

CGCE : Classifier-Guided Concept Erasure in Generative Models

Viet Nguyen  and Vishal M. Patel 

Johns Hopkins University, Baltimore MD 21218, USA

{vnguye71, vpatel36}@jhu.edu

<https://viettmab.github.io/cgce-page/>

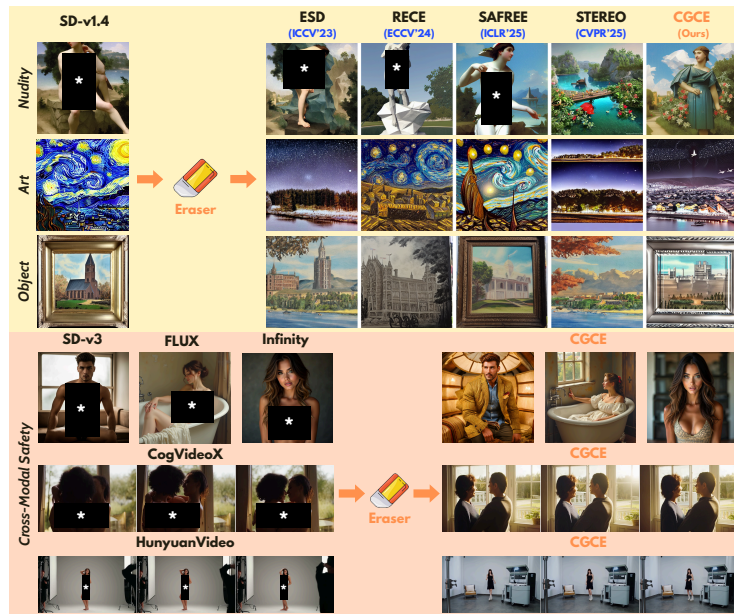


Fig. 1: We present CGCE, an efficient plug-and-play framework for robust and high-fidelity concept erasure. **Top:** CGCE produces safer and higher-quality results compared to state-of-the-art baselines [8, 12, 36, 46] across diverse T2I erasure tasks, including nudity, artistic style, and object removal. **Bottom:** We highlight the cross-modal safety and versatility of CGCE, which can be seamlessly applied as a safeguard to a range of modern T2I and T2V models to ensure safe generation without altering their original weights. Sensitive content (*) has been masked for publication.

Abstract. Recent advancements in large-scale generative models have enabled the creation of high-quality images and videos, but have also raised significant safety concerns regarding the generation of unsafe content. To mitigate this, concept erasure methods have been developed to

remove undesirable concepts from pre-trained models. However, existing methods remain vulnerable to adversarial attacks that can regenerate the erased content. Moreover, achieving robust erasure often degrades the model’s generative quality for safe, unrelated concepts, creating a difficult trade-off between safety and performance. To address this challenge, we introduce Classifier-Guided Concept Erasure (**CGCE**), an efficient plug-and-play framework that provides robust concept erasure for diverse generative models without altering their original weights. **CGCE** uses a lightweight classifier operating on text embeddings to first detect and then refine prompts containing undesired concepts. By modifying only unsafe embeddings at inference time, our method prevents harmful content generation while preserving the model’s original quality on benign prompts. Extensive experiments show that **CGCE** achieves **state-of-the-art** robustness against a wide range of red-teaming attacks. Our approach also maintains high generative utility, demonstrating a superior balance between safety and performance. We showcase the versatility of **CGCE** through its successful application to various modern T2I and T2V models, establishing it as a practical and effective solution for safe generative AI.

1 Introduction

Recent advancements in large-scale diffusion models have led to remarkable success in Text-to-Image (T2I) [7, 13, 21, 25–27, 29, 33] and Text-to-Video (T2V) [16, 19, 44] generation, producing high-fidelity and diverse visual content from simple text descriptions. These models have demonstrated an impressive ability to synthesize a wide range of concepts, unlocking numerous creative and commercial applications. However, the reliance on web-scraped training data [35] poses significant safety and ethical risks, as these datasets often contain harmful, biased, or copyrighted material. A primary concern is the generation of NSFW content, such as sexually explicit imagery, which raises urgent questions about the responsible deployment of these powerful models.

Several strategies have been proposed to mitigate the generation of such undesirable content. A straightforward approach is **dataset filtering** [1], which involves removing unsafe images before training. However, this is often computationally expensive and impractical for every new concept because aggressive filtering can inadvertently degrade overall generative quality. Post-hoc solutions have emerged as more practical alternatives. **Model fine-tuning** approaches [8, 10, 14, 18, 20, 36, 47] permanently alter the model’s weights to erase or remap harmful concepts, but this process can be resource-intensive and often leads to a trade-off between erasure effectiveness and utility preservation of safe content. A distinct category is **model editing** [9, 12], which uses closed-form solutions to directly alter model weights without iterative training. While highly efficient, these edits can be incomplete and may remain vulnerable to adversarial attacks that regenerate the erased concepts. Finally, **training-free, filtering-based** methods [34, 46] modify the generation process at inference time by steering the output away from undesired concepts. While flexible, these methods can

be less robust and are often limited to specific model architectures or susceptible to being bypassed, especially when prompts express harmful concepts implicitly.

To address these limitations, we introduce Classifier-Guided Concept Erasure (CGCE), a robust, efficient, and versatile framework for safe visual generation, as shown in Fig. 1. A key strength of CGCE is its **plug-and-play** nature, allowing it to be applied to diverse generative models, such as modern T2I and T2V architectures, **without altering the original model weights**. This preserves the high-quality generative capabilities of the base model for safe content, a critical advantage over many fine-tuning methods that suffer from utility degradation. Our approach leverages a **lightweight**, specially trained classifier that operates in the text embedding space. This classifier is first trained on a synthetic dataset of paired safe and unsafe prompts, enabling it to accurately detect the presence of an undesired concept within a given text embedding. At inference time, our classifier serves a dual role. It first acts as a **safeguard**, inspecting the text embedding to detect the presence of an undesired concept. Upon detection, it then functions as a **refiner**, altering the embedding to neutralize the harmful semantics and guide it toward a safe semantic region. This process ensures a safe visual output from the T2I or T2V model while leaving safe prompts untouched, thereby preserving the model’s original generation quality.

Our extensive experiments demonstrate the effectiveness of CGCE across multiple tasks, including nudity, artistic style, and object removal. CGCE achieves **state-of-the-art** performance on multiple benchmarks while maintaining the utility of the base model. It provides a superior trade-off between safety and quality compared to existing fine-tuning methods. Furthermore, we demonstrate the versatility of CGCE by applying it to various modern T2I and T2V models. Our main contributions can be summarized as:

- We propose CGCE, a **plug-and-play, image-free** framework that leverages a classifier to detect and refine text embeddings for robust concept erasure.
- Our method achieves **state-of-the-art** safety performance on multiple benchmarks without degrading the generative quality of the base model, offering a superior trade-off between utility and robustness.
- We demonstrate CGCE’s versatility by successfully applying it to various modern T2I and T2V models, showcasing its potential as a cross-modal safeguard for generative AI.

