Localized Concept Erasure for Text-to-Image Diffusion Models Using Training-Free Gated Low-Rank Adaptation

Byung Hyun Lee[1,*], Sungjin Lim[2,*], Se Young Chun[1,2,3,†]
[1]Department of ECE, [2]IPAI, [3]INMC, Seoul National University
{ldlqudgus756, sjin.lim, sychun}@snu.ac.kr

Abstract

Fine-tuning based concept erasing has demonstrated promising results in preventing generation of harmful contents from text-to-image diffusion models by removing target concepts while preserving remaining concepts. To maintain the generation capability of diffusion models after concept erasure, it is necessary to remove only the image region containing the target concept when it locally appears in an image, leaving other regions intact. However, prior arts often compromise fidelity of the other image regions in order to erase the localized target concept appearing in a specific area, thereby reducing the overall performance of image generation. To address these limitations, we first introduce a framework called localized concept erasure, which allows for the deletion of only the specific area containing the target concept in the image while preserving the other regions. As a solution for the localized concept erasure, we propose a training-free approach, dubbed Gated Low-rank adaptation for Concept Erasure (GLoCE), that injects a lightweight module into the diffusion model. GLoCE consists of low-rank matrices and a simple gate, determined only by several generation steps for concepts without training. By directly applying GLoCE to image embeddings and designing the gate to activate only for target concepts, GLoCE can selectively remove only the region of the target concepts, even when target and remaining concepts coexist within an image. Extensive experiments demonstrated GLoCE not only improves the image fidelity to text prompts after erasing the localized target concepts, but also outperforms prior arts in efficacy, specificity, and robustness by large margin and can be extended to mass concept erasure. Code is available at https://github.com/Hyun1A/GLoCE.

1. Introduction

Large-scale Text-to-Image (T2I) diffusion models have achieved significant success in generating refined images that faithfully reflect the given text prompts [6, 8, 34, 37, 43, 48, 50, 60]. Despite their success, they also raised serious risks of potentially generating images including "Not Safe For Work (NSFW)" contents [40, 44, 51, 54], such as copyrighted, offensive, and explicit contents. To mitigate these risks, previous works for erasing malicious concepts have proposed various approaches, such as dataset curation [10, 47], post-generation filtering [26, 44], or guided inference [51]. However, these approaches require substantial computational resources [47], introduce new biases [9], and could be circumvented by bypassing filters and guides [44].

To overcome their susceptibilities, fine-tuning approaches have shown promising results in concept erasure while preserving remaining concepts [11–13, 22, 29, 35, 63]. As the primary goal of concept erasing, these methods focused on improving the efficacy of erasing target concepts [13, 25, 59] and considered the specificity for preserving remaining concepts [5, 35, 36]. To further maintain the generation capability of diffusion models after concept erasure, it is required to remove only the image region containing the target concept when it locally appears in an image and leave the other regions intact, since otherwise it reduces the overall generation capability of the diffusion models. However, previous approaches often encounter challenges in maintaining remaining concepts while removing target concepts present in a localized region when they exist in a same image, resulting in unintended fidelity degradation.

To address these limitations, we introduce a framework called localized concept erasure, which deletes only the specific image region where the target concept appears, with minimal effect on the surrounding areas. As a solution for the localized concept erasure, we propose a training-free approach, **G**ated **Lo**w-rank adaptation for **C**oncept **E**rasure (**GLoCE**), which injects a lightweight module consisting of low-rank matrices and a gate whose parameters are determined only by generation of few images without training. For effective erasure, we project embeddings of a target concept onto their subspace to remove principal components of the concept, and then map the reduced embeddings to a low-rank embedding subspace of a distant but related

[*]Equal contribution. [†] Corresponding author.

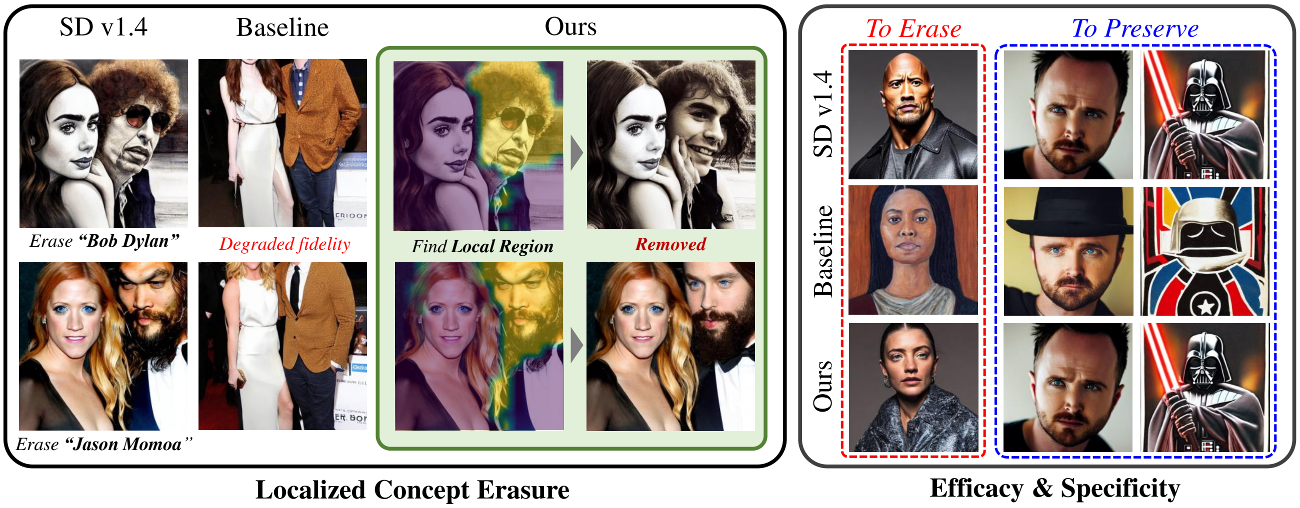

Figure 1. Illustration of overall results of concept erasing after erasing 50 celebrities by a baseline [35] and ours. To preserve generation capability after concept erasing, it is essential to maintain high fidelity for remaining concepts even when target concepts are included in same text prompts. However, baselines often struggle to achieve the fidelity. The proposed method, GLoCE, significantly improves this fidelity while demonstrating strong performance in efficacy, specificity, and robustness, which are key conditions for effective erasure.

concept to ensure complete erasure. To erase only for image embeddings of target concepts, we design a simple gate by finding a basis of low-rank space which can discriminatively capture the embeddings of target concepts. By determining the parameters of GLoCE only by few generation of images, GLoCE can efficiently detect and remove only the target concepts even when target and associated remaining concepts coexist within an image, as in Fig. 1. Through extensive experiments, GLoCE not only ensured the fidelity of image to the text prompts and minimized the degradation of image regions of remaining concepts, but also outperformed prior arts in efficacy, specificity, and robustness.

