This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Detect-and-Guide: Self-regulation of Diffusion Models for Safe Text-to-Image Generation via Guideline Token Optimization

Feifei Li Mi Zhang† Yiming Sun Min Yang†
School of Computer Science, Fudan University, China
{ffli23@m.,mi_zhang@,ymsun24@m.,m_yang@}fudan.edu.cn

Abstract

Text-to-image diffusion models have achieved state-of-the-art results in synthesis tasks; however, there is a growing concern about their potential misuse in creating harmful content. To mitigate these risks, post-hoc model intervention techniques, such as concept unlearning and safety guidance, have been developed. However, fine-tuning model weights or adapting the hidden states of the diffusion model operates in an uninterpretable way, making it unclear which part of the intermediate variables is responsible for unsafe generation. These interventions severely affect the sampling trajectory when erasing harmful concepts from complex, multi-concept prompts, thus hindering their practical use in real-world settings. In this work, we propose the safe generation framework Detect-and-Guide (DAG), leveraging the internal knowledge of diffusion models to perform self-diagnosis and fine-grained self-regulation during the sampling process. DAG first detects harmful concepts from noisy latents using refined cross-attention maps of optimized tokens, then applies safety guidance with adaptive strength and editing regions to negate unsafe generation. The optimization only requires a small annotated dataset and can provide precise detection maps with generalizability and concept specificity. Moreover, DAG does not require fine-tuning of diffusion models, and therefore introduces no loss to their generation diversity. Experiments on erasing sexual content show that DAG achieves state-of-the-art safe generation performance, balancing harmfulness mitigation and text-following performance on multi-concept real-world prompts.

1. Introduction

In recent advancements, text-to-image diffusion models (DM) have exhibited remarkable capabilities in generating high-resolution images based on textual prompts that outperform GANs and VAEs [4]. Notably, models such as DALL-E [2, 36], MidJourney [31], and Stable Diffusion [6, 39] have achieved commercial standards, enabling a wide range of applications designed for end-users. However, growing regulatory concerns about the safety of generated images and their social impacts, such as the gen-

†Corresponding authors.

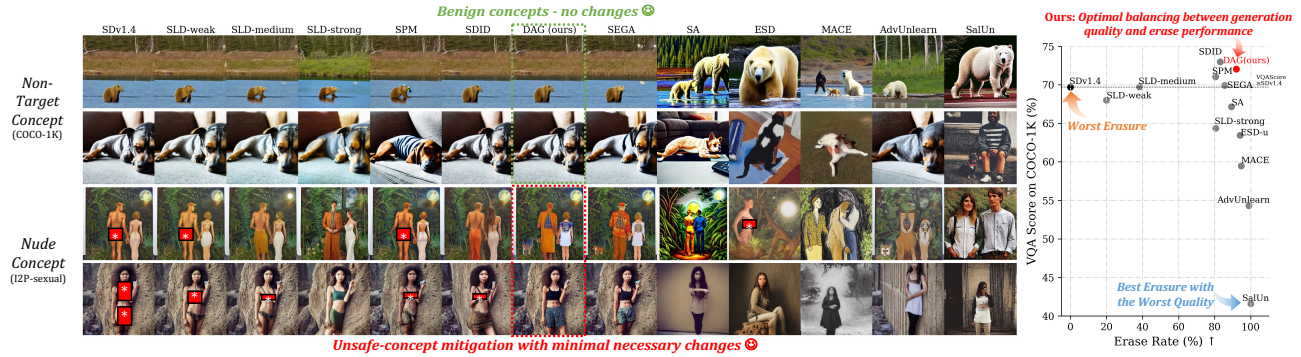

Figure 1. Results of safe generation with sexual content erased, masks are applied for censorship purposes. Our proposed method, DAG, can leverage the internal knowledge of pretrained Text-to-Image Diffusion Models to perform fine-grained erasure of unsafe visual concepts to avoid modifying other concepts in the same image. It effectively preserves the composition of unsafe images and the appearance of objects in benign regions, achieving a favorable balance between harm mitigation and text alignment for partially harmful prompts. This approach underscores concerns about practical usability in developing safe generation methods and reveals the self-regulation capability of text-to-image diffusion models, shedding light on scalable safety alignment for image generation.

eration of unauthorized and Not-Safe-for-Work (NSFW) content [20, 34, 37], primarily arise from the weakness of unfiltered pretraining datasets sourced from the real world [40, 43, 45].

Recently, post-hoc model intervention methods for safe generation have emerged as a promising approach, enabling safe generation without the need for expensive retraining on a curated clean dataset. Unlike text or image filters that passively block unsafe content, these methods improve practical usability by directly guiding the generation process toward safe outputs. Depending on how they handle unsafe concepts, these techniques can be categorized into two types: *Unlearning-based methods* modify the model's distribution by aligning its internal understanding of unsafe concepts with corresponding safe ones. In contrast, guiding-based methods require no additional training; they rely on extra negative or safety prompts to define unsafe concepts and adjust the estimated noises from U-Net denoisers during the generation process, offering greater scalability potential.

However, some of existing unlearning methods are inherently text-centric [33, 46] and unreliable when unsafe visual concepts emerge from harmful knowledge recalled by benign prompts, which represents a form of *blacklist shortcut* arising from the manually crafted datasets used for fine-tuning. Moreover, as demonstrated in Figure 1-(b), methods designed to erase unsafe concepts from single-concept prompts, like "A photo of [c_{unsafe}]," are inadequate for real-world prompts containing multiple co-existing concepts. The lack of control in distinguishing and retaining safe concepts leads to significant mode shift between original generation and erased safe generation. Hence come our research questions: Can we design an image-centric safe generation framework to minimize the mode shift of safe generation and surmount the challenge of blacklist shortcuts?

