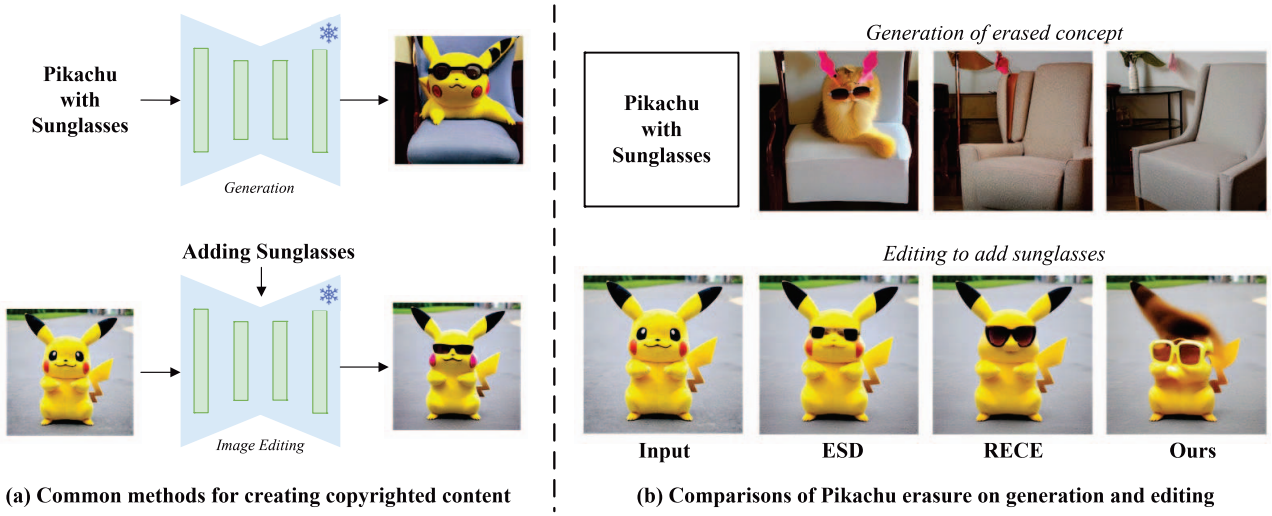
ACE: Anti-Editing Concept Erasure in Text-to-Image Models

Zihao Wang[1] Yuxiang Wei[1] Fan Li[2] Renjing Pei[2] Hang Xu[2] Wangmeng Zuo[1,3(✉)]

[1]Harbin Institute of Technology [2] Huawei Noah's Ark Lab [3] Pazhou Lab (Huangpu)



(a) Common methods for creating copyrighted content **(b) Comparisons of Pikachu erasure on generation and editing**

Figure 1. (a) Given a text-to-image (T2I) model, there are two common methods to adopt it to create undesired contents, *i.e.*, generating new images based on text prompts or editing existing images. (b) Current concept erasure methods primarily focus on preventing the generation of erased concepts but fail to protect against image editing. In contrast, our ACE method can prevent the production of such content during both generation and editing processes. As shown, after erasing Pikachu, it successfully prevents the edits involving Pikachu.

Abstract

*Recent advance in text-to-image diffusion models have significantly facilitated the generation of high-quality images, but also raising concerns about the illegal creation of harmful content, such as copyrighted images. Existing concept erasure methods achieve superior results in preventing the production of erased concept from prompts, but typically perform poorly in preventing undesired editing. To address this issue, we propose an **A**nti-Editing **C**oncept **E**rasure (**ACE**) method, which not only erases the target concept during generation but also filters out it during editing. Specifically, we propose to inject the erasure guidance into both conditional and the unconditional noise prediction, enabling the model to effectively prevent the creation of erasure concepts during both editing and generation. Furthermore, a stochastic correction guidance is introduced during training to address the erosion of unrelated concepts. We conducted erasure editing experiments with representative editing methods (i.e., LEDITS++ and MasaCtrl) to erase IP characters, and the results indicate that our ACE effectively filters out target concepts in both types of edits. Additional experiments on erasing explicit concepts and artistic styles further demonstrate that our ACE performs favorably against state-of-the-art methods. Our code will be publicly available at* https://github.com/120L020904/ACE.

1. Introduction

Recent text-to-image (T2I) diffusion models trained with large-scale datasets [49] have demonstrated an impressive ability to generate high-quality images [12, 42, 46]. Their extraordinary creative capabilities enable users to produce high-quality images, and facilitate a wide range of applications, such as image editing [4, 58] and artistic creation [13, 55, 67]. However, alongside these advancements, a significant concern has arisen regarding the potential mis-

use of these text-to-image models. For example, these models might be employed to generate unsafe content, such as copyrighted material or sexually explicit images.

To prevent the creation of unsafe content, a straightforward solution is filtering training data and retraining the model. Nonetheless, such a process is both labor-intensive and resource-consuming. Post-hoc safety checker [45, 46] and negative guidance [48] are alternative plug-and-play ways to filter undesired contents, which heavily rely on pre-trained detectors or hand-crafted prompts. More recent, concept erasure methods [14, 17, 35, 36, 68] are proposed to directly unlearn undesired concepts through model finetuning. These methods mainly focus to *precisely removing* the target concept, while *faithfully preserving* the generation of non-target concepts. For instance, ESD [14] injects the negative erase guidance into target noise prediction to guide the image away from the target concept. SPM [36] employs a lightweight adapter to eliminate concepts and further adopts latent anchoring to preserve non-target concepts.

Although these concept erasure methods can effectively prevent the generation of unsafe content giving corresponding text prompt, they can be circumvented by editing techniques. As illustrated in Fig. 1, after removing Pikachu from the model, users can still create an image of Pikachu wearing sunglasses by editing a Pikachu image using LEDIT++ [4]. This is because these methods are typically trained to remove target concept from conditional noise prediction (as shown in Fig. 2(b)), and rely on the input text (*e.g.*, "Pikachu") to trigger the guard. Therefore, when editing the image with the text "Add sunglasses" as input, the guard fails. In practice, protection from editing should also be considered in concept erasure, which we refer to as editing filtration.

To address the above issues, we propose an **A**nti-Editing **C**oncept **E**rasure method, termed **ACE**, to prevent the production of unsafe content during both generation and editing. Based on the above analysis, we explore the capabilities of CFG [20], and propose incorporating erasure guidance into both conditional and unconditional noise for anti-editing concept erasure. During erasure training, ACE additionally aligns the unconditional noise prediction of the tuned model with the proposed unconditional erasure guidance. After that, during generation or editing, the CFG prediction in the tuned model can implicitly mitigate the presence of the erased concept, thereby preventing the production of unwanted content. A prior constraint loss further adopted address the overfitting of training. Additionally, to reduce the impact of the added target concept noise guidance on the generation of non-target concepts, we further incorporate a random correction guidance with unconditional erasure guidance by subtracting randomly sampled prior concept noise guidance. With that, our ACE can thoroughly erase the target concept while preserving the generation of non-target concepts. We conducted extensive evaluations across different erasure tasks, including intellectual property (IP), explicit content, and artistic style. Our method demonstrate significant advantages in both generation and editing filtration, showcasing its effectiveness.