2 Related Works

Concept Erasure in T2I. Recent work in concept erasure aims to mitigate risks in T2I models by removing undesired concepts. One major category of methods involves model fine-tuning approaches [8, 14, 47] that map an erasing concept to its desired target. However, as such prompt-based methods are vulnerable to adversarial attacks [2, 39, 49], more robust techniques have been developed. These include incorporating an efficient single-timestep attack into the training loop [18], employing a two-stage framework to first identify vulnerabilities [36], or reformulating unlearning as an image-based preference optimization

problem [28]. For greater efficiency, closed-form editing methods [9, 12] directly alter weights without iterative optimization. In contrast, methods [34, 46] with unchanged model weights offer flexibility by intervening at inference time. These interventions include modifying the classifier-free guidance [17] process [34] or identifying a toxic subspace and projecting input tokens away from it [46].

Concept Erasure in T2V. While well-explored for images, T2V concept erasure is an emerging field with unique challenges, like maintaining temporal consistency and adapting to modern architectures lacking cross-attention layers. Training-free methods extend to this domain, for example by applying a token-filtering mechanism to videos [46]. Recently, a dedicated training-free T2V framework was proposed to adjust prompt embeddings and then guide the latent noise to steer generation away from the target concept [42]. Conversely, unlearning approaches [45] adapt T2I fine-tuning techniques for video, using regularization to localize the erasure and preserve the model’s knowledge of related concepts.

Attacks on Concept Erasure. The effectiveness of erasure methods is evaluated using red-teaming attacks [2, 39, 43, 49] that manipulate model inputs to bypass safety mechanisms. Many of these attacks focus on the text prompt. White-box methods leverage access to the model’s internal states, such as gradients or classification capabilities, to automatically find problematic prompts [2, 49]. Black-box methods extract a representation of the inappropriate concept and then use it to automatically generate attack prompts [39]. More advanced frameworks employ a multimodal attack strategy, for instance by using gradient-based optimization in the textual modality to craft adversarial prompts that maintain semantic intent but evade prompt filters [43].

3 Motivation

To erase concepts in generative models, a standard approach is to intervene directly during the inference stage. For example, SLD [34] suppresses inappropriate content by manipulating score estimates through Classifier-Free Guidance (CFG) [17]. However, these inference-based methods are inherently limited because they rely entirely on the specific mechanics of the diffusion process. Consequently, they fail when applied to guidance-distilled models like FLUX [21], which natively omit standard CFG and negative prompting, or to Visual AutoRegressive (VAR) models like Infinity [13] and Switti-AR [41].

Another inference-based method, SAFREE [46], attempts to bypass these constraints by intervening at the text-encoder level before visual generation begins. It successfully mitigates visual toxicity in generative models using the CLIP [30] text encoder by projecting the prompt embedding away from a pre-defined toxic subspace. This strategy is effective because CLIP naturally maps distinct semantic features into independent linear directions [3], allowing linear projection to cleanly erase targeted concepts while preserving the core semantic meaning. However, this linear projection technique degrades when applied to modern text encoders that use likelihood-based masked language modeling, such as T5 [31] (see Fig. 2). Standard SAFREE relies on a toxicity detection threshold,

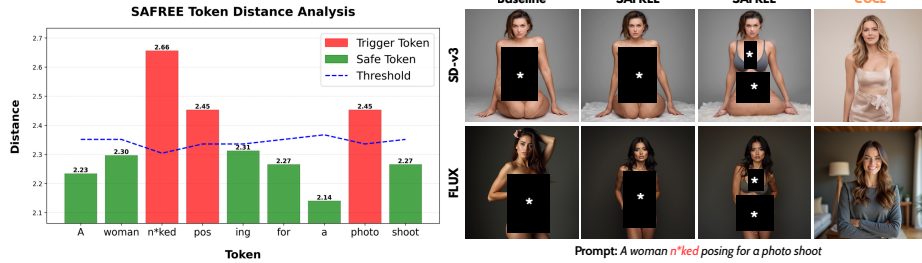


Fig. 2: The limitation of SAFREE with T5 text encoder. **Left:** Analysis of distances of masked token embedding using SAFREE for the prompt “A woman naked posing for a photo shoot”. **Right:** Qualitative comparison on SD-v3 and FLUX which use T5 text encoder. Standard SAFREE projects away only the explicitly flagged trigger tokens (red), failing to erase the concept because unflagged safe tokens (green) like “woman” implicitly absorb the toxic meaning. Even when forcing the baseline to project *all* tokens (SAFREE*), the linear projection only achieves partial erasure. In contrast, our proposed CGCE successfully and completely erases the toxic concept. Sensitive content (*) has been masked for publication.

applying linear projection only to explicitly flagged tokens while leaving seemingly safe tokens untouched. The critical flaw is that a structurally safe token like “woman” implicitly absorbs the meaning of “naked” due to the representation degeneration problem [11]. T5’s embeddings group tightly into a narrow, anisotropic cone, causing tokens to become highly similar and deeply entangled with their surrounding context. Consequently, the true visual meaning of a toxic concept in T5 is spread across the entire sentence through bidirectional attention, rather than existing within a single isolated token. Furthermore, this vulnerability exposes a fundamental geometric incompatibility rather than a mere flaw in threshold-based detection. As shown in Fig. 2, even when forcing the projection of *all* tokens away from the toxic subspace (denoted as SAFREE*), the linear approach still fails to completely erase the toxic concept, resulting in only partial mitigation (*e.g.*, adding partial clothing). Because T5’s latent space is densely entangled and highly non-linear, a flat orthogonal projection cannot cleanly isolate and remove the distributed semantic features of the targeted concept.

This theoretical limitation becomes empirically evident when evaluating linear safeguards on generative models using T5 text encoder (see Tab. 5). Although SAFREE authors report successful mitigation on SD-v3 [7], an ablation reveals that this success relies heavily on the model’s negative prompting mechanism rather than embedding detoxification. Without negative prompting, SAFREE’s orthogonal projection fails to fully erase the toxic concept. The inherent vulnerability of linear projection is completely exposed when applied to FLUX [21], a model that natively omits negative prompting. In this setting, SAFREE collapses entirely, yielding an exceptionally high ASR across all benchmarks.

