

Unlearning Concepts in Diffusion Model via Concept Domain Correction and Concept Preserving Gradient

Yongliang Wu^{1*}, Shiji Zhou^{2*†}, Mingzhuo Yang³, Lianzhe Wang², Heng Chang²,
Wenbo Zhu⁴, Xinting Hu⁵, Xiao Zhou², Xu Yang¹

¹Southeast University

²Tsinghua University

³The Chinese University of Hong Kong, Shenzhen

⁴University of California, Berkeley

⁵Nanyang Technological University

yongliangwu@seu.edu.cn, zhoushiji00@gmail.com

Abstract

Text-to-image diffusion models have achieved remarkable success in generating photorealistic images. However, the inclusion of sensitive information during pre-training poses significant risks. Machine Unlearning (MU) offers a promising solution to eliminate sensitive concepts from these models. Despite its potential, existing MU methods face two main challenges: 1) limited generalization, where concept erasure is effective only within the unlearned set, failing to prevent sensitive concept generation from out-of-set prompts; and 2) utility degradation, where removing target concepts significantly impacts the model’s overall performance. To address these issues, we propose a novel concept domain correction framework named **DoCo (Domain Correction)**. By aligning the output domains of sensitive and anchor concepts through adversarial training, our approach ensures comprehensive unlearning of target concepts. Additionally, we introduce a concept-preserving gradient surgery technique that mitigates conflicting gradient components, thereby preserving the model’s utility while unlearning specific concepts. Extensive experiments across various instances, styles, and offensive concepts demonstrate the effectiveness of our method in unlearning targeted concepts with minimal impact on related concepts, outperforming previous approaches even for out-of-distribution prompts.

Code — <https://github.com/yongliang-wu/DoCo>

Introduction

The development of diffusion models for text-to-image synthesis has progressed rapidly, showcasing a remarkable ability to generate photorealistic images (Liu, Sun, and Mu 2024). This has led to the creation of applications such as Stable Diffusion (Rombach et al. 2022) and Midjourney (Borji 2022). However, the pre-training datasets often include copyrighted materials, personal photos (Carlini et al. 2023), and unsafe information (Schramowski et al. 2023), resulting in generated images that pose copyright, privacy,

and safety concerns. This situation raises significant risks of sensitive data leakage (Wu et al. 2022), directly conflicting with the growing legislative emphasis on the “right to be forgotten” (Rosen 2011).

A naive approach is to filter out sensitive information and then retrain the entire model from scratch (Rombach et al. 2022; Rao 2023; Kim et al. 2025). However, given that these generative models are typically trained on large-scale datasets (Schuhmann et al. 2022), this method can be prohibitively costly. Another strategy involves using a Safety Checker to detect the presence of inappropriate content in generated images (Rombach et al. 2022; Liu et al. 2024b; Wang et al. 2024; Gao et al. 2024c; Yoon et al. 2024). However, this approach relies heavily on the performance of the detector and is limited by the biases.

In addressing these challenges, Machine Unlearning (MU) (Xu et al. 2023; Gao et al. 2024b,a; Zhou et al. 2024) has emerged as a potentially promising solution. Fundamentally, MU is a method that enables models to unlearn the memory of sensitive information contained within their training datasets, thereby preventing the generation of images involving sensitive concepts through crafted prompts. Recently, a series of studies have been introduced that focus on the unlearning concepts in diffusion models (Gandikota et al. 2023; Zhang et al. 2023; Gandikota et al. 2024; Kumari et al. 2023; Lyu et al. 2023; Zhang et al. 2024d; Guo, Wu, and Wan 2024; Liu et al. 2023; Zhang et al. 2024a; Truong, Dang, and Le 2024; ?). For instance, ESD (Gandikota et al. 2023) fine-tunes the model to predict in the opposite direction of classifier-free guidance, effectively aligning the targeted concept with that of an empty string. ConAbl (Kumari et al. 2023) proposes linking a target concept to a pre-defined anchor concept by minimizing the L2 distance between the predicted noises for these two concepts. SPM (Lyu et al. 2023) employs a Latent Anchor approach to eliminate concepts and uses a similarity-based retention loss to preferentially weight surrogate concepts, thereby preserving non-target concepts.

However, when applying MU for concept unlearning, two primary issues arise. First, there is the challenge of generalizability: ensuring that a model completely unlearns a tar-

*These authors contributed equally.

†Corresponding author

Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

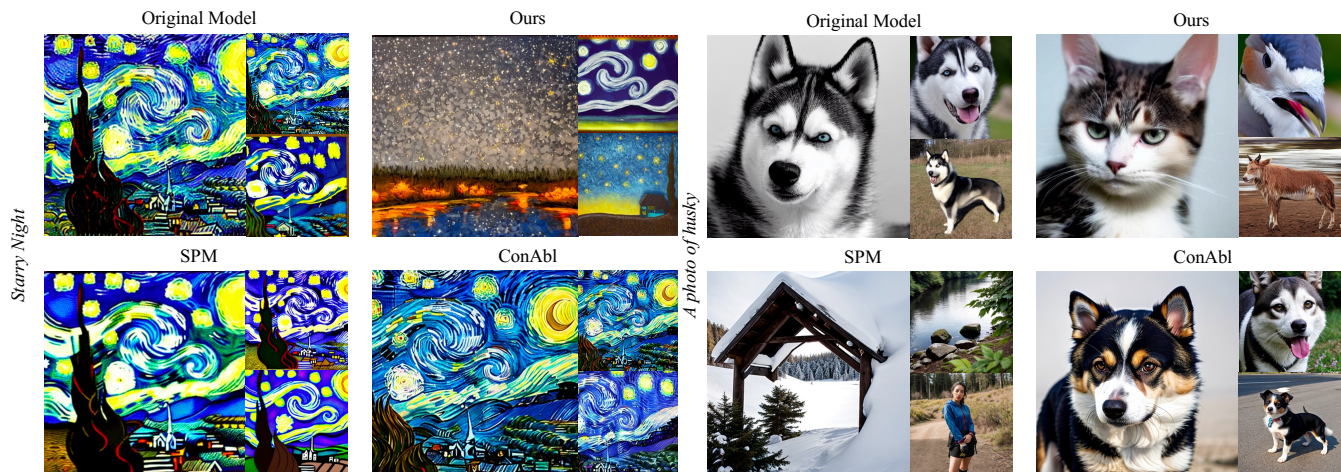


Figure 1: When transferred to strongly related out-of-distribution prompts, previous methods fail to unlearn successfully, whereas our method achieves this generalization. Left: unlearning “Van Gogh”. Right: unlearning “Dog”.