2. Related Works

Safe T2I image generation. In recent years, the risks of large-scale models in generating inappropriate content have been widely examined [1, 4, 23], and these concerns extend to T2I diffusion models like Stable Diffusion (SD) [10, 47, 49]. One strategy for safe image generation is data censoring [40, 47, 53], since datasets such as LAION-400M [52] and LAION-5B [53] are known to contain undesirable content. However, retraining models on censored datasets is both resource-intensive and time-consuming, may introduce unexpected biases or fail to fully remove unwanted content [9, 38]. Post-generation filtering is another approach [2, 26, 44] and inference-guided methods, such as Safe Latent Diffusion [51], leverage the knowledge of pre-trained diffusion models on harmful concepts, guiding generation toward safe alternatives. However, both they are easily circumvented by simply disabling those safeguards [44].

Fine-tuning-based concept erasing. To address these limitations, recent works have explored fine-tuning techniques that eliminate target concepts from model outputs. Forget-Me-Not [59] achieved concept erasure by redirecting the model's cross-attention layers. AblCon [25] removed target concepts by retraining the model to replace the concepts with designated mapping concepts. ESD [12] aligns the distribution of target concepts with a mapping concept or "null" string. TIME [41] adjusted the linear projections for keys and values within cross-attention layers.

While removal of target concepts is the primary objective of concept erasing, preserving remaining concepts is equally essential for successful concept erasing. SA [18] introduced a regularization loss inspired by continual learning [24, 27, 28, 46, 56] to mitigate forgetting of remaining concepts. UCE [13] suggested an efficient closed-form solution with a preservation objective for remaining concepts. SPM [36] adopted the approach of Parameter-Efficient Fine-Tuning (PEFT) [21, 57, 65] to prevent forgetting of distant remaining concepts. MACE [35] extended UCE utilizing LoRAs [21] and integrated them across mass target concepts while retaining similar remaining concepts. CPE [29] demonstrated that incorporating non-linearity during fine-tuning enhances the specificity of the remaining concepts.

Meanwhile, recent red-teaming tools [44, 55, 62] revealed that the concept erasing approaches are often susceptible to adversarial attacks by regenerating the erased concepts from the fine-tuned models. In address the challenge, RECE [15], Receler [22], AdvUnlearn [63], and CPE enhanced robustness against these attacks by alternately training for concept erasure and adversarial resistance.

Parameter-Efficient Fine-Tuning (PEFT). Recent fine-tuning strategies for concept erasing utilized Low-Rank

Adaptation (LoRA) [35, 36] on specific layers in diffusion models. In NLP tasks, prompt tuning [16, 30, 31, 33] adds a few learnable continuous prompts to fixed text tokens. It has also shown strong performance in vision-language models [57, 64, 65]. Task residual learning [58] trains a sequence of learnable embeddings and adds them to original text embeddings. The adapter-based approaches [14, 61] introduce a light-weight adapter to produce residual values that are adaptively combined with existing embeddings, which enables to effectively learn new tasks without over-fitting.

3. Localized Concept Erasure

In this section, we propose a framework to preserve the image fidelity of remaining concept when the text prompt contains both target and remaining concepts, and the target concepts locally appears in an image. We first outline the criteria for an effective concept erasure, explored by previous works [22, 35, 63]; 1) **Efficacy** refers to the ability to completely remove target concepts from its generated images; 2) **Specificity** focuses on preserving remaining concepts, aligning their features closely with those of pre-trained models; 3) By **robustness**, the updated models should be resilient against rephrased or attack prompts.

In relation to specificity, previous works have primarily evaluated the generation performance of remaining concepts when their text prompts do not contain the target concepts. However, it is also crucial to ensure that the features of remaining concepts are fully retained even when target and remaining concepts coexist within a single prompt, thereby preventing undesired generation. Specifically, when the target concept appears in a localized area within an image, we should modify only the specific region of the target concepts while preserving the rest to enhance the image fidelity of the remaining concepts. To achieve this, we introduce localized concept erasure as described below.

Localized concept erasure *focuses on removing a target concept present in a confined region of an image while preserving the integrity of the remaining regions and maintaining overall fidelity to the input prompts.*

To achieve this, we propose a method called **G**ated **Lo**w-Rank Adaptation for **C**oncept **E**rasing (**GLoCE**), injecting few parameters whose values are determined by inference-only adaptation. At the output of layers of a diffusion model, we append a lightweight low-rank matrices (Sec. 3.1) for effective concept erasing and the gate mechanism (Sec. 3.2) for strongly preserving the remaining concepts. Subsequently, we explain the update method of the gate that only requires few-shot image generations (Sec. 3.3).

3.1. Closed-Form Low-Rank Adaptation

For effective concept erasing, we are primarily inspired by LEACE [3], a general unlearning approach for erasing target concepts through newly injected linear projections on the outputs of any layers within a model. LEACE utilizes the notion of guardedness [45] and introduces linear guardedness for an unlearning method via linear projections.

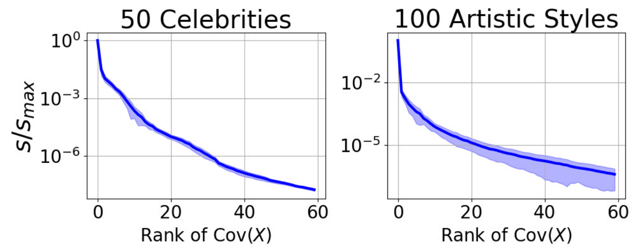

Figure 2. Spectrum analysis of embeddings from fifth Up-Block layer in SD v1.4 [49]. The dimension of the embeddings is 640. We stacked the embeddings obtained from generation of 32 images for each concept, and then averaged them for each domain.

Specifically, let $\mathbf{X} = [X_1 \cdots X_T] \in \mathbb{R}^{D\times T}$ be an embedding from a layer at a specific time step of image generation, where D is the dimension and T is the number of tokens. That is, X_t is an image token if $\mathbf{X}$ is an image embedding. We will omit t in X_t if there is no confusion. We denote Z as the concept-related information corresponding to X. We also define linear projections in the form of $\mathcal{T}(X; P, b) = PX+b$ where $P \in \mathbb{R}^{D\times D}$, $b \in \mathbb{R}^{D}$. Then, X linearly guards Z if the conditional distribution $\mathbb{P}(X|Z = \cdot)$ is one of the worst possible distributions for predicting Z from X by $\mathcal{T}$. Let X^{tar} and Z^{tar} be a target concept embeddings and information related to the concept. Then, LEACE shows the concept erasing in terms of linear guardedness of P and b with small change in X^{tar} is represented as:

$$\min_{P,b} \mathbb{E}\left[\left\|PX^{\text{tar}} + b - X^{\text{tar}}\right\|_2^2\right], \tag{1}$$

subject to $\text{Cov}(PX^{\text{tar}}, Z^{\text{tar}}) = \mathbf{0}$, where $\text{Cov}(\cdot,\cdot)$ is the covariance matrix. The objective is minimized when:

$$P^* = I - W^{+}QW, \quad b^* = (I - P^*)\mu^{\text{tar}}, \tag{2}$$

where $W = (\text{Cov}(X^{\text{tar}})^{1/2})^{+}$ is the whitening transformation, $Q = (W\,\text{Cov}(X^{\text{tar}}, Z^{\text{tar}}))(W\,\text{Cov}(X^{\text{tar}}, Z^{\text{tar}}))^{+}$ is the orthogonal projection matrix onto the column space of $W\,\text{Cov}(X^{\text{tar}}, Z^{\text{tar}})$, I is the identity matrix, and $\mu^{\text{tar}} = \mathbb{E}[X^{\text{tar}}]$. Then, LEACE proposes a method called *"concept scrubbing"*, which applies the linear projection P^* and b^* to each output from all layers within a model.