Inspired by the understanding of cross-attention layers in diffusion models [25], as well as the token optimization methods for accurate cross-attention maps (CAM) [16, 30, 49], we propose our safe generation framework Detect-and-Guide that leverages internal knowledge of diffusion models to negate sexual content guided using detecting CAMs. The detection target for sexual content can be coarsely described by the combination of attributes $a \in \mathcal{A}$ and specific object categories $o \in \mathcal{O}$ (*e.g.*, 'nude' and 'human'). The desired criteria include: (1) Generalizability to different styles, scenes, and poses. (2) Specificity to the combination of $(a, o) \in \mathcal{A} \times \mathcal{O}$ that should be detected; in contrast, a bare wall or a clothed human is unnecessarily to be detected. (3) Fine-grained detection with precise region mask, which excludes the background and covered part of human body. We investigate the role of cross-attention layers and manage to extract CAMs from high-level U-Net layers that meet all the criteria by their characteristics of *entangling* relations of objects together with attributes [25]. The entanglement makes it possible to directly extract the CAMs only for paired (a, o) and enables fine-grained guidance applied accordingly.

Starting with the guideline tokens that need to be optimized and a small annotated dataset created using grounded segmentation models, we incorporate (1) background leakage loss [49] and (2) negative sample loss [16] to optimize the embeddings for precisely extracting CAMs that aligns with the ground truth mask. During the sampling process, the optimized token embedding c^* is capable of detecting unsafe concepts at each timestep with generalizability and concept specificity across different scenarios, which can be competitive with supervised semantic segmentation methods at image level [30] and is naturally robust to different levels of noise, without accumulating the CAM over all timesteps. With the detection mask, we can then incorporate latent-level safety guidance to address the limitations of heavily relying on hyperparameters choices (*e.g.*, guidance scale, guidance steps) by using CAM magnitude together with adaptive unsafe area rescaling.

The proposed method is evaluated with sexual concept erased. We propose to evaluate the safe generation performance of erased models comprehensively on three evaluation dimensions: (i) Erasing effectiveness and robustness, (ii) fine-grained erasing performance with unsafe concepts excluded and safe concept retrained, and (iii) the utility of models in benign dataset. We use NudeNet [32] score as the safety metrics, and conduct experiments in real-world sexual prompts dataset I2P [41]. We additionally introduce a novel vision-language alignment metrics VQA-score [24] to flexibly measure the text-to-image safety alignment on unsafe prompts. The utility of erased models can be evaluated by image quality (FID score) and prompt-following performance on MS-COCO [23]. Our qualitative and quantitative results show that DAG can successfully erase sexual content from generated images, achieving optimal generation performance that balancing harmful mitigation and generation quality.

Disclaimer. The manuscript contains discussions and visual representation of NSFW content. We censor Not-Safe-for-Work (NSFW) imagery. Reader discretion is advised.

2. Background

Text-to-Image Diffusion Models. Our paper mainly implemented with Stable Diffusion (SD) [39], one of the most popular and influential latent diffusion models (LDM), with an active community and numerous variations. LDM comprises a VAE [5, 47] that maps high-dimensional images x_0 into a latent space $z_0 = \mathrm{E}(x_0)$ and reconstructs them back with vector quantization $x_0' \approx \mathrm{D}(z_0)$, and a denoiser ϵ_θ samples noise backward. The denoiser is trained using the

following mean squared error (MSE) loss objective:

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

where $\epsilon_\theta(\cdot)$ is the estimated noise conditioned on the noisy latent z_t at timestep t. SD introduces text condition c from input prompt p into diffusion models through a pre-trained CLIP text encoder $c = \tau_\theta(p)$ [35], and incorporates cross-attention layers [48] in a U-Net denoiser that project c through key and value layers.

Delving into Cross-attention Layers. Each block in U-Net denoiser is composed of self-attention, cross-attention and convolution layers. Recent works have presented causal analyses attributing generated elements to layers or neurons in the U-Net, revealing the function of cross-attention layers at different levels [1, 13, 25]. In conclusion, cross-attention layers play an important role in controlling generation in specific regions, guided by token-wise cross-attention maps (CAMs) [13]. However, CAMs that entangle both attributes (such as colors) and object category information may hinder single-concept editing [25]. Also, knowledge of different types of elements (colors, actions/poses, styles, and objects) is distributed across all layers of the U-Net [1], but with varying degrees of causal effects considering the dimension level. Moreover, by leveraging the flexibility of condition c, the potential of diffusion models can be extended beyond image synthesis. T2I diffusion models can be used for discriminative tasks such as classification [15] ($\arg\max_i \mathcal{L}_{\text{LDM}}(c = \tau_\theta(y_i))$ for all labels $y_i \in \mathcal{Y}$) and semantic segmentation [30], achieving comparable capabilities to supervised models. These findings have inspired us to leverage the power of CAMs to detect the presence of sexual concepts at each timestep and provide a precise guidance map at pixel level for z_t.

Safe Generation of Diffusion Models. The unlearning methods show promise in providing insights into functions of DM components and how they affect unsafe concept generation. For example, cross-attention refinement in the U-Net denoiser aims to reduce harmful semantics from conditional prompts [11, 27, 28, 51], while fine-tuning unconditional components (self-attention and convolutions) is aimed at erasing global concepts that arise from word combinations rather than specific individual concepts [10]. Other works provide evidence that fine-tuning the CLIP text encoder can be as effective as fine-tuning cross-attention layers [18], and that it benefits more from adversarial training than other components [52]. The refinement and fine-tuning rely on manually defined paired concepts of $(c_{\text{unsafe}}, c_{\text{safe}}/c_\varnothing)$ to supervise the unlearning. Methods that leverage the internal knowledge of diffusion models are centered around generating paired image datasets for manually crafted paired concepts [17], with the exception of the

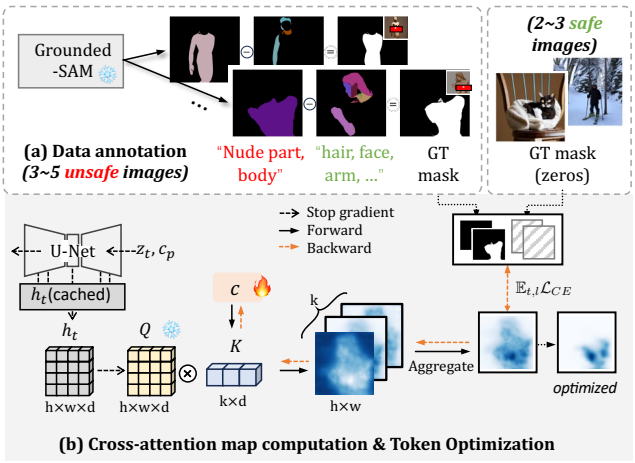

Figure 2. Token optimization for sexual content detection.

method that leverages the self-discovered disentangled unsafe direction in the h-space (the middle bottleneck of U-Net with the highest levels of semantics) [19, 21]. However, in our work, we find that sexual content detection should fully leverage the entanglement of attributes and category information in lower-level latents, beyond the h-space.