get concept with only a limited number of training samples. Second, there is the need for concept preservation: ensuring that while unlearning target concepts, the model’s utility on other concepts remains unaffected. Existing methods have not fully resolved these issues. Specifically, for unlearning generalization, some approaches focus on reducing the distance between sample pairs generated from the target and anchor concepts (Kumari et al. 2023) or minimize the value and key in the cross-attention layers between target text embeddings and anchor text embeddings (Gandikota et al. 2024). These methods are effective only for sampled or related prompts and fail to achieve complete concept elimination. For concept preservation, some methods (Kumari et al. 2023; Lyu et al. 2023; Gandikota et al. 2024) incorporate regularization strategies that, while attempting to retain other concepts, overlook the inherent contradiction between unlearning and utility preservation, thus failing to maintain overall effectiveness.

To address these challenges, we introduce a novel unlearning framework named **DoCo (Domain Correction)**. This framework employs a discriminator to simulate a membership inference attack (MIA) (Shokri et al. 2017), distinguishing between the output domains of target and anchor concepts. Through adversarial training against MIA, we align these output domains, thereby aiming for the complete elimination of the target concept in a distributional sense. To the best of our knowledge, we are the first to focus on the generalizability of unlearning concepts in text-to-image diffusion models using adversarial training. We find that DoCo performs well with certain prompts strongly associated with a concept, even if not explicitly mentioned, termed Out-of-Distribution (OOD) prompts. For instance, “Van Gogh” and “Starry Night” are linked. As shown in Figure 1, our method successfully unlearns “Starry Night” and “Husky” without it being in the training set, outperforming other methods.

To tackle the issue of concept preservation, we propose a concept-preserving gradient approach. This approach trims

the conflicting parts of the unlearning gradient with utility regularization, ensuring that each optimization iteration does not compromise the model’s utility. This allows us to achieve both concept unlearning and preservation simultaneously. Through comprehensive experiments across a wide range of instances, styles, and offensive concepts, we demonstrate that our method effectively unlearns targeted concepts while having a minimal impact on closely related surrounding concepts that should be preserved.

Our contributions are summarized as follows: (1) We formulate a novel framework DoCo for unlearning concepts in diffusion models. By employing a distribution classifier in an adversarial manner, we align the output domains of the target concepts to be forgotten with those of the anchor concepts. This approach enhances the generalizability of the unlearning effect. (2) We propose a concept-preserving gradient surgery method. By trimming the contradictory parts between the unlearning gradient and the retraining gradient, our method aims to optimize the unlearning objectives while minimally affecting the model’s performance on other concepts. (3) Our extensive experiments demonstrate that DoCo can successfully unlearn specific concepts with minimal impact on related concepts. Compared to state-of-the-art methods, our approach even generalizes to strongly related out-of-distribution prompts.

Related Work

Text-to-image diffusion models (Rombach et al. 2022; Saharia et al. 2022; Nichol et al. 2022; Zhang et al. 2024c; Cheng et al. 2024; Yang et al. 2025; Singh et al. 2024) have recently shown impressive potential in generating high-quality images, typically trained on large-scale, web-crawled datasets such as LAION-5B (Schuhmann et al. 2022). However, these uncurated datasets often include harmful or copyrighted content, which can lead the models to generate sensitive or Not Safe For Work (NSFW) images. The research community has initiated various efforts

to address this issue, which can generally be classified into three main categories: dataset filtering, model fine-tuning, and post-generation classification.

Dataset Filtering

Dataset filtering involves excluding specific images from the training dataset. For instance, Adobe Firefly uses licensed and public-domain materials to ensure suitability for commercial use (Rao 2023). Similarly, Stable Diffusion v2.0 (Rombach et al. 2022) employs an NSFW filter to eliminate inappropriate content from the LAION-5B dataset. While this approach has demonstrated some effectiveness, it presents notable challenges. Firstly, it can impose significant operational burdens, as it may necessitate retraining the model from scratch. Secondly, the accuracy and inherent biases of these filtering systems can lead to unreliable exclusion of sensitive content. Additionally, detecting abstract copyright elements, such as an artist’s unique style, remains challenging for detection methods.

Model Fine-tuning

Model fine-tuning methods focus on adjusting the model’s weights to prevent the generation of harmful content (Zhang et al. 2025, 2024b; Park et al. 2024; Ko et al. 2024; Kim, Min, and Yang 2025). ESD (Gandikota et al. 2023) aligns the probability distributions of the targeted concept with that of an empty string. However, this technique may lead to a collapse problem due to the lack of explicit control. Forget-Me-Not (Zhang et al. 2023) employs attention re-steering to identify and modify attention maps associated with specific concepts within the cross-attention layers of the diffusion U-Net architecture. However, both methods lack a mechanism for preserving the integrity of other non-target concepts. UCE (Gandikota et al. 2024) uses a closed-form solution to adjust cross-attention weights, recalibrating them to induce deliberate changes in the keys and values associated with specific text embeddings for the concepts to be edited, while aiming to limit the impact on a separate set of concepts to be retained. ConAbl (Kumari et al. 2023) suggests associating a target concept with a predefined anchor concept by minimizing the L2 distance between the predicted noises for these two concepts, and introduces a regularization loss to preserve the integrity of the anchor concept. However, these approaches do not fully resolve the tension between unlearning and retention objectives, potentially impacting the optimization efficiency of both goals. SPM (Lyu et al. 2023) introduces a novel Latent Anchoring fine-tuning strategy combined with a similarity-based retention loss to differentially weight surrogate concepts that are distant in the latent space. While this approach can alleviate the contradictions between the two objectives, it does not directly resolve their conflicts. Our method directly employs a gradient surgery technique to manipulate the gradients of the two conflicting objective functions. Furthermore, previous approaches heavily depend on the configuration of the training set, such as forming pairs of anchor and target concepts or directly manipulating concepts, which can lead to decreased generalization performance when applied beyond the train-

ing scope. In contrast, our method modifies the concept at the distribution level, effectively mitigating this problem.