The contributions of this work can be summarized as:

- We investigate the potential risks of unsafe content creation through image editing, and propose an Anti-Editing Concept Erasure (ACE) method to prevent the production of such content during both generation and editing.
- A unconditional erasure guidance is proposed for anti-editing concept erasure, along with concept preservation mechanism to ensure the generation of non-target concepts.
- Extensive experiments demonstrate that our ACE can successfully erase target concepts and exhibits superior filtration capabilities during both generation and editing.

2. Related Work

2.1. Concept Erasure in T2I Models

The concept erasure [9, 11, 15, 16, 18, 21, 23–25, 28–30, 33, 39, 41, 43, 48, 51, 52, 59, 62–64, 71] in T2I models has been the subject of numerous studies. Fine-tuning models are an important method in concept erasure. ESD [14] suggests integrating negative guidance into target concept noise through training. SPM [36] proposes prior correction based on the cosine similarity of text and utilizes a comparable Lora approach to train the model. MACE [35] leverages a closed-form solution to amalgamate multiple erasure Lora weights. RECE [17] employs analytical methods to search inappropriate text embedding and integrates it into erasure closed-form solution. AdvUnlearn [68] incorporate adversarial training to improve the robustness of the erasure method. To the best of our knowledge, current fine-tuning methods lack consideration for editing filtration, thus rendering them ineffective in preventing customized editions to target concept images.

2.2. Text-driven Image Editing

Due to the broad generative capacities inherent in text-to-image DMs, the employment of DMs for image editing [3, 5, 7, 8, 26, 27, 32, 37, 38, 40, 47, 50, 54, 56, 57, 60, 65, 70] has progressively garnered traction. MasaCtrl [6] introduces source image data into the image editing process by substituting keys and values in the self-attention layer, thus modifying the actions of objects in the image. LEDITS++[4] uses inference guidance and attention masks from DM to confine editing regions while using DDPM inversion for enhanced restoration of source image. Image editing enables users to customize images to meet their specific requirements using only a single image, posing new challenges in terms of security for generative models.

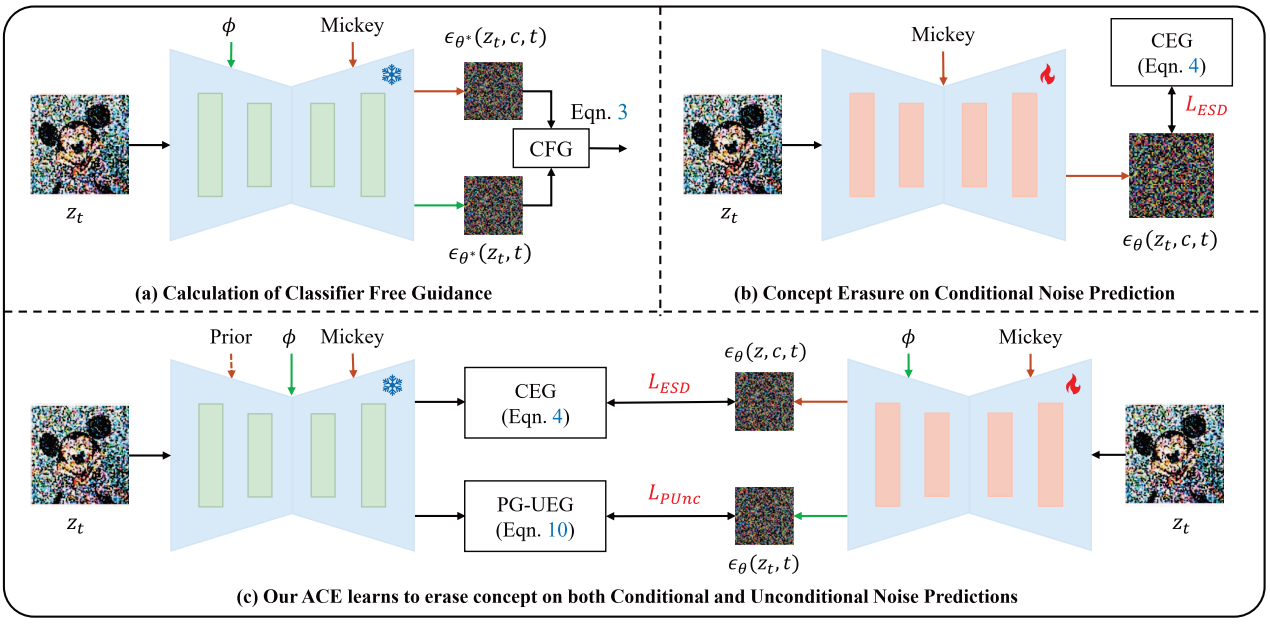


Figure 2. **Overview of our proposed ACE. (a)** In CFG, both conditional noise and unconditional noise are adopted to generate high-quality images. **(b)** ESD [14] unlearns the target concept (*e.g.*, Mickey) by aligning conditional noise prediction with conditional erasure guidance (CEG). **(c)** During the fine-tuning, our ACE injects erasure guidance into both conditional and unconditional noise prediction, preventing the production of unsafe content during both generation and editing. PG-UEG denotes the prior-guided unconditional erasure guidance calculated following Eqn 9.

2.3. Attacks in T2I Models

As research on concept erasure in T2I models advances, red team studies focusing on the robustness of detection erasure methods are also increasingly emerging. P4D [10] processes a method of inserting adversarial text into regular input text to facilitate the production of insecure images using the T2I model. Ring-A-Bell [53] extracts the discrepancy vector between the embeddings of insecure concept text and secure concept text and employs it to derive the attack text embedding. UnlearnDiff [69] employs Projected Gradient Descent (PGD) to tackle the optimization challenge inherent in adversarial attacks and maps the optimized text embeddings onto discrete tokens.

3. Proposed Method

Given a target concept (*e.g.*, Pikachu), concept erasure task [14, 36] aims to unlearn it from pre-trained text-to-image (T2I) models, preventing the illegal use of these models to create copyrighted content. However, existing methods can be circumvented and fail to prevent users from producing new undesirable images through image editing, which raises new concerns. To address this, we propose an **A**nti-Editing **C**oncept **E**rasure (**ACE**) method, as illustrated in Fig. 2, to prevent the production of undesirable content through both generation and editing. In this section, we will first introduce the prior knowledge of our method (Sec. 3.1), including employed T2I model and concept erasure method. To address the editing issue, we further propose to erase the target concept from both conditional and unconditional prediction for anti-editing erasure (Sec. 3.2). Finally, to preserve the generation of non-target concepts, a prior concept preservation mechanism is introduced (Sec. 3.3).