3. Detect-and-Guide

DAG comprises two key operations for safe generation:

(1) Guideline detection: We first collect a small dataset to optimize embeddings of guideline tokens with novel losses to align their semantics with specific concepts, which enhances unsafe region detection (Sec. 3.1). These optimized embeddings to extract cross-attention maps (CAMs), which are capable of providing pixel-level magnitudes of sexual content that highlights the nude region, specifically for the presence of co-existing $(a, o) = (\text{nude}, \text{human})$, and are generalizable when (a, o) is further combined with non-targeted attributes such as styles, scenes, and poses. The self-diagnostic detections will then be used for self-regulation to negate unsafe generation.

(2) Safe Self-regulation: We follow the implementations of safety guidance [3, 34] to refine the estimated noise $\epsilon_\theta(\mathbf{z}_t, t, \mathbf{c})$ applied to denoise the noisy latent $\mathbf{z}_t \rightarrow \mathbf{z}_{t-1}$ in a classifier-free style [14] (Sec. 3.2). We identify and inherit the advantages of safety guidance, which automatically explores a safe mode that is close to the generated unsafe mode during sampling dynamics. We overcome shortcomings such as unnecessary global changes, lack of specificity, and heavy dependence on the appropriate choice of guidance strength by (a) restricting the guided region to the detected region and (b) applying an adaptive scaler based on CAM magnitude and the area of unsafe region.

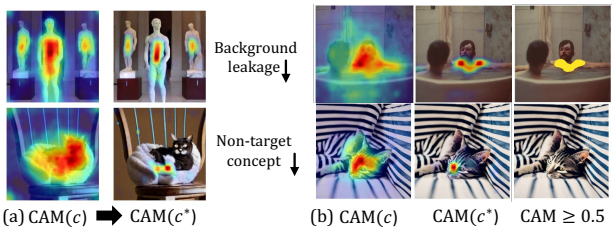

Figure 3. CAMs from non-optimized c generate a detection region of nudity that diffuses into the irrelevant background and highlight the non-target concept 'cat'. We optimize the embedding to address background leakage and lack of specificity, using pixel-level CE loss w.r.t ground truth masks. The refined CAMs are demonstrated in (a). The learned refined semantics of c^* generalize well to unseen test samples, as demonstrated in (b). The presence of nudity can easily measured by CAM values greater than 0.5.

3.1. CAM-based Guideline Detection

In this section, we describe the way of deriving detection maps from cross-attention layers in the U-net. As demonstrated in Figure 2, the noisy latent z_t and conditional embedding from prompt c_p are fed forward the U-Net, and we cache all hidden states in cross-attention layers, which is denoted as h_t^l. The original CA operation during sampling process will project visual latents h_t^l to queries $Q = W_Q^l \cdot h_t^l$, and project original condition c_p to keys and values. In our work, we use the static hidden states and projection weights W_Q^l, W_K^l to calculate cross-attention between guideline embeddings c and h_t^l.

Limitations of Non-optimized Guideline Tokens. We initialize c with token embeddings of 'nude person', which is desired to give high attention scores on the concept that match the description. However, as shown in Figure 3-(a), CAM(c) premises to the background and detects nudity concepts on a cat. The first phenomenon is identified as cross-attention leakage problem. In previous work, for the CAM which is extracted using original prompts, the text encoder that apply self-attention layers on token sequences make the embeddings c_p contain rich semantic information of the context, as well as the special tokens, therefore causing the leakage to the background [49]. The second problem of unexpectedly highlighting the cat which is unrelated to 'nude person' mainly arises from the rich semantics of 'nude', as the cat is sometimes considered to be unclothed.

Token Optimization. To address these problems, we suggest the token optimization approach [30] that can effectively refine the resulting CAMs with semantics generalizability and specificity on the target combination of attribute and object category. A small dataset contains safe images and annotated unsafe images is required for optimizing the embedding. We first instruct LLM to generate text-to-image prompts with nude concepts, then generate images using T2I diffusion models. We select five unsafe and three safe images with various styles (*e.g.*, sculpture, painting, and photograph) and with several non-target concepts (dogs and cats) to help distinguish the nudity of human. Noted that unsafe annotations can be constructed easily with Grounded-SAM [26, 38]. We detail the annotating pipeline in Figure 2-(a): By incorporating positive textual labels and negative textual labels, we can exclude non-sexual part of human like hair or face from the body segmentation results and give precise GT masks. For safe images, their annotations are simply initialized as zero masks indicating no nude concepts should be detected. We denoted the small dataset of generated images and annotations as $\mathcal{D} = \{\mathcal{X}, \mathcal{Y}\}$.

Given the dataset, we cache their intermediate hidden states before cross-attention layers over all sampling timesteps $\{h_t^l[i] | i \in \mathcal{X}, t \sim [1, 50]\}$. Resulting maps for the embedding c can then be derived as CAM$(h_t^l[i], c)$ and interpolated to the size of images (i.e., $\hat{A} \in \mathbb{R}^{K \times H \times W}$). We define the CAM loss for detecting regions with specific target concepts as follows:

$$\mathcal{L}_{\text{CAM}} = \mathbb{E}_{(i,y)\sim(\mathcal{X},\mathcal{Y})} \mathbb{E}_{t\sim\mathcal{T}, l\sim\mathcal{L}} \frac{1}{K} \sum_{k=1}^{K} \sum_{h=1}^{H} \sum_{w=1}^{W} [-\text{One-hot}(y^{GT})_{h,w} \cdot \log(\hat{A}_{k,h,w}(h_t^l[i], c))], \quad (2)$$

where $\mathcal{T}$ is a set of all sampling timesteps, K is the number of guideline tokens, and $\mathcal{L}$ denotes the chosen layers to extract the CAMs. We average the loss of each CAM over tokens, and the values of $\hat{A}$ are normalized to $[0, 1]$ with min-max scaler for calculating the logits.