Post-generation Classification

Post-generation classification methods employ a safety checker (Rombach et al. 2022) to detect images after the generation, determining whether harmful content has been produced. This approach shares a similar drawback with dataset filtering, as it heavily relies on the performance of the classifier, thus impacting the effectiveness of content filtering. Additionally, it struggles to adequately protect against copyright-protected content, which requires specially designed and trained detectors.

Method

Diffusion Models

Diffusion Models consist of two distinct Markov chains: the forward chain, which incrementally converts the original data into a noisy state, and the reverse chain, which meticulously reconstructs the original data from this noisy state (Song, Meng, and Ermon 2020; Ho, Jain, and Abbeel 2020). In the diffusion model forward process, noise ϵ is progressively added to the input image across multiple timesteps $t \in [0, T]$. The noisy image at timestep t can be generated according to the following equation:

$$x_t = \sqrt{\alpha_t}x_0 + \sqrt{1 - \alpha_t}\epsilon, \quad (1)$$

where x_0 represents the original clean image and α_t modulates the noise level.

The denoising network $\epsilon_\theta(x_t, c, t)$ is adeptly trained to reverse the noise addition, reconstructing the image at timestep $t - 1$ from the noisy image at timestep t . This network can also incorporate conditions from other modalities, such as a caption c , to guide the denoising process (Rombach et al. 2022). This process can be formulated as follows:

$$p_\theta(\mathbf{x}_0, \dots, \mathbf{x}_T) = p(\mathbf{x}_T) \prod_{t=1}^T p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t), \quad (2)$$

where $p(\mathbf{x}_T)$ is the distribution of the data after adding noise over T timesteps, and $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$ represents the conditional probability of recovering the data at the timestep $t - 1$ from the noisy data at timestep t .

Concept Unlearning Formulation

We define the process of unlearning concepts as the removal of a specific concept c^* from the pre-trained model M_{init} . Assuming an unlearning algorithm \mathcal{U} that eliminates c^* from M_{init} , the parameters of M_{init} will be altered as a result of this operation. The modified model is denoted as \hat{M} , and the process can be represented by the following formula:

$$\mathcal{U}(M_{\text{init}}, c^*) = \hat{M}. \quad (3)$$

Previous methods (Kumari et al. 2023; Gandikota et al. 2023; Lyu et al. 2023) typically align the target concept c^* explicitly with an anchor concept c (which could also be an empty string). This can be mathematically expressed as:

$$\min_M \text{dis}(\hat{M}(c^*), M_{\text{init}}(c)) = 0. \quad (4)$$

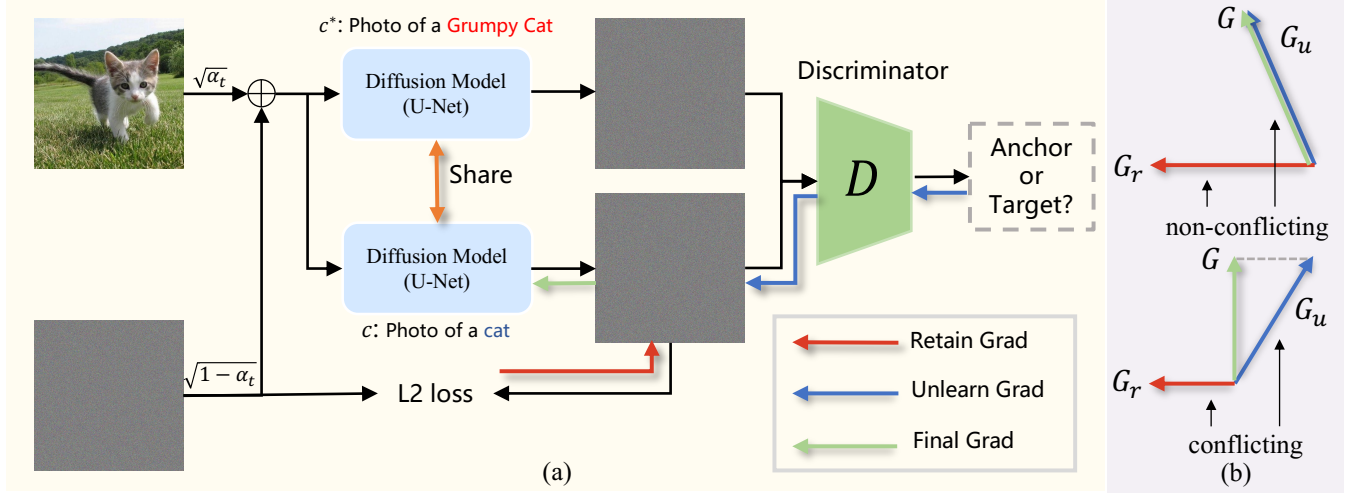


Figure 2: (a) The overall architecture of **DoCo**, which updates the model parameters through an adversarial training process. This process compels the diffusion model (acting as the generator) to produce denoised results that the discriminator cannot reliably classify as being associated with the target concept, such as “Grumpy Cat”, or the anchor concept, such as “Cat”. (b) If the unlearning gradient G_u does not conflict with the retraining gradient G_r , we update the parameters in the direction of G_u . If G_u conflicts with G_r , we mitigate the contradictory gradient between them.

Since previous methods optimize the above problem by minimizing the loss between training prompts containing the target concept c^* and the anchor concept c , this strategy often heavily depends on the design of the training prompt and lacks the ability to generalize beyond the specific prompt template used during training. Ideally, we aim to achieve an alignment of distributions rather than merely minimizing the distance for a specific training prompt template. This unlearning objective can be formulated as:

$$\min_{\hat{M}} \text{dis}(P(\hat{M}(c^*)), P(M_{\text{init}}(c))). \quad (5)$$

Simultaneously, we wish to maintain the distribution of the original anchor and other unrelated concepts unchanged. Thus, the following retaining objective should be satisfied:

$$P(\hat{M}(c)) \sim P(M_{\text{init}}(c)). \quad (6)$$

Therefore, Equation 5 can also be approximated as:

$$\text{dis}(P(\hat{M}(c^*)), P(\hat{M}(c))) \approx 0. \quad (7)$$

We assume that an unlearning algorithm should satisfy both Equation 5 and Equation 7 simultaneously.

Concept Domain Correction

To address the first optimization objective outlined in Equation 5, we propose a Concept Domain Correction framework, based on adversarial training (Goodfellow et al. 2020). The goal is to align the target concept domain with the anchor concept domain.

