Compensation-free Machine Unlearning in Text-to-Image Diffusion Models by Eliminating the Mutual Information

Xinwen Cheng, Jingyuan Zhang, Zhehao Huang, Yingwen Wu, Xiaolin Huang[*]

Institute of Image Processing and Pattern Recognition
School of Automation and Intelligent Sensing
Shanghai Jiao Tong University
{xinwencheng,tonyzhang666,kinght_h,yingwenwu,xiaolinhuang}@sjtu.edu.cn

Abstract

The powerful generative capabilities of diffusion models have raised growing privacy and safety concerns regarding generating sensitive or undesired content. In response, machine unlearning (MU) – commonly referred to as concept erasure (CE) in diffusion models – has been introduced to remove specific knowledge from model parameters meanwhile preserving innocent knowledge. Despite recent advancements, existing unlearning methods often suffer from excessive and indiscriminate removal, which leads to substantial degradation in the quality of innocent generations. To preserve model utility, prior works rely on *compensation*, *i.e.*, re-assimilating a subset of the remaining data or explicitly constraining the divergence from the pre-trained model on remaining concepts. However, we reveal that generations beyond the compensation scope still suffer, suggesting such post-remedial compensations are inherently insufficient for preserving the general utility of large-scale generative models. Therefore, in this paper, we advocate for developing *compensation-free* concept erasure operations, which precisely identify and eliminate the undesired knowledge such that the impact on other generations is minimal. In technique, we propose to **MiM-MU**, which is to unlearn a concept by minimizing the mutual information with a delicate design for computational effectiveness and for maintaining sampling distribution for other concepts. Extensive evaluations demonstrate that our proposed method achieves effective concept removal meanwhile maintaining high-quality generations for other concepts, and remarkably, without relying on any post-remedial compensation for the first time.

1 Introduction

Diffusion models (DM) have made remarkable strides in recent years, showing a great ability to generate realistic images. However, this powerful generative capability raises pressing privacy and safety concerns, particularly regarding the potential generation of undesirable content, such as Not Safe For Work (NSFW) images [1, 2, 3], copyright-infringing pictures [4, 5, 6] and training data replication [7, 8]. To effectively and thoroughly disable these generations without the need to retrain the model, *"machine unlearning"* (MU), which is also known as *"concept erasing"* (CE) for diffusion models, has emerged as a critical approach. MU aims at rapidly and seamlessly removing concept-related information from model parameters meanwhile maintaining the model performance on other generations.
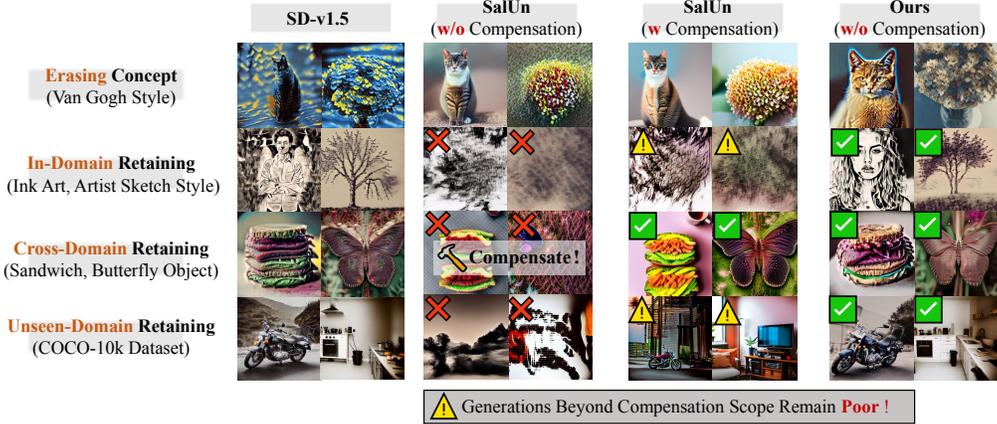
---

[*]Corresponding Author.

Figure 1: **Retainability across different concept domains of SalUn [1] (compensation-dependent) and our method (comepnsation-free) when unlearning "Van Gogh" style, revealing the failure of post-remedial compensation to restore generations beyond explicitly compensated concepts.** Images in the first three rows are generated with concepts in UnlearnCanvas benchmark (which is elaborated in Sec. 4.1) in the form of *"A {object} in {artist} style"* and images in the last row are respectively generated with prompts from COCO-10k dataset *"The shiny motorcycle has been put on display"* and *"A kitchen filled with furniture and a stove top oven"*. SalUn with compensation refers to compensating concepts in the cross-domain of UnlearnCanvas (*i.e.*, 20 objects including "Sandwich" and "Butterfly") by re-assimilating corresponding data. However, as can be seen, such compensation failed to restore concepts beyond the compensated scope, *i.e.*, "Ink Art" and "Artist Sketch" style in the same style domain as "Van Gogh" and "kitch" and "stove" in the COCO-10k dataset.

A variety of competitive methodologies have been developed to manipulate the unlearned model's behavior on the erasing concept to unlearn. The manipulations can be broadly categorized into three types: (1) **Retargeting** model outputs of erasing concept to that of an anchor concept [1, 9, 10, 11, 12], (2) **Repelling** from the pre-trained model's behavior [13, 14] and (3) **Suppressing** intermediate activations (*e.g.*, the cross-attention maps) to reduce responses [15, 16, 17]. However, recent evaluations on the comprehensive UnlearnCanvas benchmark [18] have revealed that above erasures significantly degrade innocent generations due to their indiscriminate and aggressive removal [9, 1, 13, 12, 14], necessitating additional maintenance on the remaining data to preserve model utility.

Although such compensations achieve some effectiveness and are widely adopted and allowed in previous works, we raise the attention that their power is not as satisfactory as desired. Prior work typically assesses the retainability of the unlearned model within the same scope as the compensated concepts. However, in **Fig.**1, we illustrate the retainability of our method and SalUn [1] (which achieves the best performance in UnlearnCanvas benchmark [18]) across different concept domains, revealing that generations regarding concepts *outside* the explicitly maintained scope remain significantly degraded. Specifically, we highlight two **fundamental limitations** inherent to post-remedial compensations: (1) The inadvertent damages introduced by unlearning are usually difficult to diagnose, potentially creating subtle but cumulative performance degradation; (2) Compensation is typically restricted to a narrowly presupposed scope, whereas generative models are expected to handle a vast and diverse range of concepts so that generations beyond compensation might remain poor. These underscore the urgent need for developing *compensation-free* unlearning approaches for large-scale generative models, which could effectively eliminate undesired concepts meanwhile minimally impacting other generations when erasing.

Since there is no chance to relearn, a compensation-free unlearning necessitates precise identification and removal of target knowledge. To achieve this goal, we propose to minimize

$$p(y|x), \forall x \sim \mathcal{S}_{\theta_U},$$

where y denotes the erasing concept and $\mathcal{S}_{\theta_U}$ denotes the sampling distribution of the unlearned model θ_U. The idea is that when $p(y|x) \to 0$, the generations by the unlearned model are devoid of any semantic associated with the erasing concept y and thus the probability of being identified as y is very small. By Bayes' rule, minimizing $p(y|x)$ is equivalent to minimizing the likelihood ratio $p(x|y)/p(x)$, for $p(y)$ is a constant independent of the generated x. This likelihood ratio quantifies *mutual information* between the textual concept y and the generated image x, *i.e.*, $\mathcal{I}(x,y) = \log p(x|y) - \log p(x)$. Our method is hence named Mutual Information Minimization (**MiM-MU**).

As theoretically elucidated by Kong et al. [19], a pre-trained diffusion model allows an *exact* density estimation for $p(x)$ and $p(x|y)$. This enables MiM-MU to leverage the pre-trained diffusion model as a competitive discriminator to quantify concept-related information in images generated by the unlearned model, followed by back-propagating to the unlearned model to minimize such mutual information. Beyond the core principle, we also address two critical technical issues: (1) *Efficiency:* to facilitate the optimization, we analyze the back-propogated gradient flow and identify that the Jacobian of the pre-trained model could be reasonably omitted to reduce computational overhead; (2) *Minimal Interference:* ensure the unlearning minimally affects innocent generations, we propose that the unlearned sampling distribution should remain as close as possible to that of the pre-trained model while minimizing mutual information and identify the one as the marginal distribution of the pre-trained model. We evaluate style and object unlearning on a comprehensive benchmark UnlearnCanvas [18], including 50 styles and 20 objects. Remarkably, our unlearning method achieves favorable concept erasure meanwhile well preserving the general utility without relying on any post-remedial compensation for the first time.

We summarize our key contributions as follows:

❶ We provide a principled formulation of the concept of erasure objective in diffusion models from an information-theoretic perspective, by quantifying the mutual information between textual concepts and unlearned sampling distribution with the pre-trained diffusion model.

❷ To preserve the model's general utility during unlearning, we propose to align the sampling distribution of the unlearned model with the marginal distribution of the pre-trained model, which is identified as the closest concept-irrelevant distribution to the original.

❸ We reveal that existing post-remedial compensation strategies exhibit limited recovery and fail for generations beyond the compensation scope, advocating for more benign erasure rather than excessive removal and unreliable compensations. In contrast, our method achieves faithful concept erasure while preserving the general model utility without any compensation for the first time.

2   Related Work

Machine unlearning for Text-to-image (T2I) Diffusion Models. The powerful generation capabilities of T2I models also bring about safety and privacy concerns of undesirable contents, including NSFW generations [2, 3], artistic copyright [4, 5, 6], and training data replication [8, 7]. While certain undesired generations can be mitigated through filtering mechanisms or modified inference-time guidance [3, 20, 16, 21, 22], these strategies are easily circumvented. Machine Unlearning (MU), also referred to as concept erasing (CE) in diffusion models, offers a more thorough solution by permanently removing undesirable knowledge from the model's parameters while preserving knowledge unrelated to the target concept. Existing erasing ideas can be broadly categorized into 3 types: retargeting, repelling, and suppressing. Retargeting retargets model outputs on the erasing concept to that of a neighborhood anchor concept [9, 23, 10, 11, 12, 1], which should be sufficiently distinct from the erasing concept to achieve effective unlearning, meanwhile not too divergent to damage benign generations severely, requiring delicate design and inspection to trade-off the unlearning and maintaining. Re-pelling steers the classifier-free-guidance (CFG) term away from the original denoising trajectory, thereby avoiding the search for a suitable anchor concept [13, 14]. Suppressing aims to locate the erasing concept associated knowledge and obliviate them on purpose, mainly focusing on diminishing activations of the cross-attention map [15, 16, 17].

Model utility maintenance in current T2I MU. Since there are usually multiple concepts in an image, erasing inadvertently affects unrelated concepts, leading to significant degradation in model utility. Current work either reduces interference with the remaining concepts when unlearning or compensates for damage after unlearning to preserve model utility. The former enhances the locality of the forgetting operation [14, 24], and the latter constrains the divergence

Table 1: The constraints and manipulated objective to maintain the performance on the remaining concepts.

Manner	Objective	Method
None	-	ESD [13], FMN [15], SDD [23]
Reduce Interference to Enhance Locality	Gradient Prediction	DoCo[24] SepME [14]
Encourage Alignment to Repair Performance	Attention Map	SEOT [16]
	Text Embedding	CA [9], UCE [10], REFACT[25], MACE [12], RECE [26]
	Predicted Noise	SalUn[1], EDiff [27], SHS [17], AdvUnlearn [28], SafeGen[29],

before and after unlearning on the remaining concepts [9, 10, 25, 16, 12, 26]. **Tab.**1 summarizes common maintenance measures. However, even with explicit maintenance, most of the existing methods still fail to preserve the model utility, as revealed by a comprehensive benchmark UnlearnCanvas [18].

In this paper, we raise the attention that the additional compensation is limited and insufficient for a large-scale generative task, and a practical concept erasing method should prioritize a non-interfering erasure over excessive and indiscriminate removal.

3 MUMI: Machine Unlearning by Eliminating the Mutual Informatoin

3.1 Problem Statement

MU in T2I diffusion models (more detailed preliminaries can be found in Appendix A) aims to prevent generating any erasing concept-related content meanwhile preserving model utility on other generations. A thorough erasure demands that any generation $\boldsymbol{x}$ by the unlearned model should not contain any semantics of the erasing concept. Mathematically, the probability of $\boldsymbol{x}$ being classified into the erasing concept y by an oracle classifier approaches 0, *i.e.*, $\forall \boldsymbol{x} \sim \mathcal{S}_{\theta_U}$, $p(y|\boldsymbol{x}) \rightarrow 0$ where $\mathcal{S}_{\theta_U}$ denotes the sampling distribution of the unlearned model θ_U:

$$\underset{\theta_U}{\text{minimize}} \ \mathcal{D}_{\text{forget}} := \mathbb{E}_{\boldsymbol{x} \sim \mathcal{S}_{\theta_U}} \left[\log p(y|\boldsymbol{x}) \right]. \tag{1}$$

There are two main challenges when optimizing Eq. 1: (1) $p(y|\boldsymbol{x})$ requires an oracle classifier to identify the quantity of the erasing concept y related semantics in generated image $\boldsymbol{x}$; (2) Directly back-propogating Eq. 1 to optimize the whole sampling distribution of the unlearned model $\mathcal{S}_{\theta_U}$ might be computationally expensive and impractical.

In this section, we first show that diminishing $p(y|\boldsymbol{x})$ is to diminish the mutual information between textual concept y and corresponding image $\boldsymbol{x}$, *i.e.*, $\log p(\boldsymbol{x}|y) - \log p(\boldsymbol{x})$, which could be well estimated by the pre-trained diffusion model from an information-theoretic perspective. Instead of straightforwardly optimizing the sampling distribution of the unlearned model through gradient descent, we characterize the family of distributions that have minimal mutual information and propose to approach the one that is *closest* to the pre-trained model to avoid excessive damage to model utility.