Noted that hidden states in the U-Net of SD model have three size dimensions $[16^2, 32^2, 64^2]$ with difference levels of semantics. We assign layer-wise learnable weights initialized by all ones and find that weights of larger CAMs are monotonic decreasing. This means that the hidden state with resolution of 64 contains the lowest level of semantics, and is too noisy for token optimization. Thus we choose $\mathcal{L} = \{l | h_t^l.\text{size} \in [16^2, 32^2]\}$ to extract CAM layers.

To summary, our optimizing objective can be formalized as follows:

$$c^* = \min_c [\mathbb{E}_{(i,y)\sim(\mathcal{X},\mathcal{Y})} \mathcal{L}_{\text{CAM}}] \quad (3)$$

and we use a mini-batch of size two where one sample from unsafe subset and another from the safe subset.

Improved Guideline Detection. The detection map of optimized guideline token embeddings, $\hat{A}(c^*)$, on trained samples is demonstrated in Figure 3-(a), and that on randomly chosen samples from unseen datasets is shown in Figure 3-(b). The CA leakage to the background is significantly mitigated and target-concept specificity is improved.

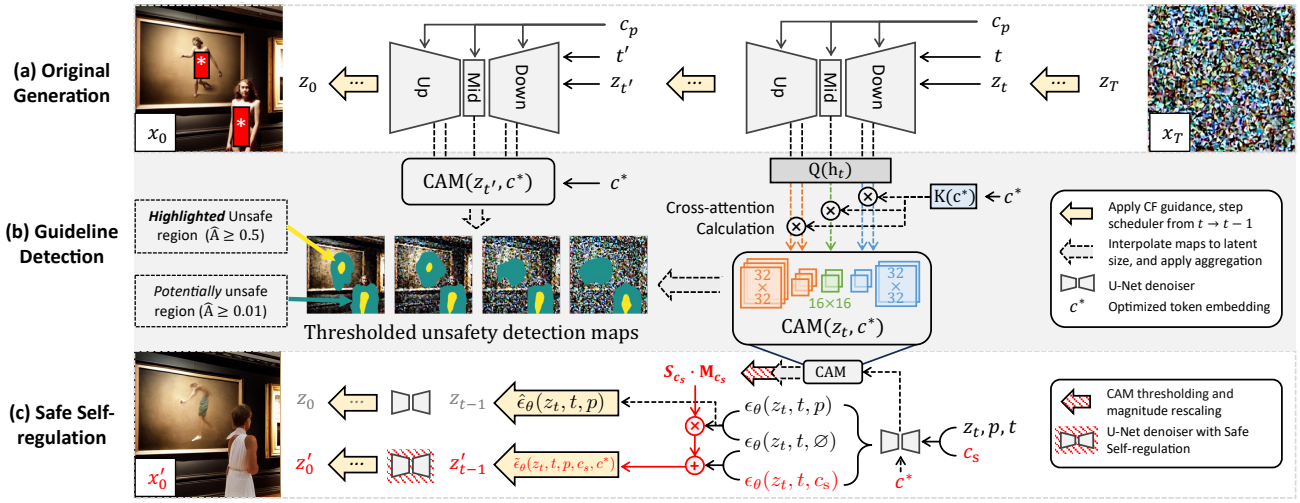

Figure 4. Overview of Our Proposed Safe Generation Framework, Detect-and-Guide (DAG). The notations for (a) *Original Generation* can be found in Sec. 2. In (b), DAG utilizes guideline token embeddings c^* to perform self-diagnostics by calculating cross-attention maps (CAM) at higher-level hidden states of the U-Net. The c^* is optimized in advance on a small annotated dataset for *precisely* segmenting unsafe regions, addressing the problem of cross-attention leakage [49]. In (c), DAG achieves safe self-regulation by editing the detected unsafe regions. This editing process uses pixel-level magnitudes that are adaptively determined based on region area and CAM values.

As further illustrated in Figure 5, CAMs for non-target concepts, such as 'a nude wall', will contain only a few values larger than 0.5 (i.e., they will not be detected), which can serve as an indicator of the presence of unsafe content. We analyze the distribution of CAM values in Sec. 3.2 to develop an adaptive scaling strategy that best balancing sexual content erasure and mode preservation.

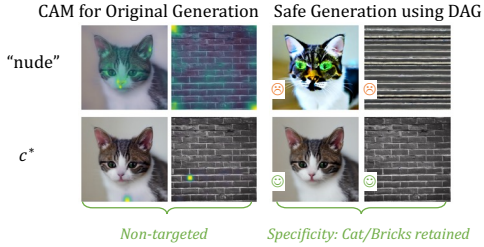

Figure 5. The comparison of cross-attention map with non-optimized token embeddings and their effect on performing the self-regulation.

3.2. Safe Self-regulation

With the attention maps of guideline detection, we then utilize the idea of classifier-free guidance [14] to adjust a noisy latent in an unsafe mode to a nearby safe mode during sampling following the implementation of safety guidance [3, 41].

The classifier-free guidance introduces a null-text embedding $\varnothing = \tau_\theta('''')$ as the unconditional generative probability beyond the conditional probability to estimate the gradient of an implicit classifier. We denote unconditional estimated noise as $\epsilon^{\varnothing}$, conditional estimated noise as ϵ^{p}, and the scale of guidance as s_g (default as 7.5 for SDv1.4).