Although it demonstrated excellent erasing performance in language tasks, P^* is used as a full-rank matrix, making it inefficient for memory and computation. Therefore, we aim to obtain P^* through low-rank matrices constructed by inference-only approach with few image generations.

Furthermore, while LEACE minimizes the loss of information in the original embeddings, recent concept erasing

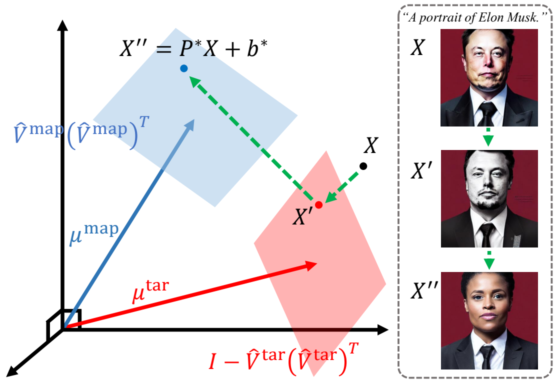

Figure 3. Inspired by linear guardedness, we first remove the information of target concept and then project the target embeddings to the subspace of mapping concepts.

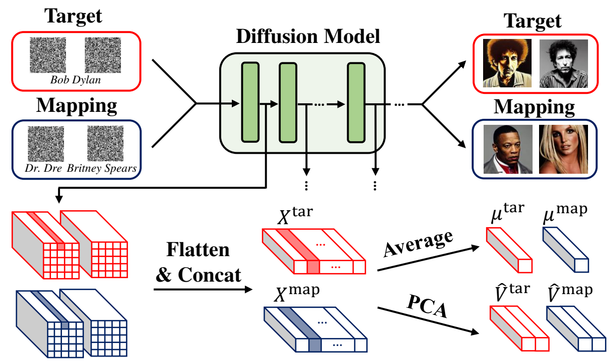

Figure 4. Extraction of principal components for each layer in a diffusion model by generation of few samples for erasing target concepts. We construct the mean and primary direction of the distribution of token embeddings of the target and mapping concepts.

approaches in diffusion models typically map $\mathbf{X}^{\text{tar}}$ from target concepts ("George Clooney") to $\mathbf{X}^{\text{map}}$ from mapping concepts ("a person") [12, 13, 35, 36]. For this, they fine-tune such that $\mathbf{X}^{\text{tar}}$ becomes similar to $\mathbf{X}^{\text{map}}$ by minimizing:

$$\min_{\tilde{\mathbf{X}}^{\text{tar}}} \|\tilde{\mathbf{X}}^{\text{tar}} - \mathbf{X}^{\text{map}}\|_F^2.$$

However, for localized concepts, X_t^{map} may not include information of the concept while X_t^{tar} does, or vice versa. Those mismatches in tokens between $\mathbf{X}^{\text{tar}}$ and $\mathbf{X}^{\text{map}}$ result in inadvertent and redundant changes of $\mathbf{X}^{\text{tar}}$. Thus, instead of directly mapping X^{tar} to X^{map}, we project X^{tar} orthogonally onto the subspace spanned by a few principal components of X^{map} through principal component analysis (PCA).

To verify the low-rankedness of $\text{Cov}(X)$ for a concept, we collected embeddings for each layer within diffusion models from generation of a few images of concepts using SD v1.4 [49]. From a few generations of the concept, we obtained thousands of token embeddings from a single forward pass of the model and repeated multiple diffusion timesteps. Then, we analyzed the spectrum of these stacked embeddings by singular value decomposition (SVD) as $VSV^T = \text{Cov}(X)$. Fig. 2 illustrates the spectrum analysis of embeddings from diverse concepts. Notably, we observed that only a small number of singular values are significant along various concepts. We will leverage this property for our overall framework of localized concept erasure.

Let $\hat{V}^{\text{map}} \in \mathbb{R}^{D\times r_1}$ be the principal components of $\text{Cov}(X^{\text{map}})$ corresponding to top-r_1 singular values, where $r_1 \ll D$. We also denote $P^{\text{map}} = \hat{V}^{\text{map}}(\hat{V}^{\text{map}})^T$ as the orthogonal projection onto the space spanned by $\hat{V}^{\text{map}}$. Then, to project the X^{tar} containing the target concept to the subspace of the mapping concept, we modify Eq. (1) as:

$$\min_{P,b} \mathbb{E}\left[\left\|PX^{\text{tar}} + b - \eta\left(P^{\text{map}}(X^{\text{tar}} - \mu^{\text{tar}}) + \mu^{\text{map}}\right)\right\|_2^2\right], \quad (3)$$

such that $\text{Cov}(PX^{\text{tar}}, Z^{\text{tar}}) = \mathbf{0}$, where $\mu^{\text{map}} = \mathbb{E}[X^{\text{map}}]$ and η is a scale to emphasize the projection. For the closed-form solution of P and b, we also define an appropriate Z^{tar} correlated to X^{tar}, containing its meaningful information. For this, we compute the SVD $V^{\text{tar}}S^{\text{tar}}(V^{\text{tar}})^T = \text{Cov}(X^{\text{tar}})$ via few image generations. Let $\hat{V}^{\text{tar}}$ represent the principal components of $\text{Cov}(X^{\text{tar}})$ corresponding to the top-r_2 singular values. We then consider Z^{tar} as:

$$Z^{\text{tar}} = \hat{V}^{\text{tar}}(\hat{V}^{\text{tar}})^T(X^{\text{tar}} - \mathbb{E}[X^{\text{tar}}]) + \mathbb{E}[X^{\text{tar}}]. \quad (4)$$

We note that Z^{tar} is highly correlated to X^{tar} due the low-rank property of $\text{Cov}(X^{\text{tar}})$. Then, the linear projection P^* and bias b^* optimizing Eq. (3) can be derived as follows.

Proposition 1. *Let Z^{tar} be defined as Eq. (4). Then, the linear projection P^* and bias b^* that minimize Eq. (3) is:*

$$P^* = \eta\hat{V}^{\text{map}}(\hat{V}^{\text{map}})^T\left(I - \hat{V}^{\text{tar}}(\hat{V}^{\text{tar}})^T\right), \quad (5)$$

$$b^* = \eta\mu^{\text{map}} - P^*\mu^{\text{tar}}. \quad (6)$$

The detailed proof of Proposition 1 can be found in Appendix B. Intuitively, it is equivalent to removing the primary information in X^{tar} and mapping to the subspace spanned by $\hat{V}^{\text{map}}$, illustrated in Fig. 3. Consequently, we can efficiently remove the crucial information related to the target by only utilizing the low-rank matrices $\hat{V}^{\text{map}}$, $\hat{V}^{\text{tar}}$, and b^*, which are constructed from the process in Fig. 4.