An intuitive approach is to fine-tune the pre-trained model M_{init} , which acts as a *generator*, and to train a *discriminator* D alternately until the *discriminator* can no longer distinguish between $\hat{M}(c)$ and $\hat{M}(c^*)$. However, the Markov

chain nature of Diffusion Models makes direct manipulation of the generated images quite challenging (Rombach et al. 2022). Therefore, we propose to align the noise distribution directly within the latent space. By adopting this approach, the task of the discriminator shifts from assessing the “authenticity” of images (Goodfellow et al. 2020) to determining the text condition of the currently predicted noise.

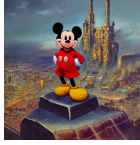
As shown in Figure 2 (a), assume we have a target concept c^* that we wish to unlearn and an anchor concept c . We can generate samples x by drawing from the distribution $p(x|c)$, where x represents an image generated by the pre-trained diffusion model M_{init} . Then, the noised samples $\{x_i\}_{i=0}^t$ are obtained by applying the process outlined in Equation 1.

The denoising network, represented as ϵ_θ , is tasked with estimating the noise at a specific timestep t , conditioned on two distinct conceptual conditions, c^* and c . After predicting the noise, we perform a denoising step to obtain x_{t-1} . We then construct a discriminator, denoted as D , to simulate a membership inference attack (MIA) (Wu et al. 2022), with the goal of distinguishing between the denoised results associated with these two concepts. The parameters of the denoising network are updated to ensure that the discriminator cannot discern which concept the denoised result is conditioned on. We iteratively update the parameters of both the generator and the discriminator, which can be formulated as a min-max objective:

$$\mathcal{L}_u = \min_{\epsilon_\theta} \max_D V(\epsilon_\theta, D) = \mathbb{E}_{x \sim p(x|c)} [\log(D(x_{t-1}(c)))] + \mathbb{E}_{x \sim p(x|c^*)} [\log(1 - D(x_{t-1}(c^*)))]. \quad (8)$$

$x_{t-1}(c)$ and $x_{t-1}(c^*)$ represent the denoised results obtained from the conditions c and c^* , respectively.

A captivating oil painting of *Mickey Mouse*, dressed in a vintage explorer's outfit complete with a leather aviator's jacket and goggles, standing atop a gently sloping hill at the edge of an ancient, sprawling metropolis. The city behind him blends futuristic skyscrapers with classical architecture, creating a unique juxtap.



A cute *Pikachu* exploring a lush green grass field, with wildflowers around and a clear blue sky above.

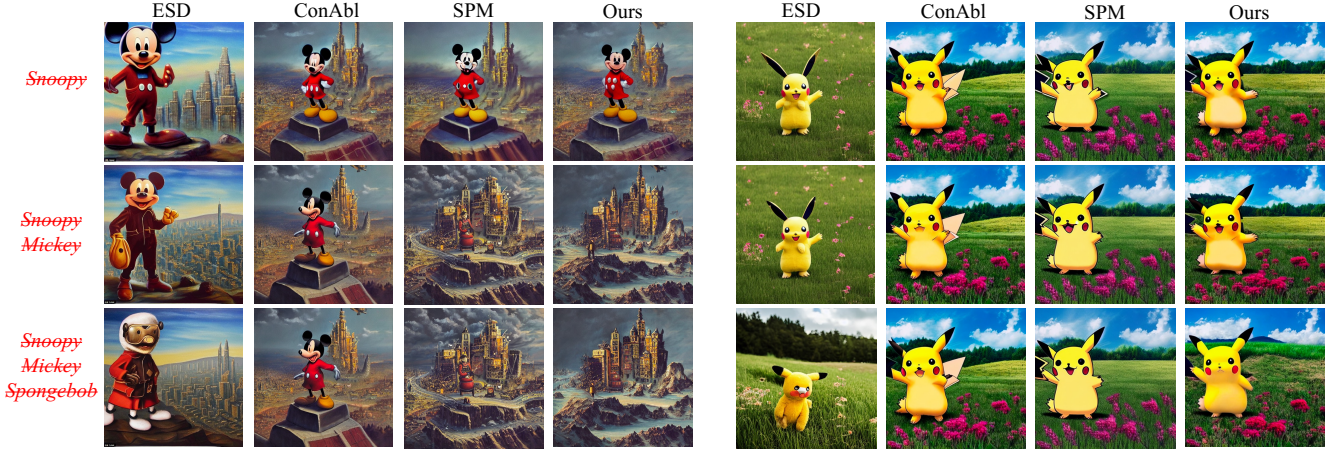
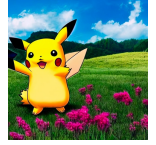


Figure 3: Visualization examples of instance unlearning. The prompts for image generation are displayed at the top. Concepts that have been unlearned are indicated in red text on the left side of the images.

Theoretically, the framework of adversarial training ensures alignment between the target concept domain and the anchor concept domain (Ganin et al. 2016). Since this alignment is achieved at the domain level without directly manipulating the target and anchor concepts, it is anticipated that the generalization capability of concept unlearning will surpass that of prior approaches.

Concept Preserving Gradient

Ensuring the retention of non-target concepts after a specific concept has been unlearned is a crucial evaluation metric for unlearning methods (Nguyen et al. 2022). This forms the second optimization objective, as illustrated in Equation 6. Previous methods focus on utilizing a relearning strategy, which involves adding the standard diffusion loss (Kumari et al. 2023) to anchor concepts or creating a preservation set to ensure retention:

$$\mathcal{L}_r = \|\epsilon - \epsilon_\theta(x_t, c, t)\|_2^2 \quad (9)$$

Although this approach has proven partially feasible, it can still lead to a conflict between the objectives of unlearning and retaining. We adopt an effective gradient surgery paradigm (Yu et al. 2020; Zhu et al. 2023) to directly mitigate the conflicts between the two objectives. As illustrated in Figure 2 (b), we formally define the gradients of the unlearning and retaining objectives as \mathbf{G}_u and \mathbf{G}_r , respectively. The training process presents two scenarios: (1) If the angle between \mathbf{G}_u and \mathbf{G}_r is less than 90° , the optimization for the unlearning loss aligns with the retaining objective, so we align the updated gradient \mathbf{G} with \mathbf{G}_u . (2) If the angle exceeds 90° , a conflict arises, and optimizing along \mathbf{G}_u would interfere with non-target concepts. To resolve this, we project \mathbf{G}_u onto the orthogonal direction of

\mathbf{G}_r , thus avoiding interference with the retaining goal. This concept-preserving gradient approach can be succinctly expressed through the following mathematical formulation:

$$\mathbf{G} = \begin{cases} \mathbf{G}_u, & \text{if } \mathbf{G}_u \cdot \mathbf{G}_r \geq 0 \\ \mathbf{G}_u - \lambda \cdot \frac{\mathbf{G}_u \cdot \mathbf{G}_r}{\|\mathbf{G}_r\|^2} \mathbf{G}_r, & \text{otherwise.} \end{cases} \quad (10)$$

Where λ is a hyper-parameter used to control the extent of gradient surgery, which we set to a default value of 1.

