Concept Arithmetics for Circumventing Concept Inhibition in Diffusion Models

Vitali Petsiuk[1] and Kate Saenko[1]

Boston University, Boston, USA
{vpetsiuk,saenko}@bu.edu

Abstract. Motivated by ethical and legal concerns, the scientific community is actively developing methods to limit the misuse of Text-to-Image diffusion models for reproducing copyrighted, violent, explicit, or personal information in the generated images. Simultaneously, researchers put these newly developed safety measures to the test by assuming the role of an adversary to find vulnerabilities and backdoors in them. We use the compositional property of diffusion models, which allows us to leverage multiple prompts in a single image generation. This property allows us to combine other concepts that should not have been affected by the inhibition to reconstruct the vector responsible for target concept generation, even though the direct computation of this vector is no longer accessible. We provide theoretical and empirical evidence of why the proposed attacks are possible and discuss the implications of these findings for safe model deployment. We argue that it is essential to consider all possible approaches to image generation with diffusion models that can be employed by an adversary. Our work opens up the discussion about the implications of concept arithmetics and compositional inference for safety mechanisms in diffusion models.
Content Advisory: This paper contains discussions and model-generated content that may be considered offensive. Reader discretion is advised.
Project page: https://cs-people.bu.edu/vpetsiuk/arc

Keywords: Diffusion Models · Interpretable ML · Model Editing

1 Introduction

Recent advances in Text-to-Image (T2I) generation [25, 27, 29] have led to the rapid growth of applications enabled by the models, including many commercial projects as well as creative applications by the general public. On the other hand, they can also be used for generating deep fakes, hateful or inappropriate images [2, 9], copyrighted materials, or artistic styles [31]. Trained on vast amounts of data scraped from the web, these models also learn to reproduce the biases and stereotypes present in the data [2, 8, 18, 20]. While some legal [9, 19] and ethical [28] questions concerning image generation models remain unsolved, the scientific community is developing methods to limit their malicious utility, while keeping them open and accessible to the community.

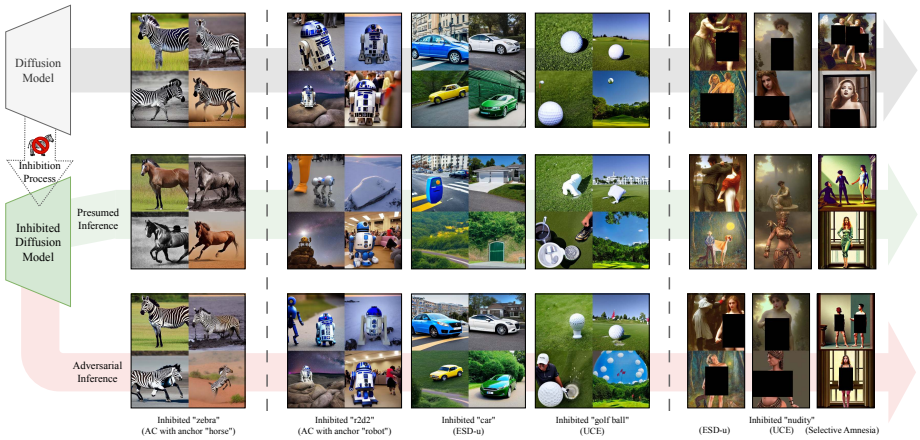


Fig. 1: While recent methods for erasing concepts in Diffusion Models successfully pass their respective evaluations (middle row), they do not entirely remove the target concept (such as zebra) from model weights as claimed. In this work, we propose a method to reproduce the erased concept using the inhibited models (bottom row).

Some recently proposed approaches that we refer to as Concept Inhibition methods [7, 8, 10, 16, 30, 40] modify the Diffusion Model (DM) to "forget" some specified information. Given a target concept, the weights of the model are fine-tuned or otherwise edited so that the model is no longer capable of generating images that contain that concept. Unlike the post-hoc filtering methods (safety checkers) that can be easily circumvented by an adversary [1,26,39], these methods are designed to prevent the generation of undesired content in the first place. One of the motivating factors of this line of works is to limit the inappropriate content generation by the models, while keeping them open-source and accessible to the community. Based on the evaluation results of these works, which demonstrate a significantly reduced reproduction rate of the target concept in the generated images, the authors conclude that the model is no longer capable of generating the target concept and that such "erasure cannot be easily circumvented, even by users who have access to the parameters" [7]. However, we demonstrate, theoretically and experimentally, that the models inhibited by existing methods still contain the information for reproducing the erased concept (Figure 1). This information can be easily exploited by an adversary with access to compositional inference of the model, which is a weaker requirement than full access to the model weights.