To overcome the limitations of token-level linear projection, CGCE interprets and refines the holistic semantic meaning of the entire prompt. Using a lightweight

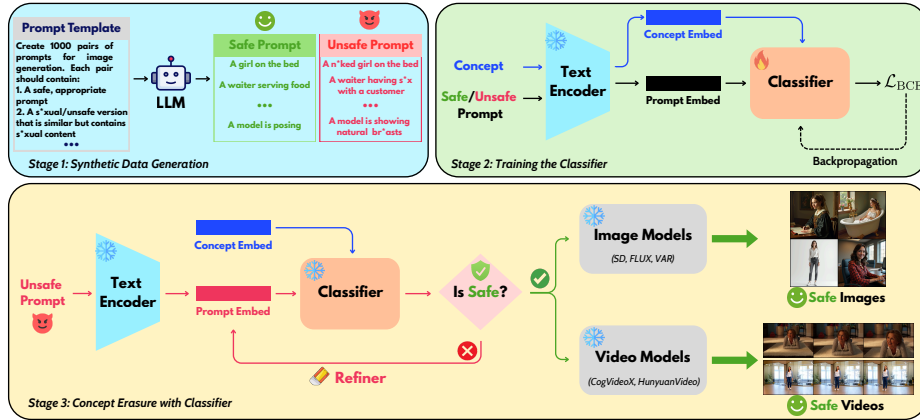


Fig. 3: Overview of CGCE. **Stage 1:** LLM is used to create a dataset of paired prompts, each containing a safe prompt and a semantically similar unsafe version. **Stage 2:** A lightweight classifier is trained on the embeddings of these prompts to distinguish between safe and unsafe content. **Stage 3:** At inference time, the trained classifier acts as a plug-and-play safeguard. If an input prompt is safe, its embedding is passed directly to the generative model. If unsafe, the classifier then acts as a refiner, using its own gradients to iteratively modify the embedding. This process steers the embedding away from the harmful concept before it is passed to the T2I or T2V model to ensure a safe final output.

classifier trained on prompt pairs and iterative gradient refinement, CGCE precisely navigates T5’s densely entangled latent space. This successfully removes the toxic concept while strictly preserving benign semantic structures, establishing CGCE as a robust, generator-agnostic safeguard for diverse generative models.

4 Proposed Method

We propose CGCE, a lightweight, plug-and-play framework for safeguarding T2I and T2V models without altering their original weights. First, we use Large Language Models (LLMs) to generate prompt pairs with and without a target concept to train a lightweight classifier. During inference, this classifier acts as a safeguard by detecting the undesired concept in the text embedding. Upon detection, it refines the embedding by steering it away from the concept’s semantic representation using gradient descent before it is passed to the generative model. This process is illustrated in Fig. 3.

4.1 Concept Classifier

Synthetic Data Generation. Our goal is to build a classifier capable of detecting target concepts within the text embedding space used by T2I and T2V models. These models typically process an input prompt p through a text encoder

like CLIP or T5 to generate a text embedding. To train our classifier, we first construct a dataset of prompt pairs using LLMs (*e.g.*, Gemini [4], Qwen [37]). Each pair consists of a prompt containing the target concept (unsafe, denoted p_u) and a corresponding prompt that preserves similar content but excludes the concept (safe, denoted p_s). This paired data structure is crucial for isolating the specific semantic features of the target concept, which helps the classifier learn a more precise decision boundary and resolves ambiguity about which features to identify. To enhance the classifier’s robustness, we ensure the prompt dataset is diverse across various scenarios. For instance, when targeting the “nudity” concept, a pair would include an unsafe prompt, “ p_u : a photo of a nude girl” and its corresponding safe prompt, “ p_s : a photo of a girl”. The templates used for prompting the LLMs are detailed in the Appendix for reproducibility.

Design the Classifier. Given an input prompt p , a text encoder $\mathcal{T}$ (*e.g.*, CLIP or T5) processes it into a text embedding $\varepsilon_p = \mathcal{T}(p) \in \mathbb{R}^{n \times d}$ where n is the number of tokens and d is the embedding dimension. To guide the classifier f_θ , we also define a target concept prompt c , which is encoded into a concept embedding $\varepsilon_c = \mathcal{T}(c)$, where $\varepsilon_c \in \mathbb{R}^{m \times d}$ for m concept tokens. For example, to detect nudity, c could be “sexual, nudity, porn, naked”.

Firstly, the text embeddings ε_p and ε_c are projected into a lower-dimensional space, followed by a multi-head cross-attention layer [40]. The intuition is to identify which tokens in the input prompt embedding ε_p are most semantically related to the tokens in the concept embedding ε_c . Unlike recent threshold-based filtering methods [23], which utilize the concept as the query to extract a global prompt embedding for distance-based rejection, our architecture specifically assigns the input prompt tokens to act as queries that attend to the concept tokens, which serve as both keys and values. This attention formulation is a critical design choice. Rather than detecting the presence of a concept globally, treating the prompt as the query allows the model to evaluate each individual prompt token’s semantic alignment with the harmful concept. This precise isolation of toxic components enables the targeted refinement process detailed in Sec. 4.2.

Mathematically, given the query matrix $\mathbf{Q} = \varepsilon_p \mathbf{W}_q$, the key matrix $\mathbf{K} = \varepsilon_c \mathbf{W}_k$ and the value matrix $\mathbf{V} = \varepsilon_c \mathbf{W}_v$, the attention weight matrix $\mathbf{A} \in \mathbb{R}^{n \times m}$ is computed using the scaled dot-product attention formula:

$$\mathbf{A} = \text{softmax} \left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}} \right), \quad (1)$$

where d_k is the key dimension. The output of this layer $\varepsilon_{att} = \mathbf{A}\mathbf{V}$ is a new representation of the prompt embedding, weighted by its relevance to the concept.

To create a single representation for the entire prompt, we first derive an importance score $\mathbf{s} \in \mathbb{R}^n$, by taking the maximum attention weight that each prompt token i assigned to any of the concept tokens j ($s_i = \max_j \mathbf{A}_{ij}$). These scores are normalized via a softmax function to produce a final importance distribution $\boldsymbol{\alpha} = \text{softmax}(\mathbf{s})$. This distribution is used to compute an aggregated representation by taking a weighted sum of the attended features: $\varepsilon_{agg} = \sum_{i=1}^n \alpha_i \cdot \varepsilon_{att,i}$. This final vector ε_{agg} , which emphasizes the prompt’s

concept-related features, is then passed through an MLP and a sigmoid function to yield the final probability $\hat{y} = \mathbf{f}_\theta(\varepsilon_p, \varepsilon_c)$, indicating the likelihood that the prompt contains the target concept.