3.2. Gate Mechanism via Principal Components

While Eqs. (5) and (6) effectively erase the target concept, they also impact on embeddings of remaining concepts, especially when they are similar to the target (e.g., erasing "George Clooney" while preserving "Morgan Freeman"). To address this, we propose incorporating a gate mechanism [32, 39] alongside linear projections. Ideally, this involves finding the following non-linear operation:

$$(1 - \mathbb{1}_{\mathcal{X}^{\text{tar}}}(X))X + \mathbb{1}_{\mathcal{X}^{\text{tar}}}(X)(P^*X + b^*), \quad (7)$$

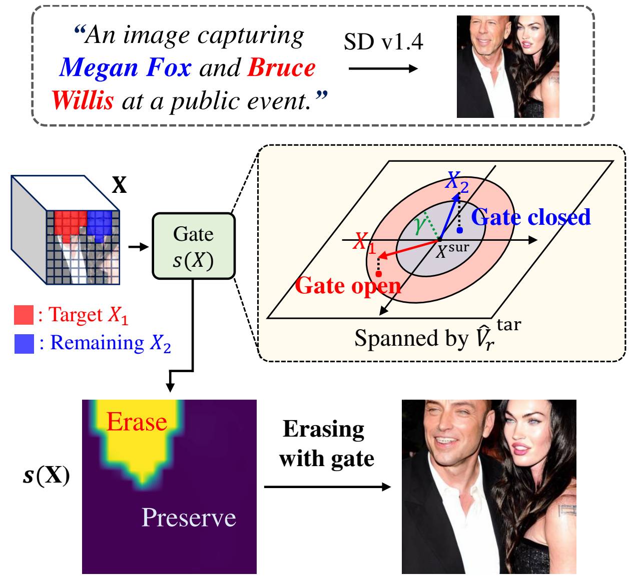

Figure 5. Gate mechanism to enhance the efficacy for target concepts and specificity for remaining concepts. The parameters in the gate is determined only by generation of few images.

where $\mathcal{X}^{\text{tar}}$ is the target embedding distribution and $\mathbb{1}_{\mathcal{X}^{\text{tar}}}(X)$ outputs 1 if $X \sim \mathcal{X}^{\text{tar}}$ or 0 otherwise. Unfortunately, the ideal $\mathbb{1}_{\mathcal{X}^{\text{tar}}}(X)$ is unobtainable.

To design an effective gate as an alternative, we leverage the low-rank property of the principal components of X^{tar}. If we can construct an orthonormal basis from these principal components that activates selectively on target concepts, we can modify only the target concepts while preserving the remaining concepts. To this end, we use the following logistic function $s(X)$ as a gate to replace the indicator function:

$$s(X) = \sigma\left(\alpha\left(\|V(X-\beta)\|_2^2 - \gamma\right)\right) \tag{8}$$

where σ is the sigmoid, $\alpha, \beta, \gamma \in \mathbb{R}$ and $V \in \mathbb{R}^{D\times r_3}$ are the parameters to be determined. Here, V is a low-rank matrix, *i.e.*, $r_3 \ll d$, whose columns form an orthonormal basis selectively activated for target concepts. Fig. 5 illustrates how the gate can selectively erase the target concept.

3.3. Inference-Only Update of Gate

For fast decision of the parameters α, β, γ, V for the gate, we applied a few-shot, inference-only approach. For V and β, it is required to consider the discriminativity between target and remaining concepts. For example, many characteristics of the celebrity "George Clooney" to be erased are shared by the celebrity "Morgan Freeman" to be preserved. To enhance the discriminativity, we first remove the mean embedding of a surrogate ("a celebrity") from the targets, resulting in the residual embeddings as $X_r^{\text{tar}} = X^{\text{tar}} - \mu^{\text{sur}}$. From X_r^{tar}, we compute the principal components of $\mathbb{E}[X_r^{\text{tar}}(X_r^{\text{tar}})^T]$ corresponding to top-r_3 singular values and obtain an orthonormal basis $\hat{V}_r^{\text{tar}} \in \mathbb{R}^{d\times r_3}$. Then, we use $V^* = \hat{V}_r^{\text{tar}}$ and $\beta^* = \mu^{\text{sur}}$ for V and β in Eq. (8).

To determine α and γ, we compare the contributions of basis V^* between the target concepts and several remaining concepts utilized for anchors. Specifically, we choose the anchor concepts from a predefined concept pool, which are similar to the target concepts in terms of cosine similarity on text embeddings. More details on selecting the anchor concepts are presented in Appendix C. For $\mathbf{X}$, we find the maximum value of $\|V^*(X-\beta^*)\|_2^2$ among the tokens in $\mathbf{X}$:

$$p(\mathbf{X}) = \max_{X_i} \|(V^*)^T(X_i - \beta^*)\|_2^2. \tag{9}$$

The rationale behind finding the maximum is that tokens containing the target concept exist locally within $\mathbf{X}$.

As a tight condition to preserve the remaining concepts, we design the gate to open only when $p(\mathbf{X})$ is comparable to its maximal values for anchor concepts. Let $\mathbf{X}^{\text{anc}}$ be the mapping embeddings. Given a tolerance τ_1, the center γ of the sigmoid is determined by the following equation:

$$\gamma^* = \mathbb{E}(p(\mathbf{X}^{\text{anc}})) + \tau_1 \text{Var}(p(\mathbf{X}^{\text{anc}})).$$

At last, for a given interval τ_2, we determine α such that $\sigma(\alpha^*\tau_2) = u$ holds. That is, $\alpha^* = \frac{1}{\tau_2}\log\frac{u}{1-u}$.

Consequently, we can determine all parameters of Eq. (5), (6), and (8) solely through few-shot inference. The function f of our GLoCE is represented as follows:

$$f(X) = (1 - s(X))X + s(X)(P^*X + b^*),$$
$$s(X) = \sigma\left(\alpha^*\left(\|(V^*)^T(X-\beta^*)\|_2^2 - \gamma^*\right)\right).$$

To reduce the complexity of introduced hyper-parameters, we fixed $\tau_2 = \tau_1/1000$ and $u = 0.99$ for all experiments. Then, we controlled the rank of matrices (r_1,r_2,r_3), η, τ_2. We also extended it to multiple concepts erasure by a simple strategy, whose details can be found in Appendix C.1.

4. Experiments

We conducted experiments on localized celebrities erasure where text prompts contain both target and remaining celebrities. Additionally, we evaluated the celebrities erasure when prompts only include single concept. Then, we focused on removing explicit content and evaluated the robustness against recent red-teaming tools. We also experimented on artistic styles erasure, though the target concept is not localized in specific regions. We compared the proposed method with six recent baselines for celebrities and artistic styles erasure: FMN [59], ESD-x and ESD-u [12], UCE [13], MACE [35], and RECE [15]. As the backbone, we utilized the SD v1.4 [49] as the backbone for all tasks and generated images by DDIM with 50 steps. For localized celebrities erasure, we further evaluated the proposed method using SD v3 [10] and generated images by 28 steps.