3.1. Preliminaries

Stable Diffusion. In this work, we adopt Stable Diffusion 1.4 [46] as text-to-image model, which is one of the representative T2I diffusion models. It first employs a variational autoencoder (VAE) to transform real images x into an image latent z. Then, a text-conditioned diffusion model ϵ_θ is trained on the latent space to predict latent codes, and mean-squared loss is adopted,

$$\mathcal{L}_{\text{LDM}} = \mathbb{E}_{z_t,t,c,\epsilon\sim\mathcal{N}(0,I)}\left[\|\epsilon - \epsilon_\theta(z_t, c, t)\|_2^2\right], \quad (1)$$

where ϵ denotes the unscaled noise and c is the text embedding encoded by text encoders. z_t is the latent noised to time t. During inference, a random Gaussian noise z_T is iteratively denoised to z_0, and decoded to final image.

Classifier-Free Guidance. To improve the quality of generated images, classifier-free guidance [20] is adopted during diffusion inference. Based on Tweedie's formula and the principles of diffusion model, we have:

$$\nabla_{z_t} \log p(c|z_t) = -\frac{1}{\sigma_t}(\epsilon_\theta(z_t, c, t) - \epsilon_\theta(z_t, t)). \quad (2)$$

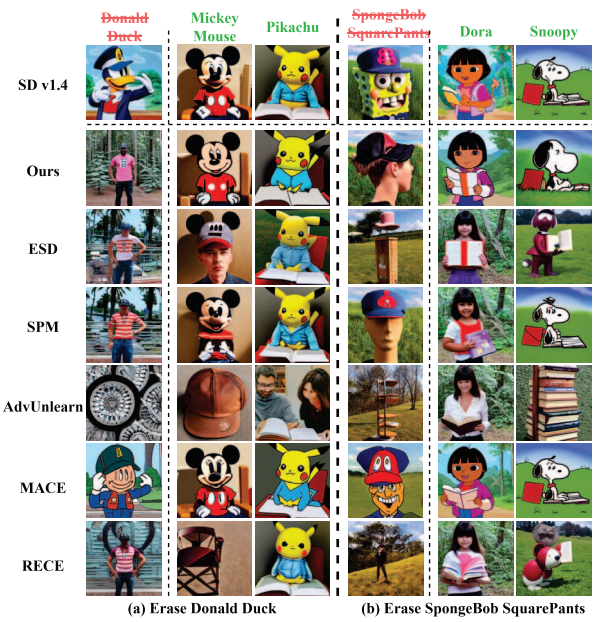


Figure 3. **Qualitative comparisons of IP character removal.** Our ACE effectively erases the target concept while generating other concepts successfully.

Here, σ_t is a constant. To increase the probability of text condition c appearing in the final image, the final noise prediction is the composition of noise prediction from both conditional and unconditional texts,

$$\tilde{\epsilon} = \epsilon_\theta(z_t, t) + \omega(\epsilon_\theta(z_t, c, t) - \epsilon_\theta(z_t, t)), \tag{3}$$

where $\epsilon_\theta(z_t, t)$ denote the unconditional noise prediction, and ω is a hyperparameter controlling the guidance scale.

Concept Erasure. Given a target concept indicated by text c (*e.g.*, Pikachu), concept erasure task finetunes the model to reduce the probability of generating images containing this concept. For example, ESD [14] removes the target concept from the conditional noise prediction, and a conditional erasure guidance (CEG) is defined as:

$$\tilde{\epsilon}_c = \epsilon_{\theta^\star}(z_t, t) - \eta_c(\epsilon_{\theta^\star}(z_t, c, t) - \epsilon_{\theta^\star}(z_t, t)), \tag{4}$$

where $\epsilon_{\theta^\star}(\cdot)$ represents the original T2I model, and z_t is the encoded latent image contains target concept c. η_c is a control scale hyperparameter. During training, ESD aligns the noise prediction of the target concept in tuned model $\epsilon_\theta(z_t, c, t)$ with the above CEG,

$$\mathcal{L}_{\text{ESD}} = \mathbb{E}_{z_t,t,c}\left[\|\epsilon_\theta(z_t, c, t) - \tilde{\epsilon}_c\|_2^2\right]. \tag{5}$$

After the training, the erasure guidance $-\nabla_{z_t} \log p(c|z_t)$ is introduced into conditional noise prediction of the target concept. Therefore, the prediction of tuned model will be guided away from the erased concept, preventing the generation of images containing the erased concept.

3.2. Anti-Editing Concept Erasure

Editing Filtration. Although existing erasure methods can successfully prevent the generation of an erased concept through text prompts, they can be easily circumvented by editing techniques. As shown in Fig. 1, when utilizing tuned ESD model to add sunglasses on an image of Pikachu using LEDITS++ [4], it successfully produces an image of Pikachu with sunglasses, raising potential copyright concerns. This is because these methods are typically trained to erase the concept from the noise prediction of the target concept (as shown in Fig. 2 (b)), and rely on inputting concept text (*e.g.*, "Pikachu" or "Mickey") to trigger the guard. However, during the editing process, the target concept may not necessarily be used in the text prompt. Therefore, these erasure methods fail to prevent the reconstruction of the erased concept. In practice, the erasure model should also have the ability to prevent the creation of undesired concepts through image editing, a feature we refer to as editing filtration.

Unconditional Erasure Guidance. As we all know, current generation and editing methods heavily rely on classifier-free guidance [20] (CFG) to improve the quality of generated images, where unconditional noise prediction performs an important role. To address the issue of editing filtration, we further propose to erase the target concept from both conditional and unconditional noise prediction, thereby preventing edited images from containing target concepts. Specifically, similar to ESD, we define the unconditional erasure guidance (UEG) as,

$$\tilde{\epsilon}_{\text{u}} = \epsilon_{\theta^\star}(z_t, t) + \eta_{\text{u}}(\epsilon_{\theta^\star}(z_t, c, t) - \epsilon_{\theta^\star}(z_t, t)). \tag{6}$$

During training, we additionally align the unconditional noise prediction of the tuned model with the UEG,

$$\mathcal{L}_{\text{Unc}} = \mathbb{E}_{z_t,t,c}\left[\|\epsilon_\theta(z_t, t) - \tilde{\epsilon}_{\text{u}}\|_2^2\right]. \tag{7}$$

When fine-tuned unconditional noise (our UEG) is subtracted in the CFG process, the erased concept guidance will be subtracted, thereby reducing the probability of the erased concept appearing regardless of the input text prompt. Then, the CFG noise prediction during inference will move away from the target concept regardless of any text input, thereby effectively preventing the production image containing the target concept. As erasure models are usually trained on a small dataset, they are prone to be overfitting, where the erasure guidance is introduced into the noise prediction for other conditional text prompts. This weakens the erasure effects and leads to incomplete erasures. To address the issue of overfitting, we introduce a prior constraint loss during the training process. Specifically, we regularize the prediction of the prior concept in the new model to be consistent with that of the original model:

$$\mathcal{L}_{\text{Cons}} = \mathbb{E}_{z_t,t,c_p\in\mathcal{C}_p}\left[\|\epsilon_\theta(z_t, c_p, t) - \epsilon_{\theta^\star}(z_t, c_p, t)\|_2^2\right], \tag{8}$$

