ for $\mathbf{e}_m \otimes \mathcal{S} = 0$;
  Initialize learnable variables $\mathbf{w}_l$ with zero matrices.
  $\Delta\theta_{j,dm} = \begin{cases} (\mathbf{w}_l \otimes (\beta\mathcal{S}))^{\top} & II \\ 0 & else, \end{cases}$
end
/**Concept-irrelevant unlearning for optimizing $\mathbf{w}_l$.*/
for $n, \mathbf{x}_0 \in D$ **do**
  Randomly select a sampling step t;
  $\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, \epsilon \in \mathcal{N}(0, \mathbf{I})$;
  $\epsilon_{c_f} = \epsilon(\mathbf{x}_t, c_f; \theta_{dm})$; $\epsilon_{c_\emptyset} = \epsilon(\mathbf{x}_t, c_\emptyset; \theta_{dm})$;
  for $c_{j,f} \in c_f$ **do**
    $\epsilon'_{c_{j,f}} = \epsilon(\mathbf{x}_t, c_{j,f}; \theta_{dm} + \sum_{i\in[1,\mathrm{N}]} \Delta\theta_{i,dm})$;
    $\epsilon'_{c_\emptyset} = \epsilon(\mathbf{x}_t, c_\emptyset; \theta_{dm} + \sum_{i\in[1,\mathrm{N}]} \Delta\theta_{i,dm}))$;
    Obtain $\mathcal{L}_{cor}(c_{j,f}, \sum_{i\in[1,\mathrm{N}]} \Delta\theta_{i,dm}))$ with Eq. (13);
    Calculate η_j using Eq. (17);
  end
  Calculate $\mathcal{L}_{mom}^n$ using Eq. (19);
  if $\mathcal{L}_{mom}^n \leq \tau$ **then**
    break;
  end
  $\min_{\mathcal{W}} \mathcal{L}_{SRS-ME}^{\dagger} = \sum_{j=1}^{\mathrm{N}} \eta_j \mathcal{L}_{cor}(c_{j,f}, \sum_{i\in[1,\mathrm{N}]} \Delta\theta_{i,dm}) + \lambda\|\sum_{i\in[1,\mathrm{N}]} \Delta\theta_{i,dm})\|_p$
end

SRS-ME‡. SRS-ME and SRS-ME† are specifically designed to preserve model watermarks and flexibly manipulate concepts, allowing only the image-independent layers to be fine-tuned. However, special concepts such as "Nudity" should not be recovered, and researchers can manipulate arbitrary model weights when watermark preservation is not considered. To erase concepts under such scenarios, we introduce SRS-ME‡ and formulate it as follows:

$$\min_{\Delta\theta_{dm}} \mathcal{L}_{SRS-ME^{\ddagger}} = \sum_{i=1}^{\mathrm{N}} \eta_i \mathcal{L}_{cor}(c_{i,f}, \Delta\theta_{dm}) + \lambda\|\Delta\theta_{dm}\|_p, \quad s.t.\ \mathcal{L}_{mom} \leq \tau, \tag{20}$$

where $\mathcal{L}_{mom}$ at the n-th iteration is expressed as

$$\mathcal{L}_{mom}^n = \alpha\mathcal{L}_{mom}^{n-1} + (1-\alpha)\sum_{i=1}^{\mathrm{N}} \eta_i \mathcal{L}_{cor}(c_{i,f}, \Delta\theta_{dm}), \tag{21}$$

where $\eta_i = \frac{\|\mathcal{L}_{cor}(c_{1,f}, \Delta\theta_{dm})\|_2}{\|\mathcal{L}_{cor}(c_{i,f}, \Delta\theta_{dm})\|_2}$.

The comparison between our methods is depicted in Table 1.

Table 2: Quantitative results (FID/ACC/LPIPS) of SRS-ME on style unlearning. ORI denotes the results of original DMs.

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
$Scene_1$	t_0	208.8/12.4/.338	219.9/36.8/.388	199.8/6.40/.336	274.9/26.4/.329	202.7/38.0/.312	216.5/15.2/.596	56.0/95.2/.153	134.6/98.2/.125	110.0/99.2/.126
	t_1	0.75/98.0/.00	2.12/90.8/.00	199.3/6.00/.336	276.7/26.8/.328	203.9/38.3/.313	209.0/25.2/.526	47.1/98.8/.096	119.3/100/.102	101.1/99.2/.107
	t_2	0.78/98.4/.00	3.28/90.4/.00	1.25/98.8/.00	276.5/26.8/.329	203.1/38.0/.312	208.7/25.6/.526	44.7/99.6/.097	99.9/100/.080	85.8/100/.066
$Scene_2$	t_0	208.8/12.4/.338	219.9/36.8/.388	199.8/6.40/.336	274.9/26.4/.329	202.7/38.0/.312	216.5/15.2/.596	56.0/95.2/.153	134.6/98.2/.125	110.0/99.2/.126
	t_1	208.7/12.4/.339	219.7/38.0/.388	199.5/6.40/.336	0.18/96.0/.00	2.83/99.6/.00	218.7/16.0/.596	75.1/86.4/.192	179.8/98.8/.141	116.6/98.8/.141
	t_2	208.5/12.4/.338	219.6/38.0/.388	199.6/6.40/.336	0.14/96.0/.00	2.08/100/.00	2.11/98.4/.00	61.0/92.4/.168	94.7/100/.076	114.9/98.4/.151
$Scene_3$	t_0	255.0/15.6/.349	95.9/86.2/.153	218.0/2.00/.351	170.2/64.5/.151	182.3/49.2/.308	148.3/94.4/.435	57.6/95.6/.134	97.2/100/.090	145.6/84.0/.239
	t_1	1.04/98.0/.00	67.4/90.4/.125	1.01/98.8/.00	39.5/96.0/.026	181.9/48.0/.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	205.7/66.8/.393	1.15/98.8/.00	279.7/29.2/.325	182.0/48.0/.308	59.8/98.0/.264	46.6/97.6/.096	103.0/100./.086	99.5/99.2/.123
	t_3	1.02/98.0/.00	67.4/90.4/.125	0.98/98.8/.00	279.9/29.2/.326	182.2/48.0/.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	279.7/28.8/.326	182.1/48.0/.307	209.9/.527/25.2	52.2/97.2/.116	88.5/100/.063	83.0/99.6/.061
$Scene_4$	t_0	188.3/36.0/.256	218.6/37.2/.395	214.2/5.20/.365	183.5/60.4/.245	96.7/98.8/.182	52.5/97.2/.189	65.1/95.6/.160	50.8/100./.016	108.3/99.2/.147
	t_1	0.77/98.0/.00	218.6/37.2/.395	214.6/5.20/.364	175.6/64.0/.211	96.0/99.6/.166	45.4/96.8/.147	49.7/98.4/.096	49.9/100/.016	112.6/100/.156
	t_2	0.81/98.0/.00	218.6/37.6/.395	215.0/5.20/.365	365.4/5.60/.463	246.1/19.6/.379	135.4/59.6/.402	77.4/86.0/.176	61.0/100/.024	112.6/99.6/.151
	t_3	0.76/98.0/.00	3.82/90.4/.00	214.9/5.20/.364	354.6/2.80/.458	219.4/28.0/.340	131.5/58.2/.395	73.3/90.0/.166	52.0/100/.016	100.5/98.4/.123
	t_4	0.87/98.0/.00	2.82/90.4/.00	214.4/5.20/.365	354.4/2.80/.458	219.2/28.4/.340	288.4/11.6/.623	82.1/85.6/.185	77.0/100/.045	103.1/98.4/.123
	t_5	0.921/98.0/.00	4.11/90.4/.00	1.06/98.8/.00	274.3/26.4/.343	230.9/28.0/.368	277.5/17.6/.615	68.4/90.8/.154	78.3/100/.046	91.8/99.6/.098

4 EXPERIMENTS

4.1 Experimental Settings

Implementation Details. We follow prior works [11, 25] to unlearn concepts from Stable Diffusion [40]. 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 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. (19), β in Eq. (18), λ in Eq. (16) and τ are set to 0.9, 1e-4, 1e-6, 1e-4 respectively. $\|\theta\|_p = \frac{\|\theta\|_1}{\mathrm{M}}$, 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/SepCE4MU-7B0C/README.md.

Evaluation Metrics. ① Fréchet Inception Distance (FID) [16] between images generated by unlearned and original DMs;

② Classification accuracy (ACC) of pre-trained classification models $\mathcal{M}$ for images generated by DMs;

③ Perceptual Image Patch Similarity (LPIPS) [53] 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 [15] is optimized with 20 epochs. The cyclical learning rate [45] is employed with the maximum learning rate of 1e-2.

For the object classification network, we directly utilize the pre-trained ResNet50 [15]. The object unlearning evaluation utilizes categories in Imagenette [19] following [11, 25], 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'.

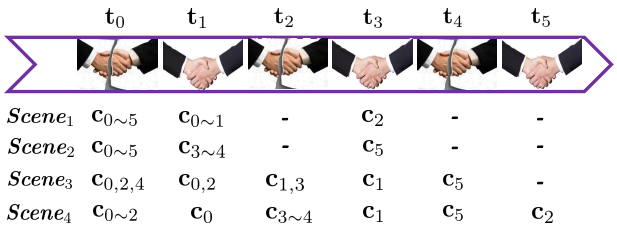

Figure 7: Experimental scenario clarification.

Baselines. We use advanced unlearning methods as baselines, including FMN [52], ESD [11] and Ablation [25].

Evaluation Data. We yield 250 images using DMs for each concept, *i.e.*, 50 seeds per concept and 5 images 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 $Scene_4$.

SRS-ME can handle these cases with the same setting. At each timestamp, $\mathbf{e}_m$ contains all other previously erased or recovered concepts. For instance, $\Delta\theta_1$ at t_0 in $Scene_1$ is obtained using

$$\mathbf{e}_m = [\mathbf{e}(c_\emptyset), \mathbf{e}(c_0), \mathbf{e}(c_{2\sim5})], \tag{22}$$

and $\Delta\theta_1$ at t_2 in $Scene_3$ is calculated based on

$$\mathbf{e}_m = [\mathbf{e}(c_\emptyset), \mathbf{e}(c_0), \mathbf{e}(c_{2\sim4})]. \tag{23}$$

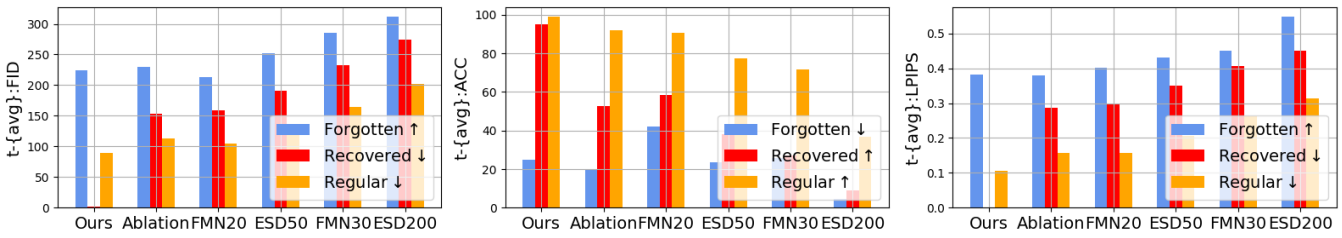

Figure 8: Performance comparison on scene$_1$ of style unlearning.