3.2 Mutual Information for Quantifying Concept Semantics

Through Bayes' rule, we have $p(y|\boldsymbol{x}) = p(\boldsymbol{x}|y)p(y)/p(\boldsymbol{x})$, where $p(y)$ is fixed as the probability of erasing concept y. Therefore, the goal of MU is to fine-tune the pre-trained model so that $p(\boldsymbol{x}|y)/p(\boldsymbol{x})$ should be as small as possible for any generated image $\boldsymbol{x}$ by the unlearned model.

Notice that $p(\boldsymbol{x}|y)/p(\boldsymbol{x})$ is the exponent of the mutual information between textual concept y and image $\boldsymbol{x}$, *i.e.*, $\mathcal{I}(\boldsymbol{x}, y) = \mathbb{E}_{p(\boldsymbol{x}, y)}[\log p(\boldsymbol{x}|y) - \log p(y)]$. It is well known that the diffusion model learns data distribution by maximizing the ELBO bound (Eq. A1), which is a lower bound of sample density $p(\boldsymbol{x})$. Interestingly, Kong et al. [19] establishes a rigorous connection between the pre-trained diffusion models θ_P and *exact* density estimations with the help of information theory. They elucidate that the pre-trained diffusion model can well estimate the density $p(\boldsymbol{x})$ and $p(\boldsymbol{x}|y)$ with the integral of optimal noise reconstruction error at different noise levels α:

$$-\log p(\boldsymbol{x}) = \frac{1}{2} \int_0^\infty \mathbb{E}_{\boldsymbol{\epsilon}} \left[\|\boldsymbol{\epsilon} - \hat{\epsilon}_{\theta_P}(\boldsymbol{x}_\alpha)\|_2^2 \right] d\alpha + \text{const}, \tag{2}$$

$$-\log p(\boldsymbol{x}|y) = \frac{1}{2} \int_0^\infty \mathbb{E}_{\boldsymbol{\epsilon}} \left[\|\boldsymbol{\epsilon} - \hat{\epsilon}_{\theta_P}(\boldsymbol{x}_\alpha|y)\|_2^2 \right] d\alpha + \text{const}. \tag{3}$$

The derivations are introduced in Appendix B.1 for completeness and comprehension. We summarize two crucial insights to understand this relationship here: (1) The information quantity in a Gaussian noise channel is exactly the Minimum Mean Square Error (MMSE) of optimal noise reconstruction, *i.e.*, $\mathcal{I}(\boldsymbol{x}_\alpha, \boldsymbol{x}) = \|\boldsymbol{\epsilon} - \hat{\epsilon}_{\theta_P}(\boldsymbol{x}_\alpha)\|_2^2$. (2) The density estimation $p(\boldsymbol{x})$ contains a Gaussian density term (*i.e.*, the *constant* term) and a *correction* term (*i.e.*, the *first* term) that measures how much better we can denoise the target distribution than we could with optimal denoiser for Gaussian source data $p_G(\boldsymbol{x}) \sim \mathcal{N}(0, \mathbb{I})$. Such an information-theoretic perspective demonstrates that a pre-trained model's ability to successfully reconstruct the noise from a noisy channel reflects its acquired semantic information in the images. Thus the pre-trained diffusion model is not only a powerful denoiser but also a valuable repository of semantic information. From this perspective, diminishing the generative capacity of *semantic information* is equivalent to degrading their *noise reconstruction* ability.

Non-negative mutual information. By substituting Eq. 2 and Eq. 3 and applying orthogonality principle [30], we have a non-negative formulation of mutual information as follows:

$$\mathcal{I}(\boldsymbol{x}, y) = [\log p(\boldsymbol{x}|y) - \log p(\boldsymbol{x})] = \frac{1}{2} \int_0^\infty \mathbb{E}_{\boldsymbol{\epsilon}} \left[\|\hat{\epsilon}_\alpha(\boldsymbol{x}_\alpha) - \hat{\epsilon}_\alpha(\boldsymbol{x}_\alpha|y)\|_2^2 \right] d\alpha. \tag{4}$$

4

The derivation is referred to Appendix B.2. This non-negative expression of the mutual information has been indicated to be effective in attributing the semantics in the generated image to corresponding textual words in prompts [31, 32], locating the emergence of specific semantics during generation.

3.3 Diminishing the Mutual Information to Unlearn

The above information-theoretic view of diffusion model highlights that the pre-trained diffusion model is a continuous density model with competitive log-likelihood estimation, facilitating discriminating the quantity of erasing concept in generated images. Consequently, the images generated by the unlearned model should exhibit minimal mutual information with the erasing concept when inspected by the pre-trained model. This procedure is analogous to training a generative adversarial network (GAN) [33], but the discriminator here is fixed as the pre-trained diffusion model. The gradient flow in our framework is as follows (timestep t is used to indicate noise level α hereafter):

$$\tilde{\boldsymbol{x}}_t, \hat{\epsilon}_{\theta_U} \xrightarrow{\text{Generate}} \boldsymbol{x} \xrightarrow{\text{Add Gaussian Noise}} \boldsymbol{x}_t \xrightarrow{\text{Noise Reconstruction Prediction}} \hat{\epsilon}_{\theta_P}(\boldsymbol{x}_t), \hat{\epsilon}_{\theta_P}(\boldsymbol{x}_t|y) \xrightarrow{\text{Compute}} \mathcal{I}(\boldsymbol{x}, y). \tag{5}$$

The diminish of the mutual information requires back-propagating through two forward paths of the generator θ_U and discriminator θ_P, which is computationally expensive and impractical for large foundation models. In this section, we analyze the gradient flow of the above optimization and propose an alternative objective.

We denote the mutual information at timestep t as $\mathcal{I}_t(\boldsymbol{x}, y) := \frac{1}{2} \|\hat{\epsilon}_{\theta_P}(\boldsymbol{x}_t|y) - \hat{\epsilon}_{\theta_P}(\boldsymbol{x}_t)\|_2^2$, and then minimizing the unlearning objective in Eq. 1 corresponds to minimize each $\mathcal{I}_t(\boldsymbol{x}, y)$. The gradient of $\mathcal{I}_t(\boldsymbol{x}, y)$ w.r.t. the unlearned model θ_U is as the following:

$$\frac{\partial \mathcal{I}_t(\boldsymbol{x}, y)}{\partial \theta_U} = \frac{\partial \mathcal{I}_t(\boldsymbol{x}, y)}{\partial \hat{\epsilon}_{\theta_P}} \cdot \frac{\partial \hat{\epsilon}_{\theta_P}}{\partial \boldsymbol{x}_t} \cdot \frac{\partial \boldsymbol{x}_t}{\partial \boldsymbol{x}} \cdot \frac{\partial \boldsymbol{x}}{\partial \theta_U}$$

$$= \mathbb{E}_{\boldsymbol{\epsilon}} \left[w(t) \cdot \underbrace{(\hat{\epsilon}_{\theta_P}(\boldsymbol{x}_t|y) - \hat{\epsilon}_{\theta_P}(\boldsymbol{x}_t))}_{\text{Pre-trained CFG}} \cdot \underbrace{\left(\frac{\partial \hat{\epsilon}_{\theta_P}(\boldsymbol{x}_t|y)}{\partial \boldsymbol{x}_t} - \frac{\partial \hat{\epsilon}_{\theta_P}(\boldsymbol{x}_t)}{\partial \boldsymbol{x}_t} \right)}_{\text{Pre-trained U-Net Jacobian}} \cdot \underbrace{\frac{\partial \hat{\epsilon}_{\theta_U}(\tilde{\boldsymbol{x}}_t|y)}{\partial \theta_U}}_{\text{Unlearned U-Net Gradient}} \right]. \tag{6}$$

The derivation is referred to Appendix C.1 and $w(t)$ is a coefficient. The obtained gradient in Eq. 6 can be decomposed into 3 components: (1) CFG term of the pre-trained diffusion model, (2) U-Net Jacobian of the pre-trained diffusion model and (3) U-Net gradient of the unlearned model.

Kullback-Leibler (KL) divergence minimization by omitting the U-Net Jacobian. The Jacobian term in Eq. 6 is computationally expensive to evaluate and poorly conditioned at low noise levels, as it approximates the scaled Hessian of the marginal density. Following common practice in Score Distillation Sampling (SDS) [34, 35], we omit this term and obtain an approximate gradient:

$$\frac{\partial \mathcal{I}_t(\boldsymbol{x}, y)}{\partial \theta_U} \approx \mathbb{E}_{\boldsymbol{\epsilon}} \left[w(t) \cdot \underbrace{(\hat{\epsilon}_{\theta_P}(\boldsymbol{x}_t|y) - \hat{\epsilon}_{\theta_P}(\boldsymbol{x}_t))}_{\text{Pre-trained CFG}} \cdot \underbrace{\frac{\partial \hat{\epsilon}_{\theta_U}(\tilde{\boldsymbol{x}}_t|y)}{\partial \theta_U}}_{\text{Unlearned U-Net Gradient}} \right]. \tag{7}$$

Then we show that the approximate gradient in Eq. 7 corresponds to minimizing the KL divergence between the unconditional and conditional latent distributions at timestep t:

$$\underset{\theta_U}{\text{minimize}} \ \mathcal{D}_{\theta_P}^{\text{KL}}(\boldsymbol{x}) := \text{KL}(p_{\theta_P}(\boldsymbol{x}_t) \| p_{\theta_P}(\boldsymbol{x}_t|y)) = \mathbb{E}_{\boldsymbol{\epsilon}} \left[\log \frac{p_{\theta_P}(\boldsymbol{x}_t)}{p_{\theta_P}(\boldsymbol{x}_t|y)} \right]. \tag{8}$$

The derivation is referred to Appendix C.2. Intuitively, the unlearning objective in Eq. 1 is converted to demanding the unlearned model to shift its generation *at any noise level* towards a density region less associated with the concept y, where relevance is inspected by the pre-trained diffusion model.

3.4 Deviating Least from the Pre-trained Model to Preserve

Directly minimizing the KL divergence in Eq. 8 w.r.t. the model parameters θ_U overlooks the preservation of innocent generations, generally resulting in a severe performance degradation. To address this, we seek the minimums of Eq. 8 for the one with favorable model utility. The minimum is attained when the expected log-ratio $\log p_{\theta_P}(\boldsymbol{x}_t|y) / \log p_{\theta_P}(\boldsymbol{x}_t) = 0$. The optimal sampling distribution of the unlearned model is the one that removes concept-related generations meanwhile maintaining other generations, thus, we prioritize degrading the conditional distribution $p_{\theta_U}(\boldsymbol{x}|y)$. The collection of concept-independent conditional distribution for Eq. 8 is defined as:

$$\mathcal{Q}^*(t, y) := \{ q_{\theta_U}(\boldsymbol{x}|y) \mid p_{\theta_P}(\boldsymbol{x}_t|y) = p_{\theta_P}(\boldsymbol{x}_t), \ \forall \boldsymbol{x} \sim q_{\theta_U}(\boldsymbol{x}|y) \}. \tag{9}$$

When erasing the undesired concept, we should best preserve the generation quality over the rest of the prompt space. *Therefore, we seek a $q_{\theta_U}(\boldsymbol{x}|y) \in Q^*$ that is closest to $p_{\theta_P}(\boldsymbol{x}|y)$ to avoid excessive forgetting.* In information theory, among all the distributions that are independent of y, the one that has minimal KL divergence to $p_{\theta_P}(\boldsymbol{x}_t|y)$ is the marginal distribution $p_{\theta_P}(\boldsymbol{x}_t)$ [36]:

$$\underset{q_{\theta_U}(\boldsymbol{x}_t|y) \in Q^*}{\text{minimize}} \quad \mathbb{E}_{\boldsymbol{x} \sim p_{\theta_U}} \left[KL(q_{\theta_U}(\boldsymbol{x}_t|y) \| p_{\theta_P}(\boldsymbol{x}_t|y)) \right] \implies q_{\theta_U}^*(\boldsymbol{x}_t|y) = p_{\theta_P}(\boldsymbol{x}_t). \tag{10}$$

In this way, we can have minimal impact on other generations when effectively erasing the concept. Therefore, we fine-tune the unlearned model to approach its conditional generation $p_{\theta_U}(\boldsymbol{x}|y)$ to approach the marginal distribution of the pre-trained diffusion model $p_{\theta_P}(\boldsymbol{x})$. Mathematically, aligning these two distributions is equivalent to aligning the conditional score of the unlearned model with the unconditional one of the pre-trained model:

$$\underset{\theta_U}{\text{minimize}} \ \mathcal{D}_{\text{MI}}(\boldsymbol{x}) := \text{KL}(q_{\theta_U}^*(\boldsymbol{x}_t|y) \| q_{\theta_U}(\boldsymbol{x}_t|y)) = \mathbb{E}_{\boldsymbol{\epsilon}} \left[\| \hat{\epsilon}_{\theta_U}(\boldsymbol{x}_t|y) - \hat{\epsilon}_{\theta_P}(\boldsymbol{x}_t) \|_2^2 \right] \tag{11}$$

The derivation is referred to Appendix C.3. In practice, we use a fixed set of concept-related images $\mathcal{X}_f$ to stand for the conditional sampling distribution of unlearned model $q_{\theta_U}(\boldsymbol{x}|y)$, *i.e.*, $\boldsymbol{x}_t = \sqrt{\bar{\alpha}_t}\boldsymbol{x} + \sqrt{1 - \bar{\alpha}_t}\boldsymbol{\epsilon}, \boldsymbol{x} \in \mathcal{X}_f$. *From the information-theoretic perspective, such alignment degrades the conditional denoiser's ability to reconstruct semantic information from noisy channels, diminishing the acquired information quantity in model parameters.*