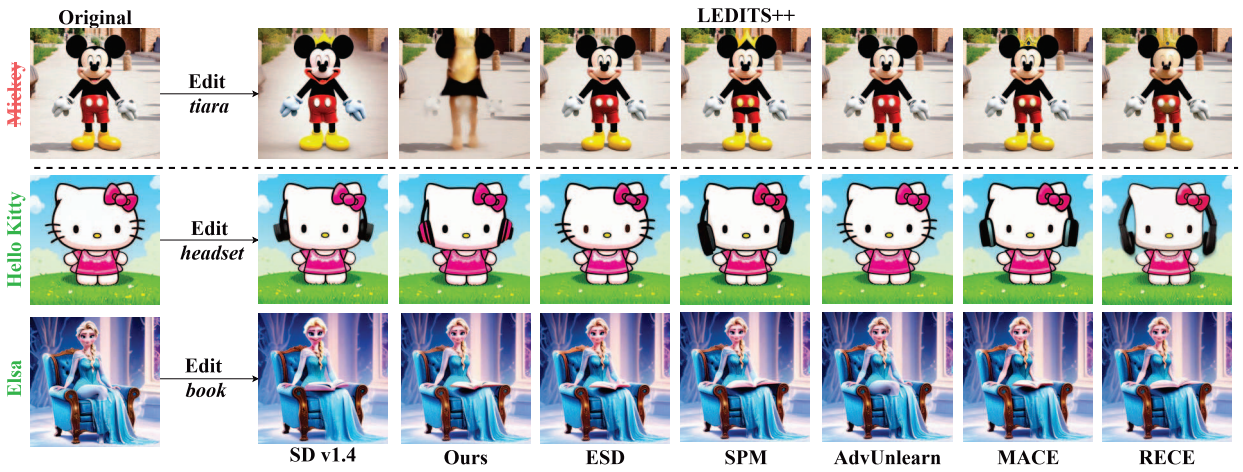


Figure 4. **Comparison of our ACE method with other methods in terms of editing filtering.** After erasing Mickey Mouse, our method filtered out edits involving Mickey Mouse while not affecting edits related to other IP characters. In contrast, the competing methods either fail to prevent editing (*e.g.*, ESD, SPM, RECE, and MACE) or cannot perform editing on non-target concepts (*e.g.*, AdvUnlearn).

Figure 5. **Qualitative results of nudity removal.** Figure (a) shows the results of explicit editing using SD-Inpainting, while Figure (b) displays images generated using text with explicit label. Static adversarial text is used for editing text, while dynamic adversarial attacks are employed for generation. It can be observed that our method effectively reduces exposure in both editing and generation tasks. Moreover, our method maintains its effectiveness when editing and generating using adversarial text, indicating its robustness.

where c_p represents prior concept, and $\mathcal{C}_p$ represents the set of prior concepts. Intuitively, the larger the set of priors, the better it helps mitigate overfitting. However, it is challenging to traverse all the prior concepts as the pre-trained models have a large general semantic space. Our goal is to preserve the concepts more likely to be affected, thus minimizing the influence to other concepts. We assume that these concepts are semantic-related concepts to the erased concept and use LLM [1] to obtain them. By adding this loss, it ensures that the erasure guidance introduced during training aligns with our conceptualization in the Eqn. 7.

3.3. Prior Concept Preservation

In practice, training with the method proposed in Sec. 3.2 affects the generation prior of relevant concepts (see Sec. 4.4). This is because incorporating UEG not only decreases the probability of producing erased concepts, but also decreases probability of adjacent concepts. Therefore, we reverse mechanism of UEG by subtracting the guidance of prior concepts from the unconditional noise, which prevents the probability reduction of these concepts and minimizes concept forgetting. The prior concepts are sampled from the semantic-related concepts obtained using LLM,

	Method				(a) Generation Prevention						(b) Editing Filtration					
					Erase Concept		Prior Concept		Overall		Erase Concept		Prior Concept		Overall	
	ESD	Unc	Cons	Cor	$\text{CLIP}_e \downarrow$	$\text{LPIPS}_e \uparrow$	$\text{CLIP}_p \uparrow$	$\text{LPIPS}_p \downarrow$	$\text{CLIP}_d \uparrow$	$\text{LPIPS}_d \uparrow$	$\text{CLIP}_e \downarrow$	$\text{LPIPS}_e \uparrow$	$\text{CLIP}_p \uparrow$	$\text{LPIPS}_p \downarrow$	$\text{CLIP}_d \uparrow$	$\text{LPIPS}_d \uparrow$
(1)	✓				**0.171**	0.440	0.246	0.286	0.075	0.153	0.301	0.060	**0.305**	**0.050**	0.004	0.011
(2)	✓	✓			0.166	**0.551**	0.283	0.236	0.117	**0.315**	0.285	0.149	**0.305**	0.057	0.019	0.092
(3)	✓	✓	✓		0.159	0.507	0.254	0.337	0.095	0.170	0.274	0.168	0.300	0.077	0.026	0.091
(4)		✓	✓	✓	0.211	0.303	0.293	0.199	0.082	0.104	**0.273**	**0.175**	0.301	0.079	0.028	0.096
(5)	✓	✓	✓	✓	0.175	0.397	**0.295**	**0.196**	**0.120**	0.201	0.274	0.168	0.303	0.070	**0.029**	**0.097**

Table 1. **Quantitative Evaluation of generation and editing after ablation.** The best results are highlighted in bold. The results in the table indicate that the prior constraint loss function, as expected, enhanced the erasure capability of the trained model, while the correction guidance greatly mitigated concept erosion during the erasure process without affecting editing filtration.

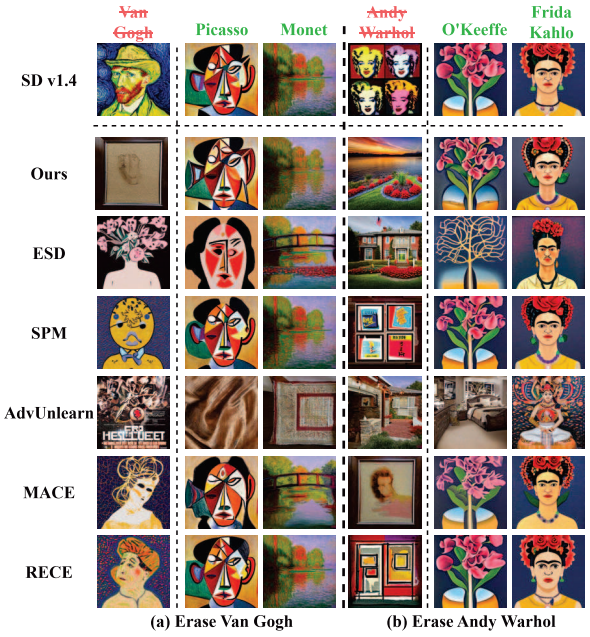


Figure 6. **Qualitative results of artistic style removal.** Our method erases the target style effectively and has minimal impact on other artistic styles.