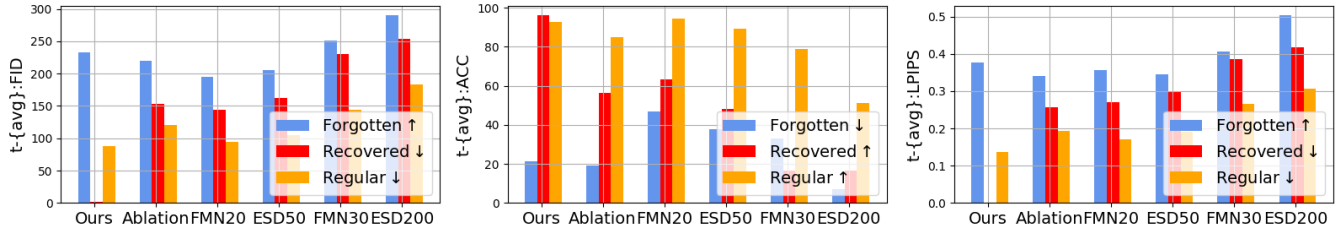

Figure 9: Performance comparison on scene$_4$ of style unlearning.

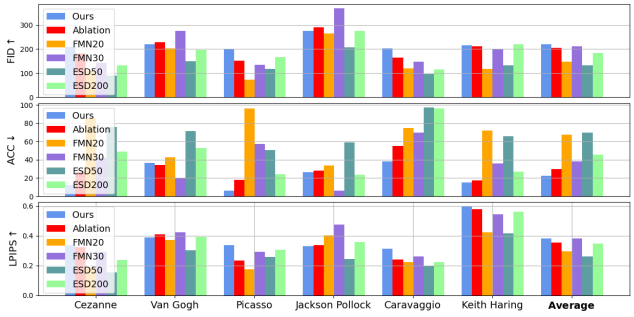

Figure 10: Performance comparison of single style unlearning for models combined to erase multiple styles. The x-axis indicates the erased styles.

Notably, in iterative concept erasure, we do not recalculate 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 $\Delta\theta_{5,dm}$ for erasing c_5, and directly utilize $\theta_{dm} + \sum_{i\in[2,3,4,5]} \Delta\theta_{i,dm}$ as the unlearned model weights.

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. Additionally, the DM unlearning at all timestamps does not affect the generation performance on the blank prompt. This is because we utilize the blank prompt as a constant vector in $\mathbf{e}_m$ of Eq. (9). This demonstrates the feasibility of our SRS-ME in preserving model watermarks, *i.e.*, utilizing watermark prompts as constant vectors in $\mathbf{e}_m$; (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. 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. 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. 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 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

Figure 11: Visual examples of SRS-ME on style unlearning.

Table 3: Quantitative results (FID/ACC/LPIPS/) of SRS-ME on object unlearning. ORI denotes the results of original DMs.

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	331.3/1.2/.331	215.6/48.8/.386	260.7/2.0/.443	166.5/12.4/.358	316.5/2.8/.479	325.9/0.4/.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	262.2/1.6/.443	167.0/12.8/.358	317.0/2.8/.479	325.6/0.4/.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	166.3/12.4/.358	315.0/2.8/.479	325.3/0.4/.393	20.0/98.8/.255	27.3/92.8/.228	14.6/97.6/.219
Scene$_2$	t_0	331.3/1.2/.331	215.6/48.8/.386	260.7/2.0/.443	166.5/12.4/.358	316.5/2.8/.479	325.9/0.4/.393	18.2/98.0/.195	75.4/73.6/.411	57.3/85.2/.347
	t_1	333.2/1.2/.331	215.9/48.4/.386	262.0/2.8/.443	0.24/82.0/.000	0.36/85.2/.000	325.4/0.4/.392	19.7/96.0/.184	55.3/84.4/.343	24.5/95.2/.293
	t_3	332.4/1.2/.331	215.9/49.2/.386	261.8/2.0/.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	286.3/10.4/.319	34.1/90.0/.206	270.9/3.20/.397	35.1/68.8/.183	189.2/30.8/.340	26.8/82.8/.147	8.51/96.8/.038	45.1/86.8/.241	9.80/100/.150
	t_1	0.729/90.4/.000	15.2/81.2/.088	0.471/60.4/.177	26.7/70.8/.339	187.9/30.8/.061	11.6/92.8/.061	4.68/97.2/.018	13.8/92.4/.049	18.4/97.2/.182
	t_2	0.788/90.4/.000	268.2/28.8/.390	0.440/60.0/.000	228.2/9.20/.412	189.1/30.4/.340	17.2/91.2/.113	18.0/93.2/.134	23.3/92.0/.131	10.3/99.6/.145
	t_3	0.597/91.2/.000	15.2/81.2/.088	0.496/59.6/.000	228.1/.096/.412	189.3/29.6/.340	15.4/93.6/.112	7.53/96.4/.039	23.3/90.4/.140	13.8/98.4/.165
	t_4	0.625/91.2/.000	15.2/81.2/.088	.571/60.0/.000	227.7/9.20/.412	188.6/30.4/.340	312.2/0.00/.371	10.8/96.8/.058	19.6/92.0/.101	13.3/99.2/.179
Scene$_4$	t_0	244.3/22.4/.276	169.0/48.0/.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	1.118/91.2/.000	168.8/48.0/.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	0.586/91.2/.000	168.9/47.2/.362	89.9/40.0/.291	216.8/7.20/.502	237.3/8.00/.431	50.6/83.2/.357	35.1/89.2/.267	30.6/89.2/.213	27.4/99.6/.444
	t_3	0.663/92.4/.000	0.231/79.6/.000	90.0/40.4/.291	209.0/5.60/.414	206.2/11.2/.371	77.0/69.2/.428	35.5/88.4/.242	44.9/83.2/.245	19.2/100/.333
	t_4	0.621/90.8/.000	0.220/79.6/.000	90.1/40.0/.291	209.0/6.40/.414	205.8/11.2/.371	303.9/2.80/.431	33.2/91.6/.265	31.0/88.4/.210	14.7/100/.290
	t_5	0.625/91.2/.000	15.2/81.2/.088	0.571/60.0/.000	227.7/9.20/.412	188.6/30.4/.340	312.2/0.00/.371	10.8/96.8/.058	19.6/92.0/.101	13.3/99.2/.179

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 Ablation, 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.

Additionally, we present visual examples of SRS-ME in Figure 11. As observed, the proposed SRS-ME effectively removes undesirable concepts while preserving the overall layout for most images.

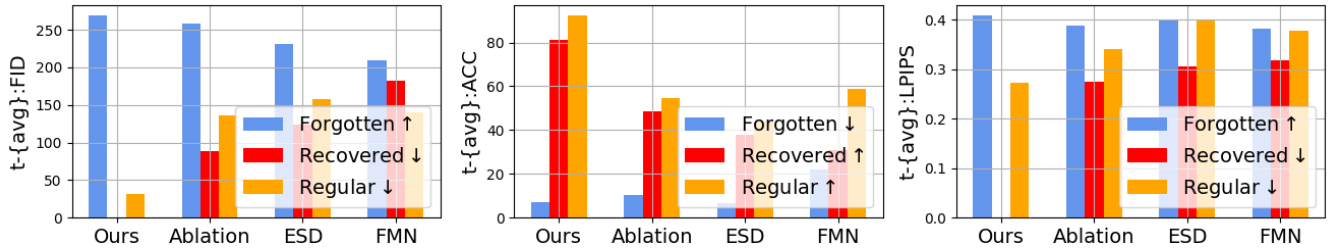

Figure 12: Performance comparison on scene$_1$ of object unlearning.

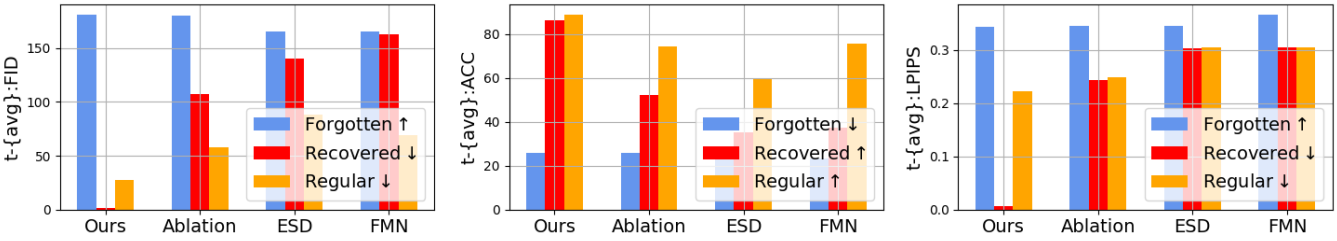

Figure 13: Performance comparison on scene$_4$ of object unlearning.

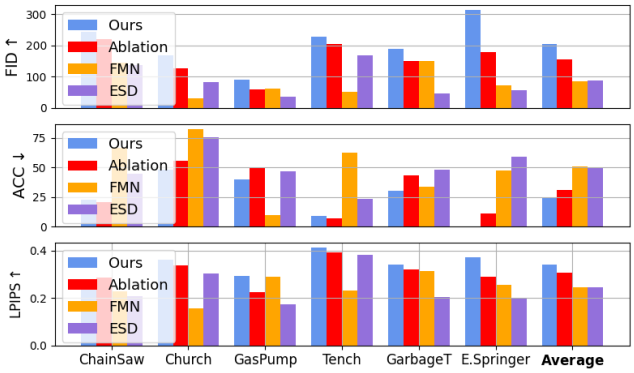

Figure 14: Performance comparison of single object unlearning for models combined to erase multiple objects. The x-axis indicates the erased objects.

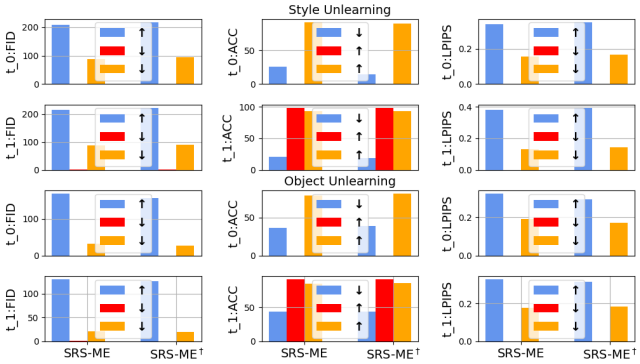

Figure 15: Performance comparison at scene$_4$ between our SRS-ME and SRS-ME†. , , and denote results of forgotten, recovered and regular concepts, respectively.