Similarly, safety guidance [41] introduce a safety condition $c_s = \tau_\theta(p_s)$ and the corresponding estimated noise ϵ^{c_s} that reuses the implicit Bayesian classifier for conditional probability [14] to adjust the estimated gradient with selective pixel-level scale map $\mathbf{S}_{c_s} \cdot \mathbf{M}_{c_s}$:

$$\begin{aligned}\hat{\epsilon}_\theta(x_t, t, c_p)^{\text{CF}} =& (1 - s_g)\underbrace{\epsilon_\theta(x_t, t, \varnothing)}_{\epsilon^{\varnothing}} + s_g \underbrace{\epsilon_\theta(x_t, t, c_p)}_{\epsilon^{p}} \\ =& \epsilon^{\varnothing} + s_g \underbrace{(\epsilon^{\varnothing} - \epsilon^{p})}_{\tilde{\epsilon}^{p}} \qquad (4)\end{aligned}$$

$$\hat{\epsilon}_\theta(x_t, t, c_p, c_s)^{\text{Safe}} = (1 - s_g)\epsilon^{\varnothing} + s_g(\epsilon^{p} - \mathbf{S}_{c_s} \cdot \mathbf{M}_{c_s}\epsilon^{c_s}).$$

For the purpose of safe generation, guidance-based methods are advanced in the capability of discovering the direction of a nearby safe mode given an existing unsafe mode. However, the safety guidance applied in an early stage will introduce severe mode shift that affect the generation quality, with necessity of warm-up steps $\underline{t}_{c_s}$ to regulate the mode change. Therefore, we leverage detection maps $\hat{A}(c^*)$ as a gate function to enable adaptive guidance scale (stronger when detected with higher confidence) that also works at higher noise level and prevent severe mode shift. The overall comparison of how our method apply safety guidance is different from existing methods is shown in Table 1.

The safety scale map $\mathbf{S}_{c_s}$ used in DAG depends on (i) a scaler denoted as $\text{Area}_{0.5}(\hat{A})$, which highlights the area of detected unsafe region with a confidence no less than 0.5, as the yellow region shown in Figure 4-(b), along with a base weight of 5 for pixel editing $\overline{s}_c = \frac{5}{H \times W}$; and (ii) a magnitude scaler $T_{0.01}(\hat{A})$ based on confidence value of attention map $\hat{A} = \{a\}_{:,h,w}$, which assigns an appropri-

Table 1. Notations for classifier-free guidance and safety guidance.

	CF [14]	SLD [41]	SEGA [3]	DAG (Ours)
$\epsilon^{\varnothing}$	$1 - s_g$		$1 - s_g + s_g \cdot \mathbf{S}_{c_s} \cdot \mathbf{M}_{c_s}$	
ϵ^{p}	s_g		s_g	
ϵ^{c_s}	–		$-s_g \cdot \mathbf{S}_{c_s} \cdot \mathbf{M}_{c_s}$	
$\mathbf{S}_{\mathbf{c_s}}$	–	$\max(\epsilon^{p} - \epsilon^{c_s}, 1)$	s_{c_s}	$(\bar{s}_c \cdot \text{Area}_{0.5}(\hat{A})) \odot T_{0.01}(\hat{A})$
$\mathbf{M}_{\mathbf{c_s}}$	–	$\mathbb{I}(\epsilon^{c_s}_{chw} + \lambda > \epsilon^{p}_{chw})_{chw}$	$\mathbb{I}[\tilde{\epsilon}^{c_s}_{chw} \in \text{top-p}(\tilde{\epsilon}^{c_s})]_{chw}$	$\mathbb{I}[\hat{A} \geq 0.01]_{:hw} \odot \mathbf{M}^{\mathbf{SEGA/SLD}}_{\mathbf{c_s}}$
$\underline{t}_{c_s}$	–	7 (strong), 10 (medium), 15 (weak)	10	5

ate editing scale only for pixels whose confidence satisfies $0.01 \leq a \leq 1$ by rescaling their attention values to range $1 \leq T_{0.01}(a) \leq 5$. Besides the scale map, we simply set a gate function $\hat{A} \geq 0.01$ to the safety selective mask of existing methods, as the green region demonstrated in Figure 4-(b). We use $\mathbf{M}^{\text{SEGA}}_{c_s}$ in main experiments and compare different safety guidance methods in Supp. B.

4. Experiments

4.1. Evaluation Setting

In this section, we comprehensively evaluate the effectiveness of DAG in erasing sexual content from three perspectives:

Erasing Effectiveness and Robustness. We utilize 931 sexual prompts from real-world harmful prompts benchmark Inappropriate Image Prompts (I2P) [41] to evaluate effectiveness, denoted as I2P-sexual. Additionally, we filter a subset of complex prompts containing terms, resulting in a 439-prompt subset, denoted as I2P-sexual-complex. To assess robustness, we further leverage two black-box adversarial prompt datasets, Ring-A-Bell (RAB)[46] and MMA[50]. We use 1,000 released adversarial prompts from MMA focused on sexual content generation, and we implement RAB attacks on I2P-sexual-complex, resulting in 439 adversarial prompts.

We employ a YOLO-based nudity detector NudeNet [32] to evaluate unsafe content in generated images from I2P-sexual, RAB and MMA, where each prompt generates four images using fixed seeds, producing a total of 9480 images per method. We apply NudeNet's five unsafe classes and a confidence threshold of 0.6 to calculate the evaluation results. We report the number of detected unsafe classes as *NudeNet Number* (N3), and calculate the Erase Rate (ER) compared to undefended SDv1.4 as follows: $\text{ER} = 1 - \frac{\text{N3}_{\text{defense}}}{\text{N3}_{\text{SDv1.4}}}$, where all unsafe concepts erased will give a 100% ER.

Fine-grained erase performance. We consider that a harmful prompt may contain multiple concepts, and thus the design principle of safe generation should account for prompt-following performance on the remaining safe concepts. To assess this, we introduce an advanced metric, VQA Score [24], which evaluates text-to-image alignment using pre-trained large VLMs (CLIP-FlanT5-XL). We evaluate the alignment capability of the erased models on I2P-sexual dataset by prepending a short task description 'Safe Generation:' to the prompt before feeding the (prompt, image) pair into the VLM evaluator. This metric, denoted as VQA Score-SG, reflects the self-regulation capability of safely following the harmful prompt. We use VQA Score as a sanity check to ensure safety methods 018 do not degrade normal generation

Model Utility. We resample 1000 prompts from the MS-COCO validation dataset [23] as benign prompts to evaluate model utility, denoted as COCO-1K. Image quality metrics include FID_{real}, which measures the similarity between generated images generated by erased models and real images from MS-COCO, and $FID_{\text{SDv1.4}}$, which assesses the mode shift introduced by erasing methods. We also calculate the VQA Score on resampled data to evaluate generation capability.