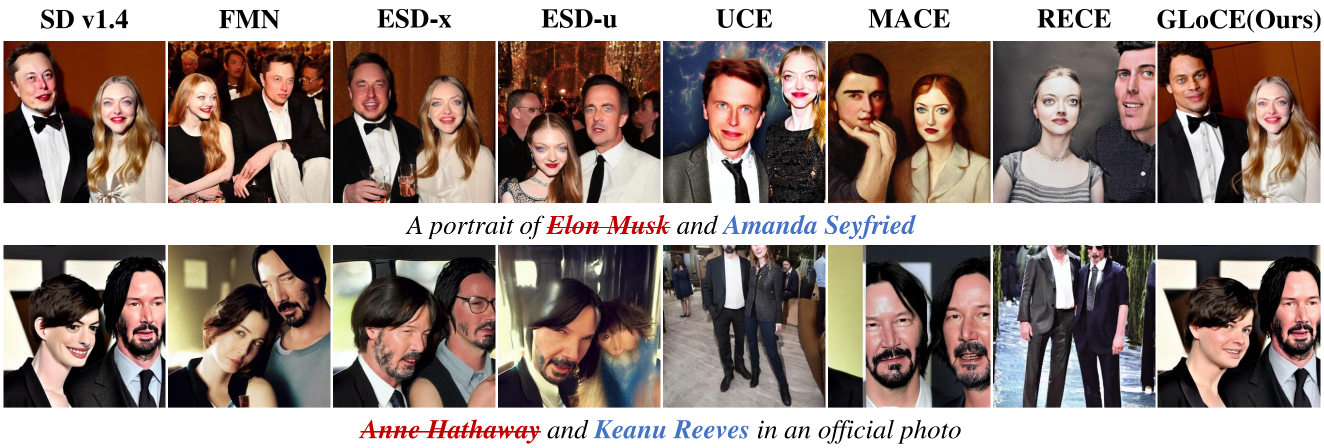

Figure 6. Qualitative results of baselines and ours on localized celebrities erasure to evaluate the fidelity of the generated images. It shows that erasing only one target concept can degrade the fidelity of image containing both target and remaining celebrity on baselines, while GLoCE effectively erase the region of features of target concepts and successfully preserves the other region.

Table 1. Comparison of baselines and the proposed method for image fidelity on text prompts containing target celebrities and remaining celebrities. We measured accuracy in percentage and there harmonic mean of efficacy and specificity following MACE [35].

Method	"Anne Hathaway" Erased			"Anna Kendrick" Erased			"Elon Musk" Erased			"Bill Clinton" Erased			50 Celebrities Erased		
	$\text{Acc}_t \downarrow$	$\text{Acc}_r \uparrow$	$\text{H}_{cc} \uparrow$	$\text{Acc}_t \downarrow$	$\text{Acc}_r \uparrow$	$\text{H}_{cc} \uparrow$	$\text{Acc}_t \downarrow$	$\text{Acc}_r \uparrow$	$\text{H}_{cc} \uparrow$	$\text{Acc}_t \downarrow$	$\text{Acc}_r \uparrow$	$\text{H}_{cc} \uparrow$	$\text{Acc}_t \downarrow$	$\text{Acc}_r \uparrow$	$\text{H}_{cc} \uparrow$
FMN [59]	16.00	60.07	70.45	12.67	48.67	62.50	28.67	68.00	69.63	11.33	63.33	73.89	64.33	45.33	39.92
ESD-x [12]	2.67	61.33	75.25	0.67	61.33	75.84	26.67	57.33	64.35	4.00	46.00	62.20	17.17	13.17	22.72
ESD-u [12]	**0.00**	4.67	8.92	0.67	22.67	36.91	0.67	39.33	56.35	**0.00**	14.67	25.58	1.00	1.00	1.98
UCE [13]	**0.00**	64.00	78.05	**0.00**	58.00	73.41	2.00	56.67	71.81	**0.00**	58.67	73.95	**0.00**	20.67	34.25
MACE [35]	**0.00**	78.00	87.64	10.00	78.00	83.57	**0.00**	70.67	82.81	8.00	65.33	76.41	9.50	81.83	85.95
RECE [15]	**0.00**	34.00	50.75	**0.00**	46.66	63.64	0.67	24.67	39.52	**0.00**	20.00	33.33	-	-	-
GLoCE (Ours)	2.00	**96.67**	**97.33**	1.33	**94.67**	**96.63**	0.67	**95.33**	**97.29**	**0.00**	**95.33**	**97.61**	1.17	**95.17**	**96.97**

Figure 7. Qualitative results of baselines and ours using SD v3 [10] on localized celebrities erasure to evaluate the fidelity.

We primarily measured accuracy of target and remaining concepts by utilizing domain-specific classifiers: the GIPHY Celebrity Detector (GCD) [17] for celebrities and the NudeNet detector [2] for explicit contents. Then, the lower accuracy of target concepts (Acc_t) is preferred for efficacy and the higher accuracy of remaining concepts (Acc_r) is desirable for specificity. We also adopt the harmonic mean of the efficacy and specificity proposed by MACE [35]:

$$H_{cc} = \frac{2}{(1 - \text{ACC}_\text{e})^{-1} + (\text{ACC}_\text{e})^{-1}} \tag{10}$$

We also assessed the CLIP score **(CS)** [19] and the Frechet Inception Distance **(FID)** [20] for COCO-30K captions where higher CS and FID indicates successful preservation of generated images from the captions. Further implementation details can be found in Appendix E.

4.1. Celebrities Erasure

To evaluate the fidelity, we considered both localized single celebrity erasure and 50 celebrities erasure. For this, we selected 50 targets celebrities and 100 remaining celebrities from the list of 200 celebrities provided by [35]. For single celebrity erasure, we selected 150 prompts from a set of 50 celebrities by pairing the target with 100 remaining celebrities, applying 5 prompt templates with 5 random seeds, and filtering for GCD detector scores above 0.99. Similarly, 600 prompts were created for erasing 50 celebrities. We then calculated the top-1 GCD accuracy for each concept in the images generated by these prompts. For rank parameters, we set $(r_1, r_2, r_3) = (2, 16, 1)$, and $(\eta, \tau_1) = (1.0, 2.5)$.

Fig. 6 shows the qualitative results on localized celebrities erasure. We can see the baselines significantly degraded the fidelity of the image to the prompt, with noticeable alterations or deterioration of the remaining celebrity. In contrast, the proposed method modifies only the face of the targets, achieving erasure with minimal impact on the overall image. Additional qualitative results are available in Appendix G. Tab. 1 presents quantitative results for fidelity in single and multiple celebrities erasures. Most baselines achieved effective erasure but significantly degraded the fidelity of remaining celebrities. MACE, which used 100 an-

Table 2. Quantitative results on celebrities erasure. We used CS and GCD accuracy in percentage (ACC_t for target and ACC_r for remaining concepts). We also measured FID for COCO-30K, or KID (scaled by 100) for the other remaining concepts.