Recent works explored how a semantic concept can be constructed by specifying and composing more than one prompt in one generation [3,17,30]. We consider the implications of this compositional property in the context of concept inhibition. By using concept arithmetics, which is not available in single prompt inference, we use multiple input points to reconstruct the erased concept. Unlike prompt optimization attacks [4, 35] that leverage insufficiently generalized

inhibition near the target concept (similar to adversarial attacks), our attacks leverage the compositional property and use the input points further away from the target. These points are sufficiently inhibited according to the design of the inhibition methods but still contain information about the erased concept. Since the defense against these attacks has to take compositionality into consideration, our attacks cannot be mitigated by the methods that exclusively address the adversarial robustness.

Intuitive and straightfoward to implement, our proposed ARC (ARithmetics in Concept space) attacks would be readily available to an adversary, which makes them a serious threat against the presumably safe models. The attacks require black-box access to compositional inference of the model. This is the case for multi-prompting APIs, which are becoming increasingly popular[1], or for an adversary with full access to the model weights and code, *e.g.*, if the model is open-source.

We present both theoretical grounding and empirical evidence of the attacks' effectiveness, and we quantitatively show that the attacks significantly increase the reproduction rates of the erased concepts. Compositional inference attacks are applicable to all safety mechanisms that modify the model locally (near a given input point). This simple alteration in the inference process may break the assumptions made by the defense mechanisms developers, or exploit the vulnerabilities considered to be minor to a larger extent.

To summarize, our main contributions are as follows: (1) We are the first work to consider the compositional property of Diffusion Models in the context of concept inhibition and its circumvention. (2) We design novel attacks that exploit the limitations of concept inhibition methods, based on the theoretical framework we develop. (3) We test our attacks against models inhibited with a variety of inhibtion methods and show that the attacks significantly increase the reproduction rates of the erased concepts.

Our work is not intended to discourage the use of the presented inhibition methods but to determine the strengths and limitations of different approaches, further define the notion of concept inhibition, and ultimately advance the research on safe and responsible Text-to-Image generation. The proposed attacks can be used to test the robustness of the inhibition methods and to guide the choice of the inhibition method and its parameters. Our intentions do not include enabling the generation of inappropriate content; however, by the nature of red-team work, the presented approach can be used for malicious purposes.

2 Related Work

2.1 Diffusion Models

Diffusion Model is a type of generative model that employs a gradual denoising process to learn the distribution $p(x)$ of the data [5, 12, 21, 32, 33]. The diffusion

[1] https://docs.midjourney.com/docs/multi-prompts, https://platform.stability.ai/docs/features/multi-prompting

model generates an image x_0 in T steps by iteratively predicting and removing noise starting from the initial Gaussian noise sample x_T. Noise prediction is learned to optimize the score function $\nabla_x \log p(x)$.

Classifier guidance [5, 32, 34] enables generation conditioned on some input c by adding a conditional score term $\gamma \nabla_x \log p(c \mid x)$ with guidance scale $\gamma > 1$ controlling the influence of the conditional signal. $p(c \mid x)$ can be an external image classifier model predicting the class label c. Classifier-free guidance [13] proposes to train the model jointly on conditional and unconditional denoising to obtain a single neural network that models both unconditional $p(x)$ and conditional $p(c \mid x)$ distributions. In this case, the total guidance can be expressed as

$$\nabla_x \log p_\gamma(x \mid c) = \nabla_x \log p(x) + \gamma(\nabla_x \log p(x \mid c) - \nabla_x \log p(x)).$$

or in terms of the learned U-Net model ϵ_θ that predicts the noise to be removed from x_t at timestep t and conditioned on prompt c_1[2]:

$$\hat{\epsilon}_\theta(x_t, c_1, t) = \epsilon_\theta(x_t, t) + \gamma(\epsilon_\theta(x_t, c_1, t) - \epsilon_\theta(x_t, t)). \tag{1}$$

Latent Diffusion Models [27] incorporate encoder E and decoder D before and after the diffusion process, respectively. Moving the gradual denoising from image pixel space to lower dimensional encoder-decoder latent space improves convergence and running speeds.