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 14 displays a comparison of single object unlearning performance for models combined to erase multiple objects. Additionally, Figures 12 and 13 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 also offer visual examples of SRS-ME on object unlearning in Figure 16. 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 15. 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‡

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 [42] and categorize images exposing body parts into different nudity classes with Nudenet [1]. For evaluating model preservation performance, we adopt 1859 images generated using prompts from 1000 different categories listed in the ImageNet

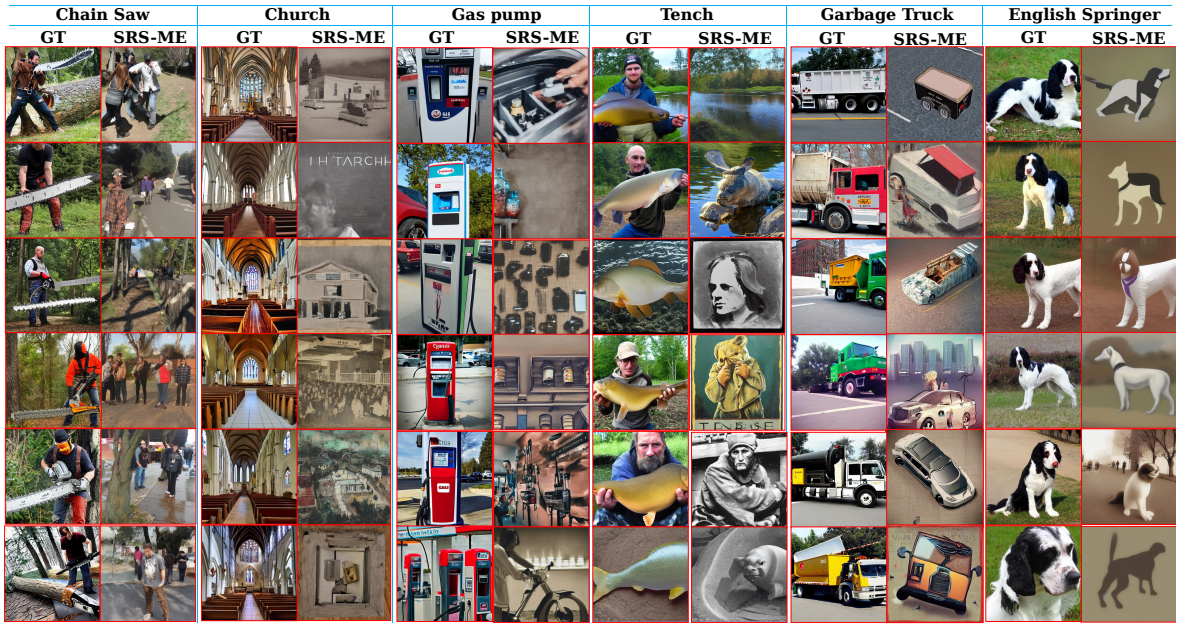

Figure 16: Visual examples of SRS-ME on object unlearning.

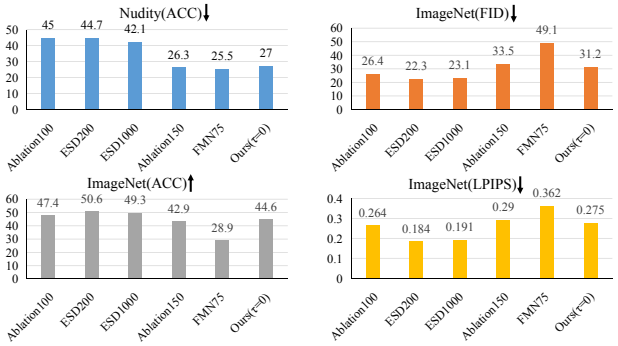

Figure 17: Comparative experiments on 'Nudity' unlearning.

dataset [4]. Results in Figure 17 show that our approach exhibits superior model preservation performance compared to previous methods under comparable erasure performance, *e.g.*, the results of Ablation150, FMN75, and our 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]https://github.com/rohitgandikota/erasing/blob/main/data

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		t_0	**220.9/24.9**/.379	0.00/0.00/.00	146.8/92.0/.167
	Scene$_1$	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		t_0	220.4/20.0/**.383**	0.00/0.00/.00	**100.2/97.5/.135**
	Scene$_1$	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**

4.6.2 Impact of the momentum statistic. Figure 18 shows that the unlearning loss exhibits instability, and the integration of momentum statistics significantly mitigates 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 20 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 21 illustrates that increasing β accelerates the unlearning optimization but also causes more modifications to the model weights.

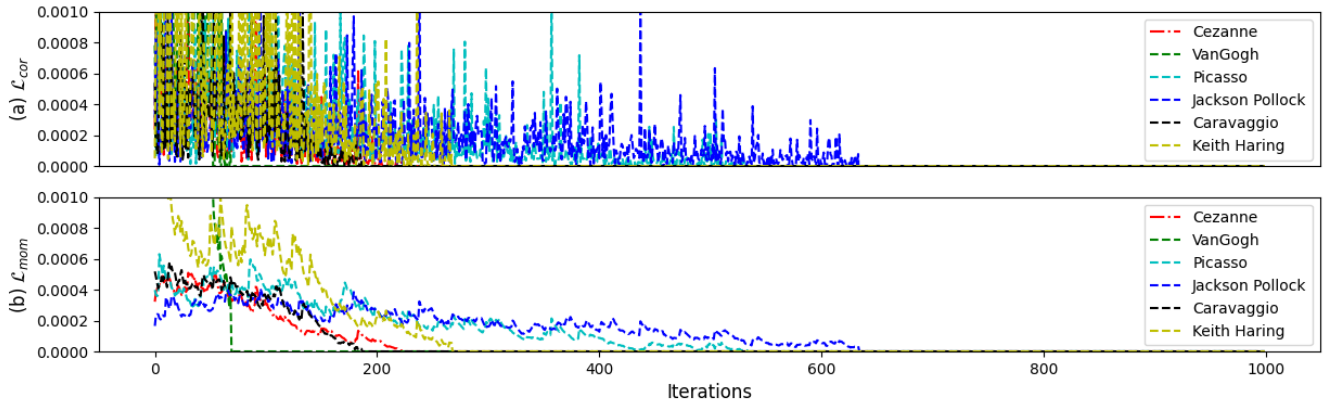

Figure 18: The impact of momentum statistics.

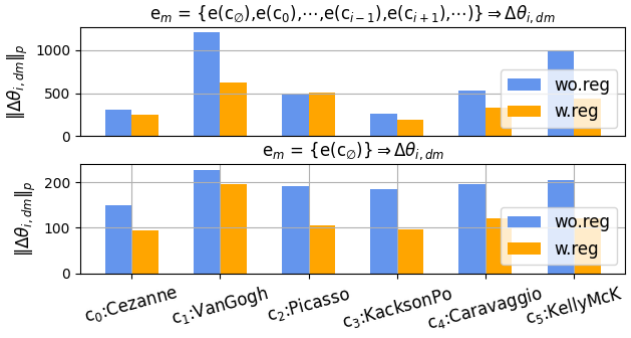

Figure 19: Influence of the weight regularization on SRS-ME.

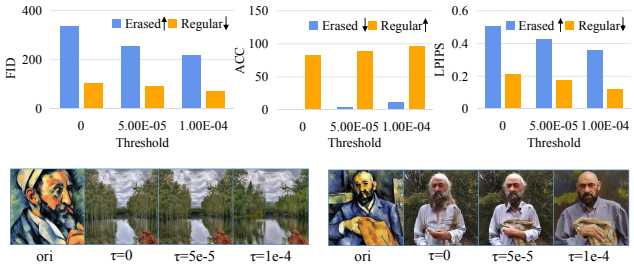

Figure 20: Impact of the threshold τ on SRS-ME.

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 $\mathbf{e}_m$. Experimental results are depicted in Figure 22. As observed, the closer the decoupled concept aligns with the forgotten concept, the more challenging the erasure becomes.

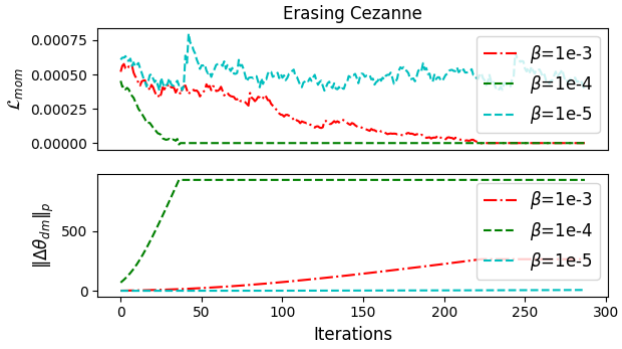

Figure 21: Impact of the hyperparameter β on SRS-ME. We set both τ and λ to 0.

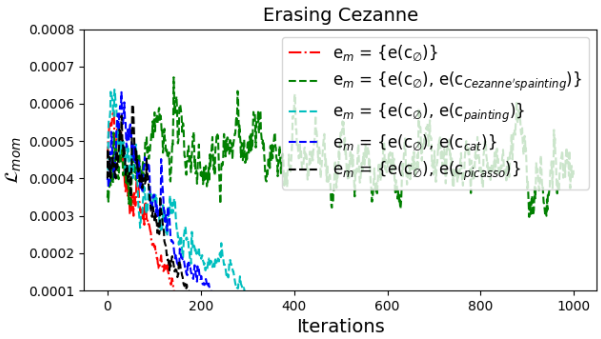

Figure 22: Impact of similar concepts on weight decoupling.

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

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

[3] Zhi-Yi Chin, Chieh-Ming Jiang, Ching-Chun Huang, Pin-Yu Chen, and Wei-Chen Chiu. 2023. Prompting4debugging: Red-teaming text-to-image diffusion models by finding problematic prompts. *arXiv preprint arXiv:2309.06135* (2023).

[4] Rishav Chourasia and Neil Shah. 2023. Forget unlearning: Towards true data-deletion in machine learning. In *International Conference on Machine Learning*. PMLR, 6028–6073.

[5] Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. 2023. Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2023).

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

[7] Xin Ding, Yongwei Wang, Zuheng Xu, William J Welch, and Z Jane Wang. 2020. Ccgan: Continuous conditional generative adversarial networks for image generation. In *International conference on learning representations*.

[8] Jinhao Duan, Fei Kong, Shiqi Wang, Xiaoshuang Shi, and Kaidi Xu. 2023. Are diffusion models vulnerable to membership inference attacks? *arXiv preprint arXiv:2302.01316* (2023).

[9] Jan Dubiński, Antoni Kowalczuk, Stanisław Pawlak, Przemyslaw Rokita, Tomasz Trzciński, and Paweł Morawiecki. 2024. Towards More Realistic Membership Inference Attacks on Large Diffusion Models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 4860–4869.

[10] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. 2022. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618* (2022).

[11] Rohit Gandikota, Joanna Materzynska, Jaden Fiotto-Kaufman, and David Bau. 2023. Erasing concepts from diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

[12] Songwei Ge, Taesung Park, Jun-Yan Zhu, and Jia-Bin Huang. 2023. Expressive text-to-image generation with rich text. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 7545–7556.

[13] Aditya Golatkar, Alessandro Achille, Ashwin Swaminathan, and Stefano Soatto. 2023. Training Data Protection with Compositional Diffusion Models. *arXiv preprint arXiv:2308.01937* (2023).

[14] Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. 2022. Vector quantized diffusion model for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10696–10706.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[16] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems* 30 (2017).

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

[18] Seunghoo Hong, Juhun Lee, and Simon S Woo. 2023. All but One: Surgical Concept Erasing with Model Preservation in Text-to-Image Diffusion Models. *arXiv preprint arXiv:2312.12807* (2023).