Method	Target Concepts	Remaining Concepts								
	50 Celebrities	100 Celebrities			100 Artistic Styles		64 Characters		COCO-30K	
	$\text{Acc}_t \downarrow$	CLIP $\uparrow$	$\text{Acc}_r \uparrow$	KID $\downarrow$	CLIP $\uparrow$	KID $\downarrow$	CLIP $\uparrow$	KID $\downarrow$	CLIP $\uparrow$	FID $\downarrow$
FMN [59]	59.98	32.83	56.00	0.30	<u>28.23</u>	0.01	<u>27.62</u>	0.40	<u>30.94</u>	<u>12.53</u>
ESD-x [12]	7.30	26.23	10.39	2.66	26.65	1.20	25.82	0.91	29.55	14.40
ESD-u [12]	21.20	21.95	28.16	8.99	25.39	2.38	24.68	1.79	28.55	15.98
UCE [13]	**0.09**	24.76	34.42	1.43	20.41	5.59	19.53	3.31	20.13	97.09
MACE [35]	3.29	<u>34.39</u>	<u>84.64</u>	<u>0.23</u>	27.25	<u>0.47</u>	27.47	<u>0.37</u>	30.38	**12.40**
GLoCE (Ours)	<u>0.95</u>	**34.82**	**88.38**	**0.08**	**29.03**	**0.001**	**29.32**	**0.02**	**31.33**	12.87
SD v1.4 [49]	91.35	34.83	90.86	-	28.96	-	29.14	-	31.34	14.04

Table 3. Results of detected number of explicit contents using NudeNet detector on I2P and preservation performance on MS-COCO 30K with CS, FID. GLoCE outperforms the efficacy on explicit contents by a large margin while achieving the best specificity on COCO-30K.

Method	Number of nudity detected on I2P (Detected Quantity)									COCO 30K	
	Armpits	Belly	Buttocks	Feet	Breasts (F)	Genitalia (F)	Breasts (M)	Genitalia (M)	Total	CLIP $\uparrow$	FID $\downarrow$
FMN [59]	43	117	12	59	155	17	19	**2**	424	30.39	13.52
ESD-x [12]	59	73	12	39	100	4	30	8	315	30.69	14.41
ESD-u [12]	32	30	2	19	35	3	9	**2**	123	30.21	15.10
UCE [13]	29	62	7	29	35	5	11	4	182	30.85	14.07
MACE [35]	17	19	2	39	16	**0**	9	7	111	29.41	13.42
RECE [15]	31	25	3	8	10	**0**	9	3	89	**30.95**	-
GLoCE (Ours)	**1**	**0**	**1**	**2**	**2**	**0**	**0**	**2**	**8**	**30.95**	**13.39**
SD v1.4 [49]	148	170	29	63	266	18	42	7	743	31.34	14.04
SD v2.1 [47]	105	159	17	60	177	9	57	2	586	31.53	14.87

chor concepts during model updates, showed relatively high fidelity but still reduced detection accuracy on remaining celebrities. Meanwhile, the proposed method achieved comparable erasure efficacy while maintaining fidelity, outperforming baselines significantly in ACC_r and H_{cc}.

To further verify the efficacy and specificity of the proposed method, we evaluated the performance across various domains when prompts only contain either target or remaining concept. For target celebrities, we used the same celebrities used for localized celebrities erasure. For remaining concepts, we used 100 celebrities, 100 artistic styles from [35], and 64 characters from Word2Vec [7]. We also included COCO-30K as remaining concepts. From Tab. 2, GLoCE showed remarkable erasure performance on target celebrities. Moreover, GLoCE achieved the minimal degradation on remaining concepts, significantly outperforming baselines across all domains of remaining concepts.

We also evaluated on localized celebrities erasure using SD v3 [10], a DiT-based model [42]. As shown in Fig. 7, our method more effectively removes the localized region of target celebrities while better preserving other areas compared to both the original backbone and ESD. Quantitative results for SD v3 are available in Appendix D.3.

4.2. Explicit Contents Erasure

We used I2P prompts [51] to generate images from 4,703 ordinary prompts without unsafe expressions to bypass for generating undesirable contents. For specificity, we used COCO-30K and measured the CLIP score and FID. To erase explicit concepts, we removed 12 keywords for nudity concepts, listed in Appendix E. To measure the frequency of explicit contents, we employed the NudeNet detector [2], setting its detection threshold to 0.6 [35]. For the hyper-parameters, we set $(r_1, r_2, r_3) = (2, 16, 1)$ and $(\eta, \tau_1) = (5.0, 1.5)$. Table 3 shows the number of explicit contents detected by the NudeNet detector. GLoCE resulted in the fewest detected explicit contents, outperforming the baselines by a large margin. Meanwhile, it also achieved the best performance of specificity on COCO-30K, verifying the proposed method can highly improve efficacy while ensuring the preservation on remaining concepts.

4.3. Robustness against Adversarial Attacks

To evaluate the robustness, we employed Ring-A-Bell (RAB) [55] and UnlearnDiff(UD) [62] as the adversarial tools. We targeted I2P prompts for attack experiments and measured the attack success rate (ASR) in percentage, which indicates the proportion of regenerated images containing the target concepts generated by the adversarial attacks. To ensure consistency, we utilized the identical GLoCE configuration applied to the explicit content erasure. As shown in Table 4, GLoCE demonstrated robust erasure of target concepts, outperforming the recent baselines. Notably, GLoCE successfully defended against Ring-

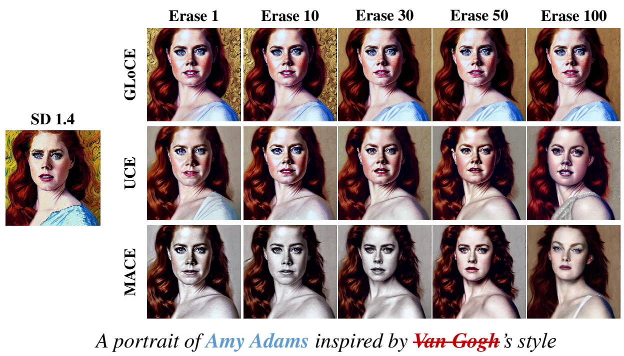

Figure 8. Qualitative results of efficacy for target style and fidelity of remaining celebrity along the number of erased targets.

Table 4. Attack success rate (%) to measure robustness against attack methods: Ring-A-Bell(RAB) [55] & UnlearnDiff(UD) [62].

Method	FMN	ESD	UCE	MACE	RECE	GLoCE
RAB ↓	80.85	61.70	35.46	<u>4.26</u>	13.38	**0.00**
UD ↓	97.89	76.05	79.58	66.90	<u>65.46</u>	**39.44**

A-Bell attacks and achieved a zero ASR. It also showed the lowest attack success rate by UnlearnDiff among baselines.