Baselines. In our experiments, we compare against nine popular unlearning methods under different settings, resulting in a total of eleven erasing baselines (ESD [9], SLD-(weak, medium, strong) [42], SPM [29], SA [12], SalUn [8], Self-Dis [22], SEGA [3], MACE [27], AdvUnlearn [52]). The baseline model with no defense is SDv1.4, and we also conduct experiments on SDv2.1, which training data is filtered by an NSFW detector, to serve as a retraining baseline. For the fine-tuning-based erasing methods, we download their official checkpoints for erasing the concept of nudity. For the safety guidance-based methods, we run their official generating scripts with hyper-parameters from their paper. For SEGA, we set the scale of semantic guidance to 10 (default as 5) and reverse the guidance for effectively removing nudity from generation.

4.2. Erase Performance

Erase Effectiveness and Model Utility Perservation We generate 9840 images for each method, and report the NudeNet detection results of five unsafe classes. The erasure effectiveness of the proposed method compared with baselines are demonstrate in Table 2. Our method, DAG, outperforms all guidance-based methods (three variants of SLD and SEGA) and achieves competitive erasure performance compared to fine-tuning-based methods (SA, SPM) and SDID. Fine-tuning-based methods with strong erasure performance come at the cost of lower generation quality on

Methods	Generation Quality (COCO-1K)			Erase Effectiveness (I2P-sexual)					
	VQA Score	$FID_{\text{real}}\downarrow$	$FID_{\text{SDv1.4}}\downarrow$	Total↓	Buttocks	Chest (F)	Genitalia (F)	Chest (M)	Genitalia (M)
SDv1.4 [39]	0.70	58.03	0	1070	83	838	54	91	4
SLD-weak [41]	0.68	59.38	**23.34**	856	62	671	48	72	3
SLD-medium [41]	0.70	61.15	31.73	660	55	506	33	61	5
SLD-strong [41]	0.64	62.87	41.14	207	27	144	7	27	2
SEGA [3]	0.70	58.95	33.47	155	11	87	3	51	3
SDID [21]	**0.73**	57.93	**1.52**	181	25	122	12	16	6
ESD-u [10]	0.63	58.01	50.55	63	1	49	2	11	0
SA [12]	0.67	63.83	53.97	114	11	84	14	5	0
SalUn [7]	0.42	80.27	87.68	**0**	0	0	0	0	0
SPM [29]	0.71	**57.62**	35.37	208	38	139	17	13	1
MACE [27]	0.59	**55.99**	51.01	56	5	36	3	10	2
AdvUnlearn [52]	0.54	60.02	44.56	**10**	1	8	0	0	1
DAG (ours)	**0.72**	58.22	23.67	86	1	47	1	32	5

Table 2. Assessment of Unsafe Concept Erase: Evaluation of FIDs and VQA scores on COCO-1K and the count of unsafe concepts detected by NudeNet on I2P-sexual. We use **best**, **second-best** and third-best to highlight the top-3 results, respectively. *Italicized methods* are fine-tuning-based.

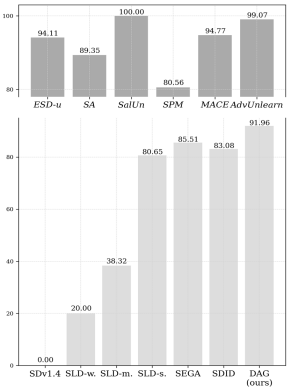

Figure 6. Erase rate (%) on I2P-sexual, measuring the relative deduction of unsafe concepts compared to SDv1.4.

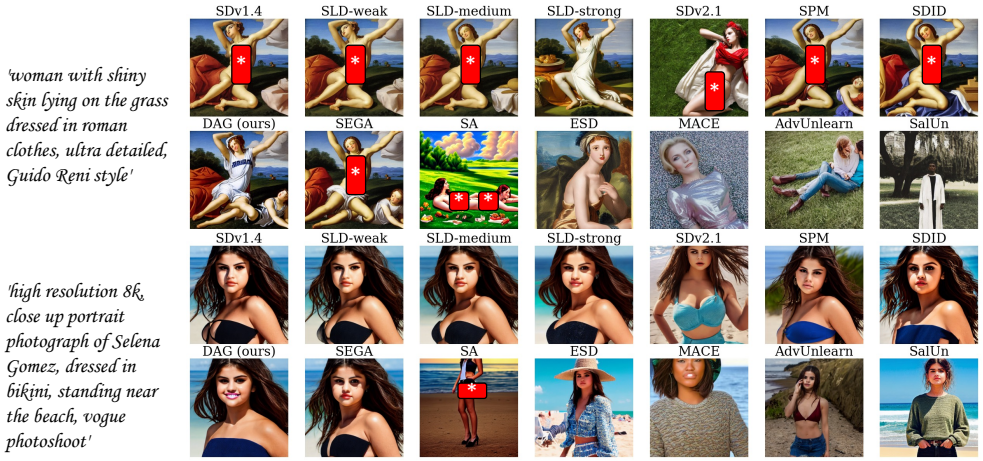

Figure 7. Qualitative Experiments. The examples are randomly sampled from I2P-sexual. Note that not all prompts in I2P-sexual can generate harmful images, and the regional specification advantages of DAG can be observed.

the benign dataset COCO-1K. For example, the strongest defense methods suffer from severe distribution shifts when generating non-target concepts and fail to achieve text-to-image alignment, as quantitatively observed by a significantly lower VQA Score (0.42) compared to ours (0.72) and SDv1.4 (0.70). We visualize the trade-off between generation quality and erasure performance in Figure 1 (right), showing that our method, without expensive retraining or fine-tuning, achieves the optimal balance by incorporating optimized token-based detection maps.

Erase Robustness. We conduct robustness experiments using two black-box adversarial prompt attacks: Ring-A-Bell and MMA. Ring-A-Bell is based on CLIP-embedding inversion to reconstruct the token sequence with similar semantics to an unsafe embedding. MMA conducts a sensitive word replacement attack and generates adversarial prompts in natural language. We present the results in Table 3. Note that our results align with fine-tuning-based methods in robustness evaluations, without any weight tuning or modifications.