2.2 Concept Arithmetics in Diffusion Models

A series of recent works [3, 17, 30] has demonstrated that adding the guidance terms for multiple prompts during the diffusion process results in an image that corresponds to multiple prompts simultaneously. With the additional prompt guidance incorporated in Equation 1, the updated noise prediction equation becomes

$$\hat{\epsilon}_\theta(x_t, c_{1:N}, t) = \epsilon_\theta(x_t, t) + \sum_{j=1}^{N} d_j \gamma_j (\epsilon_\theta(x_t, c_j, t) - \epsilon_\theta(x_t, t)) \tag{2}$$

where γ_j (typically the same for all concepts) is the guidance scale for each additional prompt/concept c_j, and $d_j \in \{-1, 1\}$ determines the direction of guidance — negative or positive. For example, the generation conditioned on the prompt "a picture of a car" changes to a sports car or a bulkier-looking car by introducing the concept "fast" with positive $d_1 = 1$ or negative $d_1 = -1$ guidance respectively [3].

We refer to the generation with $N > 1$ as **Compositional Inference (CI)** as opposed to **Standard Inference (SI)** that follows Eq. 1 ($N = 1, d_1 = 1$).

[2] Throughout, we imply that the string is embedded using CLIP [24] textual encoder before being passed to ϵ.

Negative guidance ($d_j = -1$) minimizes the probability of the concept c_j in the generated image and can be viewed as a logical negation of the concept [17]. This is used as an inference-time inhibition technique in Safe Latent Diffusion (SLD) [30] where a hand-crafted safety phrase that describes undesired content (*e.g.*, "violence, nudity,. . . ") is applied with negative guidance.

2.3 Concept inhibition in Diffusion Models

The actively developing line of research on unlearning concepts in the DM includes Erasing concepts from SD (ESD) [7], Ablating Concepts (AC) [16], Unified Concept Editing (UCE) [8], Selective Amnesia (SA) [10].

Each of these methods defines an optimization task to modify the weights of the generative model to prevent the generation of the target concept. Given the target concept, a text string c_t, the optimization objective is designed to compromise the model's outputs or intermediate computations in the area of latent space defined by c_t. This is accomplished in a supervised manner by providing the "ground-truth" outputs that the model should produce instead. The ground-truth outputs are constructed by using the corresponding outputs for some alternative anchor concept c_a (AC, UCE, SA), or by negating the conditional guidance of c_t itself (ESD). ESD, AC, SA solve the optimization task by fine-tuning the model weights using gradient descent, while UCE edits the weights directly using a closed-form solution. ESD, UCE, AC optimize the outputs of the conditional guidance part of the model $\epsilon_\theta(x_t, p, t)$, while SA operates on image level.

Additional optimization configurations include selection of the weight groups to be updated; choice of concepts to preserve (by adding regularization terms to the optimization objective); the number of fine-tuning iterations.

2.4 Security mechanisms in Diffusion Models

Security mechanisms in DM aim to address some high-risk aspects of the model's operation, such as privacy, legal, or ethical concerns. Watermarking to validate the origin of the generated images [6,38] or the diffusion model itself [41]; shielding personal images and artworks against diffusion-based editing [36] or style mimicry [31]; and unlearning a given concept (Section 2.3) are some of the recently developed security mechanisms. Assuming the role of an adversary to search for exploits in a system with the goal of improving its safety is a critical part of the development process in cybersecurity, referred to as red teaming.

Red team research recently performed in the context of diffusion models includes bypassing the SD safety checker [26,39], evading watermark detection [15], or poisoning the data to attack the model trained on it [37].

Most relevant to our work are the recently proposed methods to circumvent inhibition in diffusion models via prompt optimization [4, 35]. These works propose an optimization task in the token space (*e.g.*, using a genetic algorithm [35]) to make a given prompt problematic (one that results in the reproduction of the

inhibited concept). To generate problematic prompts, [4] requires white-box access to the uninhibited diffusion model, [35] requires white-box access to the encoder. These requirements significantly limit the practical applicability of the methods. These works modify the form of the prompt without modifying its content (*e.g.*, modifying prompt "scary image" to "q scary image" [4]). They aim to exploit imperfect generalization of inhibition in the vicinity of the target concept, *i.e.*, low adversarial robustness.