Re-interpret Safe-Self-Distillation(SDD) [23]. At the first sight, our proposed Eq. 11 is quite similar to SDD [23] in formulation. SDD aligns the unlearned model's conditional score with its *own* unconditional one:

$$\underset{\theta_U}{\text{minimize}} \ \mathcal{D}_{\text{SDD}}(\boldsymbol{x}) := \mathbb{E}_t \left[\| \hat{\epsilon}_{\theta_U}(\tilde{\boldsymbol{x}}_t|y) - \hat{\epsilon}_{\theta_U}(\tilde{\boldsymbol{x}}_t) \|_2^2 \right], \quad \tilde{\boldsymbol{x}}_t = \boldsymbol{x}_T + \sum_{k=T}^{t+1} \hat{\epsilon}_{\theta_{\text{EMA}}}(\tilde{\boldsymbol{x}}_k|y). \tag{12}$$

where $\boldsymbol{x}_T \sim \mathcal{N}(0, \mathbb{I})$ and θ_{EMA} is an exponential moving average (EMA) updated version of the unlearned model. Regrettably, this intuitively reasonable alignment did not receive much attention. Now by comparing SDD with Eq. 7 and our Eq. 11, we found the reasons that might lead to sub-optimal performance. In SDD, the gradient of $\mathcal{D}_{\text{SDD}}$ w.r.t. the unlearned model is:

$$\frac{\partial \mathcal{D}_{\text{SDD}}(\boldsymbol{x})}{\partial \theta_U} = \mathbb{E}_{\boldsymbol{\epsilon}} \left[\underbrace{(\hat{\epsilon}_{\theta_U}(\tilde{\boldsymbol{x}}_t|y) - \hat{\epsilon}_{\theta_U}(\tilde{\boldsymbol{x}}_t))}_{\text{Unlearned CFG}} \frac{\partial \hat{\epsilon}_{\theta_U}(\tilde{\boldsymbol{x}}_t|y)}{\partial \theta_U} \right]. \tag{13}$$

Here, the CFG term is provided by the unlearned model, whose discriminability of the erasing concept might gradually become deficient with the gone of generation capability regarding the erasing concept. Besides, to preserve model utility on other generations, the unlearned distribution should deviate least from the original one when erasing. While the distillation teacher in SDD is $p_{\theta_U}(\boldsymbol{x})$, which drifts further and further away from $p_{\theta_P}(\boldsymbol{x}|y)$ as unlearning proceeds. In Sec. 4.3, we empirically verify that the model utility of the SDD unlearned model gradually breaks down when ceaselessly self-distilling from its own unconditional distribution.

4 Experiments

4.1 Experiment Setups

Evaluation models and datasets. There are mainly two kinds of concept erasing: (1) forget an artistic painting *style* for the purpose of copyright protection; (2) forget an *object* for the degeneration of models. We evaluate on a comprehensive benchmark, **UnlearnCanvas** [18]. It includes 70 concepts in total, which are divided into two main domains: a style domain comprising 50 artist styles and an object domain with 20 common objects. Each image in the dataset is a stylized object, *e.g.*, "A Dog in Van Gogh Style". The variety of concepts enables a comprehensive evaluation over a rich unlearning target bank, and the stylized object allows a distinct inspection on the retainability of innocent knowledge from both in-domain and cross-domain perspectives. The pre-trained model in UnlearnCanvas benchmark is stable-diffusion-v1-5 fine-tuned on UnlearnCanvas dataset.

Evaluation metrics. Machine unlearning in generative models should be assessed from three perspectives: the *completeness* of the erasure, the *retainability* of innocent knowledge, and the *quality* of generated images. We use Unlearning Accuracy (**UA**), *i.e.*, the proportion of generated images classified as the erasing concept, to indicate unlearning completeness. For retainability, we indicate with In-domain Retain Accuracy (**IRA**) and Cross-domain Retain Accuracy (**CRA**). For

example, when unlearning "Van Gogh" style, IRA refers to successful generations of other styles (*e.g.*, "Monet") and CRA refers to successful generations of all the objects (*e.g.*, "Dogs"). We indicate the distributional quality of image generation with Fréchet Inception Distance (**FID**). Furthermore, most existing erasing methods rely on explicitly compensating for a presupposed scope of remaining concepts, where the retainability is typically evaluated. However, we argue that retainability should also be assessed in an out-of-distribution (O.O.D) concept domain–*i.e.*, concepts unseen during unlearning–to ensure the *general* utility of the unlearned model.

Unlearning baselines. Original UnlearnCanvas benchmark contains 9 most recently proposed MU methods for diffusion models, including (1) **ESD** (Erased Stable Diffusion) [13], (2) **FMN** (Forget-Me-Not) [15], (3) **CA** (Ablating Concepts) [9], (4) **UCE** (Unified Concept Editing) [10], (5) **SalUn** (saliency Unlearning) [1], (6) **SEOT** (Suppress EOT) [16], (7) **SHS** (ScissorHands) [17], (8) **EDiff** (EraseDiff) [27], and (9) **SPM** (concept-SemiPermeable Membrane) [37]. We add (10) **SDD** (Safe Self-distilling) for further comparison due to its benign retainability.

4.2 Unlearning Performance on UnlearnCanvas Benchmark

Tab. 2 demonstrates performance of style and object unlearning in UnlearnCanvas benchmark. To better distinguish the locality of existing erasing operations, we divide existing methods into two groups based on whether explicit compensations for innocent concepts are applied, with the bottom group representing *compensation-free* methods. It is evident in **Tab**. 2 that existing methods face great challenges with either incomplete unlearning (*e.g.*, CA, SEOT, SPM, FMN) or severe damage to innocent knowledge (*e.g.*, ESD, FMN, UCE, SHS, EDiff) as highlighted by red color. In terms of accuracy metrics, SalUn achieves the best total average accuracy (92.77%), appearing to strike a good balance between unlearning and retention. However, we reveal in our subsequent experiment that re-assimilating the remaining data in SalUn is limited, and SalUn exhibits poor retainability in concepts out of the compensation scope.

Among the 3 existing compensation-free methods, SDD achieves a total average accuracy of 81.00%, surpassing all the other compensation-dependent methods except for SalUn and EDiff. While our method further improves upon SDD with a substantial accuracy gain of 8.42%. Noteworthily, our IRA and CRA consistently exceed 90%, indicating a favorable retainability. Moreover, our method achieves the lowest FID at 49.14, while that of SalUn and SDD are 61.05 and 70.40 respectively. The significant gap between SalUn and MiM-MU in FID suggests that post-remedial compensations might have limited ability to fully restore the original generation quality.

Table 2: Quantitative performance overview of different DM unlearning methods on UnlearnCanvas benchmark [18]. Methods are divided into two categories depending on whether there is explicit maintenance for the remaining concepts. Each unlearning request targets either a single style or a single object, and the results are averaged across 50 styles and 20 objects, respectively. Best values are marked in green, second-best values are marked in yellow for Avg.Acc and FID columns, and underperforming ones are marked in red for each column.

Method	Style Unlearning				Object Unlearning				Total	FID ↓
	UA ↑	IRA ↑	CRA ↑	Avg. Acc ↑	UA ↑	IRA ↑	CRA ↑	Avg. Acc ↑	Avg. Acc ↑	
CA [9]	60.82%	96.01%	92.70%	83.84%	46.67%	90.11%	81.97%	72.92%	78.38%	54.21
UCE [10]	98.40%	60.22%	47.71%	68.78%	94.31%	39.35%	34.67%	56.11%	62.45%	182.01
SEOT [16]	56.90%	94.68%	84.31%	78.63%	23.25%	95.57%	82.71%	67.18%	72.91%	62.38
SalUn [1]	86.26%	90.39%	95.08%	90.58%	86.91%	96.35%	99.59%	94.95%	92.77%	61.05
SPM [37]	60.94%	92.39%	84.33%	79.22%	71.25%	90.79%	81.65%	81.23%	80.23%	59.79
EDiff [27]	92.42%	73.91%	98.93%	88.42%	86.67%	94.03%	48.48%	76.39%	82.41%	81.42
SHS [17]	95.84%	80.42%	43.27%	73.18%	80.73%	81.15%	67.99%	76.63%	74.91%	119.34
ESD [13]	98.58%	80.97%	93.96%	91.17%	92.15%	55.78%	44.23%	64.05%	77.61%	65.55
FMN [15]	88.48%	56.77%	46.60%	63.95%	45.64%	90.63%	73.46%	69.91%	66.93%	131.37
SDD [23]	83.79%	75.28%	77.38%	78.82%	84.31%	79.71%	85.51%	83.18%	81.00%	70.40
MiM-MU	80.12%	93.99%	93.18%	89.10%	81.14%	91.41%	96.65%	89.73%	89.42%	49.14

The accuracy-related metrics reported above primarily indicate the presence or absence of a concept, while a more detailed evaluation of concept generation quality requires examining the generated images. **Fig.** 2 showcases generations of "Monet" style unlearned methods (More visualizations are referred to Appendix D.1). It can be observed that SalUn fails to generate certain objects when requested alongside the "Monet" style in UA, indicating its excessive degradation. Moreover, its generations for the remaining concepts exhibit noticeable distortions and color oversaturation. Both implies that Salun's damage on image generation lacks of specificity and re-assimilating the remaining

data have limited effectiveness in restoring fine-grained details. In contrast, MiM-MU faithfully generates the object without any painting style in UA and produces high-quality images (*i.e.*, nearly identical to those generated by the pre-trained model) in IRA and CRA, demonstrating a more nuanced removal. **We think this might be an encouraging progress which demonstrates that refraining from invasive interference during erasure is a more effective and reliable strategy for preserving the generation quality of large-scale generative models than post-remedial compensations.**
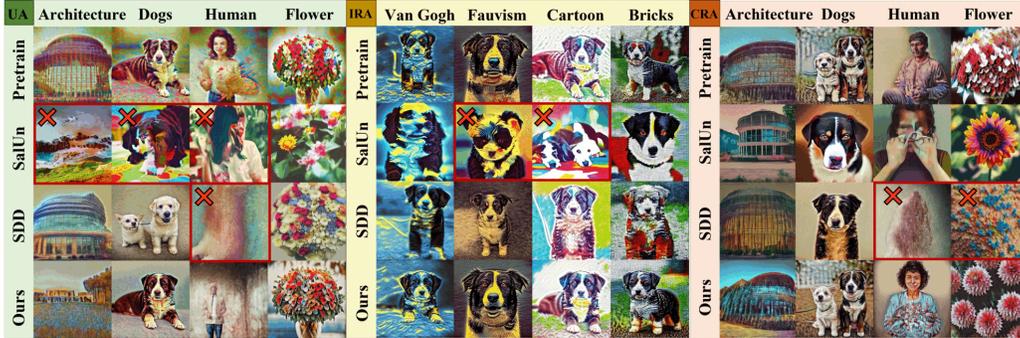

Figure 2: **Qualitative performance overview of different DM unlearning methods on "Monet" style.** For unlearning completeness (**UA**), we showcase the erasing concept with 4 concepts from the cross domain, *i.e.*, 4 different objects in "Monet" style. For the retainability, we use "Dogs" object and "Seed Image" style to combine with 4 remaining in-domain (**IRA**) and cross-domain concepts (**CRA**). Generations of the unlearned model should be *different* from the pre-trained model in terms of the erasing concept in **UA** and should *approach* the pre-trained model in **IRA** and **CRA**.

4.3 Compensation-free vs. Compensation-dependent MU Methods

While SalUn achieves superior performance on accuracy-related metrics in the UnlearnCanvas benchmark, its generation quality (FID) is notably inferior to ours. This suggests that its compensation mechanisms fail to fully recover the original generation quality. Furthermore, we highlight two fundamental **limitations** that critically undermine the reliability of post-remedial compensations: (1) The inadvertent damage caused by indiscriminate removal is difficult to diagnose, potentially creating subtle but **cumulative performance degradation that remains unrepaired**. (2) While generative models must handle a vast and diverse range of concepts, current compensation is typically restricted to a presupposed scope, which is relatively minimal so that **generations beyond the scope might remain poor**. We empirically validate these limitations through: (1) sequential unlearning tasks and (2) retainability assessment on out-of-distribution (O.O.D.) concept domain.

Sequential unlearning (SU). In practice, unlearning requests might arrive sequentially, demanding multiple executions of unlearning methods. Extensive evaluations in UnlearnCanvas [18] benchmark (Tab. A7 in original paper) demonstrate that **none of** existing MU methods could achieve resilient unlearning performance during SU. There are 3 universal deficiencies: **(i)** *degraded retainability*, **(ii)** *unlearning rebound effect*, and **(iii)** *catastrophic retaining failure*. These phenomena imply that the undesired knowledge is not essentially removed and post-remedial compensation is disordered, exposing fragile knowledge management (*i.e.*, removal and retention) of existing compensation-dependent methods. We sequentially unlearn 6 artist styles (in line with UnlearnCanvas) and track unlearning and retaining performances for each unlearning request in **Tab.** 3 (while visualizations are provided in Appendix D.2). We highlight with red that the unlearned model's performance on previous erased concept stages a recovery in SalUn, which implies that the erased concept is merely temporarily concealed. In contrast, our UA of each unlearning request stays at a high level throughout SU, indicating superior resilience of MiM-MU. Moreover, without any compensation, our RA is consistently higher than SalUn during SU. The resilient unlearning and benign retention collectively demonstrate the promising knowledge management of the compensation-free method, MiM-MU.

Deteriorated general utility in O.O.D. Domain Concepts of the compensation-dependent method. We empirically validate our concern about O.O.D. retainability by comparing MiM-MU with SalUn on COCO-10k dataset. Quantitative metrics are reported in **Tab.** 4 and qualitative visualizations are showcased in **Fig.** 3. The "Pretrain" in **Tab.** 4 refers to the stable-diffusion-v1-5 fine-tuned on UnlearnCanvas

Table 4: Generation quality in COCO-10k datasets.

Method	FID $\downarrow$	CLIP $\uparrow$
Pretrain	37.31	0.2780
SalUn	49.74	0.2761
Ours	**34.42**	**0.2780**

Table 3: Quantitative performance of MiM-MU and SalUn in sequential unlearning (SU) task. Each column represents a new unlearning request, denoted by $\mathcal{T}_i$, where $\mathcal{T}_1$ is the oldest. Each row represents the UA for a specific unlearning objective or the retaining accuracy (RA), given by the average of IRA and CRA. Results indicating *unlearning rebound* effect (where ΔUA $\geq 5\%$) are highlighted in red.