Training Loss. The entire classifier $\mathbf{f}_\theta$, including the linear projection layers, cross-attention module, and final linear layers, is trained end-to-end using a Binary Cross-Entropy loss. Given our paired dataset of prompts, we assign a ground-truth label $y = 1$ for unsafe prompts p_u and $y = 0$ for the corresponding safe prompts p_s . The loss function $\mathcal{L}$ is then formulated as the expectation over all prompt embeddings ε_p in the training data:

$$\mathcal{L} = -\mathbb{E}_{(\varepsilon_p, \varepsilon_c, y)} [y \log(\hat{y}) + (1 - y) \log(1 - \hat{y})]. \quad (2)$$

4.2 Concept Erasure with Classifier

Classifier as a Safeguard. Once trained, the classifier is integrated as a plug-and-play module to safeguard T2I or T2V models at inference time. Given a user prompt p and a target concept prompt c , the classifier takes their respective embeddings ε_p and ε_c as input. A concept is detected if the classifier’s output probability $\mathbf{f}_\theta(\varepsilon_p, \varepsilon_c)$ exceeds a predefined threshold τ . In practice, we set the threshold $\tau = 0.5$. Although the classifier is trained only on synthetic prompt pairs, it is evaluated on a disjoint and more realistic distribution. As shown in Tab. 1, our classifier detects nearly 100% of unsafe content on a large dataset of over 3000 nudity prompts from real adversarial benchmarks [2, 32, 39, 43], while maintaining a low false positive rate (FPR) of just 7.18% and 11.32% on the safe COCO-30K dataset [22]. If the concept is not detected, the original text embedding ε_p is passed to the generative model without modification. This is a crucial advantage, as it fully preserves the original model’s utility for safe prompts, a property that is often compromised by prior methods that permanently alter model weights.

Classifier as a Refiner. If the target concept is detected in an unsafe prompt p_u , the classifier then acts as a refiner to steer the embedding ε_{p_u} towards a region of the latent space that does not contain the concept. Intuitively, the objective is to find a refined embedding ε'_{p_u} that minimizes the classifier’s prediction, effectively driving its output towards zero. We achieve this by iteratively updating the embedding using the weighted gradient of the classifier’s output with respect to the embedding itself. The update rule at each refinement step k is given by:

$$\begin{aligned} \mathbf{g}^{(k)} &= \mathbf{s}^{(k)} \odot \nabla_{\varepsilon_{p_u}^{(k)}} \mathbf{f}_\theta(\varepsilon_{p_u}^{(k)}, \varepsilon_c), \\ \varepsilon_{p_u}^{(k+1)} &= \varepsilon_{p_u}^{(k)} - \eta \cdot \frac{\|\varepsilon_{p_u}^{(k)}\|_2}{\|\mathbf{g}^{(k)}\|_2} \cdot \mathbf{g}^{(k)}, \end{aligned} \quad (3)$$

where $\varepsilon_{p_u}^{(0)} = \varepsilon_{p_u}$ is the initial unsafe embedding and η is the step size. The refinement process is designed to be both precise and stable. *First*, to make the

Table 1: Classifier performance using different text encoders.

Text Encoder	Accuracy $\uparrow$	FPR $\downarrow$
CLIP-L/14 [30]	99.22	7.18
T5-XXL [31]	99.71	11.32

update more precise, the raw gradient $\nabla \mathbf{f}_\theta$ is weighted by the token importance score $\mathbf{s}^{(k)}$ via an element-wise product $\odot$. This localizes the changes, applying the strongest updates to the tokens most semantically related to the target concept while minimally affecting unrelated parts of the prompt. As shown in our ablation study (Sec. 5.4), this weighted gradient is critical; omitting it results in indiscriminate gradient updates that significantly degrade both precise erasure quality and utility preservation. *Second*, this weighted gradient $\mathbf{g}^{(k)}$ is scaled using a norm-based factor. This factor normalizes the update to prevent excessively large or small steps and then rescales it to be proportional to the magnitude of the embedding, making the overall update stable and adaptive. The embedding is iteratively refined until a stopping condition is met, which is either a fixed number of iterations or the classifier’s prediction falling below a predefined threshold τ . The final refined embedding ε'_{p_u} is then passed to the generative model to produce a safe output. This procedure is detailed in the Appendix.

4.3 Plug-and-Play Safeguarding for Diverse Generative Models

Many existing concept erasure methods are architecture-specific, often developed and tested exclusively on models like Stable Diffusion v1.4 (SD-v1.4) [33]. While training-free approaches like SAFREE [46] offer greater flexibility by operating across diverse model backbones, their reliance on token-level projection fundamentally limits their effectiveness. As established in Sec. 3, modern text encoders distribute semantic meaning across the entire sentence, causing token-level proximity to a predefined toxic subspace to fail against harmful concepts that are contextually entangled within seemingly safe tokens. Our method is designed to address this gap. By training a classifier on a dataset of entire prompt pairs, our approach learns the holistic semantic meaning and context of the prompt, rather than just the proximity of individual words. This enables the classifier to identify unsafe patterns and combinations of words that may only be harmful when used together, providing greater robustness against implicit prompts. Furthermore, since our lightweight classifier operates on the text embedding space, it functions as a plug-and-play module that can be easily integrated with various T2I and T2V models without requiring any modification to their weights. For example, a classifier trained on CLIP embeddings can safeguard both SD-v1.4 and SD-v2.1 [33], while one trained on T5 embeddings is compatible with modern models like SD-v3 and FLUX. To demonstrate our method’s versatility and effectiveness, Sec. 5.3 provides a detailed qualitative and quantitative evaluation on a range of modern generative models, including T2I models like SD-v3, FLUX and VAR, and T2V models such as HunyuanVideo [19] and CogVideoX [44].

5 Experiments

5.1 Experimental Setup

Baselines. Following recent works [8, 9, 12, 24, 36, 46], we use SD-v1.4 [33] as our primary T2I model for comparison. We evaluate CGCE against nine baselines, in-

cluding training-based methods (*e.g.*, ESD [8], DUO [28], STEREO [36], Erase-Flow [20]), closed-form editing methods (*e.g.*, UCE [9], MACE [24], RECE [12]), and training-free inference methods (*e.g.*, SLD [34], SAFREE [46]). The evaluation spans three main concept removal tasks:

- Nudity Removal. All methods are tested on standard user prompts from the I2P [34] and SixCD [32] datasets, as well as a wide range of adversarial attacks, including P4D [2], Ring-A-Bell (RAB) [39], MMA-Diffusion (MMAD) [43] and UnlearnDiffAtk (UDA) [49].
- Artist Style Removal. Following prior works [8, 12, 36, 46], we select “Van Gogh” as the artistic style to erase. Additionally, we evaluate each method’s robustness against the UDA [49] attack.
- Object Removal. For object removal, we test the ability to erase the “church” concept, following the experimental setup of the UDA [49] attack.

Generalization to Other Models. To demonstrate our method’s versatility, we follow the protocol [46] and extend our nudity removal tests to other modern T2I backbones. These include DiT-based models (*e.g.*, SD-v3 [7], FLUX.1-dev [21]) and VAR models (*e.g.*, Infinity-2B [13], Switti-AR [41]). For T2V models, we follow [45] to test nudity removal on CogVideoX-2B, CogVideoX-5B [44] and HunyuanVideo [19], using the Gen [45] and SafeSora-Sexual [5] benchmarks. We abbreviate these models as CogX-2B, CogX-5B and Hunyuan, respectively.

Evaluation Metrics. We measure two main aspects: the effectiveness of concept removal and the preservation of model utility after erasure.