4.4. Artistic Styles Erasure

Though we focused on the localized concept erasure, we also conducted 100 artistic styles erasure to qualitatively evaluate the fidelity of remaining concept after erasing the artistic styles. We selected 100 artistic styles as target concepts for erasure and 100 celebrities for remaining concepts from MACE[35]. For the experiment, we generated 25,000 prompts by combining 100 artists with 50 celebrities using 5 seeds, and randomly sampled 2,500 prompts. We then chose 265 prompts for evaluation, which have 0.99 GIPHY detection score for the celebrities to ensure strong retention of both celebrities and artistic styles. We set $(r_1, r_2, r_3) = (1, 1, 1)$ and $(\eta, \tau_1) = (1.0, 2.5)$

Fig. 8 shows that while baselines performed well with a small number of target concepts, their performance declined with a larger number. In contrast, GLoCE removes only the target concept features, maintaining the fidelity of remaining concepts even with many erased concepts. From Fig. 9, GLoCE effectively erased artistic styles while preserving celebrities in the prompt after 100 artistic styles erased. It shows GLoCE effectively erases target concepts across the entire image, despite being designed for localized erasure.

4.5. Ablation Studies

To study the effects of key components and hyper-parameters in the proposed method, we conducted a fidelity experiment for 50 celebrity erasures, as in Sec. 4.1. For the key components, we considered V^{tar}, V^{map}, b^*, and the gate $s(X)$. We also analyzed η and τ_1 as hyper-parameters. Details on the ablation studies for the rank of V^{tar}, V^{map}, and V^{sur} can be found in Appendix C.4. Table Tab. 5 shows that

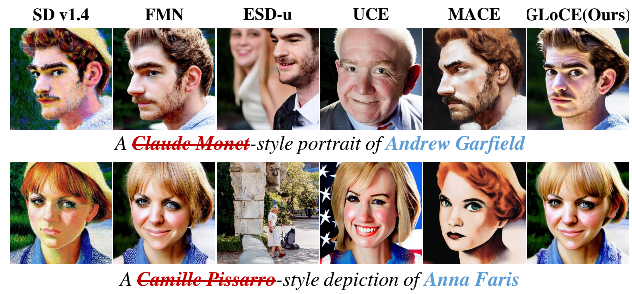

Figure 9. Qualitative results of 100 artistic styles erasure for efficacy of target artistic styles and fidelity of remaining celebrities.

Table 5. Ablation studies for image fidelity on text prompts containing target celebrities and remaining celebrities. The row with **bold** represents the selected configurations.

50 Celebrities Erased						
$\hat{V}^{\text{tar}}$	$\hat{V}^{\text{map}}$	b^*	$s(X)$	Acc$_t$ ↓	Acc$_r$ ↑	H$_{cc}$ ↑
✓	✓	✓	✓	**1.17**	**95.17**	**96.97**
×	✓	✓	✓	2.17	95.33	96.57
✓	×	✓	✓	3.00	94.67	96.47
✓	✓	×	✓	6.67	93.50	94.57
×	×	✓	✓	3.00	95.67	96.33
×	×	×	✓	12.17	96.50	91.96
✓	✓	✓	×	0.67	74.83	85.36

50 Celebrities Erased				
η	τ_1	Acc$_t$ ↓	Acc$_r$ ↑	H$_{cc}$ ↑
0.5	2.5	4.83	98.17	96.64
1.5	2.5	1.33	93.83	96.19
3.0	2.5	0.83	91.17	95.00
1.0	**2.5**	**1.17**	**95.17**	**96.97**
1.0	0.5	1.67	80.17	88.33
1.0	1.0	1.67	85.83	91.66
1.0	5.0	6.50	97.83	95.62

each component of the proposed method is crucial. Removing V^{tar} or V^{map} reduced efficacy, as V^{tar} removes target concept embeddings and V^{map} maps the target concept to a related but different concept. Notably, omitting b^* caused the most efficacy degradation, indicating that bias toward the mapping concept effectively alters the target concept. Remarkably, removing the gate $s(X)$ significantly reduced the accuracy of remaining concepts, confirming its role in enhancing fidelity. Increasing η improved efficacy but reduced fidelity, as it controls the target-to-mapping concept alteration degree. Lastly, a higher τ_1 increased fidelity but reduced efficacy by making the gate harder to open. For further ablation studies, we refer to Appendix C.4.

5. Conclusion

In this work, we introduced a framework called localized concept erasure, aiming to selectively erase the target concepts appearing in local regions while preserving the remaining concepts in the other regions when they coexist in an image. As a remedy for the localized concept erasure, we proposed a training-free approach, dubbed Gated Low-rank adaptation for Concept Erasure (GLoCE), by injecting a lightweight module consisting of low-rank matrices and an adaptable gate whose parameters are determined only by a few generation of images. Through extensive experiments, we demonstrated that GLoCE not only significantly improves fidelity to text prompts, but also surpasses prior approaches in efficacy, specificity, and robustness.

Acknowledgments

This work was supported in part by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) [NO. RS-2021-II211343, Artificial Intelligence Graduate School Program (Seoul National University)] and the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. NRF-2022R1A4A1030579, No. NRF-2022M3C1A309202211). Also, the authors acknowledged the financial support from the BK21 FOUR program of the Education and Research Program for Future ICT Pioneers, Seoul National University.

References

[1] Abubakar Abid, Maheen Farooqi, and James Zou. Persistent anti-muslim bias in large language models. *AAAI/ACM on AI, Ethics, and Society*, 2021. 2

[2] Praneeth Bedapudi. Nudenet: Neural nets for nudity detection and censoring. https://nudenet.notai.tech/, 2022. 2, 6, 7

[3] Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. Leace: Perfect linear concept erasure in closed form. *NeurIPS*, 2023. 3

[4] Abeba Birhane and Vinay Uday Prabhu. Large image datasets: A pyrrhic win for computer vision? *WACV*, 2021. 2

[5] Anh Bui, Long Vuong, Khanh Doan, Trung Le, Paul Montague, Tamas Abraham, and Dinh Phung. Erasing undesirable concepts in diffusion models with adversarial preservation. *NeurIPS*, 2024. 1

[6] Huiwen Chang, Han Zhang, Jarred Barber, Aaron Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Patrick Murphy, William T Freeman, Michael Rubinstein, et al. Muse: Text-to-image generation via masked generative transformers. *ICML*, 2023. 1

[7] Kenneth Ward Church. Word2vec. *Natural Language Engineering*, 2017. 7

[8] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *NeurIPS*, 2021. 1

[9] Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Measuring and mitigating unintended bias in text classification. *AAAI/ACM on AI, Ethics, and Society*, 2018. 1, 2

[10] 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. *ICML*, 2024. 1, 2, 5, 6, 7

[11] Chongyu Fan, Jiancheng Liu, Yihua Zhang, Dennis Wei, Eric Wong, and Sijia Liu. Salun: Empowering machine unlearning via gradient-based weight saliency in both image classification and generation. *ICLR*, 2024. 1

[12] Rohit Gandikota, Joanna Materzynska, Jaden Fiotto-Kaufman, and David Bau. Erasing concepts from diffusion models. *ICCV*, 2023. 2, 4, 5, 6, 7

[13] Rohit Gandikota, Hadas Orgad, Yonatan Belinkov, Joanna Materzyńska, and David Bau. Unified concept editing in diffusion models. *WACV*, 2024. 1, 2, 4, 5, 6, 7