		MiM-MU							**SalUn**					
Metrics		$\mathcal{T}_1$	$\mathcal{T}_1 \sim \mathcal{T}_2$	$\mathcal{T}_1 \sim \mathcal{T}_3$	$\mathcal{T}_1 \sim \mathcal{T}_4$	$\mathcal{T}_1 \sim \mathcal{T}_5$	$\mathcal{T}_1 \sim \mathcal{T}_6$	Metrics	$\mathcal{T}_1$	$\mathcal{T}_1 \sim \mathcal{T}_2$	$\mathcal{T}_1 \sim \mathcal{T}_3$	$\mathcal{T}_1 \sim \mathcal{T}_4$	$\mathcal{T}_1 \sim \mathcal{T}_5$	$\mathcal{T}_1 \sim \mathcal{T}_6$
				Unlearning Request							Unlearning Request			
UA ↑	$\mathcal{T}_1$	98.00%	99.00%	98.00%	99.00%	100.00%	100.00%	UA ↑	84.00%	79.00%	78.00%	65.00%	67.00%	64.00%
	$\mathcal{T}_2$	-	93.00%	96.00%	96.00%	97.00%	100.00%		-	81.42%	75.00%	72.00%	69.00%	61.00%
	$\mathcal{T}_3$	-	-	97.00%	97.00%	100.00%	99.00%		-	-	90.00%	85.00%	84.00%	87.00%
	$\mathcal{T}_4$	-	-	-	98.00%	100.00%	100.00%		-	-	-	84.00%	86.00%	81.00%
	$\mathcal{T}_5$	-	-	-	-	100.00%	100.00%		-	-	-	-	79.00%	81.00%
	$\mathcal{T}_6$	-	-	-	-	-	92.00%		-	-	-	-	-	89.00%
RA ↑		90.90%	86.56%	80.58%	77.27%	70.34%	67.47%		85.43%	80.32%	71.42%	65.41%	63.24%	60.19%

"A small kitchen with low a ceiling." "A cat peering into a wooden bowl which is sitting on a table." "A woman preparing food inside of a kitchen." "A white plane is going low to make a landing." "A white toilet is at the end of a trail." "A street filled with advertising signs hanging from the sides buildings." "A fire hydrant stands alone in the middle of the concrete."
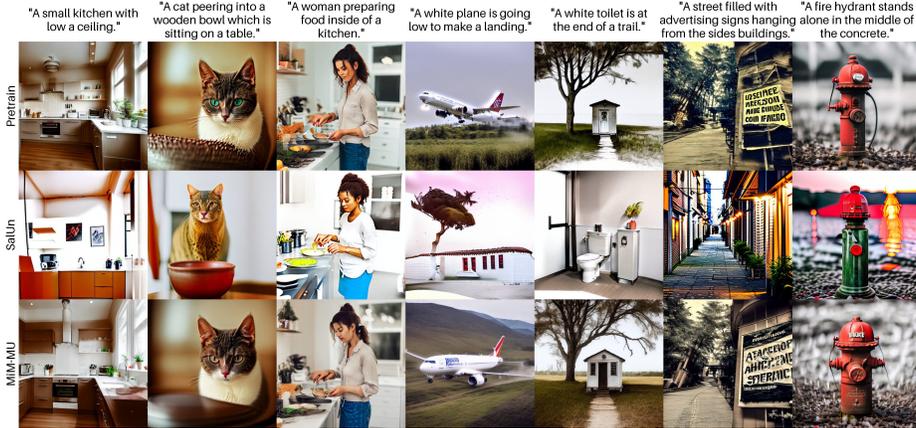
Figure 3: **Qualitative visualizations of MiM-MU and SalUn unlearned model on COCO-10k dataset generations.** Generations of SalUn unlearned model exhibit distortions and misalignment with the requested prompts. In contrast, MiM-MU produces high-quality and textually aligned images on COCO-10k dataset without utilizing additional maintenance to preserve model utility.

dataset, which has worse FID than vanilla sd-v1.5 due to learning stylized images. Notably, SalUn yields a significantly higher FID than the pre-trained model, indicating its poor retainability in COCO-10k generations. In contrast, MiM-MU achieves an even lower FID than the pre-trained model, suggesting that its erasure not only preserves utility but also successfully recovers the quality degradation introduced by fine-tuning. Visualizations in **Fig.** 3 demonstrate that MiM-MU generates more textually aligned images with clear textures, whereas generations of SalUn suffer from text-image misalignment and pronounced distortions. This result underscores the advantage of a nuanced erasure to ensure the *general* utility after unlearning.

Compensation-free method exhibits advantage in fine-grained erasure. Although UnlearnCanvas benchmark represents a significant step forward compared to earlier datasets, it still adopts relatively coarse-grained concept definitions, where concepts are relatively independent (*e.g.*, Dogs and Cats). However, in real-world settings, semantic boundaries are often highly entangled. To further assessing the granularity of existing concept erasure methods, we perform fine-grained erasure over 3 fine-grained datasets, Standford Dogs [38], Oxford Flowers [39] and CUB-200 [40]. To emphasize the potential risk of existing careless removal, we only compensate (*i.e.*, replaying corresponding data) generations of neighborhood classes in SalUn, while leaving remaining classes (termed *Other Concepts*) uncompensated, revealing their unintended and unrevealed degradation to other concepts.

Performances are referred to **Tab.** 5. Across all 3 datasets, Mim-MU consistently achieves more thorough forgetting and SalUn fails in Oxford Flower dataset, with UA only 66.67% while Mim-MU is 100.0%. The retainability include 2 perspectives in fine-grained erasure: Neighborhood Retain Accuracy (NRA) and Other-class Retain Accuracy (ORA). For near classes, Mim-MU fails to preserve generation of neighboorhood classes in CUB-200, with an NRA of 47.50% while that of SalUn is 78.33%. In the rest 2 datasets, Mim-MU achieves comparable retainability with accuracy floating within $\pm 2\%$. It is worth noting that we explicitly compensate the neighborhood classes for SalUn, therefore, we regard it credible for Mim-MU to achieve comparable retainability. For ORA, Mim-MU achieves better retainability in these, which are not explicitly compensated in SalUn as

Table 5: Quantitative comparison between SalUn and Mim-MU on 3 fine-grained classification datasets, Standford Dogs [38], Oxford Flowers [39] and CUB-200 [40]. Best values are marked in green and under-performing ones are marked in red for each column.

Dataset	Method	UA ↑	NRA ↑	ORA ↑	Avg.Acc ↑	N.FID ↓	O.FID ↓
Dogs	SalUn	95.83%	35.83%	65.61%	65.76%	108.19	54.06
	Mim-MU	100.00%	37.50%	70.00%	69.17%	92.46	42.42
Flowers	SalUn	66.67%	44.17%	67.81%	59.55%	131.83	64.93
	Mim-MU	100.00%	43.33%	68.54%	70.62%	130.99	58.51
CUB-200	SalUn	91.67%	78.33%	68.30%	79.43%	52.56	42.04
	Mim-MU	95.83%	47.50%	67.42%	70.25%	56.61	16.30

well. Consistently, Mim-MU exhibits lower FID scores in both neighborhood and other concepts across 3 datasets, indicating its stronger capability to maintain high-quality image generation and prevent undesired artifacts.

Fig.4 presents an illustrative comparison between SalUn and MiM-MU across 3 fine-grained datasets. Generations by SalUn would loss background environment details (*e.g.*, the 1st, 3rd, 4th images in UA column of Standford Dog, the 4th image in ORA column of CUB-200) and exhibit blurred object edges (*e.g.*, the 3rd, 4th, 5th images in UA column of Oxford Flowers) as well as overly smooth textures (*e.g.*, the 3rd, 4th images in UA column of CUB-200,). In addition, many of SalUn's retained-concept generations display over-saturation—colors look unnaturally vibrant artifact (*e.g.*, 3rd image in ORA column of CUB-200). These suggest that SalUn's erasure causes collateral degradation in image fidelity. By contrast, MiM-MU unlearned model generations exhibit more clear textures and natural colors, achieving higher fidelity.

Figure 4: **Qualitative performance overview of Mim-MU and SalUn on 3 fine-grained datasets.** For unlearning completeness (**UA**), we showcase the erasing concept with 5 seeds. For the retainability, we indicate 5 neighborhood concepts (**NRA**) and 5 other concepts (**ORA**). Generations of the unlearned model should be *different* from the pre-trained model in terms of the erasing concept in **UA** and should *approach* the pre-trained model in **NRA** and **ORA**.

5 Limitations of Existing MU Methods

In this section, we examine the limitations of existing concept erasure approaches to highlight some advantages of Mim-MU. We first investigate the resilience of unlearned models produced by SalUn, SDD, and Mim-MU when subjected to subsequent fine-tuning, revealing their vulnerability to concept resurgence. We then empirically reveal the failure of SalUn in handling multi-concept unlearning scenarios and the performance breakdown of SDD during its self-distillation process.

Unlearning resilience of Mim-MU to subsequent fine-tuning. Suriyakumar et al. [41] reveals that subsequent fine-tuning on the unlearned model with even seemingly unrelated data, can inadvertently cause the model to relearn or resurge the previously erased concepts. Such vulnerability underscores the fragility of existing unlearning methods, *i.e.*, the unlearned model would become unsafe again after further updates. To investigate the resilience of Mim-MU to such fine-tuning, we perform similar experiments for MiM-MU, SDD, and SalUn in **Tab.**6. We observed that MiM-MU exhibits minor signs of concept recovery after additional fine-tuning, while SalUn and SDD stages an obvious recovery, especially when fine-tuned with a random subset of other remaining data.

Table 6: UA of "Abstractionism Style" when subsequently training (fine-tuning) the unlearned model on the remaining data. Epoch-i refers to UA of the unlearned model after fine-tuning i epochs. **Class-wise** refers to fine-tuning the unlearned model with "Objects in Seed Images Style" data, and **Random-subset** refers to fine-tuning the unlearned model with a random subset from UnlearnCanvas benchmark without the erasing concept. Δ UA is computed as UA(Epoch-0) - UA(Epoch-8), reflecting the degradation in erasure performance (*i.e.*, recovery of the erasing concept) after fine-tuning.

Fine-tune Data	Method	Epoch-0	Epoch-2	Epoch-4	Epoch-6	Epoch-8	Δ UA $\downarrow$
Class-wise	SalUn	89.00%	86.00%	86.00%	82.00%	82.00%	+7.00%
	SDD	100.00%	78.00%	87.00%	84.00%	90.00%	+10.00%
	Mim-MU	97.00%	89.00%	87.00%	93.00%	90.00%	+7.00%
Random-subset	SalUn	89.00%	11.00%	8.00%	7.00%	13.00%	+76.00%
	SDD	100.00%	42.00%	31.00%	1.00%	23.00%	+77.00%
	Mim-MU	97.00%	90.00%	92.00%	86.00%	86.00%	+11.00%

Failure of SalUn in multi-concept scenario. We conducted experiments to simultaneously erase the same 6 styles as in **Tab.** 3 and present the results in **Tab.** 7. We appreciate SalUn's pioneering contribution in introducing the weight saliency map to unlearning, however, we identified two practical

Table 7: Performances of SalUn and Mim-MU on multi-concepts unlearning.

Method	UA↑	IRA↑	CRA↑	RA↑	Avg.Acc↑
SalUn	10.83%	98.56%	98.92%	98.74%	69.44%
Mim-MU	98.33%	69.44%	92.75%	81.10%	86.17%

limitations of this mechanism in the multi-concept setting: **(i)** *Incomplete forgetting*: Even unlearning with 30 epochs, UA of SalUn is only 10.83%, while that of Mim-MU is 98.33%, which indicates the failure of SalUn in simultaneous concept erasure. **(ii)** *Escalating computational overhead*: SalUn requires computing the prediction loss of the remaining data and regularizes it (add as a regularization term) to preserve model utility. Consequently, the total time grows to roughly twice the number of forgetting data. Also, it requires computing the saliency mask with forgetting data. As the forgetting set expands, both mask computation and model fine-tuning scale almost linearly, leading to a steep runtime increase. In contrast, Mim-MU neither requires mask computation nor utility compensation, exhibiting an obvious time advantage.

As for the failure of SalUn in success removal when unlearning simultaneously, we would like to attribute to two possible reasons: **(1) Gradient Direction Cancellation**. SalUn constructs a single weight-saliency mask by aggregating gradients from all "forget" concepts. When those prompts represent heterogeneous concepts, their gradients often conflict and summation partially cancel. As a result, they remain frozen, leaving key parameters untouched during unlearning and resulting only in superficial erasure. **(2) Fixed edit-budget bottleneck**. With the increase of target concepts, the pool of highly effective parameters would increase. A fixed threshold of saliency mask limits the number of editable weights, causing each concept to receive fewer dedicated updates. In contrast, Mim-MU localises concepts through a mutual-information mechanism, which does not rely on weight attribution sparsity to reduce interference, scaling more naturally in the multi-concept setting.

Performance breakdown as self-distillation in SDD. As elucidated in Sec. 3.4, the distillation teacher in SDD, which is the unconditional distribution of the unlearned model, will gradually deviate from the pre-trained model as unlearning proceeds. We empirically validate that continually distilling from it will induce model performance breakdown by unlearning "Flame" object and "Abstractionism" style in **Fig.** 5. As unlearning progresses, UA of MiM-MU and SDD both exhibit a progressive decline. However, the IRA and CRA of SDD experience a marked and rapid decline, while ours consistently converges to a high retainability. This highlights the instability of SDD's self-distillation, which progressively erodes innocent knowledge, while MiM-MU could accurately locate and remove the undesired knowledge by identifying the mutual information.