Through the approach described above, the conflicting component of the unlearning gradient is eliminated. Consequently, the optimization direction will not impair the utility of the model, allowing us to efficiently unlearn concepts while avoiding damage to other content.

Experiments

Implement Details

We train the model for 2,000 iterations, using a batch size of 8 and a learning rate of $6e-6$. During the initial 1000 iterations, we employ a warm-up strategy that updates only the parameters of the discriminator. For the discriminator, we utilize a PatchGAN architecture (Zhu et al. 2017), which is composed of 5 convolutional layers. Each layer employs convolutional kernels 4×4 with a stride of 2. All experiments are carried out on two A100 GPUs.

Evaluation metrics

We utilize the CLIP Score (CS), CLIP Accuracy (CA) and FID as metrics to evaluate the performance on the unlearning of instances and styles (Hessel et al. 2021; Kumari et al. 2023; Heusel et al. 2017; Li et al. 2024a; Liu et al. 2024a; Li et al. 2024b). A concept to be evaluated is incorporated into

	Snoopy			Mickey			Spongebob			Van Gogh			Picasso			Rembrandt		
	CS	CA	FID	CS	CA	FID	CS	CA	FID	CS	CA	FID	CS	CA	FID	CS	CA	FID
SD	74.29	93.01	-	72.07	93.25	-	72.80	90.38	-	74.31	100.00	-	69.67	100.00	-	72.46	100.00	-
<i>Erasing Snoopy</i>									<i>Erasing Van Gogh</i>									
	CS↓	CA↓	FID↑	CS↑	CA↑	FID↓	CS↑	CA↑	FID↓	CS↓	CA↓	FID↑	CS↑	CA↑	FID↓	CS↑	CA↑	FID↓
ESD	<u>50.77</u>	<u>38.62</u>	<u>150.22</u>	54.88	31.70	120.55	56.55	39.21	125.55	50.64	<u>33.58</u>	195.76	63.48	88.72	94.88	65.10	79.02	93.35
ConAbl	63.47	74.87	88.79	69.23	83.37	51.77	70.12	<u>86.37</u>	58.08	54.60	41.26	180.47	62.83	97.34	95.93	65.96	96.50	87.54
SPM	55.15	40.25	110.22	<u>71.63</u>	92.12	26.45	72.53	81.00	<u>31.22</u>	<u>51.80</u>	42.19	<u>198.65</u>	68.96	100.00	35.39	<u>70.53</u>	<u>97.85</u>	<u>56.12</u>
Ours	49.10	35.77	204.37	71.84	<u>89.24</u>	<u>33.21</u>	<u>71.18</u>	87.20	30.43	49.55	31.08	215.76	<u>66.21</u>	100.00	<u>65.44</u>	71.22	98.13	53.16
<i>Erasing Snoopy and Mickey</i>									<i>Erasing Picasso</i>									
	CS↓	CA↓	FID↑	CS↓	CA↓	FID↑	CS↑	CA↑	FID↓	CS↑	CA↑	FID↓	CS↓	CA↓	FID↑	CS↑	CA↑	FID↓
ESD	<u>50.30</u>	46.05	<u>177.82</u>	<u>49.54</u>	<u>30.15</u>	<u>197.80</u>	50.25	17.75	167.28	67.65	74.41	94.43	57.45	<u>32.38</u>	170.59	69.00	76.12	81.24
ConAbl	61.64	71.25	106.92	62.29	63.87	106.27	69.78	80.37	69.00	66.70	94.23	119.26	55.45	37.06	210.29	69.85	92.65	82.06
SPM	54.75	<u>39.50</u>	111.07	54.00	28.87	129.97	71.20	<u>80.21</u>	<u>37.81</u>	<u>73.55</u>	<u>95.11</u>	43.70	<u>49.22</u>	38.96	<u>269.58</u>	<u>71.22</u>	<u>97.05</u>	<u>53.89</u>
Ours	46.53	38.11	210.88	47.35	27.77	212.06	<u>70.88</u>	85.44	34.52	73.87	97.81	42.23	48.91	32.15	303.44	72.30	98.09	41.58
<i>Erasing Snoopy, Mickey and Spongebob</i>									<i>Erasing Rembrandt</i>									
	CS↓	CA↓	FID↑	CS↓	CA↓	FID↑	CS↓	CA↓	FID↑	CS↑	CA↑	FID↓	CS↑	CA↑	FID↓	CS↓	CA↓	FID↑
ESD	<u>49.50</u>	43.25	<u>201.10</u>	<u>47.50</u>	<u>28.95</u>	<u>200.94</u>	<u>45.27</u>	12.33	<u>193.63</u>	64.83	83.95	95.26	66.14	80.51	66.74	34.48	30.19	220.91
ConAbl	61.42	69.87	107.11	62.12	63.75	106.01	68.47	78.62	61.14	65.02	94.64	101.18	65.81	98.20	62.75	53.53	39.76	133.64
SPM	54.61	<u>40.50</u>	112.33	54.12	29.75	128.80	52.36	<u>23.12</u>	152.16	<u>73.13</u>	100.00	46.89	<u>69.26</u>	100.00	34.26	32.69	36.98	275.29
Ours	47.21	35.75	210.54	46.07	27.88	209.69	45.22	31.65	200.21	73.50	100.00	44.20	69.35	100.00	<u>40.50</u>	<u>33.05</u>	<u>31.05</u>	<u>260.20</u>

Table 1: Quantitative results of instance unlearning and artist style unlearning. The top-performing results are highlighted in bold, and the second-best results are underlined.