[19] Jeremy Howard and Sylvain Gugger. 2020. Fastai: A layered API for deep learning. *Information* 11, 2 (2020), 108.

[20] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685* (2021).

[21] Chi-Pin Huang, Kai-Po Chang, Chung-Ting Tsai, Yung-Hsuan Lai, and Yu-Chiang Frank Wang. 2023. Receler: Reliable Concept Erasing of Text-to-Image Diffusion Models via Lightweight Erasers. *arXiv preprint arXiv:2311.17717* (2023).

[22] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. 2023. Imagic: Text-based real image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6007–6017.

[23] Amirhossein Kazerouni, Ehsan Khodapanah Aghdam, Moein Heidari, Reza Azad, Mohsen Fayyaz, Ilker Hacihaliloglu, and Dorit Merhof. 2022. Diffusion models for medical image analysis: A comprehensive survey. *arXiv preprint arXiv:2211.07804* (2022).

[24] Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. 2022. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2426–2435.

[25] Sanghyun Kim, Seohyeon Jung, Balhae Kim, Moonseok Choi, Jinwoo Shin, and Juho Lee. 2023. Towards safe self-distillation of internet-scale text-to-image diffusion models. *arXiv preprint arXiv:2307.05977* (2023).

[26] Yunji Kim, Jiyoung Lee, Jin-Hwa Kim, Jung-Woo Ha, and Jun-Yan Zhu. 2023. Dense text-to-image generation with attention modulation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 7701–7711.

[27] Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013).

[28] Nupur Kumari, Bingliang Zhang, Sheng-Yu Wang, Eli Shechtman, Richard Zhang, and Jun-Yan Zhu. 2023. Ablating concepts in text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

[29] Bowen Li, Xiaojuan Qi, Philip Torr, and Thomas Lukasiewicz. 2020. Lightweight generative adversarial networks for text-guided image manipulation. *Advances in Neural Information Processing Systems* 33 (2020), 22020–22031.

[30] Hang Li, Chengzhi Shen, Philip Torr, Volker Tresp, and Jindong Gu. 2023. Self-Discovering Interpretable Diffusion Latent Directions for Responsible Text-to-Image Generation. *arXiv preprint arXiv:2311.17216* (2023).

[31] Rui Liu, Yixiao Ge, Ching Lam Choi, Xiaogang Wang, and Hongsheng Li. 2021. Divco: Diverse conditional image synthesis via contrastive generative adversarial network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 16377–16386.

[32] Yugeng Liu, Zheng Li, Michael Backes, Yun Shen, and Yang Zhang. 2023. Watermarking diffusion model. *arXiv preprint arXiv:2305.12502* (2023).

[33] Zhi-Song Liu, Wan-Chi Siu, and Li-Wen Wang. 2021. Variational autoencoder for reference based image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 516–525.

[34] Ninareh Mehrabi, Palash Goyal, Christophe Dupuy, Qian Hu, Shalini Ghosh, Richard Zemel, Kai-Wei Chang, Aram Galstyan, and Rahul Gupta. 2023. Flirt: Feedback loop in-context red teaming. *arXiv preprint arXiv:2308.04265* (2023).

[35] Zixuan Ni, Longhui Wei, Jiacheng Li, Siliang Tang, Yueting Zhuang, and Qi Tian. 2023. Degeneration-tuning: Using scrambled grid shield unwanted concepts from stable diffusion. In *Proceedings of the 31st ACM International Conference on Multimedia*. 8900–8909.

[36] Taesung Park, Jun-Yan Zhu, Oliver Wang, Jingwan Lu, Eli Shechtman, Alexei Efros, and Richard Zhang. 2020. Swapping autoencoder for deep image manipulation. *Advances in Neural Information Processing Systems* 33 (2020), 7198–7211.

[37] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.

[38] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125* 1, 2 (2022), 3.

[39] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *International Conference on Machine Learning*. PMLR, 8821–8831.

[40] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10684–10695.

[41] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 2023. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 22500–22510.

[42] Patrick Schramowski, Manuel Brack, Björn Deiseroth, and Kristian Kersting. 2023. Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 22522–22531.

[43] Shawn Shan, Jenna Cryan, Emily Wenger, Haitao Zheng, Rana Hanocka, and Ben Y Zhao. 2023. Glaze: Protecting artists from style mimicry by text-to-image models. *arXiv preprint arXiv:2302.04222* (2023).

[44] Huajie Shao, Shuochao Yao, Dachun Sun, Aston Zhang, Shengzhong Liu, Dongxin Liu, Jun Wang, and Tarek Abdelzaher. 2020. Controlvae: Controllable variational autoencoder. In *International Conference on Machine Learning*. PMLR, 8655–8664.

[45] Leslie N Smith. 2017. Cyclical learning rates for training neural networks. In *2017 IEEE winter conference on applications of computer vision (WACV)*. IEEE, 464–472.

[46] Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. 2023. Diffusion art or digital forgery? investigating data replication in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6048–6058.

[47] Jiaming Song, Chenlin Meng, and Stefano Ermon. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502* (2020).

[48] Xingqian Xu, Zhangyang Wang, Gong Zhang, Kai Wang, and Humphrey Shi. 2023. Versatile diffusion: Text, images and variations all in one diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 7754–7765.

[49] Yuhui Xu, Lingxi Xie, Xiaotao Gu, Xin Chen, Heng Chang, Hengheng Zhang, Zhensu Chen, Xiaopeng Zhang, and Qi Tian. 2023. Qa-lora: Quantization-aware

low-rank adaptation of large language models. *arXiv preprint arXiv:2309.14717* (2023).
[50] Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. 2023. Diffusion models: A comprehensive survey of methods and applications. *Comput. Surveys* 56, 4 (2023), 1–39.
[51] Yijun Yang, Ruiyuan Gao, Xiaosen Wang, Nan Xu, and Qiang Xu. 2023. MMA-Diffusion: MultiModal Attack on Diffusion Models. *arXiv preprint arXiv:2311.17516* (2023).
[52] Eric Zhang, Kai Wang, Xingqian Xu, Zhangyang Wang, and Humphrey Shi. 2023. Forget-me-not: Learning to forget in text-to-image diffusion models. *arXiv preprint arXiv:2303.17591* (2023).
[53] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 586–595.
[54] Yimeng Zhang, Jinghan Jia, Xin Chen, Aochuan Chen, Yihua Zhang, Jiancheng Liu, Ke Ding, and Sijia Liu. 2023. To generate or not? safety-driven unlearned diffusion models are still easy to generate unsafe images... for now. *arXiv preprint arXiv:2310.11868* (2023).
[55] Yunqing Zhao, Tianyu Pang, Chao Du, Xiao Yang, Ngai-Man Cheung, and Min Lin. 2023. A recipe for watermarking diffusion models. *arXiv preprint arXiv:2303.10137* (2023).
[56] Yufan Zhou, Ruiyi Zhang, Changyou Chen, Chunyuan Li, Chris Tensmeyer, Tong Yu, Jiuxiang Gu, Jinhui Xu, and Tong Sun. 2022. Towards language-free training for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 17907–17917.

A APPENDIX

A.1 Proof for Eq. (5)

We summarize Eq. (2) as follows

$$\begin{aligned} &\mathcal{DM}(\mathbf{z}, c_{j,f}; \theta_1) = \mathcal{DM}(\mathbf{z}, c_{j,f}; \theta_2) \\ &\Rightarrow \mathbf{x}_0^{\Diamond} = \mathbf{x}_0^{\square}, \end{aligned} \tag{24}$$

where $\mathbf{x}_0^{\Diamond}$ and $\mathbf{x}_0^{\square}$ represent latent representations produced by the DMs with parameters θ_1 and θ_2, respectively. The index '0' denotes the final schedule time. $c_{j,f}$ is a specified concept.

Taking DDIM [47] as an example, given $\mathbf{x}_T \sim \mathcal{N}(0, \mathbf{I})$, $\mathbf{z} \sim \mathcal{N}(0, \mathbf{I})$, the sampling process is expressed as

$$\mathbf{x}_{t-1} = \sqrt{\overline{\alpha}_{t-1}} \frac{\mathbf{x}_t - \sqrt{1-\overline{\alpha}_t}\epsilon(\mathbf{x}_t)}{\sqrt{\overline{\alpha}_t}} + \sqrt{1-\overline{\alpha}_{t-1}-\delta_t^2}\epsilon(\mathbf{x}_t) + \delta_t \mathbf{z}. \tag{25}$$

For clarity, we simply Eq. (25) as

$$\mathbf{x}_{t-1} = \lambda_1 \mathbf{x}_t - \lambda_2 \epsilon(\mathbf{x}_t) + \lambda_3 \mathbf{z}, \tag{26}$$

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. (24) and (26), we have

$$\begin{aligned} &\mathbf{x}_0^{\Diamond} = \mathbf{x}_0^{\square} \\ &\Rightarrow \lambda_1 \mathbf{x}_1^{\Diamond} - \lambda_2 \epsilon(\mathbf{x}_1^{\Diamond})^{\Diamond} = \lambda_1 \mathbf{x}_1^{\square} - \lambda_2 \epsilon(\mathbf{x}_1^{\square})^{\square}. \end{aligned} \tag{27}$$

Eq. (27) could be satisfied when

$$\begin{aligned} &\mathbf{x}_1^{\Diamond} = \mathbf{x}_1^{\square} \\ &\epsilon(\mathbf{x}_1^{\Diamond})^{\Diamond} = \epsilon(\mathbf{x}_1^{\square})^{\square}. \end{aligned} \tag{28}$$

Similarly, $\mathbf{x}_t^{\Diamond} = \mathbf{x}_t^{\square}$ can be satisfied when

$$\begin{aligned} &\mathbf{x}_{t+1}^{\Diamond} = \mathbf{x}_{t+1}^{\square} \\ &\epsilon(\mathbf{x}_{t+1}^{\Diamond})^{\Diamond} = \epsilon(\mathbf{x}_{t+1}^{\square})^{\square}. \end{aligned} \tag{29}$$

Notably, $\mathbf{x}_T^{\Diamond} = \mathbf{x}_T^{\square} = \mathbf{x}_T$. Hence, the sufficient condition for $\mathbf{x}_0^{\Diamond} = \mathbf{x}_0^{\square}$ can be formulated as

$$\forall_{t\in[0,T]} \epsilon(\mathbf{x}_t^{\Diamond})^{\Diamond} = \epsilon(\mathbf{x}_t^{\square})^{\square}. \tag{30}$$

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

$$\epsilon(\mathbf{x}) = \epsilon(\mathbf{x}, c_\emptyset) + \lambda_4(\epsilon(\mathbf{x}, c_{j,f}) - \epsilon(\mathbf{x}, c_\emptyset)) \tag{31}$$

Combined with Eq. (31), Eq. (30) could be satisfied when

$$\begin{aligned} &\forall_{t\in[0,T]} \epsilon(\mathbf{x}_t^{\Diamond}, c_\emptyset)^{\Diamond} = \epsilon(\mathbf{x}_t^{\square}, c_\emptyset)^{\square}, \\ &\forall_{t\in[0,T]} \epsilon(\mathbf{x}_t^{\Diamond}, c_{j,f})^{\Diamond} = \epsilon(\mathbf{x}_t^{\square}, c_{j,f})^{\square}. \end{aligned} \tag{32}$$