which is mentioned in the previous section. We call this new guidance prior-guided unconditional erasure guidance (PG-UEG), which is defined as:

$$\tilde{\epsilon}_{\text{pu}} = \epsilon_{\theta^\star}(z_t, t) + \eta_{\text{u}}(\epsilon_{\theta^\star}(z_t, c, t) - \epsilon_{\theta^\star}(z_t, t)) - \eta_{\text{p}}\gamma_p(\epsilon_{\theta^\star}(z_t, c_p, t) - \epsilon_{\theta^\star}(z_t, t)), \quad (9)$$

where γ_p represents the guidance control term related to the prior retained concept. c_p refers to the same prior concept in $\mathcal{L}_{\text{Cons}}$ which are obtained through random sampling from the set $\mathcal{C}_p$. We calculate γ_p using the CLIP model to measure the relevance of different prior concepts to the target concept image and then compare it to the relevance of the target concept text to its image. Specifically, $\gamma_p = \frac{\text{CLIP}(x, c_p)}{\text{CLIP}(x, c)}$. The new loss for our ACE is:

$$\mathcal{L}_{\text{PUnc}} = \mathbb{E}_{z_t, t, c, c_p \in \mathcal{C}_p}\left[\|\epsilon_\theta(z_t, t) - \tilde{\epsilon}_{\text{pu}}\|_2^2\right]. \quad (10)$$

The final training loss for our ACE is summarized as: $\mathcal{L}_{\text{ACE}} = \lambda_{\text{PUnc}}\mathcal{L}_{\text{PUnc}} + \lambda_{\text{Cons}}\mathcal{L}_{\text{Cons}} + \lambda_{\text{ESD}}\mathcal{L}_{\text{ESD}}$.

In our implementation, we adopt LORA [22] for parameter-efficient tuning, and the training process follows [14]. More details are provided in Suppl.

4. Experiments

We conduct experiments on various tasks to evaluate our ACE, including IP characters erasure, artistic styles erasure, and nudity erasure. ESD [12], SPM [36], AdvUnlearn [68], MACE [35], and RECE [17] are adopted as competing methods. Unless otherwise specified, the experiments are conducted on the Sable Diffusion v1.4.

4.1. IP Character Removal

Experiment Setup. To access our ACE on IP character removal, we employ ten iconic IP characters as examples, including Hello Kitty, Snoopy, Mickey Mouse, Elsa, Donald Duck, Dora the Explorer, Winnie the Pooh, Sonic the Hedgehog, Elsa, and Pikachu. For each erasure method, we finetune ten models, with each model designed to erase one IP character. Following [14, 17], we adopted CLIP [44] score and LPIPS [66] score as metrics for evaluation. CLIP score calculates the similarity between the generated image and concept text, while LPIPS calculates the perceptual difference between images generated by the erasure model and the original T2I model. CLIP_e calculates the CLIP similarity between images generated with erased concept text and their corresponding text, where lower value indicates more thorough erasure. CLIP_p calculates the relevance under prior concepts, and higher value indicates better prior preservation. LPIPS_e calculates the LPIPS similarity between images generated with erased concept text by the trained model and the original model, and higher value indicates more thorough erasure. LPIPS_p calculates the similarity under prior concepts, in which lower values indicate better prior preservation. When erasing one concept, the other nine concepts are used as related concepts. Following RECE [17], we further calculate the overall scores between erased and related characters to measure the trade-off between the concept erasure and prior preservation, where

Method	(a) Generation Prevention: Erase Concept CLIP$_e$ ↓	Erase Concept LPIPS$_e$ ↑	Prior Concept CLIP$_p$ ↑	Prior Concept LPIPS$_p$ ↓	Overall CLIP$_d$ ↑	Overall LPIPS$_d$ ↑	(b) Editing Filtration: Erase Concept CLIP$_e$ ↓	Erase Concept LPIPS$_e$ ↑	Prior Concept CLIP$_p$ ↑	Prior Concept LPIPS$_p$ ↓	Overall CLIP$_d$ ↑	Overall LPIPS$_d$ ↑
SD v1.4 [46]	0.301	0.000	0.301	0.000	0.000	0.000	0.308	0.063	0.308	0.063	0.000	0.000
ESD [14]	0.227	0.331	0.276	0.255	0.049	0.076	0.306	0.042	0.307	0.041	0.001	0.000
SPM [36]	0.239	0.288	0.296	**0.107**	0.056	0.181	0.302	0.061	0.303	0.056	0.001	0.005
AdvUnlearn [68]	**0.166**	**0.468**	0.209	0.403	0.043	0.065	0.310	0.011	**0.311**	**0.010**	0.001	0.001
MACE [35]	0.250	0.317	**0.298**	0.134	0.048	0.184	0.303	0.056	0.304	0.054	0.001	0.002
RECE [17]	0.176	0.426	0.257	0.270	0.081	0.156	0.300	0.066	0.303	0.054	0.003	0.012
Ours	0.175	0.397	0.295	0.196	**0.120**	**0.201**	**0.274**	**0.168**	0.303	0.070	**0.029**	**0.097**

Table 2. **Quantitative comparisons of IP character erasure.** The best two results are highlighted with **bold** and underline.

	Buttocks	Breast (F)	Genitalia (F)	Breast (M)	Genitalia (M)	Feet	Armpits	Belly	Total↓	FID30k↓	CLIP30k↑
SD v1.4 [46]	61	204	37	38	16	70	241	183	850	14.07	0.313
ESD [14]	15	29	5	11	10	37	68	36	211	13.80	0.304
SPM [36]	14	29	7	2	12	41	53	28	186	14.63	**0.312**
AdvUnlearn [68]	4	6	**2**	**0**	**8**	13	12	**7**	52	15.35	0.293
MACE [35]	7	24	8	10	9	35	61	35	189	**12.60**	0.294
RECE [17]	14	20	7	16	10	39	45	35	186	14.45	0.309
Ours	**3**	**2**	3	4	9	**6**	**5**	**7**	**39**	14.69	0.308

Table 3. **Exposure detection of generated images in the I2P dataset.** The best two results are highlighted with **bold** and underline.

	Erase Concept CLIP$_e$ ↓	Erase Concept LPIPS$_e$ ↑	Relate Concept CLIP$_p$ ↑	Relate Concept LPIPS$_p$ ↓	Overall CLIP$_d$ ↑	Overall LPIPS$_d$ ↑
SD v1.4 [46]	0.310	0.000	0.310	0.000	0.000	0.000
ESD [14]	0.216	0.444	0.296	0.241	0.080	0.202
SPM [36]	0.266	0.268	0.308	0.074	0.042	0.195
AdvUnlearn [68]	0.186	**0.476**	0.229	0.410	0.043	0.066
MACE [35]	0.228	0.366	0.298	0.196	0.069	0.169
RECE [17]	0.253	0.307	**0.309**	**0.051**	0.057	0.255
Ours	**0.160**	0.471	0.303	0.126	**0.143**	**0.345**

Table 4. **Quantitative evaluation of artist style erasure.** The best two results are highlighted with **bold** and underline. Our ACE performs better in terms of thorough erasure and also demonstrates comparable prior preservation.