Figure 5: **Unlearning and retention during MiM-MU and SDD unlearning.** SDD exhibits a performance breakdown while ours well preserve the model utility as unlearning proceeds.

Summary. In summary, Mim-MU consistently preserves erasure effectiveness across different settings. Collectively, these findings suggest that Mim-MU provides a more reliable and practically applicable solution for concept erasure compared with existing approaches.

6 Discussion

Further improvement on fine-grained concept erasure. Although Mim-MU could achieve a comparable retention in fine-grained erasure without explicitly compensations, we also realize that there exists room for further improvement in mitigating unintended damage to other entangled concepts. In deed, fine-grained erasure refers to two concepts y_1 and y_2 are not independent with high correlations. Formally, this manifests as $p(Y_1, Y_2) \neq p(Y_1)p(Y_2)$. We think this challenge might be further improved using tools from information theory. Specifically, DiffusionPID [32] disentangles semantic dependencies by decomposing mutual information into unique, shared, and synergistic components, allowing precise attribution of each word's unique contribution to the generated content. As visualized in their work, this decomposition provides a highly accurate alignment between different information components and concept locations in the generated image. Also, such decomposition from text space is convenient because this does not require access to the remaining data to compensate explicitly. In general, the entanglement of concepts could be formalized as statistically non-independent in our theoretical framework, leaving a good starting point for future study.

7 Conclusion

In this paper, we demonstrate that the widely adopted post-remedial compensations in existing concept erasure methods are inherently limited and insufficient for large-scale generative models. We identify the concept-related knowledge from the information-theoretic perspective of the diffusion model and propose to diminish the mutual information between the textual concept and the image to erase a concept. Our proposed MiM-MU minimally impacts other generations during removal by demanding the least deviations of the unlearned model from the pre-trained model. Extensive experiments demonstrate that MiM-MU can successfully remove undesired generations meanwhile preserving the general model utility without any compensation.

Bibliography

[1] Chongyu Fan, Jiancheng Liu, Yihua Zhang, Dennis Wei, Eric Wong, and Sijia Liu. Salun: Empowering machine unlearning via gradient-based weight saliency in both image classification and generation. In *International Conference on Learning Representations (ICLR)*, 2023.

[2] Javier Rando, Daniel Paleka, David Lindner, Lennart Heim, and Florian Tramèr. Red-teaming the stable diffusion safety filter. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[3] Patrick Schramowski, Manuel Brack, Björn Deiseroth, and Kristian Kersting. Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[4] Shawn Shan, Jenna Cryan, Emily Wenger, Haitao Zheng, Rana Hanocka, and Ben Y Zhao. Glaze: Protecting artists from style mimicry by text-to-image models. In *32nd USENIX Security Symposium (USENIX Security 23)*, 2023.

[5] Zhenting Wang, Chen Chen, Lingjuan Lyu, Dimitris N Metaxas, and Shiqing Ma. Diagnosis: Detecting unauthorized data usages in text-to-image diffusion models. In *International Conference on Learning Representations (ICLR)*, 2024.

[6] Shengwei An, Lu Yan, Siyuan Cheng, Guangyu Shen, Kaiyuan Zhang, Qiuling Xu, Guanhong Tao, and Xiangyu Zhang. Rethinking the invisible protection against unauthorized image usage in stable diffusion. In *33rd USENIX Security Symposium (USENIX Security 24)*, 2024.

[7] Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Understanding and mitigating copying in diffusion models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

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

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

[10] Rohit Gandikota, Hadas Orgad, Yonatan Belinkov, Joanna Materzyńska, and David Bau. Unified concept editing in diffusion models. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2024.

[11] Alvin Heng and Harold Soh. Selective amnesia: A continual learning approach to forgetting in deep generative models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[12] Shilin Lu, Zilan Wang, Leyang Li, Yanzhu Liu, and Adams Wai-Kin Kong. Mace: Mass concept erasure in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[13] Rohit Gandikota, Joanna Materzynska, Jaden Fiotto-Kaufman, and David Bau. Erasing concepts from diffusion models. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

[14] Mengnan Zhao, Lihe Zhang, Tianhang Zheng, Yuqiu Kong, and Baocai Yin. Separable multi-concept erasure from diffusion models. *arXiv preprint arXiv:2402.05947*, 2024.

[15] Gong Zhang, Kai Wang, Xingqian Xu, Zhangyang Wang, and Humphrey Shi. Forget-me-not: Learning to forget in text-to-image diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[16] Senmao Li, Joost van de Weijer, Taihang Hu, Fahad Shahbaz Khan, Qibin Hou, Yaxing Wang, and Jian Yang. Get what you want, not what you don't: Image content suppression for text-to-image diffusion models. In *International Conference on Learning Representations (ICLR)*, 2024.

[17] Jing Wu and Mehrtash Harandi. Scissorhands: Scrub data influence via connection sensitivity in networks. In *European Conference on Computer Vision (ECCV)*, 2024.

[18] Yihua Zhang, Yimeng Zhang, Yuguang Yao, Jinghan Jia, Jiancheng Liu, Xiaoming Liu, and Sijia Liu. Unlearncanvas: A stylized image dataset to benchmark machine unlearning for diffusion models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[19] Xianghao Kong, Rob Brekelmans, and Greg Ver Steeg. Information-theoretic diffusion. *International Conference on Learning Representations (ICLR)*, 2023.

[20] Manuel Brack, Felix Friedrich, Dominik Hintersdorf, Lukas Struppek, Patrick Schramowski, and Kristian Kersting. Sega: Instructing text-to-image models using semantic guidance. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[21] Jaehong Yoon, Shoubin Yu, Vaidehi Patil, Huaxiu Yao, and Mohit Bansal. Safree: Training-free and adaptive guard for safe text-to-image and video generation. In *International Conference on Learning Representations (ICLR)*, 2025.

[22] Anubhav Jain, Yuya Kobayashi, Takashi Shibuya, Yuhta Takida, Nasir Memon, Julian Togelius, and Yuki Mitsufuji. Trasce: Trajectory steering for concept erasure. *arXiv preprint arXiv:2412.07658*, 2024.

[23] Sanghyun Kim, Seohyeon Jung, Balhae Kim, Moonseok Choi, Jinwoo Shin, and Juho Lee. Towards safe self-distillation of internet-scale text-to-image diffusion models. *International Conference on Machine Learning (ICML) Workshops*, 2023.

[24] Yongliang Wu, Shiji Zhou, Mingzhuo Yang, Lianzhe Wang, Heng Chang, Wenbo Zhu, Xinting Hu, Xiao Zhou, and Xu Yang. Unlearning concepts in diffusion model via concept domain correction and concept preserving gradient. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2025.

[25] Dana Arad, Hadas Orgad, and Yonatan Belinkov. Refact: Updating text-to-image models by editing the text encoder. In *Annual Conference of the North American Chapter of the Association for Computational Linguistics – Human Language Technologies (NAACL-HLT)*, 2024.

[26] Chao Gong, Kai Chen, Zhipeng Wei, Jingjing Chen, and Yu-Gang Jiang. Reliable and efficient concept erasure of text-to-image diffusion models. In *European Conference on Computer Vision (ECCV)*, 2024.

[27] Jing Wu, Trung Le, Munawar Hayat, and Mehrtash Harandi. Erasediff: Erasing data influence in diffusion models. *arXiv preprint arXiv:2401.05779*, 2024.

[28] Yimeng Zhang, Xin Chen, Jinghan Jia, Yihua Zhang, Chongyu Fan, Jiancheng Liu, Mingyi Hong, Ke Ding, and Sijia Liu. Defensive unlearning with adversarial training for robust concept erasure in diffusion models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[29] Xinfeng Li, Yuchen Yang, Jiangyi Deng, Chen Yan, Yanjiao Chen, Xiaoyu Ji, and Wenyuan Xu. SafeGen: Mitigating Sexually Explicit Content Generation in Text-to-Image Models. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 2024.

[30] Steven M Kay. *Fundamentals of Statistical Signal Processing: Estimation Theory*. Prentice-Hall, Inc., 1993.

[31] Xianghao Kong, Ollie Liu, Han Li, Dani Yogatama, and Greg Ver Steeg. Interpretable diffusion via information decomposition. In *International Conference on Learning Representations (ICLR)*, 2024.

[32] Shaurya Dewan, Rushikesh Zawar, Prakanshul Saxena, Yingshan Chang, Andrew Luo, and Yonatan Bisk. Diffusion pid: Interpreting diffusion via partial information decomposition. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[33] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11): 139–144, 2020.

[34] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *International Conference on Learning Representations (ICLR)*, 2023.

[35] Haochen Wang, Xiaodan Du, Jiahao Li, Raymond A Yeh, and Greg Shakhnarovich. Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[36] Shun-ichi Amari. *Information Geometry and its Applications*, volume 194. Springer, 2016.

[37] Mengyao Lyu, Yuhong Yang, Haiwen Hong, Hui Chen, Xuan Jin, Yuan He, Hui Xue, Jungong Han, and Guiguang Ding. One-dimensional adapter to rule them all: Concepts, diffusion models and erasing applications. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[38] Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Li Fei-Fei. Novel dataset for fine-grained image categorization. In *First Workshop on Fine-Grained Visual Categorization, IEEE Conference on Computer Vision and Pattern Recognition*, Colorado Springs, CO, June 2011.

[39] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Indian Conference on Computer Vision, Graphics and Image Processing*, Dec 2008.

[40] Peter Welinder, Steve Branson, Takeshi Mita, Catherine Wah, Florian Schroff, Serge Belongie, and Pietro Perona. Caltech-ucsd birds 200. 2010.

[41] Vinith Menon Suriyakumar, Rohan Alur, Ayush Sekhari, Manish Raghavan, and Ashia C Wilson. Unstable unlearning: The hidden risk of concept resurgence in diffusion models. In *ICLR 2025 Workshop on Navigating and Addressing Data Problems for Foundation Models*, 2024.

[42] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

[43] Dongning Guo, Shlomo Shamai, and Sergio Verdú. Mutual information and minimum mean-square error in gaussian channels. *IEEE Transactions on Information Theory*, 51(4):1261–1282, 2005.

Appendix

A Preliminaries

Diffusion models. Generative models learn to model the true data distribution $p(x)$ by marginalize it out the latent variables and maximizing the **E**vidence **L**ower **BO**und (ELBO). Diffusion model establishes hierarchical latent variables by pre-defining the latent encoder as a linear Gaussian model, constructing a forward process $\{x_1, \ldots, x_T\}$. The latent variable at timestep t is designed as $x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon_0$, with variance schedule $\bar{\alpha}_t > 0$. The ELBO of diffusion model is:

$$\log p_\theta(x_0) \geq \mathbb{E}_{q(x_{1:T}|x_0)} \left[\log \frac{p(x_0, x_{1:T})}{q(x_{1:T} \mid x_0)} \right]. \tag{A1}$$

The training objective turns into matching the approximate denoising transition step $p_\theta(x_{t-1}|x_t)$ to ground-truth denoising transition step $q(x_{t-1}|x_t, x_0)$ as closely as possible. Through re-parameterization tricks, the training objective turns into:

$$\theta^* = \arg\min_\theta \mathbb{E}_{x \sim p(x), \epsilon, t} \left[\|\hat{\epsilon}_{\boldsymbol{\theta}}(x_t, t) - \epsilon\|_2^2 \right]. \tag{A2}$$

where $\epsilon \sim \mathcal{N}(0, \mathcal{I})$ and $t \sim \text{Uniform}(1, \ldots, T)$. This objective is equivalent to learn the score function from the score-based perspective of diffusion models. Through Tweedie's formula, the score function for the smoothed density is proportional to the predicted noise for x_t, $\nabla_{x_t} \log p(x_t) = -\dfrac{\hat{\epsilon}_{\boldsymbol{\theta}}(x_t)}{\sqrt{1 - \bar{\alpha}_t}}$. The sampling process of diffusion model initiates with a random noise $x_T \sim \mathcal{N}(0, 1)$ and updates it iteratively with the predicted score.

Text-to-image (T2I) diffusion models. T2I diffusion models aim to control the semantics of generated data through textual conditioning information y. Therefore, it learns conditional probability distribution $p(x|c) = p(c|x)p(x)/p(c)$, where $c := \mathcal{T}(y)$ is the textual embedding of textual input y and $\mathcal{T}$ denotes the text encoder. To avoid training an external classifier to provide the guidance term $p(c|x)$, [42] proposes Classifier-free Guidance (CFG) to elegantly control how much the learned conditional model cares about the conditioning information with a simple scaler γ:

$$\nabla_x \log p(x_t|c) = \gamma \nabla_x \log p(x_t|c) + (1 - \gamma)\nabla_x \log p(x_t). \tag{A3}$$

Re-targetting based concept erasure methods. A common way of unlearning is to align the behavior on an *erasing concept* with that of an *anchor concept* by the pre-trained model θ_P to alter its semantics. The anchor concept should be distinct from the erasing one, but not that divergent to severely hurts model performance. However, altering the semantics unavoidably confuses model performance on non-malicious *retain concepts*, so that additional maintenance should be applied to patch up the degradation. Mathematically, these methods follow the following framework:

$$\min_\theta \quad \underbrace{\left\| \Phi_{\theta_U}(x_t^f|c_\text{erase}) - \Phi_{\theta_P}(x_t^f|c_\text{anchor}) \right\|_2^2}_{\text{Unlearning Term} \mathcal{L}_\text{U}} + \underbrace{\left\| \Phi_{\theta_U}(x_t^r|c_\text{retain}) - \Phi_{\theta_P}(x_t^r|c_\text{retain}) \right\|_2^2}_{\text{Retaining Term} \mathcal{L}_\text{R}} \tag{A4}$$

Different methods focus on manipulating different responses $\Phi_\theta(\cdot)$ of the diffusion models, mainly including three types: (1) the input cross-attention text-embedding projection, $\Phi_\theta(x_t|c) = W^T C$; (2) the intermediate cross-attention maps, $\Phi_\theta(x_t|c) = \mathcal{A}_\theta(x_t|c)$; (3) the output predicted noise $\Phi_\theta(x_t|c) = \epsilon_\theta(x_t, t, c)$. In this paper, we focus on manipulating the output predicted noise due to its straightforward connection with the information-theoretic view of diffusion models.