According to Eq. (32), Eq. (2) can be resolved when

$$\begin{aligned} &\epsilon(\mathbf{x}_t, c_\emptyset; \theta_{dm} + \sum_{k\in S_{sub}^*} \Delta\theta_{k,dm}) = \epsilon(\mathbf{x}_t, c_\emptyset; \sum_{k\in S_{sub}} \Delta\theta_{k,dm}), \\ &\epsilon(\mathbf{x}_t, c_\emptyset; \theta_{dm} + \sum_{k\in S_{sub}} \Delta\theta_{k,dm}) = \epsilon(\mathbf{x}_t, c_\emptyset; \theta_{dm}), \\ &\epsilon(\mathbf{x}_t, c_{i,f}; \theta_{dm} + \sum_{k\in S_{sub}^*} \Delta\theta_{k,dm}) = \epsilon(\mathbf{x}_t, c_{i,f}; \sum_{k\in S_{sub}} \Delta\theta_{k,dm}), \\ &\epsilon(\mathbf{x}_t, c_{j,f}; \theta_{dm} + \sum_{k\in S_{sub}^*} \Delta\theta_{k,dm}) = \epsilon(\mathbf{x}_t, c_{j,f}; \theta_{dm}), \end{aligned} \tag{33}$$

where S_{sub} denotes the arbitrary subset of $[1, \mathrm{N}]$, $c_{j,f}$ denotes the recovered concept, S_{sub}^* represents S_{sub} that removes the index j, $i \in S_{sub}, j \in S_{sub}, j \notin S_{sub}^*, i \neq j$. From Eq. (33), the first two terms can be summarized as:

$$\forall_{S_{sub}\in[1,\mathrm{N}]} \epsilon(\mathbf{x}_t, c_\emptyset; \theta_{dm} + \sum_{k\in S_{sub}} \Delta\theta_{k,dm}) = \epsilon(\mathbf{x}_t, c_\emptyset; \theta_{dm}). \tag{34}$$

Namely, Eq. (2) is realized when the following conditions satisfied for $\forall S_{sub}$,

$$\begin{aligned} &\epsilon(\mathbf{x}_t, c_\emptyset; \theta_{dm} + \sum_{k\in S_{sub}} \Delta\theta_{k,dm}) = \epsilon(\mathbf{x}_t, c_\emptyset; \theta_{dm}), \\ &\epsilon(\mathbf{x}_t, c_{i,f}; \theta_{dm} + \sum_{k\in S_{sub}^*} \Delta\theta_{k,dm}) = \epsilon(\mathbf{x}_t, c_{i,f}; \sum_{k\in S_{sub}} \Delta\theta_{k,dm}), \\ &\epsilon(\mathbf{x}_t, c_{j,f}; \theta_{dm} + \sum_{k\in S_{sub}^*} \Delta\theta_{k,dm}) = \epsilon(\mathbf{x}_t, c_{j,f}; \theta_{dm}), \end{aligned} \tag{35}$$

where $i \neq j$.

A.2 Proof for Eq. (8)

For convenience, we omit the positions where $\Delta\theta_{k,dm} = 0$ and formulate Eq. (5) (also described as Eq. (35)) as

$$\begin{aligned} &\epsilon(\mathbf{x}_t, c_\emptyset; \mathbf{w}_{II} + \sum_{k\in S_{sub}} \Delta\mathbf{w}_{k,II}) = \epsilon(\mathbf{x}_t, c_\emptyset; \mathbf{w}_{II}), \\ &\epsilon(\mathbf{x}_t, c_{i,f}; \mathbf{w}_{II} + \sum_{k\in S_{sub}^*} \Delta\mathbf{w}_{k,II}) = \epsilon(\mathbf{x}_t, c_{i,f}; \sum_{k\in S_{sub}} \Delta\mathbf{w}_{k,II}), \\ &\epsilon(\mathbf{x}_t, c_{j,f}; \mathbf{w}_{II} + \sum_{k\in S_{sub}^*} \Delta\mathbf{w}_{k,II}) = \epsilon(\mathbf{x}_t, c_{j,f}; \mathbf{w}_{II}), \end{aligned} \tag{36}$$

where $\mathbf{w}_{II}$ indicates image-independent embedding update weights within θ_{dm}, $\mathbf{w}_{II} \in \theta_{dm}$ and $\Delta\mathbf{w}_{j,II} \in \Delta\theta_{j,dm}$. Eq. (36) means that $c_\emptyset$ and $\forall_{k\in S_{sub}} \Delta\mathbf{w}_{k,II}$, $c_{j,f}$ and $\forall_{k\in S_{sub}^*} \Delta\mathbf{w}_{k,II}$ should be uncorrelated.

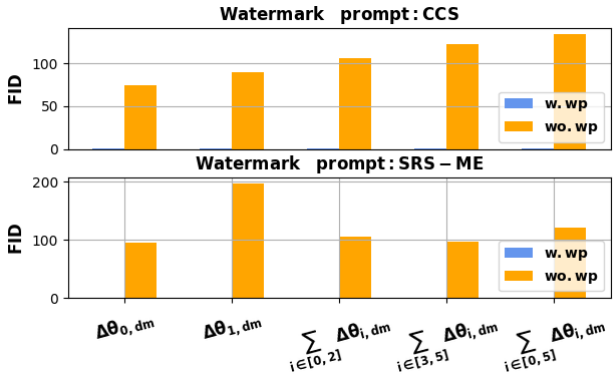

Figure 23: Watermark preservation (wp).

According to Eq. (7), Eq. (36) is expanded as

$$\begin{aligned}
&\mathbf{e}(c_\emptyset) \otimes [\mathbf{w}_{II} + \sum_{k\in S_{sub}} \Delta\mathbf{w}_{k,II}] = \mathbf{e}(c_\emptyset) \otimes \mathbf{w}_{II},\\
&\mathbf{e}(c_{i,f}) \otimes [\mathbf{w}_{II} + \sum_{k\in S^*_{sub}} \Delta\mathbf{w}_{k,II}] = \mathbf{e}(c_{i,f}) \otimes [\mathbf{w}_{II} + \sum_{k\in S_{sub}} \Delta\mathbf{w}_{k,II}],\\
&\mathbf{e}(c_{j,f}) \otimes [\mathbf{w}_{II} + \sum_{k\in S^*_{sub}} \Delta\mathbf{w}_{k,II}] = \mathbf{e}(c_{j,f}) \otimes \mathbf{w}_{II}.
\end{aligned} \tag{37}$$

For all forgotten concepts $c_{j,f} \in c_f$, Eq. (37) is further simplified as

$$\begin{aligned}
&\forall_{k\in[1,\mathrm{N}]} \mathbf{e}(c_\emptyset) \otimes \Delta\mathbf{w}_{k,II} = 0,\\
&\forall_{j,k\in[1,\mathrm{N}],j\neq k} \mathbf{e}(c_{j,f}) \otimes \Delta\mathbf{w}_{k,II} = 0.
\end{aligned} \tag{38}$$

Namely, for each $\Delta\mathbf{w}_{k,II}$ in Eq. (38), it should satisfy the condition

$$\begin{aligned}
&\mathbf{e}(c_\emptyset) \otimes \Delta\mathbf{w}_{k,II} = 0,\\
&\forall_{j\in[1,\mathrm{N}],j\neq k} \mathbf{e}(c_{j,f}) \otimes \Delta\mathbf{w}_{k,II} = 0.
\end{aligned} \tag{39}$$

For clarity, Eq. (39) is expressed as

$$\mathbf{e}_m \otimes \Delta\mathbf{w}_{k,II} = 0, \tag{40}$$

where the matrix $\mathbf{e}_m \in \mathbb{R}^{(\mathrm{N}\cdot d_{emb})\times d_{in}}$ represents

$$[\mathbf{e}(c_\emptyset)^\top; \mathbf{e}(c_{1,f})^\top; \cdots; \mathbf{e}(c_{k-1,f})^\top; \mathbf{e}(c_{k+1,f})^\top; \cdots; \mathbf{e}(c_{\mathrm{N},f})^\top]^\top. \tag{41}$$

A.3 Proof for Eq. (12)

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

$$\mathcal{DM}(\mathbf{z}, c_f, \theta_{dm}) - \mathcal{DM}(\mathbf{z}, c_\emptyset, \theta_{dm}). \tag{42}$$

For convenience, we utilize $\mathbf{x}_0^\Diamond$ and $\mathbf{x}_0^\square$ to represent latent representations of DMs for prompts c_f and $c_\emptyset$ respectively. According to Eqs. (26) and (31), we have

$$\begin{aligned}
\mathbf{x}_0^\Diamond &= \lambda_1\mathbf{x}_1 - \lambda_2(\epsilon(\mathbf{x}_1, c_\emptyset) + \lambda_4(\epsilon(\mathbf{x}_1, c_f) - \epsilon(\mathbf{x}_1, c_\emptyset))) + \lambda_3\mathbf{z},\\
\mathbf{x}_0^\square &= \lambda_1\mathbf{x}_1 - \lambda_2(\epsilon(\mathbf{x}_1, c_\emptyset) + \lambda_4(\epsilon(\mathbf{x}_1, c_\emptyset) - \epsilon(\mathbf{x}_1, c_\emptyset))) + \lambda_3\mathbf{z}.
\end{aligned} \tag{43}$$

The index '0' denotes the final schedule time. Eq. (42) is further expressed as

$$\mathbf{x}_0^\Diamond - \mathbf{x}_0^\square = \lambda_2\lambda_4(\epsilon(\mathbf{x}_1, c_f) - \epsilon(\mathbf{x}_1, c_\emptyset)). \tag{44}$$

Similarly, Eq. (44) can be generalized to arbitrary timestamp,

$$\begin{aligned}
\mathbf{x}_{t-1}^\Diamond - \mathbf{x}_{t-1}^\square &= \lambda_2\lambda_4(\epsilon(\mathbf{x}_t, c_f) - \epsilon(\mathbf{x}_t, c_\emptyset)),\\
&\propto \epsilon(\mathbf{x}_t, c_f) - \epsilon(\mathbf{x}_t, c_\emptyset).
\end{aligned} \tag{45}$$

A.4 Proofs for the answer A7

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

$$\mathcal{L}^t_{cor}(c_{i,f}, \sum_{i,k\in S_{sub},i\neq k} \Delta\theta_{k,dm}) = \mathcal{L}^t_{cor}(c_{i,f}, \mathbf{0}). \tag{46}$$

This occurs because the unlearning loss is unrelated to $\forall_{i,k\in S_{sub},i\neq k}$ $\Delta\theta_{k,dm}$, resulting in no gradient backward propagation. $\mathbf{0}$ indicates the zero matrices with the same shape as θ_{dm}.