	Unlearn Diffusion↓	P4D↓	Ring a Bell↓	Average↓
SD v1.4 [46]	100%	100%	85.21%	95.07%
ESD [14]	73.05%	74.47%	38.73%	62.08%
SPM [36]	91.49%	91.49%	57.75%	80.24%
AdvUnlearn [68]	**25.53%**	**19.15%**	4.93%	**16.54%**
MACE [35]	64.53%	66.67%	14.79%	48.66%
RECE [17]	70.92%	65.96%	26.76%	54.55%
Ours	27.65%	28.37%	**2.82%**	19.61%

Table 5. **Robustness evaluation of nudity erasure.** The best two results are highlighted with **bold** and underline. We report the attack success rates (ASR) of different adversarial methods under various erasure models. Our method achieved the second-best results without using adversarial training.

$\text{CLIP}_d = \text{CLIP}_p - \text{CLIP}_e$ and $\text{LPIPS}_d = \text{LPIPS}_e - \text{LPIPS}_p$. Higher CLIP_d and LPIPS_d indicate better trade-off.

For generation evaluation, we adopt 33 text templates for each character concept, and five images are generated for each text template using the erased model. To evaluate the effectiveness of editing filtration, we adopt the widely used LEDITS++ [4] and MasaCtrl [6] as editing methods. For each concept, we utilize Stable Diffusion 3 [12] to generate 15 images based on 3 text templates as initial images, and then perform editing on them using erased models. Each image is manipulated using 11 editing texts, such as "sunglasses". Finally, the CLIP score and LPIPS score are calculated based on edited images, concept text and original images. The final results are all reported by averaging 10 characters. More details can be found in Suppl.

Experiment Results. Fig. 3 illustrates the comparison of generation results against competing methods. One can see that, our ACE can successfully erase the target concept (*i.e.*, Donald Duck) while retaining the capability to generate related prior concepts (*e.g.*, Mickey Mouse and Pikachu). In contrast, methods such as ESD, AdvUnlearn, and RECE generate examples with noticeable concept erosion. From Table 2, our ACE demonstrates a comparable CLIP score for both the erased and related concepts. This indicates that our ACE achieves a better trade-off between target concept erasure and prior concept preservation, as further validated by the overall metrics in Table 2 (a). SPM and MACE exhibit inferior performance in thoroughly erasing the target concept. While AdvUnlearn performs well at erasing the target concept, it shows poor performance in prior preservation.

Fig. 4 further presents the comparison of editing results by LEDITS++. As shown in the figure, the competing method generates the erased concept with desired attributes after performing the editing on the given image, which is not wanted in practice. In contrast, our method can successfully hinder the editing of images containing erased concepts (*e.g.*, Mickey), while keeping the editability of non-target concepts (*e.g.*, Hello Kitty and Elsa). Table 2 (b) reports the quantitative comparisons evaluated with LEDITS++. Our method shows a significant improvement in erasing concepts, demonstrating its ability to edit filtration.

The comparison on MasaCtrl and more results can be found in Suppl.

4.2. Explicit Content Removal

Experimental Setup. To evaluate our ACE on explicit content removal, we employ "nudity" as the target concept to train the model. Following [36], we utilize the I2P dataset [48] to evaluate the performance of explicit content generation. Specifically, we select 856 text prompts with explicit labels, and each prompt generates one image. Then, Nudenet [2] is used to quantify the number of nude body parts in these generated images. Additionally, following [14, 36], we employ COCO-30k Caption dataset [31] to evaluate the conditional generation capability of erased models. Specifically, we generate one image for each caption in COCO-30k and FID [19] is calculated between generated and natural images. CLIP score is also calculated between the generated images and the captions to access the semantic alignment of generated images. For robustness evaluation, we adopt UnlearnDiff [69], P4D [10] and Ring-A-Bell [53] as adversarial tools to calculate attack success rate (ASR). Adversarial attacks were conducted on 142 sensitive texts provided by UnlearnDiff. More details can be found in Suppl.

Experiment Results. From Table 5, we can see that our method has a lower success rate in adversarial attacks when trained only for "nudity", with only AdvUnlearn performing slightly better than us with using adversarial training. As shown in Fig. 5 and Table 3, our method can effectively erase nudity content and results in fewer exposure parts. In the generation evaluation, we dynamically attack the erased models using adversarial tools. As shown in Fig. 5, our method demonstrates excellent robustness. To further showcase our method's efficacy in editing filtration, we employ SD-Inpainting [46] as an editing tool to assess the exposure levels of images after different text-guided inpainting processes. In addition to conventional text editing (*e.g.*, bikini) adversarial edited text in MMA-Diffusion [61] is also used for explicit editing. GroundingDINO [34] is used to detect clothing in the images. As shown in Fig. 5, our method successfully prevents inappropriate inpainting of exposed parts in masked areas, making it more practical for real-world applications.

More results for robustness and editing filtration evaluation can be found in Suppl.

4.3. Artistic Style Removal

Experiment Setup. To validate the performance of our model in unlearning styles, we choose ten representative artistic styles, including Leonardo da Vinci, Pablo Picasso, Michelangelo, Van Gogh, Salvador Dali, Claude Monet, Andy Warhol, Jackson Pollock, Frida Kahlo, Georgia O'Keeffe. The evaluation process and metrics are similar to the IP character removal (Sec. 4.1).

Experiment Results. Fig. 6 illustrates the results of erasing artistic styles. As shown in the figure, our method can erase the style of Van Gogh and Andy Warhol from the T2I model, while generating other styles faithfully. From Table 4, our method achieves better CLIP_e on erased concept.

4.4. Ablation Study

We further conduct the ablation study on the IP character erasure to evaluate the effectiveness of each component proposed in our ACE. Specifically, it contains the following variants: (1) *Baseline*: by only adopting the ESD loss to finetune the model. (2) *Baseline* + *Unc*: by employing unconditional erasure guidance alignment with ESD Loss to finetune the model. (3) *Baseline* + *Unc* + $\mathcal{L}_{Cons}$: by adopting ESD Loss, unconditional erasure guidance alignment, and $\mathcal{L}_{\text{Cons}}$ to finetune the model. (4) *Our method without ESD*: Ours w/o $\mathcal{L}_{ESD}$ is also effective in concept erasure and editing filtration, and performs better than ESD, indicating that our PG-UEG plays a crucial role in editing filtering. (5) *Ours full method*: by incorporating the ESD Loss, prior-guided unconditional erasure guidance alignment and $\mathcal{L}_{\text{Cons}}$ together. From Table 1, we can see that: (i) Introducing unconditional erasure guidance improves the model's editing filtration performance, indicating its effectiveness in preventing unwanted edits. (ii) We use both unconditional erasure guidance and $\mathcal{L}_{\text{Cons}}$ together leading to significant improvements in concept erasure and editing filtration performance, although it compromises the generation of related prior concepts. (iii) $\mathcal{L}_{\text{PUnc}}$ enhances the prior preservation, and without affecting editing filtration.