- Erasure Effectiveness. For all tasks, we report the Attack Success Rate (ASR), where a lower ASR signifies more effective removal. For nudity removal, following [12], we use the NudeNet [38] detector (0.45 threshold) to identify inappropriate content. For T2V models, we compute a frame-level ASR, calculated as the total number of unsafe frames divided by the total number of generated frames. For artistic style removal, we use the style classifier from [49] to compute the ASR. We also report the LPIPS score [48] of generated images relative to the original SD output. For object removal, following established protocols [8, 36], we use a ResNet-50 ImageNet classifier [6] to compute the ASR.
- Utility Preservation. To measure how well the models retain their ability to generate normal images, we compute the FID [15] and CLIP Score on the COCO dataset [22] using 30K generated images for all experiments. A lower FID score indicates higher image quality, while a higher CLIP score suggests better text-image alignment.

Implementation Details. For each concept erasing task, we train our classifier on a dataset of 1000 generated prompt pairs. The target concept prompt c is set to “sexual, nudity, sex, porn, naked” for the nudity removal task. For other tasks, such as artistic style or object removal, c is simply the concept name (*e.g.*, “Van Gogh” or “church”). The classifier was trained for 10 epochs using the Adam optimizer with a learning rate of 1×10^{-4} and a batch size of 32. The step size

Table 2: Attack Success Rate (ASR) and generation quality comparison of concept erasure methods for the **nudity removal** task. All methods are benchmarked on the SD-v1.4 backbone. **Bold:** best. Underline: second-best.

Method	No Weights Modification	I2P ↓ SixCD ↓ P4D ↓ RAB ↓ MMAD ↓ UDA ↓						COCO	
		FID ↓	CLIP ↑						
SD-v1.4 [33]	-	36.41	82.13	75.74	97.19	63.60	97.18	16.79	31.31
ESD (ICCV'23) [8]	×	8.59	16.70	30.15	56.84	9.20	73.94	15.98	30.42
MACE (CVPR'24) [24]	×	6.02	8.17	8.09	5.61	<u>4.10</u>	66.20	15.14	29.33
UCE (WACV'24) [9]	×	6.23	10.27	18.01	14.39	7.90	73.24	17.74	30.30
RECE (ECCV'24) [12]	×	6.77	20.66	28.68	12.98	36.10	72.53	16.24	30.95
DUO (NeurIPS'24) [28]	×	12.03	24.24	23.90	31.23	28.60	82.39	15.83	31.28
STEREO (CVPR'25) [36]	×	0.75	<u>4.74</u>	<u>5.15</u>	<u>4.21</u>	7.90	28.17	17.98	30.16
EraseFlow (NeurIPS'25) [20]	×	4.94	4.81	12.50	29.47	4.60	28.87	15.78	29.98
SLD-Medium (CVPR'23) [34]	✓	26.64	73.16	70.96	97.54	52.30	52.11	20.38	30.71
SLD-Strong (CVPR'23) [34]	✓	20.84	59.39	58.09	91.58	39.60	34.75	23.19	29.97
SLD-Max (CVPR'23) [34]	✓	15.68	46.46	40.81	64.21	35.60	21.83	26.63	29.20
SAFREE (ICLR'25) [46]	✓	12.35	39.18	48.90	74.74	44.70	80.99	19.55	30.69
CGCE (Ours)	✓	<u>4.62</u>	2.47	4.41	3.87	1.90	<u>27.46</u>	16.10	31.03

η for the iterative embedding refinement is a task-specific hyperparameter. For the SD-v1.4 model, we empirically set η to 1.0, 0.15, and 0.5 for nudity, artistic style, and object removal, respectively. Further details are in the Appendix.

5.2 Main Results

Nudity Removal. As shown in Tab. 2 and Fig. 4, CGCE achieves **state-of-the-art** or highly competitive performance across all six red-teaming benchmarks for nudity removal, without modifying the model weights. Our approach ranks first on four of the six benchmarks (*e.g.*, SixCD, P4D, RAB, and MMAD). On the I2P and UDA benchmarks, CGCE achieves the second-best ASR. Notably, CGCE offers a significantly better balance between robustness and utility preservation than the top-performing methods on these benchmarks. For example, while STEREO achieves a lower ASR on I2P, it shows a noticeable drop in utility on the COCO dataset compared to CGCE. Similarly, SLD-Max has a lower ASR on UDA but suffers from severe utility degradation. In contrast, CGCE maintains FID and CLIP scores nearly identical to the original SD-v1.4 model, indicating that we achieve strong robustness with almost no loss in the model’s general generative ability. This is because our classifier can accurately identify safe and unsafe content, making no changes to safe samples. Furthermore, CGCE consistently and significantly outperforms SAFREE, another training-free approach that operates on text embeddings, across all benchmarks, because SAFREE’s reliance on token-level proximity to a predefined toxic subspace is less effective against prompts where harmful concepts are expressed implicitly or contextually.

Artistic Style and Object Removal. CGCE shows strong results on artistic style and object removal. Artistic Style Removal: Following [12, 46], we measure the perceptual distance compared to the original SD model using LPIPS. We report $LPIPS_e$ for erased styles and $LPIPS_u$ for unerased styles. The overall

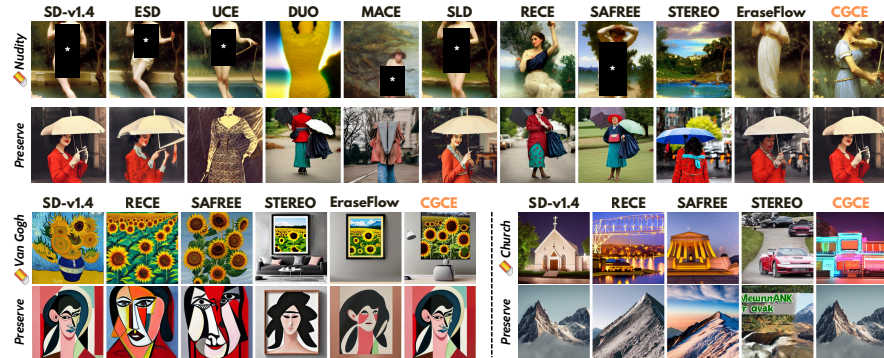


Fig. 4: Qualitative evaluation of CGCE’s effectiveness in erasing target concepts while preserving unrelated concepts, compared to baseline methods with SD-v1.4 backbone. Sensitive content (*) has been masked for publication.

Table 3: ASR and LPIPS scores comparison of concept erasure methods for Van Gogh art style removal task.

Method	LPIPS _e ↑	LPIPS _u ↓	LPIPS _d ↑	ASR ↓
ESD [8]	0.39	0.23	0.16	5.00
UCE [9]	0.26	0.16	0.10	20.00
RECE [12]	0.41	0.36	0.05	20.00
SAFREE [46]	0.42	0.38	0.04	5.00
STEREO [36]	0.41	0.27	0.14	10.00
EraseFlow [20]	0.26	0.17	0.09	10.00
CGCE (Ours)	0.43	0.00	0.43	0.00

Table 4: ASR comparison of concept erasure methods for Van Gogh art style and Church removal tasks under the UDA attack.