We omit the positions where $\Delta\theta_{k,dm} = 0$ and have

$$\begin{aligned}
&\mathcal{L}^t_{cor}(c_f, \sum_k \Delta\mathbf{w}_{k,II})\\
&= \frac{1}{h\cdot w\cdot d}\sum \mathcal{R}_{c_f,t}(\mathbf{w}_{II} + \sum_k \Delta\mathbf{w}_{k,II}) \odot \mathcal{R}_{c_f,t}(\mathbf{w}_{II}).
\end{aligned} \tag{47}$$

Combined with (47), we expand terms in Eq. (46) as

$$\begin{aligned}
\mathcal{L}^t_{cor}(c_{i,f}, \Delta_w) &= \frac{1}{h\cdot w\cdot d}\sum \mathcal{R}_{c_{i,f},t}(\mathbf{w}_{II} + \Delta_w) \odot \mathcal{R}_{c_{i,f},t}(\mathbf{w}_{II}),\\
\mathcal{L}^t_{cor}(c_{i,f}, \mathbf{0}) &= \frac{1}{h\cdot w\cdot d}\sum \mathcal{R}_{c_{i,f},t}(\mathbf{w}_{II}) \odot \mathcal{R}_{c_{i,f},t}(\mathbf{w}_{II}),
\end{aligned} \tag{48}$$

where $\Delta_w = \sum_{i,k\in S_{sub},i\neq k} \Delta\mathbf{w}_{k,II}$. Then, we have

$$\begin{aligned}
&\mathcal{L}^t_{cor}(c_{i,f}, \Delta_w) - \mathcal{L}^t_{cor}(c_{i,f}, \mathbf{0})\\
&\propto \mathcal{R}_{c_{i,f},t}(\mathbf{w}_{II} + \Delta_w) \odot \mathcal{R}_{c_{i,f},t}(\mathbf{w}_{II}) - \mathcal{R}_{c_{i,f},t}(\mathbf{w}_{II}) \odot \mathcal{R}_{c_{i,f},t}(\mathbf{w}_{II})\\
&= \left(\mathcal{R}_{c_{i,f},t}(\mathbf{w}_{II} + \Delta_w) - \mathcal{R}_{c_{i,f},t}(\mathbf{w}_{II})\right) \odot \mathcal{R}_{c_{i,f},t}(\mathbf{w}_{II})\\
&\propto \mathcal{R}_{c_{i,f},t}(\mathbf{w}_{II} + \Delta_w) - \mathcal{R}_{c_{i,f},t}(\mathbf{w}_{II})
\end{aligned} \tag{49}$$

According to Eq. (12), the right term in Eq. (49) is expressed as

$$\begin{aligned}
&\mathcal{R}_{c_{i,f},t}(\mathbf{w}_{II} + \Delta_w) - \mathcal{R}_{c_{i,f},t}(\mathbf{w}_{II})\\
&= \Big(\epsilon(\mathbf{x}_t, c_{i,f}; \mathbf{w}_{II} + \Delta_w) - \epsilon(\mathbf{x}_t, c_\emptyset; \mathbf{w}_{II} + \Delta_w)\Big) -\\
&\qquad \Big(\epsilon(\mathbf{x}_t, c_{i,f}; \mathbf{w}_{II}) - \epsilon(\mathbf{x}_t, c_\emptyset; \mathbf{w}_{II})\Big),\\
&= \Big(\epsilon(\mathbf{x}_t, c_{i,f}; \mathbf{w}_{II} + \Delta_w) - \epsilon(\mathbf{x}_t, c_{i,f}; \mathbf{w}_{II})\Big) -\\
&\qquad \Big(\epsilon(\mathbf{x}_t, c_\emptyset; \mathbf{w}_{II} + \Delta_w) - \epsilon(\mathbf{x}_t, c_\emptyset; \mathbf{w}_{II})\Big),
\end{aligned} \tag{50}$$

Since $\epsilon(\mathbf{x}_t, c; \theta_{dm} + \Delta\theta_{k,dm}) = \epsilon(\mathbf{x}_t, c; \theta_{dm})$ when $\mathbf{e}(c) \otimes \Delta\mathbf{w}_{k,II}$ $= 0$, and the conditions $\mathbf{e}(c_\emptyset) \otimes \Delta_w = 0$, $\mathbf{e}(c_{i,f}) \otimes \Delta_w = 0$, we have

$$\begin{aligned}
&\epsilon(\mathbf{x}_t, c_{i,f}; \mathbf{w}_{II} + \Delta_w) - \epsilon(\mathbf{x}_t, c_{i,f}; \mathbf{w}_{II}) = 0,\\
&\epsilon(\mathbf{x}_t, c_\emptyset; \mathbf{w}_{II} + \Delta_w) - \epsilon(\mathbf{x}_t, c_\emptyset; \mathbf{w}_{II}) = 0.
\end{aligned} \tag{51}$$

Namely, for each $c_{i,f} \in c_f$, SRS-ME fulfills Eq. (46),

$$\begin{aligned}
&\mathcal{L}^t_{cor}(c_{i,f}, \Delta_w) - \mathcal{L}^t_{cor}(c_{i,f}, \mathbf{0})\\
&\propto \mathcal{R}_{c_{i,f},t}(\mathbf{w}_{II} + \Delta_w) - \mathcal{R}_{c_{i,f},t}(\mathbf{w}_{II}) = 0.
\end{aligned} \tag{52}$$

This makes the training process of SRS-ME separable.

A.5 Others

The effect of watermark preservation is illustrated in Figure 23. We incorporate the watermark prompt into $\mathbf{e}_m$ when unlearning undesirable concepts. It can be observed that the weight decoupling can effectively preserve model watermarks.

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.

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009

Table 5: Comparative results (ACC/LPIPS/FID) on style unlearning. These objects are sequentially denoted as $c_{0\sim8}$.