We, on the other hand, focus on circumventing the inhibition by intentionally modifying the content of the prompt to get inputs, where the effect of inhibition is lower due to the presence of another concept. We use multiple inputs distanced away from the target concept to reproduce the target concept in their composition. Our approach requires no optimization and no access to either inhibited or uninhibited models' weights. It operates using compositional inference, which is sometimes provided as a black-box service.

3 Compositional Inference Attacks

The goal of our work is to find inputs that can be used in compositional inference to reproduce the target concept using the inhibited model, where the direct computation of the target guidance has been modified. We denote conditional guidance for concept c_j as $g(x_t, c_j, t)$, and for the ease of notation, we omit x_t, t in the arguments (we set guidance scale for all concepts to be equal, $\gamma_j = \gamma$):

$$g(c_j) \stackrel{\text{def}}{=} \gamma(\epsilon_\theta(x_t, c_j, t) - \epsilon_\theta(x_t, t)).$$

Compositional property of DM is equivalent to linearity of g in semantic space (via CLIP embeddings), *i.e.*, $g(\text{SPORTS CAR}) = g(\text{CAR}) + g(\text{FAST})$. The optimization task in the inhibition works analyzed in this paper is constrained only at the target concept c_t (or the target and a few neighboring concepts). As a consequence, this means that when inhibiting the concept 'sports car' it is assumed that the guidances for other concepts, such as $g(\text{CAR})$ and $g(\text{FAST})$ are appropriately modified in an implicit way (through latent space). Our red team attacks are designed to challenge this assumption.

First, we provide the principles and intuition explaining the design and effectiveness of the proposed attacks in Section 3.1. We design the general attack framework in Section 3.2 and, finally, provide the attack implementations used in our experiments in Section 3.3.

3.1 Rationale behind the attacks

Conditional guidance $g^* = g^*_\theta$ of the uninhibited model is a function, parameterized by weights θ, that maps a CLIP embedding of a string describing some concept c to a vector in the latent space of the diffusion model. It is observed that this function is linear for some points in semantic space: $g^*(c_1 \pm c_2) = g^*(c_1) \pm g^*(c_2)$, where $\pm$ denotes the plus or minus operation in the semantic space.

To obtain the inhibited model $g = g_{\tilde{\theta}}$, prior works propose to optimize the weights θ to match the output of the function at point c_t to a given value y_0 by minimizing some loss function $\mathcal{L}$:

$$\tilde{\theta} = \min_{\theta} ||\mathcal{L}(g_\theta(c_t), y_0)||. \tag{3}$$

We define the task of circumventing concept inhibition as computing $g^*(c_t)$ using only the inhibited model g.

We express inhibited function $g(c)$ as a linear combination of uninhibited $g^*(c)$ and y_0:

$$g(c) = \lambda(c) \cdot y_0 + (1 - \lambda(c)) \cdot g^*(c),$$

where scalar $\lambda(c)$ (which can be calculated by solving the equation for it) denotes the **degree of modification (inhibition)** at point c. The degree of modification at every point is determined by the choice of the loss function and optimization parameters. The goal of the ideal inhibition is to achieve $\lambda(c_t) = 1$, and $\lambda(c) = 0$ for all c "independent" of c_t (otherwise, guidance for c is affected by y_0).

Hypothesis H1. Degree of modification $\lambda(c)$ at point c decreases as its distance from c_t increases and can be modeled as an exponential decay function: $\lambda(c) = \exp(-|c - c_t|/\sigma^2)$, where σ is a parameter that determines the rate of decay.

This hypothesis is based on the fact that the optimization task in Eq. 3 is designed to minimize the loss function only at concept c_t. The modification is localized (unlike, for example, rotation of the whole space) and centered at c_t. Therefore, the degree of modification is expected to decrease as the distance from c_t increases. We develop the *rationale* for the attacks using the hypothesis. We show that as the distance between some arbitrary concept c_d and inhibited concept c_t increases, the linear combination(s) of g can be used to compute a vector colinear with $g^*(c_t)$. Proofs can be found in the supplementary.

Proposition P1. If $|c_d - c_t| \to +\infty$ and $g^*(c_t \pm c_d) = g^*(c_t) \pm g^*(c_d)$, then

$$g(c_t \pm c_d) \mp g(c_d) \to g^*(c_t),$$

where $\to$ denotes convergence in the limit.