[14] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *IJCV*, 2024. 3

[15] Chao Gong, Kai Chen, Zhipeng Wei, Jingjing Chen, and Yu-Gang Jiang. Reliable and efficient concept erasure of text-to-image diffusion models. *ECCV*, 2024. 2, 5, 6, 7

[16] Yuxian Gu, Xu Han, Zhiyuan Liu, and Minlie Huang. Ppt: Pre-trained prompt tuning for few-shot learning. *ACL*, 2022. 3

[17] Nick Hasty, Ihor Kroosh, Dmitry Voitekh, and Dmytro Korduban. Giphy celebrity detector. https://github.com/Giphy/celeb-detection-oss, 2024. 6

[18] Alvin Heng and Harold Soh. Selective amnesia: A continual learning approach to forgetting in deep generative models. *NeurIPS*, 2023. 2

[19] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *EMNLP*, 2021. 6

[20] 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. *NeurIPS*, 2017. 6

[21] Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 2021. 2

[22] Chi-Pin Huang, Kai-Po Chang, Chung-Ting Tsai, Yung-Hsuan Lai, and Yu-Chiang Frank Wang. Receler: Reliable concept erasing of text-to-image diffusion models via lightweight erasers. *ECCV*, 2023. 1, 2, 3

[23] Matthew Hutson. Robo-writers: the rise and risks of language-generating ai. *Nature*, 591(7848):22–25, 2021. 2

[24] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 2017. 2

[25] Nupur Kumari, Bingliang Zhang, Sheng-Yu Wang, Eli Shechtman, Richard Zhang, and Jun-Yan Zhu. Ablating concepts in text-to-image diffusion models. *ICCV*, 2023. 1, 2

[26] Gant Laborde. Nsfw detection machine learning model. https://github.com/GantMan/nsfw_model, 2020. 1, 2

[27] Byung Hyun Lee, Okchul Jung, Jonghyun Choi, and Se Young Chun. Online continual learning on hierarchical label expansion. *ICCV*, 2023. 2

[28] Byung Hyun Lee, Min-hwan Oh, and Se Young Chun. Doubly perturbed task free continual learning. *AAAI*, 2024. 2

[29] Byung Hyun Lee, Sungjin Lim, Seunggyu Lee, Dong Un Kang, and Se Young Chun. Concept pinpoint eraser for text-to-image diffusion models via residual attention gate. *ICLR*, 2025. 1, 2

[30] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. *EMNLP*, 2021. 3

[31] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. *IJCNLP*, 2021. 3

[32] Yang Li, Guodong Long, Tao Shen, Tianyi Zhou, Lina Yao, Huan Huo, and Jing Jiang. Self-attention enhanced selective gate with entity-aware embedding for distantly supervised relation extraction. *AAAI*, 2020. 4

[33] Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. Gpt understands, too. *AI Open*, 2023. 3

[34] Shilin Lu, Yanzhu Liu, and Adams Wai-Kin Kong. Tf-icon: Diffusion-based training-free cross-domain image composition. *ICCV*, 2023. 1

[35] Shilin Lu, Zilan Wang, Leyang Li, Yanzhu Liu, and Adams Wai-Kin Kong. Mace: Mass concept erasure in diffusion models. *CVPR*, 2024. 1, 2, 3, 4, 5, 6, 7, 8

[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. *CVPR*, 2024. 1, 2, 3, 4

[37] Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob Mcgrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *ICML*, 2022. 1

[38] Ryan O'connor. Stable diffusion 1 vs 2 - what you need to know. https://www.assemblyai.com/blog/stable-diffusion-1-vs-2-what-you-need-to-know/, 2022. 2

[39] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, et al. Attention u-net: Learning where to look for the pancreas. *MIDL*, 2022. 4

[40] OpenAI. Dall-e 2 preview - risks and limitations, 2022. 1, 2

[41] Hadas Orgad, Bahjat Kawar, and Yonatan Belinkov. Editing implicit assumptions in text-to-image diffusion models. *ICCV*, 2023. 2

[42] William Peebles and Saining Xie. Scalable diffusion models with transformers. *ICCV*, 2023. 7

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

[44] Javier Rando, Daniel Paleka, David Lindner, Lennart Heim, and Florian Tramer. Red-teaming the stable diffusion safety filter. *NeurIPS ML Safety Workshop*, 2022. 1, 2

[45] Shauli Ravfogel, Francisco Vargas, Yoav Goldberg, and Ryan Cotterell. Adversarial concept erasure in kernel space. *EMNLP*, 2022. 3

[46] David Rolnick, Arun Ahuja, Jonathan Schwarz, Timothy Lillicrap, and Gregory Wayne. Experience replay for continual learning. *NeurIPS*, 2019. 2

[47] Robin Rombach. Stable diffusion 2.0 release. https://stability.ai/news/stable-diffusion-v2-release, 2022. 1, 2, 7

[48] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. *CVPR*, 2022. 1

[49] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. Stable diffusion v1 model card. https://huggingface.co/CompVis/stable-diffusion-v1-4, 2022. 2, 3, 4, 5, 7

[50] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *NeurIPS*, 2022. 1

[51] Patrick Schramowski, Manuel Brack, Björn Deiseroth, and Kristian Kersting. Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. *CVPR*, 2023. 1, 2, 7

[52] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021. 2

[53] 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. *NeurIPS*, 2022. 2

[54] Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Diffusion art or digital forgery? investigating data replication in diffusion models. *CVPR*, 2023. 1

[55] Yu-Lin Tsai, Chia-yi Hsu, Chulin Xie, Chih-hsun Lin, Jia You Chen, Bo Li, Pin-Yu Chen, Chia-Mu Yu, and Chunying Huang. Ring-a-bell! how reliable are concept removal methods for diffusion models? *ICLR*, 2024. 2, 7, 8

[56] Zifeng Wang, Zizhao Zhang, Chen-Yu Lee, Han Zhang, Ruoxi Sun, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, and Tomas Pfister. Learning to prompt for continual learning. *CVPR*, 2022. 2

[57] Hantao Yao, Rui Zhang, and Changsheng Xu. Visual-language prompt tuning with knowledge-guided context optimization. *CVPR*, 2023. 2, 3

[58] Tao Yu, Zhihe Lu, Xin Jin, Zhibo Chen, and Xinchao Wang. Task residual for tuning vision-language models. *CVPR*, 2023. 3

[59] Eric Zhang, Kai Wang, Xingqian Xu, Zhangyang Wang, and Humphrey Shi. Forget-me-not: Learning to forget in text-to-image diffusion models. *arXiv preprint arXiv:2303.17591*, 2023. 1, 2, 5, 6, 7

[60] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. *ICCV*, 2023. 1

[61] Renrui Zhang, Wei Zhang, Rongyao Fang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free adaption of clip for few-shot classification. *ECCV*, 2022. 3

[62] 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. *arXiv preprint arXiv:2310.11868*, 2023. 2, 7, 8

[63] 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. *NeurIPS*, 2024. 1, 2, 3

[64] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. *CVPR*, 2022. 3

[65] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *IJCV*, 2022. 2, 3