Scene	t	0:Cezanne	1:VanGogh	2:Picasso	3:JacksonPo...	4:Caravaggio	5:KeithHaring	6:KellyMcK...	7:TylerEdlin	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										
Scene$_1$	t_0	208.8/12.4/.338	219.9/36.8/.388	199.8/6.40/.336	274.9/26.4/.329	202.7/38.0/.312	216.5/15.2/.596	56.0/95.2/.153	134.6/98.2/.125	110.0/99.2/.126
	t_1	0.75/98.0/.00	2.12/90.8/.00	199.3/6.00/.336	276.7/26.8/.328	203.9/38.3/.313	209.0/25.2/.526	47.1/98.8/.096	119.3/100/.102	101.1/99.2/.107
	t_2	0.78/98.4/.00	3.28/90.4/.00	1.25/98.8/.00	276.5/26.8/.329	203.1/38.0/.312	208.7/25.6/.526	44.7/99.6/.097	99.9/100/.080	85.8/100/.066
Scene$_4$	t_0	188.3/36.0/.256	218.6/37.2/.395	214.2/5.20/.365	183.5/60.4/.245	96.7/98.8/.182	52.5/97.2/.189	65.1/95.6/.160	50.8/100./.016	108.3/99.2/.147
	t_1	0.77/98.0/.00	218.6/37.2/.395	214.6/5.20/.364	175.6/64.0/.211	96.0/99.6/.166	45.4/96.8/.147	49.7/98.4/.096	49.9/100/.016	112.6/100/.156
	t_2	0.81/98.0/.00	218.6/37.6/.395	215.0/5.20/.365	365.4/5.60/.463	246.1/19.6/.379	135.4/59.6/.402	77.4/86.0/.176	61.0/100/.024	112.6/99.6/.151
	t_3	0.76/98.0/.00	3.82/90.4/.00	214.9/5.20/.364	354.6/2.80/.458	219.4/28.0/.340	131.5/58.2/.395	73.3/90.0/.166	52.0/100/.016	100.5/98.4/.123
	t_4	0.87/98.0/.00	2.82/90.4/.00	214.4/5.20/.365	354.4/2.80/.458	219.2/28.4/.340	288.4/11.6/.623	82.1/85.6/.185	77.0/100/.045	103.1/98.4/.123
	t_5	0.921/98.0/.00	4.11/90.4/.00	1.06/98.8/.00	274.3/26.4/.343	230.9/28.0/.368	277.5/17.6/.615	68.4/90.8/.154	78.3/100/.046	91.8/99.6/.098
Ablation[75,75,50,50,50,50]										
Scene$_1$	t_0	222.7/19.2/.382	246.3/9.60/.422	238.1/0.80/.337	300.4/14.8/.343	235.8/10.4/.319	229.7/12.4/.589	82.7/76.8/.194	175.5/75.6/.214	156.9/90.0/.214
	t_1	145.2/40.8/.255	219.3/33.2/.401	165.2/6.40/.252	298.2/12.8/.366	197.3/26.0/.270	227.2/19.2/.576	70.7/92.0/.158	123.8/98.8/.118	127.4/98.4/.166
	t_2	106.3/76.4/.186	182.9/50.8/.363	83.6/77.2/.188	290.6/18.0/.345	178.6/51.2/.269	200.3/29.6/.510	57.2/95.6/.123	109.2/100/.093	109.3/99.6/.128
Scene$_4$	t_0	210.5/20.0/.371	234.9/20.8/.413	198.6/0.80/.288	235.6/56.0/.266	157.7/67.2/.235	129.4/64.4/.427	64.3/96.0/.158	147.9/93.6/.151	129.5/98.4/.168
	t_1	140.5/55.2/.202	238.8/20.8/.415	173.9/6.40/.254	164.8/87.6/.178	111.5/96.0/.165	97.1/80.0/.368	57.8/96.4/.130	108.7/100/.089	103.1/98.8/.120
	t_2	175.4/39.2/.272	239.8/16.8/.415	207.7/1.60/.287	279.2/21.2/.331	213.0/23.2/.287	124.2/60.8/.409	61.7/93.6/.154	129.7/98.4/.132	119.1/98.0/.146
	t_3	129.9/82.0/.170	162.9/83.2/.296	181.7/5.60/.268	287.1/18.80/.338	194.9/28.0/.259	87.6/81.6/.313	56.7/96.0/.124	99.7/99.6/0.088	100.5/99.2/.111
	t_4	145.2/40.8/.255	219.3/33.2/.401	165.2/6.40/.252	298.2/12.80/.366	197.3/26.0/.270	227.2/19.20/.576	70.7/92.0/.158	123.8/98.8/.118	127.4/98.4/.166
	t_5	106.3/76.4/.186	182.9/50.8/.363	83.6/77.2/.188	290.6/18.00/.345	178.6/51.2/.269	200.3/29.6/.510	57.2/95.6/.123	109.2/100/.093	109.3/99.6/.128
FMN30										
Scene$_1$	t_0	344.1/0.40/.416	314.6/29.6/.517	272.3/16.4/.428	404.3/13.6/.520	340.8/15.6/.453	253.2/35.2/.558	277.9/20.2/.439	247.1/23.6/.434	209.2/76.8/.302
	t_1	247.9/10.4/.410	281.7/31.6/.469	186.8/30.0/.357	414.3/20.0/.417	241.9/35.2/.396	244.3/24.4/.564	101.9/63.2/.239	165.6/88.4/.211	151.1/93.2/.228
	t_2	201.2/19.6/.357	264.8/20.8/.447	135.6/58.8/.318	368.1/18.4/.430	190.2/54.0/.333	225.1/23.2/.544	77.7/82.4/.187	122.2/98.0/.141	125.0/97.6/.189
Scene$_4$	t_0	276.9/2.00/.369	280.9/32.4/.415	158.6/44.4/.358	323.1/27.6/.476	187.3/80.0/.338	174.8/58.5/.534	112.1/64.4/.268	195.6/56.0/.265	174.2/88.4/.268
	t_1	167.6/31.6/.323	289.6/19.2/.440	154.7/56.0/.357	281.1/32.0/.398	128.6/92.4/.280	95.6/88.0/.382	68.1/89.6/.177	105.5/99.2/.120	136.6/97.2/.198
	t_2	250.7/4.40/.391	279.4/44.0/.444	167.4/43.2/.347	391.7/17.2/.499	229.0/41.2/.385	174.3/68.4/.521	142.8/57.2/.310	207.3/.464/.281	174.9/87.6/.260
	t_3	204.2/11.6/.372	264.2/33.2/.424	168.4/42.4/.329	378.8/18.4/.457	193.5/53.6/.349	123.4/74.8/.435	75.0/82.4/.186	121.8/98.0/.139	138.4/98.0/.197
	t_4	247.9/10.4/.410	281.7/31.6/.469	186.8/30.0/.357	414.3/20.0/.417	241.9/35.2/.396	244.3/24.4/.564	101.9/63.2/.239	165.6/88.4/.211	151.1/93.2/.228
	t_5	201.2/19.6/.357	264.8/20.8/.447	135.6/58.8/.318	368.1/18.4/.430	190.2/54.0/.333	225.1/23.2/.544	77.7/82.4/.187	122.2/98.0/.141	125.0/97.6/.189
FMN20										
Scene$_1$	t_0	212.6/0.80/.382	273.8/34.0/.440	156.1/46.0/.339	371.9/25.6/.506	208.4/53.2/.349	220.6/27.2/.575	102.5/67.2/.257	195.5/67.6/.256	162.6/92.8/.246
	t_1	142.0/52.8/.265	247.8/32.0/.411	152.8/48.8/.315	328.2/26.8/.450	156.8/63.2/.292	173.1/39.2/.498	54.5/91.6/.141	96.4/100/.102	104.2/99.6/.139
	t_2	103.5/89.2/.188	174.4/53.6/.332	93.7/80.4/.249	306.0/26.4/.434	135.5/71.2/.255	141.8/54.0/.462	48.4/98.0/.111	85.4/100/.071	91.7/99.2/.097
Scene$_4$	t_0	145.2/30.8/.310	262.6/18.8/.430	113.3/71.2/.279	268.3/45.6/.395	112.6/97.6/.236	81.1/86.0/.332	55.5/94.8/.147	98.8/100/.108	123.9/98.8/.171
	t_1	95.2/98.8/.177	263.5/20.8/.423	89.6/78.4/.240	204.7/71.6/.294	84.0/99.2/.173	64.5/96.8/.268	45.7/98.4/.111	82.3/100./.065	97.1/100/.115
	t_2	155.4/32.8/.306	277.6/17.6/.428	151.8/44.4/.316	341.3/28.8/.449	163.9/69.2/.294	114.8/86.4/.418	74.6/87.6/.194	121.8/98.0/.152	134.6/97.6/.207
	t_3	116.9/83.6/.202	187.1/54.4/.342	110.4/69.2/.258	304.6/28.8/.429	142.0/70.0/.268	80.4/86.8/.338	47.1/98.4/.112	87.7/100/.077	98.0/99.6/.109
	t_4	142.0/52.8/.265	247.8/32.0/.411	152.8/48.8/.315	328.2/26.8/.450	156.8/63.2/.292	173.1/39.2/.498	54.5/91.6/.141	96.4/100/.102	104.2/99.6/.139
	t_5	103.5/89.2/.188	174.4/53.6/.332	93.7/80.4/.249	306.0/26.4/.434	135.5/71.2/.255	141.8/54.0/.462	48.4/98.0/.111	85.4/100/.071	91.7/99.2/.097
ESD200										
Scene$_1$	t_0	306.2/1.60/.465	303.4/6.80/.526	298.4/2.40/.498	398.2/0.00/.573	280.4/8.00/.482	312.9/2.00/.689	234.4/24.4/.347	248.7/8.00/.351	212.4/36.4/.366
	t_1	304.5/2.00/.442	278.3/10.0/.469	283.2/2.40/.487	386.3/0.80/.559	280.8/8.80/.474	301.2/2.40/.666	184.4/38.4/.310	256.4/8.80/.321	197.9/43.2/.343
	t_2	277.6/4.40/.430	242.8/27.2/.443	253.7/4.80/.462	365.6/1.20/.557	257.8/14.0/.454	290.3/4.40/.659	122.2/62.8/.260	210.7/34.8/.268	168.3/76.4/.265
Scene$_4$	t_0	320.4/0.00/.467	283.2/7.20/.496	262.0/4.40/.475	289.7/17.2/.386	233.8/30.8/.425	247.1/22.0/.603	107.7/68.0/.253	178.7/58.4/.233	166.5/84.4/.251
	t_1	160.0/50.4/.294	251.4/22.8/.441	214.6/9.20/.408	212.4/57.6/.265	125.5/90.8/.218	173.8/51.6/.503	65.9/89.6/.168	115.6/98.0/.123	134.3/96.4/.174
	t_2	298.5/0.00/.459	296.7/5.60/.516	277.3/3.20/.505	377.8/0.00/.573	272.5/12.4/.491	278.6/6.00/.649	164.5/46.8/.297	244.1/12.4/.315	187.5/58.4/.323
	t_3	284.6/4.00/.431	240.4/24.8/.442	264.5/4.40/.485	362.4/0.00/.538	261.6/14.4/.464	258.2/10.0/.620	97.7/71.2/.238	193.3/45.2/.249	173.5/84.8/.249
	t_4	304.5/2.00/.442	278.3/10.0/.469	283.2/2.40/.487	386.3/0.80/.559	280.8/8.80/.474	301.2/2.40/.666	184.4/38.4/.310	256.4/8.80/.321	197.9/43.2/.343
	t_5	277.6/4.40/.430	242.8/27.2/.443	253.7/4.80/.462	365.6/1.20/.557	257.8/14.0/.454	290.3/4.40/.659	122.2/62.8/.260	210.7/34.8/.268	168.3/76.4/.265
ESD50										
Scene$_1$	t_0	300.9/1.20/.462	299.9/6.40/.492	255.8/5.20/.468	349.9/6.40/.480	253.5/28.8/.448	283.8/10.4/.628	142.4/49.6/.277	225.3/23.2/.284	178.6/78.0/.266
	t_1	226.8/10.8/.384	239.5/36.4/.418	218.4/8.00/.416	307.1/15.2/.416	193.6/50.4/.343	261.9/13.6/.610	79.3/84.4/.205	143.9/83.2/.176	147.7/94.0/.195
	t_2	125.3/53.6/.233	165.0/69.6/.345	152.5/39.6/.321	282.2/26.0/.378	148.3/63.6/.283	236.4/27.2/.559	63.2/89.6/.168	119.6/96.8/.122	124.2/97.2/.160
Scene$_4$	t_0	168.8/24.0/.308	234.3/36.8/.420	164.9/28.0/.339	162.4/75.2/.207	120.7/96.0/.210	112.6/79.6/.395	56.6/94.4/.146	110.1/98.8/.106	115.2/98.0/.141
	t_1	82.4/88.4/.137	177.4/66.0/.357	135.4/45.2/.288	106.8/84.4/.125	89.4/99.6/.126	78.0/87.6/.295	46.6/98.8/.097	86.5/100/.057	95.7/98.0/.094
	t_2	207.0/14.0/.367	252.1/25.2/.434	194.0/17.6/.395	286.2/22.0/.379	182.9/58.8/.342	186.7/47.6/.500	68.5/86.8/.185	131.0/89.6/.150	138.9/94.4/.180
	t_3	138.2/53.2/.268	147.8/70.8/.323	175.7/21.2/.366	263.8/27.2/.351	154.6/63.2/.289	143.6/62.4/.455	59.4/94.0/.152	112.5/98.8/.110	118.9/97.2/.144
	t_4	226.8/10.8/.384	239.5/36.4/.418	218.4/8.00/.416	307.1/15.2/.416	193.6/50.4/.343	261.9/13.6/.610	79.3/84.4/.205	143.9/83.2/.176	147.7/94.0/.195
	t_5	125.3/53.6/.233	165.0/69.6/.345	152.5/39.6/.321	282.2/26.0/.378	148.3/63.6/.283	236.4/27.2/.559	63.2/89.6/.168	119.6/96.8/.122	124.2/97.2/.160

Table 6: Comparative results (FID/ACC/LPIPS/) on object unlearning. These objects are sequentially denoted as $c_{0\sim8}$. [v_0, v_1,v_2,v_3,v_4,v_5] denotes the number of unlearning iterations for erasing concepts $c_{0\sim5}$.