B Information-Theoretic Diffusion Model

B.1 Exact Log-likelihood Estimation by the Pre-trained Diffusion Model

The pre-trained diffusion model behaves as a noisy channel capable of denoising Gaussian noise perturbed samples, denoted as $\boldsymbol{x}_\alpha \equiv \sqrt{\sigma(\alpha)}\boldsymbol{x} + \sqrt{\sigma(-\alpha)}\boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbb{I})$ and α represents log SNR. Guo et al. [43] demonstrated that the *information* implicit in this Gaussian noise channel, denoted as $I(X; X_\alpha)$, is *exactly* the mean square error for optimal signal reconstruction. Mathematically, such information and its corresponding point-wise generalization are expressed as :

$$\frac{d}{d\alpha} I(\boldsymbol{x}; \boldsymbol{x}_\alpha) = \text{1/2} \, \text{mmse}(\alpha), \quad \frac{d}{d\alpha} D_\text{KL}[p(\boldsymbol{x}_\alpha|\boldsymbol{x})||p(\boldsymbol{x}_\alpha)] = \text{1/2} \, \text{mmse}(\boldsymbol{x}, \alpha). \tag{A5}$$

where $p(\boldsymbol{x}_\alpha) = \int p(\boldsymbol{x}_\alpha|\boldsymbol{x})p(\boldsymbol{x})d\boldsymbol{x}$ is the marginal output distribution of sample and $\text{mmse}(\boldsymbol{x}, \alpha) \equiv \mathbb{E}_{p(\boldsymbol{x}_\alpha|\boldsymbol{x})}\left[\|\boldsymbol{x} - \hat{\boldsymbol{x}}^*(\boldsymbol{x}_\alpha, \alpha)\|^2\right]$ is the pointwise MMSE. This provides an interpretation of *information-theoretic* quantities with the *estimation of optimal denoisers*.

The pre-trained diffusion model establishes a transition path between true data distribution and the standard Gaussian distribution. Kong et al. [19] apply thermodynamic variational inference along this path to recover the log likelihood for the data distribution. They consider sending samples from either the data distribution $p(\boldsymbol{x}_\alpha)$ or a standard Gaussian $p_G(\boldsymbol{x}_\alpha) = \mathcal{N}(0, I)$ through the Gaussian noise channel. The marginal output distribution with Gaussian input is $p_G(\boldsymbol{x}_\alpha) = \int p(\boldsymbol{x}_\alpha|\boldsymbol{x})p_G(\boldsymbol{x})d\boldsymbol{x}$ and corresponding MMSE for this Gaussian channel is $\text{mmse}_G(\alpha)$. Then they define the point-wise gap function $f(\boldsymbol{x}, \alpha)$ as

$$f(\boldsymbol{x}, \alpha) \equiv D_{\text{KL}}\left[p(\boldsymbol{x}_\alpha|\boldsymbol{x})||p_G(\boldsymbol{x}_\alpha)\right] - D_{\text{KL}}\left[p(\boldsymbol{x}_\alpha|\boldsymbol{x})||p(\boldsymbol{x}_\alpha)\right]. \tag{A6}$$

In the limit of zero SNR, we have $\lim_{\alpha \to 0} f(\boldsymbol{x}, \alpha) = 0$. In the high SNR limit, [19] prove that $\lim_{\alpha \to \infty} f(\boldsymbol{x}, \alpha) = \log \frac{p(\boldsymbol{x})}{p_G(\boldsymbol{x})}$. Combining this with Eq. A5, the exact log likelihood is:

$$-\log p(\boldsymbol{x}) = -\log p_G(\boldsymbol{x}) - \int_0^\infty d\alpha \frac{d}{d\alpha} f(\boldsymbol{x}, \alpha)$$

$$= -\log p_G(\boldsymbol{x}) - \frac{1}{2} \int_0^\infty d\alpha \left(\text{mmse}_G(\boldsymbol{x}, \alpha) - \text{mmse}(\boldsymbol{x}, \alpha)\right)$$

$$= -\frac{1}{2} \int_0^\infty \mathbb{E}_{p(\boldsymbol{\epsilon})}\left[\|\boldsymbol{\epsilon} - \hat{\boldsymbol{\epsilon}}_\alpha(\boldsymbol{x}_\alpha)\|^2\right] d\alpha + \text{const.} \tag{A7}$$

This represents density in terms of a Gaussian density and a correction that measures how much better we can denoise the target distribution than we could with an optimal denoiser for Gaussian source data. Furthermore, the density estimation is optimal for noise reconstruction error at different noise levels.

B.2 Non-negative Mutual Information

By substituting $p(x)$ and $p(x|)$ with Eq. 2 and Eq. 3, we have the mutual information $\mathcal{I}(\boldsymbol{x}, y)$ estimated by the pre-trained diffusion model as follows:

$$\mathcal{I}(\boldsymbol{x}, y) = [\log p(\boldsymbol{x}|y) - \log p(\boldsymbol{x})] = \left[\frac{1}{2}\int_0^\infty \mathbb{E}_{\boldsymbol{\epsilon}}\left[\|\boldsymbol{\epsilon} - \hat{\boldsymbol{\epsilon}}_{\theta_P}(\boldsymbol{x}_\alpha)\|_2^2 - \|\boldsymbol{\epsilon} - \hat{\boldsymbol{\epsilon}}_{\theta_P}(\boldsymbol{x}_\alpha|y)\|_2^2\right] d\alpha\right] \tag{A8}$$

However, above estimation is not lower bounded, which warns that model performance might break down if the optimization does not stop timely. By expanding and rearranging Eq. A8, we can have the second term in Eq. A9 equals to zero, benefiting from the orthogonality principle [30]:

$$\mathcal{I}(\boldsymbol{x}, y) = [\overbrace{\left[\frac{1}{2}\int_0^\infty \mathbb{E}_{\boldsymbol{\epsilon}}\left[\|\hat{\boldsymbol{\epsilon}}_\alpha(\boldsymbol{x}_\alpha) - \hat{\boldsymbol{\epsilon}}_\alpha(\boldsymbol{x}_\alpha|y)\|_2^2\right] d\alpha\right]}^{\mathbb{I}^+(\boldsymbol{x};y)}$$

$$+ 2\mathbb{E}_{p(y)}[\frac{1}{2}\int_0^\infty \underbrace{\mathbb{E}_{p(\boldsymbol{x}|y),\boldsymbol{\epsilon}}\left[(\hat{\boldsymbol{\epsilon}}_\alpha(\boldsymbol{x}_\alpha) - \hat{\boldsymbol{\epsilon}}_\alpha(\boldsymbol{x}_\alpha|y)) \cdot (\hat{\boldsymbol{\epsilon}}_\alpha(\boldsymbol{x}_\alpha|y) - \boldsymbol{\epsilon})\right]}_{\equiv \mathcal{O}} d\alpha] \tag{A9}$$

Compared with Eq. A8, there are two advantages of Eq. A9: (1) *Non-negativity*: It is non-negative, avoiding breakdown of model performance during minimization. (2) *Low Variance*: It avoids sampling random Gaussian noises $\boldsymbol{\epsilon}$, which will introduce variance to estimations and optimizations. Also, this non-negative expression of the mutual information has been indicated to be effective in attributing the semantics in the generated image to corresponding textual words in prompts [31, 32], locating the emergence of specific semantics during generation.

C Theoretical Derivation

C.1 Gradient of Mutual Information Minimization

We denote the mutual information at timestep t as $\mathcal{I}_t(\boldsymbol{x}, y) := \frac{1}{2}\|\hat{\boldsymbol{\epsilon}}_{\theta_P}(\boldsymbol{x}_t|y) - \hat{\boldsymbol{\epsilon}}_{\theta_P}(\boldsymbol{x}_t)\|_2^2$, and then minimizing the unlearning objective in Eq. 1 corresponds to minimize each $\mathcal{I}_t(\boldsymbol{x}, y)$. In diffusion model, we have $\boldsymbol{x}_t = \sqrt{\bar{\alpha}_t}\boldsymbol{x} + \sqrt{1 - \bar{\alpha}_t}\boldsymbol{\epsilon}, \boldsymbol{x} = \dfrac{\tilde{\boldsymbol{x}}_t - \sqrt{1 - \bar{\alpha}_t}\hat{\boldsymbol{\epsilon}}_{\theta_U}(\tilde{\boldsymbol{x}}_t|y)}{\sqrt{\bar{\alpha}_t}}$. Therefore, the gradient of

$\mathcal{I}_t(\boldsymbol{x}, y)$ w.r.t. the unlearned model θ_U is as the following:

$$
\begin{aligned}
\frac{\partial \mathcal{I}_t(\boldsymbol{x}, y)}{\partial \theta_U} &= \frac{\partial \mathcal{I}_t(\boldsymbol{x}, y)}{\partial \hat{\epsilon}_{\theta_P}} \cdot \frac{\partial \hat{\epsilon}_{\theta_P}}{\partial \boldsymbol{x}_t} \cdot \frac{\partial \boldsymbol{x}_t}{\partial \boldsymbol{x}} \cdot \frac{\partial \boldsymbol{x}}{\partial \theta_U} \\
&= \mathbb{E}_{\boldsymbol{\epsilon}} \left[(\hat{\boldsymbol{\epsilon}}_{\theta_P}(\boldsymbol{x}_t|y) - \hat{\boldsymbol{\epsilon}}_{\theta_P}(\boldsymbol{x}_t)) \cdot \left(\frac{\partial \hat{\boldsymbol{\epsilon}}_{\theta_P}(\boldsymbol{x}_t|y)}{\partial \boldsymbol{x}_t} - \frac{\partial \hat{\boldsymbol{\epsilon}}_{\theta_P}(\boldsymbol{x}_t)}{\partial \boldsymbol{x}_t} \right) \cdot \sqrt{\bar{\alpha}_t} \cdot -\frac{\sqrt{1 - \bar{\alpha}_t}}{\sqrt{\bar{\alpha}_t}} \cdot \frac{\partial \hat{\epsilon}_{\theta_U}(\tilde{\boldsymbol{x}}_t|y)}{\partial \theta_U} \right] \\
&= \mathbb{E}_{\boldsymbol{\epsilon}} \left[-\sqrt{1 - \bar{\alpha}_t} \cdot (\hat{\boldsymbol{\epsilon}}_{\theta_P}(\boldsymbol{x}_t|y) - \hat{\boldsymbol{\epsilon}}_{\theta_P}(\boldsymbol{x}_t)) \cdot \left(\frac{\partial \hat{\boldsymbol{\epsilon}}_{\theta_P}(\boldsymbol{x}_t|y)}{\partial \boldsymbol{x}_t} - \frac{\partial \hat{\boldsymbol{\epsilon}}_{\theta_P}(\boldsymbol{x}_t)}{\partial \boldsymbol{x}_t} \right) \cdot \frac{\partial \hat{\epsilon}_{\theta_U}(\tilde{\boldsymbol{x}}_t|y)}{\partial \theta_U} \right]. \\
&= \mathbb{E}_{\boldsymbol{\epsilon}} \left[w(t) \cdot \underbrace{(\hat{\boldsymbol{\epsilon}}_{\theta_P}(\boldsymbol{x}_t|y) - \hat{\boldsymbol{\epsilon}}_{\theta_P}(\boldsymbol{x}_t))}_{\text{Pre-trained CFG}} \cdot \underbrace{\left(\frac{\partial \hat{\boldsymbol{\epsilon}}_{\theta_P}(\boldsymbol{x}_t|y)}{\partial \boldsymbol{x}_t} - \frac{\partial \hat{\boldsymbol{\epsilon}}_{\theta_P}(\boldsymbol{x}_t)}{\partial \boldsymbol{x}_t} \right)}_{\text{Pre-trained U-Net Jacobian}} \cdot \underbrace{\frac{\partial \hat{\epsilon}_{\theta_U}(\tilde{\boldsymbol{x}}_t|y)}{\partial \theta_U}}_{\text{Unlearned U-Net Gradient}} \right].
\end{aligned}
$$
(A10)

The obtained gradient in Eq. 6 can be decomposed into 3 components: (1) CFG term of the pre-trained diffusion model, (2) U-Net Jacobian of the pre-trained diffusion model, and (3) U-Net gradient of the unlearned model.

C.2   Gradient of KL Divergence Minimization

In this part, we elucidate the equivalence between the approximated mutual information minimization and KL divergence minimization in Eq. 8. This could be verified by back-propagating Eq. 8 to the unlearned model:

$$
\begin{aligned}
\frac{\partial \mathcal{D}_{\theta_P}^{\mathrm{KL}}(\boldsymbol{x})}{\partial \theta_U} &= \frac{\partial \mathcal{D}_{\theta_P}^{\mathrm{KL}}(\boldsymbol{x})}{\partial \boldsymbol{x}_t} \cdot \frac{\partial \boldsymbol{x}_t}{\partial \boldsymbol{x}} \cdot \frac{\partial \boldsymbol{x}}{\partial \theta_U} \\
&= \mathbb{E}_{\boldsymbol{\epsilon}} \left[\left(\frac{\partial \log p(\boldsymbol{x}_t)}{\partial \boldsymbol{x}_t} - \frac{\partial \log p(\boldsymbol{x}_t|y)}{\partial \boldsymbol{x}_t} \right) \cdot \frac{\partial \boldsymbol{x}_t}{\partial \boldsymbol{x}} \cdot \frac{\partial \boldsymbol{x}}{\partial \theta_U} \right] \\
&= \mathbb{E}_{\boldsymbol{\epsilon}} \left[-\frac{1}{\sqrt{1 - \bar{\alpha}_t}} \cdot -(\hat{\epsilon}_{\theta_P}(\boldsymbol{x}_t|y) - \hat{\epsilon}_{\theta_P}(\boldsymbol{x}_t)) \cdot \sqrt{\bar{\alpha}_t} \cdot -\frac{\sqrt{1 - \bar{\alpha}_t}}{\sqrt{\bar{\alpha}_t}} \frac{\partial \hat{\epsilon}_{\theta_U}(\tilde{\boldsymbol{x}}_t|y)}{\partial \theta_U} \right] \\
&= \mathbb{E}_{\boldsymbol{\epsilon}} \left[-(\hat{\epsilon}_{\theta_P}(\boldsymbol{x}_t|y) - \hat{\epsilon}_{\theta_P}(\boldsymbol{x}_t)) \cdot \frac{\partial \hat{\epsilon}_{\theta_U}(\tilde{\boldsymbol{x}}_t|y)}{\partial \theta_U} \right]. \\
&= \frac{1}{\sqrt{1 - \bar{\alpha}_t}} \cdot \frac{\partial \mathcal{I}_t(\boldsymbol{x}, y)}{\partial \theta_U}
\end{aligned}
$$
(A11)

This is proportional to the gradient of the U-Net omitted mutual information minimization in Eq. 7 .