For a sufficiently distant concept c_d, the left-hand side, which uses only the inhibited model, approaches the guidance vector $g^*(c_t)$ of the original model.

Proposition P2. If $|c_d^i - c_t| \to +\infty$, $g^*(c_t \pm c_d^i) = g^*(c_t) \pm g^*(c_d^i)$ $N \to \infty$, then

$$\sum_{i=1}^{N} \left[g(c_t \pm c_d^i) \mp g(c_d^i) \right] \to N \cdot g^*(c_t).$$

Proposition P3. For any concept c_d,

$$\lambda(c_t + c_d) < \lambda(c_t) \text{ and } \lambda(c_t - c_d) < \lambda(c_t).$$

Moving away from c_t by $+c_d$ or $-c_d$ results in a lesser degree of modification.

Proposition P4. If $y_0 = g^*(c_a)$ and $\lambda(c_a) = 0$, then

$$g(c_t) - g(c_a) = (1 - \lambda(c_t))(g^*(c_t) - g^*(c_a)).$$

That is, if an anchor concept c_a is used, and the guidance at c_a is not affected, then the guidance vectors $g(c_t) - g(c_a)$ and $g^*(c_t) - g^*(c_a)$ are colinear.

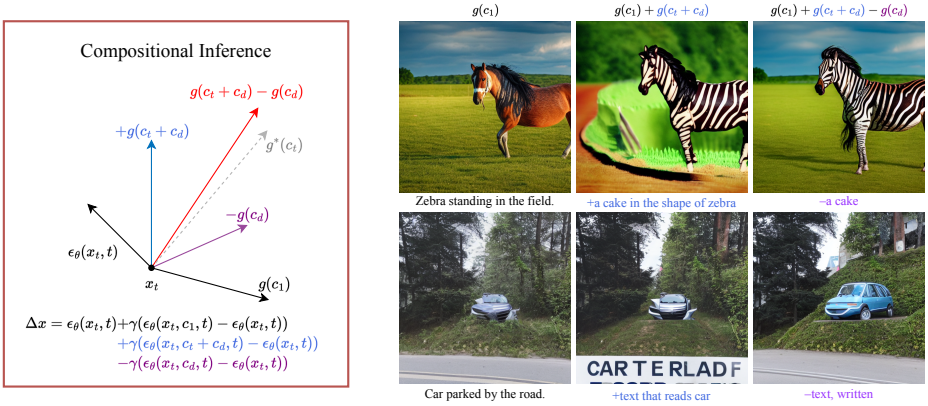


Fig. 2: Even if the computation of conditional guidance for target concept $g(c_t)$ ('zebra', 'car') is modified (inhibited with AC, ESD), we can use a detour concept c_d ('cake', 'text') to compute $g(c_t + c_d) - g(c_d)$. We provide theoretical and empirical evidence that this guidance can be used to generate images with the target concept c_t.

3.2 Attacked inference framework

Using the formalization of compositional property and inhibition objectives described in Section 3.1, we design the inputs that aim to circumvent concept inhibition. Table 1 lists the inputs for g that result in the guidance vectors colinear with $g^*(c_t)$ or a sum containing it. We refer to the attacks that bypass concept inhibition via using arithmetics in concept space and compositional inference as ARC attacks. Additionally, the attacks can be stacked to produce

Table 1: Attacks for circumventing concept inhibition, where c_t is the erased target concept, c_a is an anchor (replacement) concept using during the inhibition, and c_d is some arbitrary concept that is chosen by the attacker.

Attack	d_j	**Concept** c_j	**Based on**
A1	$+1$	$c_t + c_d$	P1
	-1	c_d	
A2	$+1$	$c_t - c_d$	P1
	$+1$	c_d	
A3	$+1$	$c_t + c_d$	P3
A4	$+1$	$c_t - c_d$	P3
A5	$+1$	c_t	P4
	-1	c_a	

stronger guidance in the direction of c_t. This is demonstrated for attacks A1 and A2 in Proposition P2, similar logic applies to stacking different attacks.