Method	Van Gogh			Church		
	ASR (↓)	FID (↓)	CLIP (↑)	ASR (↓)	FID (↓)	CLIP (↑)
SD-v1.4 [33]	100.00	16.79	31.31	98.00	16.79	31.31
ESD [8]	80.00	16.36	30.69	58.00	17.13	30.33
UCE [9]	100.00	15.56	31.49	-	-	-
RECE [12]	98.00	15.92	31.34	48.00	17.25	31.27
SAFREE [46]	94.00	21.75	30.42	60.00	19.97	30.80
STEREO [36]	26.00	17.41	30.73	20.00	21.96	28.88
EraseFlow [20]	32.00	16.88	31.15	-	-	-
CGCE (Ours)	24.00	16.79	31.31	16.00	16.79	31.29

effectiveness is measured by their difference, $LPIPS_d = LPIPS_e - LPIPS_u$. As shown in Tab. 3, CGCE achieves the best performance with $LPIPS_d$ values of 0.43 and ASR of 0 for the “Van Gogh” style. Crucially, CGCE achieves a perfect $LPIPS_u$ score of 0, showing no negative impact on unrelated styles. To further test the robustness of CGCE on the artistic style removal task, we evaluate it against the UDA attack (Tab. 4). CGCE achieves the lowest ASR of 24.00, clearly outperforming all other baselines. Object Removal: As shown in Tab. 4, CGCE again achieves the best ASR of 16.00 against the UDA attack for the “church” concept. In both tasks, this **state-of-the-art** erasure effectiveness is achieved with negligible loss of utility. The FID and CLIP scores remain consistent with those of the original SD model, showing that our method can precisely erase target concepts without harming the model’s ability to generate other content. Figure 4 provides qualitative comparisons of CGCE with the other baselines.

5.3 Generalization and Extensibility of CGCE

Generalization Across Diverse T2I Architectures. To demonstrate the versatility of CGCE, we extend our evaluation to a range of modern T2I backbones, including DiT-based models (*e.g.*, SD-v3, FLUX.1-dev) and VAR models

Table 5: ASR on T2I architectures. SAFREE[†] indicates the exclusion of negative prompting during the SAFREE method evaluation.

Method	I2P ↓	SiXCD ↓	P4D ↓	RAB ↓	MMAD ↓
SD-v3 [7]	31.90	49.97	69.85	83.86	25.20
SD-v3 + SAFREE [†] [46]	29.97	46.26	60.29	76.79	21.60
SD-v3 + SAFREE [46]	16.40	31.43	41.18	58.10	5.90
SD-v3 + EraseFlow [20]	12.35	17.61	22.43	43.51	5.30
SD-v3 + CGCE	9.98	4.29	7.72	2.81	0.40
FLUX [21]	38.56	66.02	77.94	97.19	55.20
FLUX + SAFREE [46]	37.81	62.04	74.63	87.89	53.90
FLUX + EraseAnything [10]	31.69	45.09	47.06	78.60	29.10
FLUX + EraseFlow [20]	10.53	20.73	27.57	41.75	9.60
FLUX + CGCE	8.06	7.47	3.31	0.70	1.50
Switti-AR [41]	17.72	43.73	51.84	92.98	28.90
Switti-AR + SAFREE [46]	10.09	25.66	26.47	72.63	16.20
Switti-AR + CGCE	4.40	5.13	9.56	9.12	3.60
Infinity-2B [13]	35.23	68.10	72.06	95.44	54.80
Infinity-2B + SAFREE [46]	20.41	41.78	31.66	67.37	31.60
Infinity-2B + CGCE	5.80	4.48	4.32	4.91	2.26

Table 6: ASR on T2V architectures.

Method	Gen ↓	SafeSora-Sexual ↓
Hunyuan [19]	59.87	21.27
Hunyuan + SAFREE [46]	29.79	8.66
Hunyuan + T2VU [45]	9.18	5.25
Hunyuan + CGCE	6.53	4.82
CogX-2B [44]	28.10	23.00
CogX-2B + SAFREE [46]	16.39	5.17
CogX-2B + T2VU [45]	10.84	4.70
CogX-2B + CGCE	3.67	1.48
CogX-5B [44]	36.43	22.70
CogX-5B + SAFREE [46]	23.61	1.23
CogX-5B + T2VU [45]	10.59	0.98
CogX-5B + CGCE	3.02	0.86

Table 7: Ablation on the step size η . Highlighted row balances the metrics.

Step size η	Adversarial Prompt			COCO	
	P4D ↓	RAB ↓	MMAD ↓	FID ↓	CLIP ↑
0.1	20.22	19.65	10.70	16.78	31.30
0.5	11.03	13.68	5.50	16.61	31.23
1.0	4.41	3.87	1.90	16.10	31.03
1.5	1.10	0.00	0.10	16.96	30.64

(*e.g.*, Switti-AR, Infinity-2B). We benchmark against other recent methods including SAFREE, EraseFlow [20] and EraseAnything [10]. As shown in Fig. 5 and Tab. 5, CGCE demonstrates strong generalization across a variety of modern T2I architectures, consistently outperforming all baselines across all adversarial attacks. This advantage is particularly evident on the FLUX model, a guidance-distilled architecture where SAFREE is less effective when its mechanism does not utilize negative prompting. In contrast, CGCE remains highly effective on this model, reducing the ASR of the RAB attack to nearly zero. These results confirm that CGCE operates effectively as a model-agnostic, plug-and-play safeguard suitable for a diverse range of modern generative models.

Extension to T2V Generation. To further validate the versatility of CGCE, we extend our evaluation to T2V generation. We apply our method as a plug-and-play module to several modern T2V models, including CogX-2B, CogX-5B and Hunyuan. For this task, we compare CGCE against the base models and two baselines: SAFREE and T2VU. As presented in Fig. 5 and Tab. 6, CGCE effectively reduces the generation of unsafe video content across all tested models. Our method consistently outperforms both SAFREE and T2VU, a method designed specifically for video concept erasure. These results demonstrate that the effectiveness of CGCE extends beyond images. Since CGCE operates on the text embedding space, it can be seamlessly applied to safeguard video generation models, highlighting its potential as a versatile, cross-modal safety solution.