C.3   Equivalence between KL Divergence Minimization and Noise Prediction Alignment

In this part, we elucidate the equivalence between KL divergence minimization and noise prediction alignment in Eq. 11:

$$
\mathrm{KL}(q_{\theta_U}(\boldsymbol{x}_t|y) \| q_{\theta_U}^*(\boldsymbol{x}_t|y)) = \mathbb{E}_{\boldsymbol{\epsilon}} \left[\| \hat{\epsilon}_{\theta_U}(\boldsymbol{x}_t|y) - \hat{\epsilon}_{\theta_P}(\boldsymbol{x}_t) \|_2^2 \right]
$$
(A12)

This could be verified by back-propagating the LHS and RHS respectively to examine their gradient:

$$
\begin{aligned}
\frac{\partial \mathrm{KL}(q_{\theta_U}(\boldsymbol{x}_t|y) \| q_{\theta_U}^*(\boldsymbol{x}_t|y))}{\partial \theta_U} &= \mathbb{E}_{\boldsymbol{\epsilon}} \left[\left(\frac{\partial \log p(\boldsymbol{x}_t|y)}{\partial \boldsymbol{x}_t} - \frac{\partial \log p(\boldsymbol{x}_t)}{\partial \boldsymbol{x}_t} \right) \cdot \frac{\partial \boldsymbol{x}_t}{\partial \boldsymbol{x}} \cdot \frac{\partial \boldsymbol{x}}{\partial \theta_U} \right] \\
&= \mathbb{E}_{\boldsymbol{\epsilon}} \left[-\frac{1}{\sqrt{1 - \bar{\alpha}_t}} \cdot (\hat{\epsilon}_{\theta_P}(\boldsymbol{x}_t|y) - \hat{\epsilon}_{\theta_P}(\boldsymbol{x}_t)) \cdot \sqrt{\bar{\alpha}_t} \cdot -\frac{\sqrt{1 - \bar{\alpha}_t}}{\sqrt{\bar{\alpha}_t}} \frac{\partial \hat{\epsilon}_{\theta_U}(\tilde{\boldsymbol{x}}_t|y)}{\partial \theta_U} \right] \\
&= \mathbb{E}_{\boldsymbol{\epsilon}} \left[(\hat{\epsilon}_{\theta_P}(\boldsymbol{x}_t|y) - \hat{\epsilon}_{\theta_P}(\boldsymbol{x}_t)) \cdot \frac{\partial \hat{\epsilon}_{\theta_U}(\tilde{\boldsymbol{x}}_t|y)}{\partial \theta_U} \right].
\end{aligned}
$$
(A13)

The gradient of MSE (RHS) is :

$$
\frac{\partial \mathbb{E}_{\boldsymbol{\epsilon}} \left[\| \hat{\epsilon}_{\theta_U}(\boldsymbol{x}_t|y) - \hat{\epsilon}_{\theta_P}(\boldsymbol{x}_t) \|_2^2 \right]}{\partial \theta_U} = \mathbb{E}_{\boldsymbol{\epsilon}} \left[(\hat{\epsilon}_{\theta_U}(\boldsymbol{x}_t|y) - \hat{\epsilon}_{\theta_P}(\boldsymbol{x}_t)) \cdot \frac{\partial \hat{\epsilon}_{\theta_U}(\boldsymbol{x}_t|y)}{\partial \theta_U} \right]
$$
(A14)

where $\hat{\epsilon}_{\theta_P}(\boldsymbol{x}_t)$ is distillation teacher and detached from computation graph. Therefore, we have the equivalent minimization in Eq. 11.

D   UnlearnCanvas Visualizations

In this section, we visualize the generations after unlearning to demonstrate the qualitative performance of different concept erasing methods.

D.1   Style and Object Unlearn

"Cartoon" style unlearning performance comparisons The original UnlearnCanvas benchmark provides abundant generation examples of all the 9 benchmarking methods in a case study of unlearning the "Cartoon" style. Both the successful and failure cases are demonstrated in the context of unlearning effectiveness, in-domain retainability, and cross-domain retainability. For the convenience of comparison, we copy it in **Fig.** A3. It is obvious that the original 9 benchmarking methods demonstrate less than satisfactory retainability, with evident generation failure and quality degradation of the remaining concepts. Furthermore, we visualize the generations of "Cartoon" unlearned model by SDD and MiM-MU in **Fig.** A2. Although SDD successfully prevents generating "Cartoon" style, it fails to generate "Human" when it is requested along with "Cartoon" style. Moreover, it fails to well generate "Sketch" and "Watercolore" style and "Butterfly" object. In contrast, MiM-MU effectively erases the undesired style and perfectly generates the other demanded concept. Meanwhile, it is able to generate the remaining concepts nearly as well as the pre-trained diffusion model.

"Jellyfish" object unlearning performance comparisons. We provide visualizations of "Jellyfish" unlearned models by different unlearning methods in **Fig.** A1. The conclusion still holds true that existing methods fail to achieve satisfying retainability, while MiM-MU exhibits minimal degradation on model utility across various the remaining concepts.
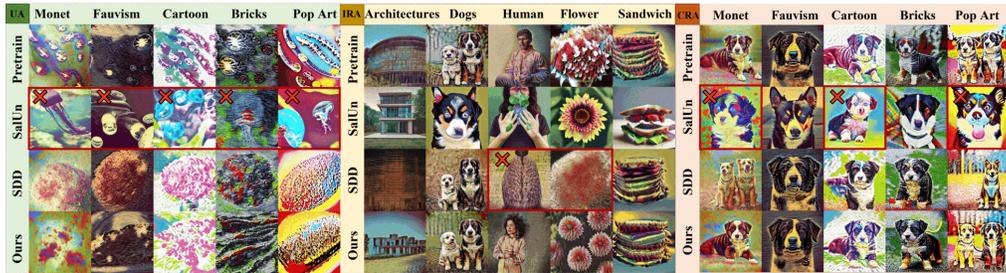

Figure A1: Visualization of the unlearning performance of MiM-MU and SDD on "Jellyfish" object. The organization this figure follows **Fig.** 2. SalUn fails to erase "Jellyfish" completely, and the painting styles of the other artists(*e.g.*, "Monet", "Cartoon", and "Pop Art") exhibit obvious degradation when compared with the pre-trained diffusion model. In contrast, MiM-MU can effectively prevent generations of "Jellyfish" meanwhile preserving high-quality generations on other concepts.
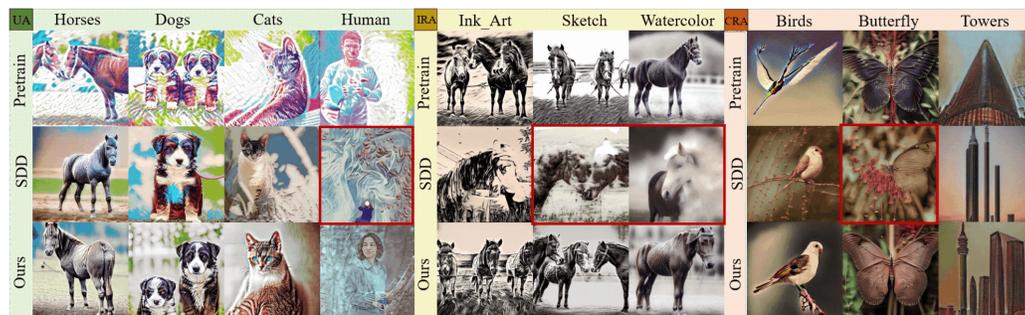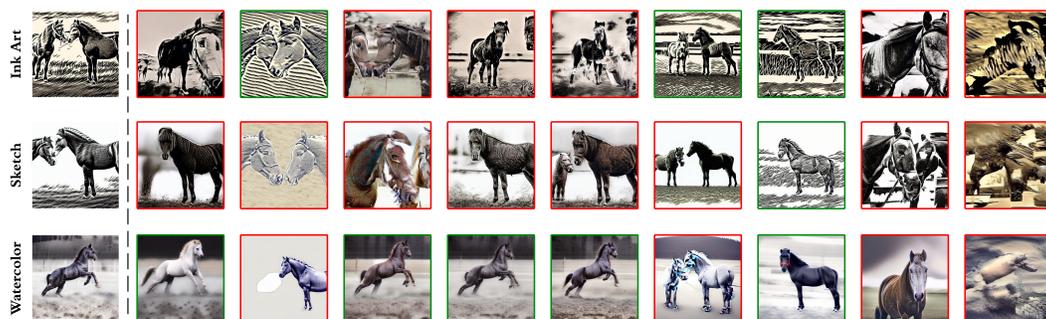

Figure A2: Visualization of the unlearning performance of MiM-MU and SDD on "Cartoon" style. The organization this figure follows **Fig.** 2. Both SDD and MiM-MU effectively erase "Cartoon" style. SDD fails to preserve generating "Sketch" and "Watercolor" styles and "Butterfly" object, while ours demonstrates pretty retainability.

Unlearning Target Concept – Cartoon style

Unlearning Effectiveness Evaluation: Test Prompt Template: "An image of {object} in Cartoon style"

In-Domain Retainability Evaluation: Test Prompt Template: "An image of Horses in {style} style"

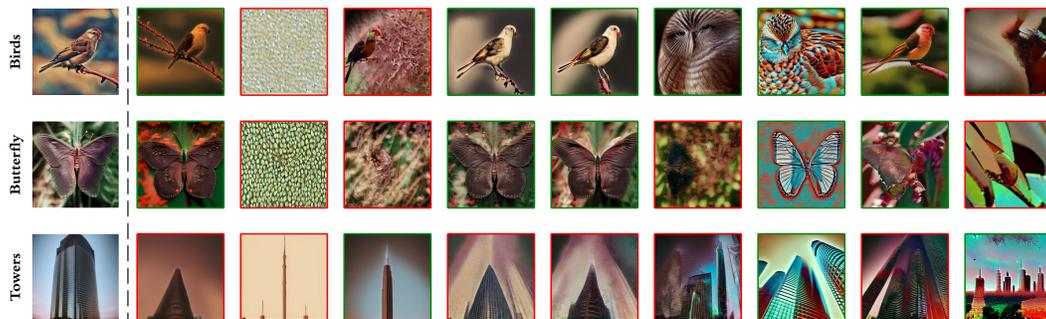**Cross-Domain Retainability Evaluation: Test Prompt Template: "An image of {object}."**

Figure A3: visualization of the unlearning performance of different methods on the task of style unlearning. Three text prompt templates are used to evaluate the unlearning effectiveness, in-domain retainability, and cross-domain retainability of each method. Images with green frame denote desirable results, while the ones with red frame denote unlearning or retaining failures.

5 Different artist styles unlearning by MiM-MU. We demonstrate the generation quality after unlearning 5 different styles by MiM-MU in **Fig.** A4. For each requested erasure, the MiM-MU unlearned model successfully removes the undesired painting style in UA. In IRA and CRA, MiM-MU well preserve the original generative capability of the remaining styles and objects.

Figure A4: Visualization of the unlearning performance of different styles by MiM-MU. Each object of the pretrained model in UA is combined with above 5 different styles successively to demonstrate the requested erasing style. From top to bottom, each row is the unlearning and retaining generations after erasing "Crayon", "Ukiyoe", "Magic Cube" (abbreviated as "Mag Cub" in figure), "Pencil Drawing" ("Pencil Draw"), "Abstractionism" ('Abstract'), respectively. The diagonal images (which are highlighted with green frameworks) in IRA indicate "Dogs" in current erasing style, therefore, it should not contain any painting style.

5 Different objects unlearning by MiM-MU. We demonstrate the generation quality after unlearning 5 different objects by MiM-MU in **Fig.** A5. For each requested erasure, the MiM-MU unlearned model successfully removes the undesired objects in UA. In IRA and CRA, MiM-MU well preserve the original generative capability of the remaining styles and objects.

Figure A5: Visualization of the unlearning performance of different objects by MiM-MU. Each object of the pretrained model in UA is combined with above 5 different styles successively to demonstrate the requested erasing style. From top to bottom, each row is the unlearning and retaining generations after erasing "Architecture", "Dogs", "Human" , "Flower", "Sandwich", respectively. The diagonal images (which are highlighted with green frameworks) in IRA indicate the erasing object in "Seed Image" style , therefore it should be empty if unlearning is successful. The third row of "Dogs" in CRA should be empty as well.