Scene	t	0:ChainSaw	1:Church	2:GasPump	3:Tench	4:GarbageT...	5:E.Springer	6:GolfBall	7:Parachute	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										
***Scene*$_1$**	t_0	331.3/1.2/.331	215.6/48.8/.386	260.7/2.0/.443	166.5/12.4/.358	316.5/2.8/.479	325.9/0.4/.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	262.2/1.6/.443	167.0/12.8/.358	317.0/2.8/.479	325.6/0.4/.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	166.3/12.4/.358	315.0/2.8/.479	325.3/0.4/.393	20.0/98.8/.255	27.3/92.8/.228	14.6/97.6/.219
***Scene*$_4$**	t_0	244.3/22.4/.276	169.0/48.0/.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	1.118/91.2/.000	168.8/48.0/.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	0.586/91.2/.000	168.9/47.2/.362	89.9/40.0/.291	216.8/7.20/.502	237.3/8.00/.431	50.6/83.2/.357	35.1/89.2/.267	30.6/89.2/.213	27.4/99.6/.444
	t_3	0.663/92.4/.000	0.231/79.6/.000	90.0/40.4/.291	209.0/5.60/.414	206.2/11.2/.371	77.0/69.2/.428	35.5/88.4/.242	44.9/83.2/.245	19.2/100/.333
	t_4	0.621/90.8/.000	0.220/79.6/.000	90.1/40.0/.291	209.0/6.40/.414	205.8/11.2/.371	303.9/2.80/.431	33.2/91.6/.265	31.0/88.4/.210	14.7/100/.290
	t_5	0.625/91.2/.000	15.2/81.2/.088	0.571/60.0/.000	227.7/9.20/.412	188.6/30.4/.340	312.2/0.00/.371	10.8/96.8/.058	19.6/92.0/.101	13.3/99.2/.179
Ablation[75,75,50,50,75,75]										
***Scene*$_1$**	t_0	363.7/2.00/.371	265.9/2.80/.409	280.2/3.20/.391	328.9/0.40/.476	297.9/9.20/.418	367.2/0.80/.418	222.7/28.4/.436	280.2/7.20/.467	302.3/14.4/.415
	t_1	189.9/29.2/.281	72.4/63.2/.300	120.3/30.8/.315	292.9/0.80/.462	187.2/30.4/.341	290.6/2.40/.357	72.2/76.4/.306	115.2/48.4/.348	66.6/80.4/.263
	t_2	158.5/41.2/.262	63.4/67.6/.282	38.8/42.8/.234	275.3/1.60/.457	166.5/29.2/.334	259.9/2.80/.348	59.1/80.0/.282	79.8/61.2/.323	20.6/93.2/.223
***Scene*$_4$**	t_0	233.5/13.6/.300	180.0/25.6/.366	86.3/30.0/.261	103.3/39.2/.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	149.9/44.0/.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
	t_2	176.8/35.6/.265	203.2/18.4/.377	118.3/22.0/.296	263.8/2.00/.454	188.8/22.4/.354	50.8/66.4/.213	61.5/77.6/.310	134.5/40.0/.359	96.5/71.6/.280
	t_3	124.1/53.6/.212	45.2/74.0/.234	87.8/37.6/.285	237.9/5.20/.422	164.4/37.6/.332	33.2/80.0/.157	34.5/89.6/.217	51.4/73.2/.243	40.3/90.8/.216
	t_4	189.9/29.2/.281	72.4/63.2/.300	120.3/30.8/.315	292.9/0.80/.462	187.2/30.4/.341	290.6/2.40/.357	72.2/76.4/.306	115.2/48.4/.348	66.6/80.4/.263
	t_5	158.5/41.2/.262	63.4/67.6/.282	38.8/42.8/.234	275.3/1.60/.457	166.5/29.2/.334	259.9/2.80/.348	59.1/80.0/.282	79.8/61.2/.323	20.6/93.2/.223
FMN[50,50,50,30,50,50]										
***Scene*$_1$**	t_0	350.3/7.60/.349	375.4/0.00/.601	254.1/0.00/.433	341.2/0.00/.483	328.9/2.40/.475	349.6/0.40/.476	259.3/38.0/.526	318.1/1.60/.470	429.2/0.40/.470
	t_1	194.9/22.8/.335	344.7/2.00/.473	174.6/5.60/.436	309.5/0.40/.462	282.8/6.80/.414	288.2/1.20/.358	92.7/76.0.393	176.2/31.6/.437	143.5/57.2/.309
	t_2	141.2/41.2/.293	187.2/42.0/.393	48.0/35.6/.292	265.9/2.40/.466	186.7/26.0/.357	205.5/12.8/.338	58.5/84.4/.386	88.1/65.2/.369	30.8/97.6/.299
***Scene*$_4$**	t_0	180.0/39.6/.273	335.8/4.00/.488	99.0/3.60/.363	126.5/45.6/.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	233.6/24.4/.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
	t_2	206.5/18.8/.331	377.1/1.20/.486	179.4/4.40/.444	308.8/0.40/.450	266.4/4.80/.403	135.7/33.2/.312	91.0/76.0/.371	219.4/16.4/.474	189.2/36.4/.321
	t_3	155.4/38.0/.319	298.0/11.2/.442	178.0/1.60/.444	303.6/1.60/.463	225.7/8.00/.383	75.3/48.8/.276	75.6/77.6/.415	156.6/36.4/.429	94.8/74.8/.300
	t_4	194.9/22.8/.335	344.7/2.00/.473	174.6/5.60/.436	309.5/0.40/.462	282.8/6.80/.414	288.2/1.20/.358	92.7/76.0.393	176.2/31.6/.437	143.5/57.2/.309
	t_5	141.2/41.2/.293	187.2/42.0/.393	48.0/35.6/.292	265.9/2.40/.466	186.7/26.0/.357	205.5/12.8/.338	58.5/84.4/.386	88.1/65.2/.369	30.8/97.6/.299
FMN[50,30,50,20,50,50]										
***Scene*$_1$**	t_0	303.7/14.4/.338	367.8/0.80/.538	205.2/2.00/.426	328.4/0.00/.477	316.1/2.00/.458	336.2/0.40/.435	164.0/63.6/.426	270.6/9.20/.487	358.0/4.40/.415
	t_1	177.7/28.8/.323	330.0/2.80/.274	166.3/4.00/.432	289.7/0.80/.464	262.4/8.00/.412	268.2/4.40/.350	86.5/74.4/.412	148.8/40.4/.421	106.1/72.8/.308
	t_2	122.1/50.4/.282	159.3/48.4/.386	48.5/38.4/.284	201.5/20.8/.433	173.2/29.6/.358	178.1/19.6/.332	54.8/86.4/.349	68.8/69.6/.349	21.0/100/.298
***Scene*$_4$**	t_0	162.8/46.8/.255	200.2/33.2/.409	79.2/7.60/.345	109.6/53.6/.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
	t_2	162.1/28.4/.315	323.4/4.40/.467	168.0/2.80/.436	243.2/12.0/.443	223.9/8.80/.389	80.3/46.8/.284	67.1/81.2/.366	150.4/38.0/.426	87.6/76.0/.285
	t_3	149.9/36.0/.311	262.0/18.0/.437	162.4/3.20/.431	254.9/6.40/.451	213.0/9.60/.381	59.3/54.8/.263	73.0/78.0/.424	125.3/50.0/.403	65.7/86.4/.300
	t_4	177.7/28.8/.323	330.0/2.80/.274	166.3/4.00/.432	289.7/0.80/.464	262.4/8.00/.412	268.2/4.40/.350	86.5/74.4/.412	148.8/40.4/.421	106.1/72.8/.308
	t_5	122.1/50.4/.282	159.3/48.4/.386	48.5/38.4/.284	201.5/20.8/.433	173.2/29.6/.358	178.1/19.6/.332	54.8/86.4/.349	68.8/69.6/.349	21.0/100/.298
ESD[100,100,50,75,100,100]										
***Scene*$_1$**	t_0	330.0/0.00/.429	314.4/0.40/.440	275.7/0.40/.444	292.5/0.00/.490	291.1/0.00/.411	295.9/0.00/.463	203.1/31.6/.475	288.0/1.60/.501	332.7/2.40/.501
	t_1	189.9/24.0/.320	143.4/32.8/.361	129.8/11.6/.336	275.9/0.00/.500	187.6/10.0/.346	234.8/8.40/.368	72.0/73.6/.368	197.5/18.0/.432	114.4/58.4/.341
	t_2	124.6/45.2/.257	84.9/64.8/.317	36.9/30.8/.235	274.3/1.20/.494	107.1/22.8/.285	179.0/13.2/.326	45.1/85.2/.319	126.4/40.8/.358	42.7/83.6/.255
***Scene*$_4$**	t_0	212.3/19.6/.305	202.7/22.8/.396	99.6/18.8/.299	144.9/32.0/.374	38.4/52.0/.213	37.6/72.4/.152	34.1/91.2/.282	134.2/33.6/.368	50.9/81.6/.244
	t_1	81.5/71.6/.153	118.2/55.2/.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
	t_2	209.8/16.4/.342	287.0/3.20/.421	169.5/6.00/.362	280.0/0.00/.498	208.0/6.00/.361	102.3/47.2/.273	86.8/65.2/.383	234.0/9.20/.473	188.3/39.2/.389
	t_3	133.1/50.0/.260	86.7/59.2/.318	91.8/20.0/.296	265.6/0.00/.493	125.4/18.4/.305	49.2/71.6/.194	48.7/82.8/.313	141.9/31.6/.377	56.6/82.8/.268
	t_4	189.9/24.0/.320	143.4/32.8/.361	129.8/11.6/.336	275.9/0.00/.500	187.6/10.0/.346	234.8/8.40/.368	72.0/73.6/.368	197.5/18.0/.432	114.4/58.4/.341
	t_5	124.6/45.2/.257	84.9/64.8/.317	36.9/30.8/.235	274.3/1.20/.494	107.1/22.8/.285	179.0/13.2/.326	45.1/85.2/.319	126.4/40.8/.358	42.7/83.6/.255

Table 7: Ablation study on the weight regulation.

Scene	t	0:Cezanne	1:VanGogh	2:Picasso	3:JacksonPo...	4:Caravaggio	5:KeithHaring	6:KellyMcK...	7:TylerEdlin	8:KilianEng
wo.reg: $\|\Delta\theta_{0,dm}\|_p$=309.5; $\|\Delta\theta_{1,dm}\|_p$=1204.4; $\|\Delta\theta_{2,dm}\|_p$=485.5; $\|\Delta\theta_{3,dm}\|_p$=256.9; $\|\Delta\theta_{4,dm}\|_p$=531.0; $\|\Delta\theta_{5,dm}\|_p$=1004.3										
	t_0	199.1/14.0/.331	214.6/41.6/.382	185.7/8.00/.314	317.8/15.6/.403	178.5/50.4/.302	229.9/20.0/.544	77.7/86.4/.193	234.5/93.6/.152	128.3/96.0/.156
$\boldsymbol{Scene}_1$	t_1	1.40/98.0/.00	3.52/90.4/.00	185.5/8.80/.313	317.9/16.0/.402	178.2/49.6/.302	229.8/20.4/.544	68.9/95.2/.140	120.1/99.6/.089	111.6/98.0/.126
	t_2	0.61/98.0/.00	3.83/90.4/.00	1.49/98.8/.00	317.8/16.0/.402	178.1/50.0/.302	229.5/19.6/.544	61.5/99.6/.159	86.5/100./.057	86.3/99.2/.072
	t_1	199.2/14.0/.331	214.5/41.2/.382	185.3/8.80/.314	0.31/96.0/.00	2.56/99.6/.00	229.6/19.6/.544	118.5/65.2/.238	253.8/88.8/.166	140.9/96.8/.192
$\boldsymbol{Scene}_2$	t_2	199.1/13.6/.331	214.5/41.2/.382	185.0/8.40/.314	0.304/96.0/.00	1.61/100/.00	1.96/98.4/.00	85.2/82.8/.207	114.9/100/.095	137.1/96.8/.201
w.reg: $\|\Delta_{0,dm}\|_p$=253.2; $\|\Delta\theta_{1,dm}\|_p$=623.2; $\|\Delta\theta_{2,dm}\|_p$=510.4; $\|\Delta\theta_{3,dm}\|_p$=190.6; $\|\Delta\theta_{4,dm}\|_p$=328.4; $\|\Delta\theta_{5,dm}\|_p$=442.1										
	t_0	208.8/12.4/.338	219.9/36.8/.388	199.8/6.40/.336	274.9/26.4/.329	202.7/38.0/.312	216.5/15.2/.596	56.0/95.2/.153	134.6/98.2/.125	110.0/99.2/.126
$\boldsymbol{Scene}_1$	t_1	0.75/98.0/.00	2.12/90.8/.00	199.3/6.00/.336	276.7/26.8/.328	203.9/38.3/.313	209.0/25.2/.526	47.1/98.8/.096	119.3/100/.102	101.1/99.2/.107
	t_2	0.78/98.4/.00	3.28/90.4/.00	1.25/98.8/.00	276.5/26.8/.329	203.1/38.0/.312	208.7/25.6/.526	44.7/99.6/.097	99.9/100/.080	85.8/100/.066
	t_1	208.7/12.4/.339	219.7/38.0/.388	199.5/6.40/.336	0.18/96.0/.00	2.83/99.6/.00	218.7/16.0/.596	75.1/86.4/.192	179.8/98.8/.141	116.6/98.8/.141
$\boldsymbol{Scene}_2$	t_2	208.5/12.4/.338	219.6/38.0/.388	199.6/6.40/.336	0.14/96.0/.00	2.08/100/.00	2.11/98.4/.00	61.0/92.4/.168	94.7/100/.076	114.9/98.4/.151