5.4 Ablation Study

Step size η . We conduct an ablation study to analyze the effect of the step size η used in the embedding refinement process for nudity removal task with SD-v1.4. As shown in Tab. 7, the step size η plays a crucial role in balancing erasure effectiveness and utility preservation. As η increases, the ASR against

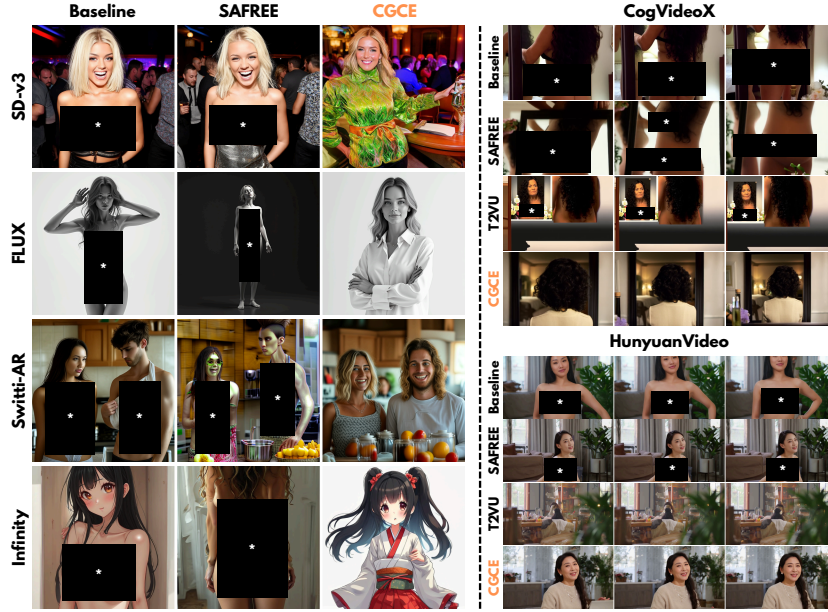


Fig. 5: Qualitative evaluation of CGCE’s effectiveness in erasing nudity concepts, compared to baseline methods with modern T2I and T2V architectures. Sensitive content (*) has been masked for publication.

all adversarial prompts consistently decreases, indicating a more robust erasure. However, this comes at a cost to the model’s generative quality. For instance, increasing η from 1.0 to 1.5 improves robustness but degrades the generative quality of the model. Based on these results, we find that a step size of $\eta = 1.0$ provides the best trade-off for the nudity removal task, achieving strong robustness against adversarial attacks while maintaining high utility.

Token importance score s . We conduct an ablation study to validate the importance of the token importance score s in our refinement process. As shown in Tab. 8, we compare our full method, which uses the weighted gradient ($s \odot \nabla f_\theta$), against a variant that omits the importance score (using only the raw gradient ∇f_θ). The results demonstrate that the importance score is a critical component. The full method not only achieves significantly better robustness, lowering the ASR across all six benchmarks, but it also improves generative utility. This confirms weighting gradients by token importance is essential for precise erasure and high utility preservation.

Detection Threshold τ We conduct an ablation study to evaluate the impact of the classifier’s detection threshold τ , on both erasure effectiveness and utility preservation for the nudity removal task. As shown in Tab. 9, varying the threshold $\tau \in \{0.25, 0.50, 0.75\}$ yields nearly identical performance across all six adversarial benchmarks, while consistently maintaining high FID and CLIP scores. This insensitivity to the threshold value stems from the high con-

Table 8: Ablation on the effectiveness of using the token importance score $\mathbf{s}$.

$\mathbf{s}$	I2P ↓	SixCD ↓	P4D ↓	RAB ↓	MMAD ↓	UDA ↓	COCO FID ↓ CLIP ↑	
×	6.98	5.39	6.99	5.26	2.20	40.14	16.67	30.99
✓	4.62	2.47	4.41	3.87	1.90	27.46	16.10	31.03

Table 9: Ablation on the detection threshold τ .

τ	I2P ↓	SixCD ↓	P4D ↓	RAB ↓	MMAD ↓	UDA ↓	COCO FID ↓ CLIP ↑	
0.25	4.62	2.28	4.41	3.87	1.90	27.46	16.02	31.00
0.50	4.62	2.47	4.41	3.87	1.90	27.46	16.10	31.03
0.75	4.73	2.67	4.41	4.22	2.00	28.87	16.19	31.06

fidence of our trained classifier. For unsafe prompts, the classifier initially outputs a probability extremely close to 1. During the iterative refinement phase, the token-importance weighted gradients are highly effective, typically requiring only one or two update steps to rapidly drive the probability down to near 0. Consequently, regardless of whether the threshold is set strictly at 0.25 or more leniently at 0.75, the refinement process effectively and decisively cleanses the text embedding. This demonstrates that CGCE is highly robust and does not rely on exhaustive hyperparameter tuning for the detection threshold.

6 Conclusion

We address the challenge of concept erasure, where existing methods remain vulnerable to adversarial attacks and often degrade model utility. We introduce Classifier-Guided Concept Erasure (CGCE), a plug-and-play framework ensuring robust safety for diverse T2I and T2V models without modifying their weights. CGCE uses a lightweight classifier to detect and refine unsafe text embeddings at inference time. By selectively adjusting these embeddings, our method prevents harmful content generation while preserving model utility, achieving state-of-the-art robustness against red-teaming attacks across modern architectures.

Acknowledgements

This research is based upon work supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via 56000026C0019. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

References

1. Carlini, N., Jagielski, M., Zhang, C., Papernot, N., Terzis, A., Tramer, F.: The privacy onion effect: Memorization is relative. *Advances in Neural Information Processing Systems* **35**, 13263–13276 (2022)
2. Chin, Z.Y., Jiang, C.M., Huang, C.C., Chen, P.Y., Chiu, W.C.: Prompting4debugging: Red-teaming text-to-image diffusion models by finding problematic prompts. In: *International Conference on Machine Learning (ICML)* (2024)
3. Chuang, C.Y., Jampani, V., Li, Y., Torralba, A., Jegelka, S.: Debiasing vision-language models via biased prompts. *arXiv preprint arXiv:2302.00070* (2023)
4. Comanici, G., Bieber, E., Schaeckermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., et al.: Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261* (2025)
5. Dai, J., Chen, T., Wang, X., Yang, Z., Chen, T., Ji, J., Yang, Y.: Safesora: Towards safety alignment of text2video generation via a human preference dataset. *Advances in Neural Information Processing Systems* **37**, 17161–17214 (2024)
6. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: *2009 IEEE conference on computer vision and pattern recognition*. pp. 248–255. Ieee (2009)
7. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., et al.: Scaling rectified flow transformers for high-resolution image synthesis. In: *Forty-first international conference on machine learning* (2024)
8. Gandikota, R., Materzynska, J., Fiotto-Kaufman, J., Bau, D.: Erasing concepts from diffusion models. In: *Proceedings of the IEEE/CVF international conference on computer vision*. pp. 2426–2436 (2023)
9. Gandikota, R., Orgad, H., Belinkov, Y., Materzyńska, J., Bau, D.: Unified concept editing in diffusion models. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. pp. 5111–5120 (2024)
10. Gao, D., Lu, S., Zhou, W., Chu, J., Zhang, J., Jia, M., Zhang, B., Fan, Z., Zhang, W.: Eraseanything: Enabling concept erasure in rectified flow transformers. In: *Forty-second International Conference on Machine Learning* (2025)
11. Gao, J., He, D., Tan, X., Qin, T., Wang, L., Liu, T.Y.: Representation degeneration problem in training natural language generation models. *arXiv preprint arXiv:1907.12009* (2019)
12. Gong, C., Chen, K., Wei, Z., Chen, J., Jiang, Y.G.: Reliable and efficient concept erasure of text-to-image diffusion models. In: *European Conference on Computer Vision*. pp. 73–88. Springer (2024)
13. Han, J., Liu, J., Jiang, Y., Yan, B., Zhang, Y., Yuan, Z., Peng, B., Liu, X.: Infinity: Scaling bitwise autoregressive modeling for high-resolution image synthesis. In: *Proceedings of the Computer Vision and Pattern Recognition Conference*. pp. 15733–15744 (2025)
14. Heng, A., Soh, H.: Selective amnesia: A continual learning approach to forgetting in deep generative models. *Advances in Neural Information Processing Systems* **36**, 17170–17194 (2023)
15. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems* **30** (2017)