D.2 Sequential Unlearning

We visualize the unlearning and retaining performance of each unlearning request during sequential unlearning in **Fig.** A6, **Fig.** A7 and **Fig.** A8. The **first column** of each figure demonstrates generations of the **pre-trained** diffusion model. From the second column to the last column, each column indicates generations of the **unlearned** model after $\mathcal{T}_1 \sim \mathcal{T}_6$ unlearning request ("Abstractionism" $(\mathcal{T}_1)$, "Byzantine" $(\mathcal{T}_2)$, "Cartoon" $(\mathcal{T}_3)$, "Cold Warm" $(\mathcal{T}_4)$, "Ukiyoe" $(\mathcal{T}_5)$, and "Van Gogh" $(\mathcal{T}_6)$) respectively. In UA, the erasing style should be *different* from the pre-trained diffusion model since the erasure is demanded. In IRA and CRA, the generation should be as similar as the pre-trained diffusion model throughout the sequential unlearning.

Fig. A6 demonstrate the unlearning performance of each unlearning request. From top to bottom, each row stands for the style of each concept, *e.g.*, **"Abstractionism"** , **"Byzantine"** , **"Cartoon"** , **"Cold Warm"**, **"Ukiyoe"**, and **"Van Gogh"** . In **Fig.** A7, from top to bottom, each row stands for a style of the remaining concepts, *e.g.*, "**Dapple**", "**Warm Smear**", "**Glowing Sunset**", "**Color Fantasy**", and "**Neon Lines**" respectively. In **Fig.** A8, we demonstrate the cross-domain retainability with 6 objects, *e.g.*, top to bottom are "**Frogs**", "**Architectures**", "**Waterfalls**", "**Flowers**", and "**Sea**". As can be seen, MiM-MU can effectively remove the undesired style once the erasure is requested, and the erased style never reappears during subsequent unlearning requests. And MiM-MU is able to generate the remaining styles and objects with high quality, *i.e.*, nearly identical to that of the pre-trained model.
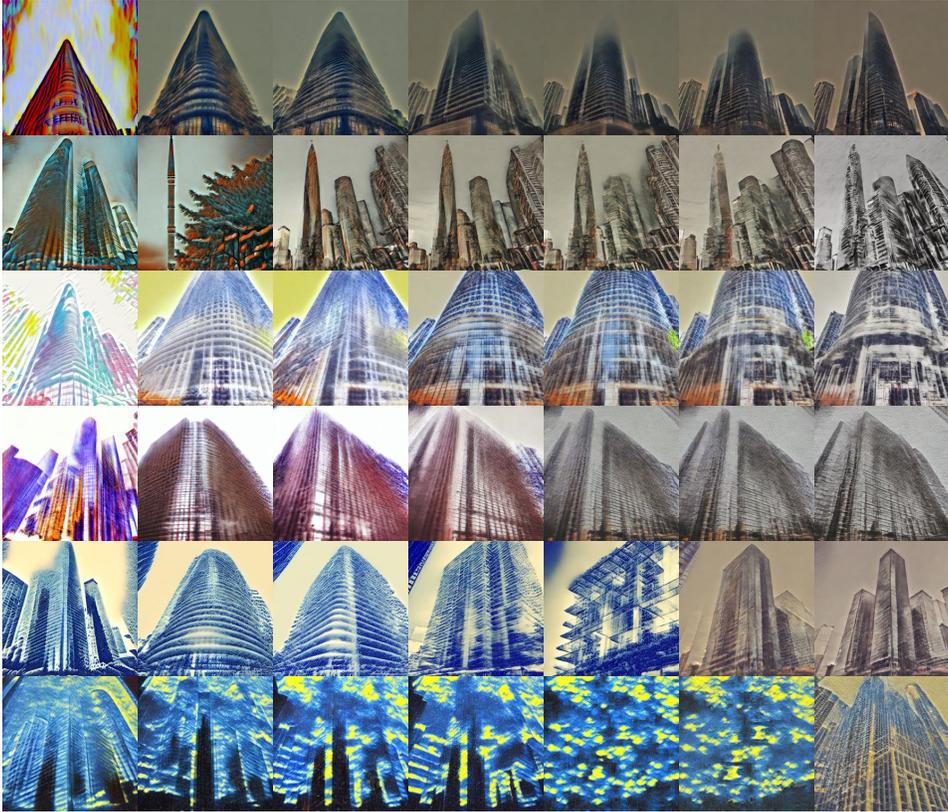

Figure A6: Visualizations of other In-domain concept generations during the sequential unlearning. From left to right, each column in sequence is the pre-trained model, followed by the models after unlearning "Abstractionism" $(\mathcal{T}_1)$, "Byzantine" $(\mathcal{T}_2)$, "Cartoon" $(\mathcal{T}_3)$, "Cold Warm" $(\mathcal{T}_4)$, "Ukiyoe" $(\mathcal{T}_5)$, and "Van Gogh" $(\mathcal{T}_6)$. From top to bottom, each row is the models' generation of "A **Towers** in **Artist** style.", where the **Artist** is the same sequence of the unlearning request. The results demonstrate that MiM-MU removes the generative capability for requested artist style thoroughly without any reemergence, indicating an exhaustive and reliable erasure operation, *i.e.*, which is complete and does not spoil previous unlearning efforts.
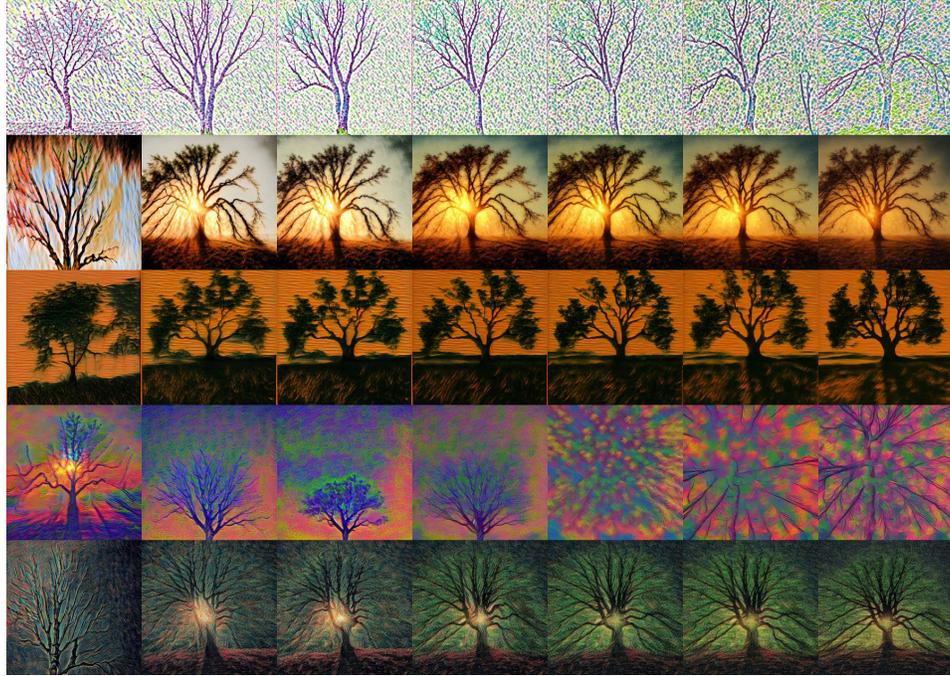
Figure A7: Visualizations of other In-domain concepts generations during the sequential unlearning. From left to right, each column in sequence is the pre-trained model, followed by the models after unlearning "Abstractionism" ($\mathcal{T}_1$), "Byzantine" ($\mathcal{T}_2$), "Cartoon" ($\mathcal{T}_3$), "Cold Warm" ($\mathcal{T}_4$), "Ukiyoe" ($\mathcal{T}_5$), and "Van Gogh" ($\mathcal{T}_6$). From top to bottom, each row is the models' generation of "A **Tree** in **Artist** style." where the **Artist** from top to bottom are "**Dapple**", "**Warm Smear**", "**Glowing Sunset**", "**Color Fantasy**", and "**Neon Lines**" respectively. The results demonstrate that our method can effectively preserve the generative capability for other artist styles during the sequential unlearning process.
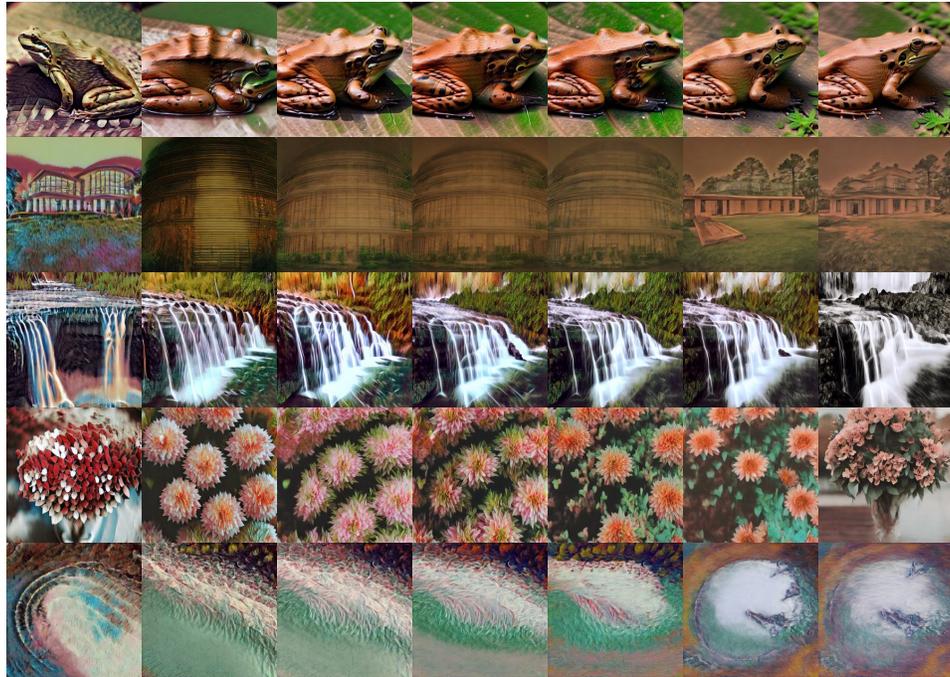

Figure A8: Visualizations of cross-domain concept (CRA) generations during the sequential unlearning process. From left to right, each column represents the pre-trained model, followed by the models after unlearning "Abstractionism" ($\mathcal{T}_1$), "Byzantine" ($\mathcal{T}_2$), "Cartoon" ($\mathcal{T}_3$), "Cold Warm" ($\mathcal{T}_4$), "Ukiyoe" ($\mathcal{T}_5$), and "Van Gogh" ($\mathcal{T}_6$). From top to bottom, each row depicts the models' generation of "A **Object** in **Seed Images** style." where the **Object** from top to bottom are "**Frogs**", "**Architectures**", "**Waterfalls**", "**Flowers**", and "**Sea**" respectively. The results indicate that our method successfully retains the generative ability for cross-domain concepts throughout the sequential unlearning process.

E Experiment Details

E.1 Training Details

For the forgetting dataset $\mathcal{X}_f$, we construct it by combining the erasing concept with each concept in the cross concept domain. For example, for unlearning "Van Gogh" style, we combine it with 20 objects respectively, *e.g.*, "A Cat in Van Gogh style.". For each combination, we use 3 images provided in the UnlearnCanvas dataset. For each unlearning request, the forget prompt y is "{Artists} Style" for style unlearning (*e.g.*, "Van Gogh style") and "{object}" for object unlearning (*e.g.*, "Dogs"). Then we fine-tune the cross-attention parameters of U-Net with our proposed loss and constructed $\mathcal{X}_f$ for 30 epochs, at a learning rate of 1×10^{-5} and batch size 1 for all the experiments. For evaluation, we evaluate with 5 seeds (188, 288, 388, 488 and 588) for all the experiments. The COCO-10k dataset used in this paper is downloaded from `https://github.com/OPTML-Group/AdvUnlearn`. All the experiments are conducted on RTX 4090 GPU.

E.2 Baselines

In particular, we adopt the following training settings to reproduce baseline methods:

- SalUn [1]: The code source of SalUn [1] is `https://github.com/OPTML-Group/UnlearnCanvas`. We follow the implementation of SalUn in UnlearnCanvas benchmark. In our implementations, we run the weight saliency analysis with 1 epoch and unlearning stage with 10 epochs. The weight saliency mask ratio is 0.5 and the learning rate is $1e-5$.
- SDD [23]: The code source of SDD [23] is `https://github.com/nannullna/safe-diffusion`. The forgetting concept for SDD is the corresponding name of artist (*e.g.*, "Van Gogh") and object (*e.g.*, "Dogs"). We run SDD for 1k5 steps with a learning rate of $1e-5$, which is in line with the original implementation.

E.3 Data Availability Statement

The datasets used in this study are all publicly available and hosted on open-access repositories:

1. The UnlearnCanvas dataset is available at `https://huggingface.co/datasets/OPTML-Group/UnlearnCanvas`.
2. Stanford Dogs dataset is available at `http://vision.stanford.edu/aditya86/ImageNetDogs/`.
3. Oxford 102 Flowers is available at `https://www.robots.ox.ac.uk/~vgg/data/flowers/102/`.
4. CUB-200 is available at `https://www.vision.caltech.edu/datasets/cub_200_2011/`.
5. The COCO-1k subset used in this work is downloaded from the COCO-10k dataset, which is available at `https://github.com/OPTML-Group/AdvUnlearn.`.

F Limitation

In this paper, we focus on addressing the poor retainability of concept erasure in diffusion models. We propose a nuanced erasure, MiM-MU, to remove the knowledge of a concept from model parameters by minimizing the mutual information between the textual concept and semantic images. Our approach aims to effectively erase specific concepts while preserving the overall utility of the pre-trained model without requiring compensatory adjustments. However, several limitations remain open for future exploration. For computational feasibility, our mutual information minimization loss omits the UNet Jacobian term, which introduces approximations. These approximations may unintentionally weaken the ability to discriminate certain concepts during minimization, potentially impacting the unlearning effectiveness. Additionally, the whole sampling distribution of diffusion model is vast. In this paper, we focus on degrading the conditional sampling distribution of a specified concept. Future work should explore more accurate and efficient approximations and investigate how to comprehensively locate and degrade all the risky sampling distributions that are possible to generate an undesired concept.