80 templates, with each template yielding 10 images. For unlearning celebrities, We use Identity Score Matching (ISM) which compute the cosine distance between the embedding of the generated face and the average face embedding of the entire clean image set (Van Le et al. 2023).

Main Results

Removal of Instance. We select ‘‘Snoopy’’ as the anchor sample and utilize the CLIP text tokenizer’s dictionary to identify the most related concepts by cosine similarity: ‘‘Mickey’’, ‘‘SpongeBob’’, and ‘‘Pikachu’’ (Kumari et al. 2023; Lyu et al. 2023).

As shown in Table 1, our method demonstrates superior performance in unlearning instance concepts, surpassing previous methods. This advantage is maintained when transfer to multi-concept cases. DoCo also exhibits superior performance in retaining non-target concepts, achieving an optimal trade-off. In contrast, other methods either show significant unlearning, such as ESD, or display inadequate unlearning capabilities while retaining non-target concepts, such as ConAbl and SPM.

We provide some visual results in Figure 3. It can be observed that ESD lacks the ability to retain other non-target concepts during unlearning, as seen in the ‘‘Pikachu’’ case. Additionally, ESD struggles with complex prompts, leading to insufficient unlearning, as demonstrated in the ‘‘Mickey’’ case. Although ConAbl maintains good retention performance, it fails to achieve effective unlearning. Both SPM

and DoCo demonstrate strong unlearning and retention capabilities. However, DoCo excels in preserving aspects such as color and texture.

Removal of Styles. We selected several representative concepts to verify the performance of unlearning styles, as detailed in Table 1. Our approach successfully unlearns specific styles while excelling in preserving non-target artistic styles. Compared to previous works, our method achieves superior results in most cases. We present several qualitative examples in Figure 4. We observe that both ESD and SPM exhibit excessive unlearning for ‘‘Van Gogh’’, resulting in aesthetically unappealing images. In contrast, our method produces more natural and visually coherent images. ESD consistently underperforms in maintaining other styles across all cases, while our approach excels in retaining performance, especially in preserving ‘‘Caravaggio’’ after unlearning ‘‘Van Gogh’’.

Removal of Celebrities. Table 2 illustrates our method’s effectiveness in unlearning celebrities such as Cristiano Ronaldo and LeBron James. Our approach consistently delivers strong results, outperforming other methods. Additionally, it maintains a good retention of other concepts after forgetting specific ones.

Ablation Study

To further demonstrate the effectiveness of the concept-preserving gradient, we present the results of employing reg-

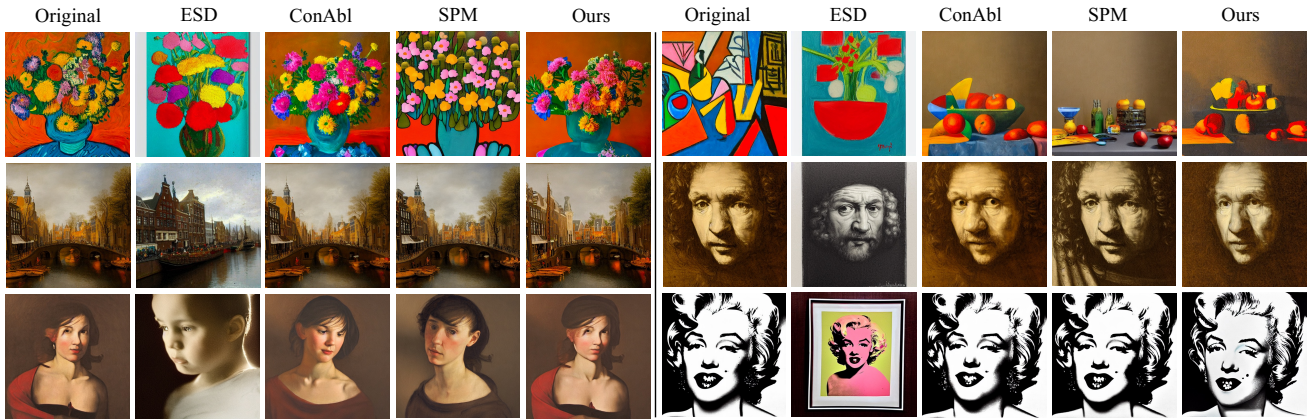


Figure 4: Visualization examples of artistic styles unlearning. Left: Unlearning "Van Gogh". Right: Unlearning "Picasso". The first row represents the forgotten style, while the subsequent rows represent other non-target concepts.

	Cristiano Ronaldo	LeBron James	Robert Downey Jr
SD	0.75	0.71	0.58
<i>Erasing Cristiano Ronaldo</i>			
	ISM \uparrow	ISM \downarrow	ISM \downarrow
ESD	0.85	0.72	0.67
ConAbl	0.75	0.70	0.62
SPM	0.89	0.77	0.64
Ours	0.89	0.69	0.62
<i>Erasing LeBron James</i>			
	ISM \downarrow	ISM \uparrow	ISM \downarrow
ESD	0.74	0.90	0.65
ConAbl	0.72	0.83	0.57
SPM	0.77	0.88	0.61
Ours	0.65	0.91	0.57

Table 2: Unlearning Celebrities: We selected three common celebrities and unlearned two of them. Our method successfully retains the other celebrities while accurately forgetting the target concept.

ularization loss (L2 Loss) as well as a baseline without any retaining operations in Table 3. It can be observed that without any retaining operations, there is a significant impact on other non-target concepts. The use of regularization loss enhances preservation but negatively affects unlearning performance. In contrast, the concept preserving gradient achieves an optimal trade-off by maintaining superior unlearning performance while ensuring effective retraining.