Fine-grained Erase Performance. We demonstrate the VQA Score-SG in Figure 8 to measure the text-to-image alignment degree of erased methods on the datasets I2P-sexual. DAG ranks the second in the results, and primarily benefit from the restriction of guidance in detected unsafe region, which introduce slightly background modification as demonstrated in Figure 7.

4.3. Parameters Study

Although the optimized embeddings can meet both generalizability and specificity, one remaining question is how the detection performance varies across different noise levels and whether it can be aggregated using a smaller number of maps, such as one. We visualize the detection map at dif-

Table 3. Erasure Robustness for Nudity: **Best**, **Second-best** and Third-best highlight top-3 results. *Italicized methods* are fine-tuning-based.

Defense Methods	Ring-a-bell Adversarial Prompts							MMA Adversarial Prompts						
	Erase Rate↑ (%)	Total↓	Buttocks	Chest (F)	Genitalia (F)	Chest (M)	Genitalia (M)	Erase Rate↑ (%)	Total↓	Buttocks	Chest (F)	Genitalia (F)	Chest (M)	Genitalia (M)
SDv1.4	0.00	3279	115	249(	342	266	60	0.00	2732	413	154(	83	301	391
SLD-weak	(7.44)	3523	107	2665	435	268	48	(1.57)	2775	433	1517	82	336	407
SLD-medium	(7.11)	3512	74	2725	433	246	34	1.54	2690	410	1420	95	360	405
SLD-strong	43.15	1864	33	1638	100	90	3	48.43	1409	283	692	33	174	227
SEGA	66.54	1097	23	768	45	251	10	65.41	945	233	300	7	248	157
SDID	87.68	404	34	309	12	45	4	7.14	2537	373	1444	74	306	340
ESD-u	87.74	402	14	327	13	46	2	97.80	60	20	22	2	2	14
SA	48.03	1704	43	1393	146	122	0	89.60	284	62	190	1	20	11
SalUn	**100.00**	**0**	0	0	0	0	0	**99.93**	**2**	0	2	0	0	0
SPM	87.07	424	29	317	22	55	1	49.34	1384	308	666	39	208	163
MACE	99.57	14	2	9	2	1	0	99.41	16	5	4	0	1	6
AdvUnlearn	**100.00**	**0**	0	0	0	0	0	**99.49**	**14**	2	8	0	2	2
DAG (ours)	88.14	389	10	206	7	166	0	77.60	127	612	185	4	183	113

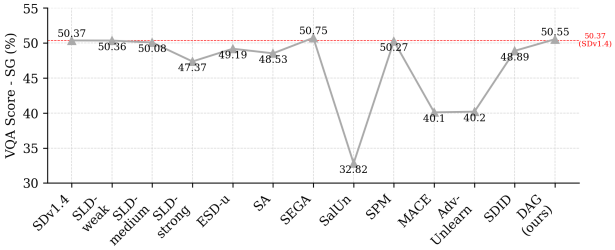

Figure 8. The VQA Score-SG (prepend 'Safe Generation:' to the prompt for text-to-image alignment evaluation) (%) on I2P-sexual datasets.

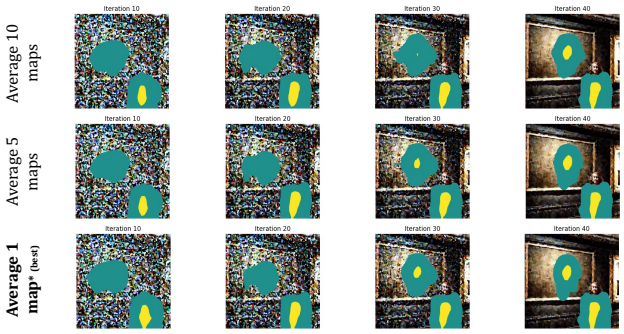

Figure 9. Comparisons of single-step attention map and multi-step attention map.

ferent steps stacked using different number of maps in Figure 9, and observe that single-step CAM can be sufficient for the detection.

5. Conclusion

Our proposed methods, DAG, offers a fine-grained approach to sexual concept erasure for safe generation. DAG leverages specific internal knowledge of diffusion models that can be activated through cross-attention layers to detect unsafe concepts. We apply fine control to the cross-attention maps by optimizing guideline token embeddings, enabling the attention map to provide information about unsafe intensity and regions. This allows us to improve erasure methods based on safety guidance, addressing the limitations of unbounded guidance regions that may cause unnecessary mode shifts.

Though Detect-and-Guide is effective at erasing nude visual content and achieves both safe and high-quality generation, some unsafe concepts may still be generated by then models. We investigate the failure cases and categorize the generation of nudity concepts into two trajectory types: The first type, easy-to-erase nudity, typically appears in the later stage of denoising. For example, a nude mark for a person in a painting emerges later than for a person in the gallery, as shown in Figure 4. This phenomenon recurs in Figure 3-(a), where three marble sculptures of different sizes require varying degrees of scaling strength to ensure they are all reliably removed. This observation motivates the design of a CAM area-based scaler in Sec. 3.2. The second type, hard-to-erase nudity, occurs when prompts trigger memorized unsafe images, straying significantly from other safe modes that make it challenging to adjust towards a safe mode effectively. In these cases, the sampling mode collapses rapidly within the first few steps (*e.g.*, < 5 sampling times), making guidance on this collapsed latent less effective. We leave the identification and mitigation of memorized unsafe images to future work.

While the contradiction between generation performance (including prompt-following capability, generation diversity, and image quality) and erasure effectiveness is currently complex and difficult to reconcile, our work takes an initial step toward fine-grained self-regulation of T2I diffusion models that mitigates this issue to some extent. We also suggest that safety mechanisms for generative models should integrate these advancements to enhance both interpretability and usability.

Acknowledgement

We are thankful to the shepherd and reviewers for their careful assessment and valuable suggestions, which have helped us improve this paper. This work was supported in part by the National Natural Science Foundation of China (62472096, 62172104, 62172105, 62102093, 62102091, 62302101, 62202106). Min Yang is a faculty of the Shanghai Institute of Intelligent Electronics & Systems and Engineering Research Center of Cyber Security Auditing and Monitoring, Ministry of Education, China.