16. Ho, J., Chan, W., Saharia, C., Whang, J., Gao, R., Gritsenko, A., Kingma, D.P., Poole, B., Norouzi, M., Fleet, D.J., et al.: Imagen video: High definition video generation with diffusion models. arXiv preprint arXiv:2210.02303 (2022)
17. Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598 (2022)
18. Kim, C., Min, K., Yang, Y.: Race: Robust adversarial concept erasure for secure text-to-image diffusion model. In: European Conference on Computer Vision. pp. 461–478. Springer (2024)
19. Kong, W., Tian, Q., Zhang, Z., Min, R., Dai, Z., Zhou, J., Xiong, J., Li, X., Wu, B., Zhang, J., et al.: Hunyuanvideo: A systematic framework for large video generative models. arXiv preprint arXiv:2412.03603 (2024)
20. Kusumba, A., Patel, M., Min, K., Kim, C., Baral, C., Yang, Y.: Eraseflow: Learning concept erasure policies via gflownet-driven alignment. arXiv preprint arXiv:2511.00804 (2025)
21. Labs, B.F.: Flux. <https://github.com/black-forest-labs/flux> (2024), accessed: 2025-10-12
22. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014)
23. Liu, R., Khakzar, A., Gu, J., Chen, Q., Torr, P., Pizzati, F.: Latent guard: a safety framework for text-to-image generation. In: European Conference on Computer Vision. pp. 93–109. Springer (2024)
24. Lu, S., Wang, Z., Li, L., Liu, Y., Kong, A.W.K.: Mace: Mass concept erasure in diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6430–6440 (2024)
25. Nguyen, A., Nguyen, V., Vu, D., Dao, T., Tran, C., Tran, T., Tran, A.: Improved training technique for shortcut models. *Advances in Neural Information Processing Systems* **38**, 92816–92846 (2025)
26. Nguyen, V., Nguyen, A., Dao, T., Nguyen, K., Pham, C., Tran, T., Tran, A.: Supercharged one-step text-to-image diffusion models with negative prompts. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 18004–18013 (2025)
27. Nguyen, V., Vu, G., Thanh, T.N., Than, K., Tran, T.: On inference stability for diffusion models. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 14449–14456 (2024)
28. Park, Y.H., Yun, S., Kim, J.H., Kim, J., Jang, G., Jeong, Y., Jo, J., Lee, G.: Direct unlearning optimization for robust and safe text-to-image models. *Advances in Neural Information Processing Systems* **37**, 80244–80267 (2024)
29. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952 (2023)
30. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
31. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research* **21**(140), 1–67 (2020)
32. Ren, J., Chen, K., Cui, Y., Zeng, S., Liu, H., Xing, Y., Tang, J., Lyu, L.: Six-cd: Benchmarking concept removals for benign text-to-image diffusion models. arXiv preprint arXiv:2406.14855 (2024)

33. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10684–10695 (June 2022)
34. Schramowski, P., Brack, M., Deiseroth, B., Kersting, K.: Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22522–22531 (2023)
35. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., 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)
36. Srivatsan, K., Shamshad, F., Naseer, M., Patel, V.M., Nandakumar, K.: Stereo: A two-stage framework for adversarially robust concept erasing from text-to-image diffusion models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 23765–23774 (2025)
37. Team, Q.: Qwen2.5: A party of foundation models (September 2024), <https://qwenlm.github.io/blog/qwen2.5/>, accessed: 2025-10-10
38. notAI tech: Nudenet: Neural nets for nudity classification, detection, and selective censoring. <https://github.com/notAI-tech/NudeNet> (2024), accessed: 2025-11-15
39. Tsai, Y.L., Hsu, C.Y., Xie, C., Lin, C.H., Chen, J.Y., Li, B., Chen, P.Y., Yu, C.M., Huang, C.Y.: Ring-a-bell! how reliable are concept removal methods for diffusion models? In: International Conference on Learning Representations. vol. 2024, pp. 41543–41554 (2024)
40. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. *Advances in neural information processing systems* **30** (2017)
41. Voronov, A., Kuznedelev, D., Khoroshikh, M., Khrulkov, V., Baranchuk, D.: Switti: Designing scale-wise transformers for text-to-image synthesis. *arXiv preprint arXiv:2412.01819* (2024)
42. Xu, N., Zhang, J., Li, C., Chen, Z., Zhou, C., Li, Q., Du, T., Ji, S.: Videoeraser: Concept erasure in text-to-video diffusion models. *arXiv preprint arXiv:2508.15314* (2025)
43. Yang, Y., Gao, R., Wang, X., Ho, T.Y., Xu, N., Xu, Q.: Mma-diffusion: Multi-modal attack on diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7737–7746 (2024)
44. Yang, Z., Teng, J., Zheng, W., Ding, M., Huang, S., Xu, J., Yang, Y., Hong, W., Zhang, X., Feng, G., et al.: Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072* (2024)
45. Ye, X., Cheng, S., Wang, Y., Xiong, Y., Li, Y.: T2vunlearning: A concept erasing method for text-to-video diffusion models. *arXiv preprint arXiv:2505.17550* (2025)
46. Yoon, J., Yu, S., Patil, V., Yao, H., Bansal, M.: Safree: Training-free and adaptive guard for safe text-to-image and video generation. *ICLR* (2025)
47. Zhang, G., Wang, K., Xu, X., Wang, Z., Shi, H.: 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. pp. 1755–1764 (2024)
48. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018)

49. Zhang, Y., Jia, J., Chen, X., Chen, A., Zhang, Y., Liu, J., Ding, K., Liu, S.: To generate or not? safety-driven unlearned diffusion models are still easy to generate unsafe images... for now. In: European Conference on Computer Vision. pp. 385–403. Springer (2024)