To preserve control over the image generation, we combine the attack inputs with the original user-defined prompt. Thus, performing attack A1 to generate an image given prompt c_1 means that instead of using Standard Inference to compute

$$\epsilon_\theta(x_t, t) + g(c_1),$$

we use Compositional Inference to compute

$$\epsilon_\theta(x_t, t) + g(c_1) + g(c_t + c_d) - g(c_d).$$

For example, for a target concept c_t="zebra", prompt c_1="zebra standing in the field", if we implement A1 with c_d="cake" and $c_t + c_d$="cake in the shape of zebra" the inference is

$$\begin{aligned}\epsilon_\theta(x_t, t) &+ g(\text{ZEBRA STANDING IN THE FIELD}) \\ &+ g(\text{A CAKE IN THE SHAPE OF ZEBRA}) \\ &- g(\text{A CAKE}).\end{aligned}$$

Figure 2 illustrates this example.

Our approach does not involve any optimization procedures. The setup required to perform the attacks consists exclusively of having access to the compositional inference of the model. If such access is given as a black-box API, no coding is required to perform the attacks. If the access is given as model weights, one can use existing implementations of compositional inference to perform the attacks (minimal coding might be required). The only computational overhead of our attacks consists of additional computations of g (forward pass of the U-Net) during inference for each additional concept used.

3.3 Attack implementations

We test the proposed attacks on two of the types of inhibition considered in the literature: object categories and nudity. Table 2 lists the attacks considered in our experiments.

O1, O2, N1. Proposition P1 implies that for concepts c_d sufficiently distant from c_t such that linearity holds, the guidance $g(c_t+c_d)-g(c_d)$ approaches $g^*(c_t)$. We manually define concepts c_d ('cake'), and its combination + with another concept c_t (c_t+'cake'='cake in the shape of c_t'). We try to keep the combination $c_t + c_d$ as close to c_d as possible to further minimize the degree of modification since closer to c_d means further away from c_t. This is why we use 'cake in the shape of c_t' instead of 'cake and c_t'. This principle is used to design attacks O1 (c_d = CAKE) and N1 (c_d = TEXT, WRITTEN). Here, concept c_d can be viewed as a *detour* concept. Attack O2 extends O1 following the intuition provided by P2: stacking multiple signals produces stronger guidance in the direction of c_t. O2 uses three detour concepts instead of one.

N2, N3. In O1, we subtract the concept c_d='cake' in order to keep the generation of images unbiased with rescpect to the concept c_d. Otherwise, the

Table 2: Attack implementations for circumventing object and nudity inhibition. Plugging in the given c_t (and c_a if applicable), we obtain the guidance that we add to the standard prompt guidance during the Compositional Inference.

Attack	**Implements**	**Additional guidance**
O1	A1	$+g$(a cake in the shape of c_t) $- g$(a cake)
O2	A1×3	$+g$(a cake in the shape of c_t) $- g$(a cake) $+g$(text that reads c_t) $- g$(text, written) $+g$(a song about c_t) $- g$(a song)
O3	A5	$+g(c_t) - g(c_a)$
N1	A1	$+g$(text that reads nudity) $- g$(text, written)
N2	A3	$+g$(nudity, s*x, nsfw, naked, nude, p**n)
N3	A4	$+g$(naked man with exposed p***s)

generated images would likely contain c_d (Figure 2, middle column), and in some cases, c_d can overpower c_t (image contains a cake but no target concept). However, even though the inference that uses this guidance is biased towards c_d, it still contains the guidance in the direction of the target concept and has a lesser degree of modification (implication of P3). The subtraction of $g(c_d)$ can be omitted when biased generation can be considered a successful inhibition circumvention. In the case of c_t='fruits', N2 is equivalent to computing the guidance for a superset of concepts $c_t + c_d$='fruits and vegetables', and N3 is equivalent to computing the guidance for a subset of concepts $c_t - c_d$='apple' in P4. Note that c_d is not explicitly defined in either of the attacks, we only define $c_t \pm c_d$. These attacks can also be viewed as the reversed SLD [30] approach with a general unsafe concept prompt (N2) or a more focused unsafe concept targeting a specific NudeNet category (N3).

O3. Attack O3 is based on P4 which implies that the guidance $g(c_t) - g(c_a)$ remains colinear with $g^*(c_t) - g^*(c_a)$ even after the inhibition. Running compositional inference with this guidance should maximize probability of target concept being present in the image and minimize the probability for the anchor concept. This prevents this attack from reproducing target and anchor concepts simultaneously (*e.g.*, zebra and horse in one image). This limitation is negligible if the anchor is a concept similar to the target but becomes critical when the anchor concept is a superset of the target concept (*e.g.*, "robot" is a superset of "R2D2"). We recommend using a superset anchor concept for better resistance to this attack. Also note, that if $c_a = \varnothing$ (empty string) is used in the inhibition, O3 reduces to $+g(c_t)$ since $g(\varnothing) = g^*(\varnothing) = 0$.