Conclusion

This paper aims to address the issues of generalization and utility drop in concept unlearning. We propose a novel concept domain correction framework named **DoCo** (**Domain Correction**). Firstly, we introduce an approach based on adversarial training for concept domain correction. This

L2 CP	Snoopy			Mickey			Spongebob		
	CS \downarrow	CA \downarrow	FID \uparrow	CS \uparrow	CA \uparrow	FID \downarrow	CS \uparrow	CA \uparrow	FID \downarrow
SD	74.29	93.01	-	72.07	93.25	-	72.80	90.38	-
	52.27	31.50	146.53	66.96	77.25	107.59	69.43	81.50	85.19
\checkmark	56.68	60.25	187.30	70.25	92.00	47.54	71.16	86.75	45.55
$\checkmark \checkmark$	49.10	<u>35.77</u>	204.37	71.84	<u>89.24</u>	33.21	71.18	87.20	30.43

Table 3: Ablation Study on Assessing the Impact of L2 Loss (L2) vs. Concept Preserving Gradient (CP). It can be observed that without any retain loss, the unlearning process fails to preserve other concepts effectively. On the other hand, directly using L2 loss leads to a degradation in unlearning performance. Our proposed CP method strikes a balance between the two, achieving better results in both unlearning and concept retention.

method achieves the purpose of generalized unlearning of the target concept by adjusting the output domains of both the target and anchor concepts. Secondly, we propose a concept-preserving gradient method based on gradient surgery. This method eliminates the conflicting parts between the unlearning gradient and the retaining gradient, ensuring that the process of unlearning minimally impacts the model's utility. The results demonstrate the ability of DoCo to achieve more generalized concept unlearning while maintaining the generative performance of other concepts.

Acknowledgments

This work is supported by the National Science Foundation of China (62206048 and 62401327), the Natural Science Foundation of Jiangsu Province (BK20220819), the Fundamental Research Funds for the Central Universities (2242024k30035), and China Postdoctoral Science Foundation (GZB20230358, 2023M741966 and 2024T170464). The Big Data Computing Center of Southeast University supports this research work. We also want to thank the Yanchuangzhixin company for its GPU support.

References

- Borji, A. 2022. Generated faces in the wild: Quantitative comparison of stable diffusion, midjourney and dall-e 2. *arXiv preprint arXiv:2210.00586*.
- Carlini, N.; Hayes, J.; Nasr, M.; Jagielski, M.; Sehwal, V.; Tramer, F.; Balle, B.; Ippolito, D.; and Wallace, E. 2023. Extracting training data from diffusion models. In *32nd USENIX Security Symposium (USENIX Security 23)*, 5253–5270.
- Cheng, W.; Mu, J.; Zeng, X.; Chen, X.; Pang, A.; Zhang, C.; Wang, Z.; Fu, B.; Yu, G.; Liu, Z.; et al. 2024. MVPaint: Synchronized Multi-View Diffusion for Painting Anything 3D. *arXiv preprint arXiv:2411.02336*.
- Gandikota, R.; Materzynska, J.; Fiotto-Kaufman, J.; and Bau, D. 2023. Erasing concepts from diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2426–2436.
- Gandikota, R.; Orgad, H.; Belinkov, Y.; Materzyńska, J.; and Bau, D. 2024. Unified concept editing in diffusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 5111–5120.
- Ganin, Y.; Ustinova, E.; Ajakan, H.; Germain, P.; Larochelle, H.; Laviolette, F.; March, M.; and Lempitsky, V. 2016. Domain-adversarial training of neural networks. *Journal of machine learning research*, 17(59): 1–35.
- Gao, C.; Gu, K.; Vosoughi, S.; and Mehnaz, S. 2024a. Semantic-Preserving Adversarial Example Attack against BERT. In *Proceedings of the 4th Workshop on Trustworthy Natural Language Processing (TrustNLP 2024)*, 202–207.
- Gao, C.; Wang, L.; Weng, C.; Wang, X.; and Zhu, Q. 2024b. Practical unlearning for large language models. *arXiv preprint arXiv:2407.10223*.
- Gao, H.; Pang, T.; Du, C.; Hu, T.; Deng, Z.; and Lin, M. 2024c. Meta-unlearning on diffusion models: Preventing relearning unlearned concepts. *arXiv preprint arXiv:2410.12777*.
- Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2020. Generative adversarial networks. *Communications of the ACM*, 63(11): 139–144.
- Guo, Y.; Wu, Y.; and Wan, C. 2024. A Multi-scale Patch Approach with Diffusion Model for Image Dehazing. In *International Conference on Intelligent Computing*, 392–402. Springer.
- Hessel, J.; Holtzman, A.; Forbes, M.; Bras, R. L.; and Choi, Y. 2021. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*.
- Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.
- Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.
- Kim, C.; Min, K.; and Yang, Y. 2025. Race: Robust adversarial concept erasure for secure text-to-image diffusion model. In *European Conference on Computer Vision*, 461–478. Springer.
- Kim, S.; Jung, S.; Kim, B.; Choi, M.; Shin, J.; and Lee, J. 2025. Safeguard Text-to-Image Diffusion Models with Human Feedback Inversion. In *European Conference on Computer Vision*, 128–145. Springer.
- Ko, M.; Li, H.; Wang, Z.; Patsenker, J.; Wang, J. T.; Li, Q.; Jin, M.; Song, D.; and Jia, R. 2024. Boosting Alignment for Post-Unlearning Text-to-Image Generative Models. *arXiv preprint arXiv:2412.07808*.
- Kumari, N.; Zhang, B.; Wang, S.-Y.; Shechtman, E.; Zhang, R.; and Zhu, J.-Y. 2023. Ablating concepts in text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 22691–22702.
- Li, B.; Liu, M.; Wang, G.; and Yu, Y. 2024a. Frame Order Matters: A Temporal Sequence-Aware Model for Few-Shot Action Recognition. *arXiv preprint arXiv:2408.12475*.
- Li, R.; Zhang, C.; Wang, Z.; Shen, C.; and Lin, G. 2024b. Self-Supervised 3D Scene Flow Estimation and Motion Prediction using Local Rigidity Prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- Liu, H.; Sun, Z.; and Mu, Y. 2024. Countering Personalized Text-to-Image Generation with Influence Watermarks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12257–12267.
- Liu, H.; Weng, Z.; Zhu, Y.; and Mu, Y. 2023. Trapdoor normalization with irreversible ownership verification. In *International Conference on Machine Learning*, 22177–22187. PMLR.
- Liu, M.; Wu, F.; Li, B.; Lu, Z.; Yu, Y.; and Li, X. 2024a. Envisioning Class Entity Reasoning by Large Language Models for Few-shot Learning. *arXiv preprint arXiv:2408.12469*.
- Liu, X.; Cui, X.; Li, P.; Li, Z.; Huang, H.; Xia, S.; Zhang, M.; Zou, Y.; and He, R. 2024b. Jailbreak Attacks and Defenses against Multimodal Generative Models: A Survey. *arXiv preprint arXiv:2411.09259*.
- Lyu, M.; Yang, Y.; Hong, H.; Chen, H.; Jin, X.; He, Y.; Xue, H.; Han, J.; and Ding, G. 2023. One-dimensional Adapter to Rule Them All: Concepts, Diffusion Models and Erasing Applications. *arXiv preprint arXiv:2312.16145*.
- Nguyen, T. T.; Huynh, T. T.; Nguyen, P. L.; Liew, A. W.-C.; Yin, H.; and Nguyen, Q. V. H. 2022. A survey of machine unlearning. *arXiv preprint arXiv:2209.02299*.
- Nichol, A. Q.; Dhariwal, P.; Ramesh, A.; Shyam, P.; Mishkin, P.; Mcgrew, B.; Sutskever, I.; and Chen, M. 2022. GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models. In *International Conference on Machine Learning*, 16784–16804. PMLR.
- Park, Y.-H.; Yun, S.; Kim, J.-H.; Kim, J.; Jang, G.; Jeong, Y.; Jo, J.; and Lee, G. 2024. Direct unlearning optimization for robust and safe text-to-image models. *arXiv preprint arXiv:2407.21035*.