More ablation results are provided in Suppl.

5. Conclusion

In this paper, we investigate the potential risks of unsafe content creation through image editing, and propose an Anti-Editing Concept Erasure (ACE) method to prevent the production of such content during both generation and editing. In addition to the conditional erasure guidance used by existing methods, we further propose an unconditional noise erasure technique to enhance anti-editing concept erasure. This guidance steers the noise prediction away from the target concept, thereby effectively preventing the production of images containing the target concept. Moreover, a concept preservation mechanism is introduced to maintain the generation prior of non-target concepts. Experiments demonstrate that our ACE can successfully erase specific concepts and exhibits superior filtration capabilities during both generation and editing compared to existing methods.

Acknowledgement. The work was supported by National Key R&D Program of China under Grant No. 2022YFA1004100.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 5

[2] P Bedapudi. Nudenet: Neural nets for nudity classification, detection and selective censoring, 2019. 8

[3] Manuel Brack, Felix Friedrich, Dominik Hintersdorf, Lukas Struppek, Patrick Schramowski, and Kristian Kersting. Sega: Instructing text-to-image models using semantic guidance. *Advances in Neural Information Processing Systems*, 36: 25365–25389, 2023. 2

[4] Manuel Brack, Felix Friedrich, Katharia Kornmeier, Linoy Tsaban, Patrick Schramowski, Kristian Kersting, and Apolinário Passos. Ledits++: Limitless image editing using text-to-image models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8861–8870, 2024. 1, 2, 4, 7

[5] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18392–18402, 2023. 2

[6] Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22560–22570, 2023. 2, 7, 3

[7] Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T Freeman, Michael Rubinstein, et al. Muse: Text-to-image generation via masked generative transformers. *arXiv preprint arXiv:2301.00704*, 2023. 2

[8] Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM Transactions on Graphics (TOG)*, 42(4):1–10, 2023. 2

[9] Die Chen, Zhiwen Li, Mingyuan Fan, Cen Chen, Wenmeng Zhou, and Yaliang Li. Eiup: A training-free approach to erase non-compliant concepts conditioned on implicit unsafe prompts. *arXiv preprint arXiv:2408.01014*, 2024. 2

[10] Zhi-Yi Chin, Chieh-Ming Jiang, Ching-Chun Huang, Pin-Yu Chen, and Wei-Chen Chiu. Prompting4debugging: Red-teaming text-to-image diffusion models by finding problematic prompts. *arXiv preprint arXiv:2309.06135*, 2023. 3, 8

[11] Anudeep Das, Vasisht Duddu, Rui Zhang, and N Asokan. Espresso: Robust concept filtering in text-to-image models. *arXiv preprint arXiv:2404.19227*, 2024. 2

[12] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis, march 2024. *URL http://arxiv. org/abs/2403.03206*, 2024. 1, 6, 7

[13] Kailai Feng, Yabo Zhang, Haodong Yu, Zhilong Ji, Jinfeng Bai, Hongzhi Zhang, and Wangmeng Zuo. Vitaglyph: Vitalizing artistic typography with flexible dual-branch diffusion models. *arXiv preprint arXiv:2410.01738*, 2024. 1

[14] Rohit Gandikota, Joanna Materzynska, Jaden Fiotto-Kaufman, and David Bau. Erasing concepts from diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2426–2436, 2023. 2, 3, 4, 6, 7, 8

[15] Rohit Gandikota, Hadas Orgad, Yonatan Belinkov, Joanna Materzyńska, and David Bau. Unified concept editing in diffusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5111–5120, 2024. 2

[16] Hongcheng Gao, Tianyu Pang, Chao Du, Taihang Hu, Zhijie Deng, and Min Lin. Meta-unlearning on diffusion models: Preventing relearning unlearned concepts. *arXiv preprint arXiv:2410.12777*, 2024. 2

[17] Chao Gong, Kai Chen, Zhipeng Wei, Jingjing Chen, and Yu-Gang Jiang. Reliable and efficient concept erasure of text-to-image diffusion models. *arXiv preprint arXiv:2407.12383*, 2024. 2, 6, 7, 4

[18] Luxi He, Yangsibo Huang, Weijia Shi, Tinghao Xie, Haotian Liu, Yue Wang, Luke Zettlemoyer, Chiyuan Zhang, Danqi Chen, and Peter Henderson. Fantastic copyrighted beasts and how (not) to generate them. *arXiv preprint arXiv:2406.14526*, 2024. 2

[19] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 8, 2

[20] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 2, 3, 4

[21] Seunghoo Hong, Juhun Lee, and Simon S Woo. All but one: Surgical concept erasing with model preservation in text-to-image diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 21143–21151, 2024. 2

[22] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 6

[23] Chi-Pin Huang, Kai-Po Chang, Chung-Ting Tsai, Yung-Hsuan Lai, Fu-En Yang, and Yu-Chiang Frank Wang. Receler: Reliable concept erasing of text-to-image diffusion models via lightweight erasers. *arXiv preprint arXiv:2311.17717*, 2023. 2

[24] Sanghyun Kim, Seohyeon Jung, Balhae Kim, Moonseok Choi, Jinwoo Shin, and Juho Lee. Safeguard text-to-image diffusion models with human feedback inversion. *arXiv preprint arXiv:2407.21032*, 2024.

[25] Nupur Kumari, Bingliang Zhang, Sheng-Yu Wang, Eli Shechtman, Richard Zhang, and Jun-Yan Zhu. Ablating concepts in text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22691–22702, 2023. 2

[26] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1931–1941, 2023. 2

[27] Fan Li, Zixiao Zhang, Yi Huang, Jianzhuang Liu, Renjing Pei, Bin Shao, and Songcen Xu. Magiceraser: Erasing any objects via semantics-aware control. *arXiv preprint arXiv:2410.10207*, 2024. 2

[28] Hang Li, Chengzhi Shen, Philip Torr, Volker Tresp, and Jindong Gu. Self-discovering interpretable diffusion latent directions for responsible text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12006–12016, 2024. 2

[29] Jia Li, Lijie Hu, Zhixian He, Jingfeng Zhang, Tianhang Zheng, and Di Wang. Text guided image editing with automatic concept locating and forgetting. *arXiv preprint arXiv:2405.19708*, 2024.