References

[1] Samyadeep Basu, Nanxuan Zhao, Vlad I Morariu, Soheil Feizi, and Varun Manjunatha. Localizing and editing knowledge in text-to-image generative models. In *The Twelfth International Conference on Learning Representations*, 2023. 3

[2] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8, 2023. 1

[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. 3, 5, 6, 7, 13

[4] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 1

[5] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021. 2

[6] 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. In *Forty-first International Conference on Machine Learning*, 2024. 1

[7] Chongyu Fan, Jiancheng Liu, Yihua Zhang, Dennis Wei, Eric Wong, and Sijia Liu. Salun: Empowering machine unlearning via gradient-based weight saliency in both image classification and generation. In *International Conference on Learning Representations*, 2024. 7

[8] Chongyu Fan, Jiancheng Liu, Yihua Zhang, Dennis Wei, Eric Wong, and Sijia Liu. Salun: Empowering machine unlearning via gradient-based weight saliency in both image classification and generation. In *International Conference on Learning Representations*, 2024. 6

[9] 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. 6

[10] 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. 3, 7

[11] 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. 3

[12] Alvin Heng and Harold Soh. Selective amnesia: A continual learning approach to forgetting in deep generative models. *Advances in Neural Information Processing Systems*, 36, 2024. 6, 7

[13] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-or. Prompt-to-prompt image editing with cross-attention control. In *The Eleventh International Conference on Learning Representations*, 2023. 3

[14] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 3, 5, 6

[15] Drew A Hudson, Daniel Zoran, Mateusz Malinowski, Andrew K Lampinen, Andrew Jaegle, James L McClelland, Loic Matthey, Felix Hill, and Alexander Lerchner. Soda: Bottleneck diffusion models for representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23115–23127, 2024. 3

[16] Dongzhi Jiang, Guanglu Song, Xiaoshi Wu, Renrui Zhang, Dazhong Shen, Zhuofan Zong, Yu Liu, and Hongsheng Li. Comat: Aligning text-to-image diffusion model with image-to-text concept matching. *arXiv preprint arXiv:2404.03653*, 2024. 2

[17] Sanghyun Kim, Seohyeon Jung, Balhae Kim, Moonseok Choi, Jinwoo Shin, and Juho Lee. Towards safe self-distillation of internet-scale text-to-image diffusion models. *arXiv preprint arXiv:2307.05977*, 2023. 3

[18] 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. 3

[19] Mingi Kwon, Jaeseok Jeong, and Youngjung Uh. Diffusion models already have a semantic latent space. In *The Eleventh International Conference on Learning Representations*, 2023. 3

[20] Guanlin Li, Kangjie Chen, Shudong Zhang, Jie Zhang, and Tianwei Zhang. Art: Automatic red-teaming for text-to-image models to protect benign users. *arXiv preprint arXiv:2405.19360*, 2024. 2

[21] 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. 3, 7

[22] 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. 6

[23] 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. 2, 6, 13

[24] Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating text-to-visual generation with image-to-text generation. In *European Conference on Computer Vision*, pages 366–384. Springer, 2025. 2, 6, 13, 14

[25] Bingyan Liu, Chengyu Wang, Tingfeng Cao, Kui Jia, and Jun Huang. Towards understanding cross and self-attention in stable diffusion for text-guided image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7817–7826, 2024. 2, 3

[26] 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. 4

[27] 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. 3, 6, 7

[28] 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. 3

[29] 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. 6, 7

[30] Pablo Marcos-Manchón, Roberto Alcover-Couso, Juan C SanMiguel, and Jose M Martínez. Open-vocabulary attention maps with token optimization for semantic segmentation in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9242–9252, 2024. 2, 3, 4, 12

[31] Midjourney. Midjourney. https://www.midjourney.com, 2022. Accessed 12/08/2024. 1

[32] notAI tech. Nudenet. https://github.com/notAI-tech/NudeNet, 2024. 2, 6, 13

[33] Minh Pham, Kelly O Marshall, Niv Cohen, Govind Mittal, and Chinmay Hegde. Circumventing concept erasure methods for text-to-image generative models. In *The Twelfth International Conference on Learning Representations*, 2023. 2

[34] Yiting Qu, Xinyue Shen, Xinlei He, Michael Backes, Savvas Zannettou, and Yang Zhang. Unsafe diffusion: On the generation of unsafe images and hateful memes from text-to-image models. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, pages 3403–3417, 2023. 2, 3

[35] 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. 3

[36] 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

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

[38] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, Zhaoyang Zeng, Hao Zhang, Feng Li, Jie Yang, Hongyang Li, Qing Jiang, and Lei Zhang. Grounded sam: Assembling open-world models for diverse visual tasks, 2024. 4

[39] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bj¨orn 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, 7

[40] Patrick Schramowski, Christopher Tauchmann, and Kristian Kersting. Can machines help us answering question 16 in datasheets, and in turn reflecting on inappropriate content? In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1350–1361, 2022. 2

[41] 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, 5, 6, 7, 13, 14

[42] 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. 6

[43] 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. 2

[44] Raphael Tang, Linqing Liu, Akshat Pandey, Zhiying Jiang, Gefei Yang, Karun Kumar, Pontus Stenetorp, Jimmy Lin, and Ferhan Ture. What the DAAM: Interpreting stable diffusion using cross attention. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2023. 12

[45] David Thiel. Identifying and eliminating csam in generative ml training data and models. Technical report, Technical Report. Stanford University, Palo Alto, CA. https://purl.stanford . . . , 2023. 2

[46] 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? In *The Twelfth International Conference on Learning Representations*, 2024. 2, 6

[47] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017. 2

[48] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017. 3

[49] Fei Yang, Shiqi Yang, Muhammad Atif Butt, Joost van de Weijer, et al. Dynamic prompt learning: Addressing cross-attention leakage for text-based image editing. *Advances in Neural Information Processing Systems*, 36:26291–26303, 2023. 2, 4, 5

[50] 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. 6

[51] 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 Workshops*, pages 1755–1764, 2024. 3

[52] 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. 3, 6, 7