- Rao, D. 2023. Responsible Innovation in the Age of Generative AI. Blog post.
- Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.
- Rosen, J. 2011. The right to be forgotten. *Stan. L. Rev. Online*, 64: 88.
- Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E. L.; Ghasemipour, K.; Gontijo Lopes, R.; Karagol Ayan, B.; Salimans, T.; et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35: 36479–36494.
- Schramowski, P.; Brack, M.; Deiseroth, B.; and Kersting, K. 2023. Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22522–22531.
- Schuhmann, C.; Beaumont, R.; Vencu, R.; Gordon, C.; Wightman, R.; Cherti, M.; Coombes, T.; Katta, A.; Mullis, C.; Wortsman, M.; et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35: 25278–25294.
- Shokri, R.; Stronati, M.; Song, C.; and Shmatikov, V. 2017. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, 3–18. IEEE.
- Singh, J.; Li, L.; Shi, W.; Krishna, R.; Choi, Y.; Koh, P. W.; Cohen, M. F.; Gould, S.; Zheng, L.; and Zettlemoyer, L. 2024. Negative Token Merging: Image-based Adversarial Feature Guidance. *arXiv preprint arXiv:2412.01339*.
- Song, J.; Meng, C.; and Ermon, S. 2020. Denoising Diffusion Implicit Models. In *International Conference on Learning Representations*.
- Truong, V. T.; Dang, L. B.; and Le, L. B. 2024. Attacks and defenses for generative diffusion models: A comprehensive survey. *arXiv preprint arXiv:2408.03400*.
- Van Le, T.; Phung, H.; Nguyen, T. H.; Dao, Q.; Tran, N. N.; and Tran, A. 2023. Anti-dreambooth: Protecting users from personalized text-to-image synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2116–2127.
- Wang, Y.; Chen, J.; Li, Q.; Yang, X.; and Ji, S. 2024. AEIOU: A Unified Defense Framework against NSFW Prompts in Text-to-Image Models. *arXiv preprint arXiv:2412.18123*.
- Wu, Y.; Yu, N.; Li, Z.; Backes, M.; and Zhang, Y. 2022. Membership inference attacks against text-to-image generation models. *arXiv preprint arXiv:2210.00968*.
- Xu, H.; Zhu, T.; Zhang, L.; Zhou, W.; and Yu, P. S. 2023. Machine unlearning: A survey. *ACM Computing Surveys*, 56(1): 1–36.
- Yang, X.; Chen, Y.; Chen, C.; Zhang, C.; Xu, Y.; Yang, X.; Liu, F.; and Lin, G. 2025. Learn to Optimize Denoising Scores: A Unified and Improved Diffusion Prior for 3D Generation. In *European Conference on Computer Vision*, 136–152. Springer.
- Yoon, J.; Yu, S.; Patil, V.; Yao, H.; and Bansal, M. 2024. Safree: Training-free and adaptive guard for safe text-to-image and video generation. *arXiv preprint arXiv:2410.12761*.
- Yu, T.; Kumar, S.; Gupta, A.; Levine, S.; Hausman, K.; and Finn, C. 2020. Gradient surgery for multi-task learning. *Advances in Neural Information Processing Systems*, 33: 5824–5836.
- Zhang, C.; Hu, M.; Li, W.; and Wang, L. 2024a. Adversarial attacks and defenses on text-to-image diffusion models: A survey. *Information Fusion*, 102701.
- Zhang, E.; Wang, K.; Xu, X.; Wang, Z.; and Shi, H. 2023. Forget-me-not: Learning to forget in text-to-image diffusion models. *arXiv preprint arXiv:2303.17591*.
- Zhang, Y.; Chen, X.; Jia, J.; Zhang, Y.; Fan, C.; Liu, J.; Hong, M.; Ding, K.; and Liu, S. 2024b. Defensive Unlearning with Adversarial Training for Robust Concept Erasure in Diffusion Models. *arXiv preprint arXiv:2405.15234*.
- Zhang, Y.; Jia, J.; Chen, X.; Chen, A.; Zhang, Y.; Liu, J.; Ding, K.; and Liu, S. 2025. To generate or not? safety-driven unlearned diffusion models are still easy to generate unsafe images... for now. In *European Conference on Computer Vision*, 385–403. Springer.
- Zhang, Z.; Wang, X.; Zhou, H.; Yu, Y.; Zhang, M.; Yang, C.; and Shi, C. 2024c. Can Large Language Models Improve the Adversarial Robustness of Graph Neural Networks? *arXiv preprint arXiv:2408.08685*.
- Zhang, Z.; Zhang, M.; Yu, Y.; Yang, C.; Liu, J.; and Shi, C. 2024d. Endowing Pre-trained Graph Models with Provable Fairness. In *Proceedings of the ACM on Web Conference 2024*, 1045–1056.
- Zhou, S.; Wang, L.; Ye, J.; Wu, Y.; and Chang, H. 2024. On the Limitations and Prospects of Machine Unlearning for Generative AI. *arXiv preprint arXiv:2408.00376*.
- Zhu, B.; Niu, Y.; Han, Y.; Wu, Y.; and Zhang, H. 2023. Prompt-aligned gradient for prompt tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15659–15669.
- Zhu, J.-Y.; Park, T.; Isola, P.; and Efros, A. A. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, 2223–2232.