[30] 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. *arXiv preprint arXiv:2402.05375*, 2024. 2

[31] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 8, 2

[32] Ming Liu, Yuxiang Wei, Xiaohe Wu, Wangmeng Zuo, and Lei Zhang. Survey on leveraging pre-trained generative adversarial networks for image editing and restoration. *Science China Information Sciences*, 66(5):151101, 2023. 2

[33] Runtao Liu, Ashkan Khakzar, Jindong Gu, Qifeng Chen, Philip Torr, and Fabio Pizzati. Latent guard: a safety framework for text-to-image generation. In *European Conference on Computer Vision*, pages 93–109. Springer, 2025. 2

[34] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 8

[35] 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*, pages 6430–6440, 2024. 2, 6, 7, 4

[36] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7559–7568, 2024. 2, 3, 6, 7, 8, 4

[37] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6038–6047, 2023. 2

[38] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4296–4304, 2024. 2

[39] Yong-Hyun Park, Sangdoo Yun, Jin-Hwa Kim, Junho Kim, Geonhui Jang, Yonghyun Jeong, Junghyo Jo, and Gayoung Lee. Direct unlearning optimization for robust and safe text-to-image models. *arXiv preprint arXiv:2407.21035*, 2024. 2

[40] Gaurav Parmar, Krishna Kumar Singh, Richard Zhang, Yijun Li, Jingwan Lu, and Jun-Yan Zhu. Zero-shot image-to-image translation. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–11, 2023. 2

[41] Minh Pham, Kelly O Marshall, Chinmay Hegde, and Niv Cohen. Robust concept erasure using task vectors. *arXiv preprint arXiv:2404.03631*, 2024. 2

[42] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 1

[43] Samuele Poppi, Tobia Poppi, Federico Cocchi, Marcella Cornia, Lorenzo Baraldi, Rita Cucchiara, et al. Safe-clip: Removing nsfw concepts from vision-and-language models. In *Proceedings of the European Conference on Computer Vision*, 2024. 2

[44] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 6

[45] Javier Rando, Daniel Paleka, David Lindner, Lennart Heim, and Florian Tramèr. Red-teaming the stable diffusion safety filter. *arXiv preprint arXiv:2210.04610*, 2022. 2

[46] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1, 2, 3, 7, 8, 4

[47] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22500–22510, 2023. 2

[48] Patrick Schramowski, Manuel Brack, Björn Deiseroth, and Kristian Kersting. Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22522–22531, 2023. 2, 8

[49] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022. 1

[50] Jing Shi, Wei Xiong, Zhe Lin, and Hyun Joon Jung. Instantbooth: Personalized text-to-image generation without test-time finetuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8543–8552, 2024. 2

[51] Zhuan Shi, Jing Yan, Xiaoli Tang, Lingjuan Lyu, and Boi Faltings. Rlcp: A reinforcement learning-based copyright protection method for text-to-image diffusion model. *arXiv preprint arXiv:2408.16634*, 2024. 2

[52] Koushik Srivatsan, Fahad Shamshad, Muzammal Naseer, and Karthik Nandakumar. Stereo: Towards adversarially robust concept erasing from text-to-image generation models. *arXiv preprint arXiv:2408.16807*, 2024. 2

[53] Yu-Lin Tsai, Chia-Yi Hsu, Chulin Xie, Chih-Hsun Lin, Jia-You Chen, Bo Li, Pin-Yu Chen, Chia-Mu Yu, and Chun-Ying Huang. Ring-a-bell! how reliable are concept removal methods for diffusion models? *arXiv preprint arXiv:2310.10012*, 2023. 3, 8

[54] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1921–1930, 2023. 2

[55] Haofan Wang, Matteo Spinelli, Qixun Wang, Xu Bai, Zekui Qin, and Anthony Chen. Instantstyle: Free lunch towards style-preserving in text-to-image generation. *arXiv preprint arXiv:2404.02733*, 2024. 1

[56] Yuxiang Wei, Yabo Zhang, Zhilong Ji, Jinfeng Bai, Lei Zhang, and Wangmeng Zuo. Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15943–15953, 2023. 2

[57] Yuxiang Wei, Zhilong Ji, Jinfeng Bai, Hongzhi Zhang, Lei Zhang, and Wangmeng Zuo. Masterweaver: Taming editability and face identity for personalized text-to-image generation. In *European Conference on Computer Vision*, pages 252–271. Springer, 2025. 2

[58] Yuxiang Wei, Yiheng Zheng, Yabo Zhang, Ming Liu, Zhilong Ji, Lei Zhang, and Wangmeng Zuo. Personalized image generation with deep generative models: A decade survey. *arXiv preprint arXiv:2502.13081*, 2025. 1

[59] Yongliang Wu, Shiji Zhou, Mingzhuo Yang, Lianzhe Wang, Wenbo Zhu, Heng Chang, Xiao Zhou, and Xu Yang. Unlearning concepts in diffusion model via concept domain correction and concept preserving gradient. *arXiv preprint arXiv:2405.15304*, 2024. 2

[60] Binxin Yang, Shuyang Gu, Bo Zhang, Ting Zhang, Xuejin Chen, Xiaoyan Sun, Dong Chen, and Fang Wen. Paint by example: Exemplar-based image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18381–18391, 2023. 2

[61] Yijun Yang, Ruiyuan Gao, Xiaosen Wang, Tsung-Yi Ho, Nan Xu, and Qiang Xu. Mma-diffusion: Multimodal attack on diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7737–7746, 2024. 8

[62] Yijun Yang, Ruiyuan Gao, Xiao Yang, Jianyuan Zhong, and Qiang Xu. Guardt2i: Defending text-to-image models from adversarial prompts. *arXiv preprint arXiv:2403.01446*, 2024. 2

[63] 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. *arXiv preprint arXiv:2410.12761*, 2024.

[64] Gong Zhang, Kai Wang, Xingqian Xu, Zhangyang Wang, and Humphrey Shi. Forget-me-not: Learning to forget in text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1755–1764, 2024. 2

[65] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 2

[66] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 6

[67] Yabo Zhang, Yuxiang Wei, Dongsheng Jiang, Xiaopeng Zhang, Wangmeng Zuo, and Qi Tian. Controlvideo: Training-free controllable text-to-video generation. *arXiv preprint arXiv:2305.13077*, 2023. 1

[68] 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. *arXiv preprint arXiv:2405.15234*, 2024. 2, 6, 7, 4

[69] Yimeng Zhang, Jinghan Jia, Xin Chen, Aochuan Chen, Yihua Zhang, Jiancheng Liu, Ke Ding, and Sijia Liu. To generate or not? safety-driven unlearned diffusion models are still easy to generate unsafe images... for now. In *European Conference on Computer Vision*, pages 385–403. Springer, 2025. 3, 8

[70] Yabo Zhang, Xinpeng Zhou, Yihan Zeng, Hang Xu, Hui Li, and Wangmeng Zuo. Framepainter: Endowing interactive image editing with video diffusion priors. *arXiv preprint arXiv:2501.08225*, 2025. 2

[71] Mengnan Zhao, Lihe Zhang, Tianhang Zheng, Yuqiu Kong, and Baocai Yin. Separable multi-concept erasure from diffusion models. *arXiv preprint arXiv:2402.05947*, 2024. 2