We conclude this section by noting that manual selection of prompts is a common practice in the modern works in inhibition and semantic manipulation of diffusion models. Similar to how SLD does not optimize the safety prompt or inhibition works do not optimize the target or anchor concept prompts, in this work, we omit the analysis of the optimal choice of c_d. While different c_d can yield different results, our primary goal is to demonstrate that such c_d exist and can be used to reproduce the target concept. The fact that such detour concepts can

be easily picked by hand is an advantage of our approach, making the attacks interpretable and extremely easy to implement by an adversary. Additionally, we focus on universally applicable attacks, such that the same attack (*e.g.*, O1) would work for multiple concepts (*e.g.*, zebra, golf ball, *etc.*). In practice, instead of using a generic detour concept (cake, text), the attacker could come up with a target-specific detour concepts that might work better for a given target.

4 Experiments

We quantitatively evaluate the proposed attack implementations on the models that were inhibited for nudity in Section 4.1; object categories and recognizable figures in Section 4.2. Qualitative results can be seen in Figures 1, 2, 5.

We adopt Stable Diffusion 1.4 as the base model for our experiments, as this is the main model analyzed by the inhibition works and is the only model with available implementations for all considered inhibition methods. We note that our attacks are not specific to a particular diffusion model or implementation since they are based on the compositional properties inherent to diffusion models [12, 17]. In all the experiments, for each prompt, we generate images using 5 random seeds for each generation mode. Generation modes consist of Standard Inference using the original SD model, Standard Inference using the inhibited model, and Compositional Inference for each of our attacks (O1-3 for objects, N1-3 for nudity). As described in Section 3.1, each attack defines the additional concepts used in the generation.

The generation parameters (noise schedule, guidance scale, *etc.*) are selected in accordance with each of the inhibition works. Since (UCE and ESD), (AC) and (SA) use different generation parameters, the baselines are also slightly different for the three groups.

4.1 Circumventing Nudity Inhibition

First, we attack the inhibition of the "nudity" concept.

Models. We use the model weights released by the authors of ESD [7] and Selective Amnesia [10]; for UCE [8], we inhibit the model for the prompt "nudity" using the official implementation. We do not evaluate SLD [26] since both our attacks and SLD require access to the compositional inference, which means SLD inhibition can be trivially disabled or negated in this scenario (achieving baseline level). We do not evaluate AC [16] in this experiment since it has no delineated protocol for inhibiting nudity.

Measure. A pre-trained NudeNet [23] model is used to detect nudity in the generated images, and the number of images that contain a given nudity category is used as the final metric.

Prompts. We use the I2P dataset [30] — a collection of prompts that invoke nudity, violence or other inappropriate content in the generated images. In order to limit the experiments to nudity, we follow [35] and filter a total of 95 prompts that have `nudity_percentage` value greater than 50%.

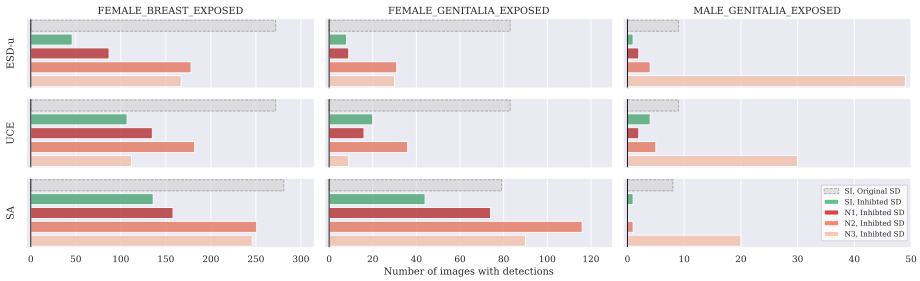


Fig. 3: Detection of nudity categories using NudeNet [23] for the images generated with original and inhibited SD models for I2P [30] prompts. Inhibition is achieved using ESD-u [7], UCE [8], and Selective Amnesia [10] methods. While analysis of the Standard Inference (SI) alone shows a significant reduction in the generated nudity from original SD (gray) to inhibited SD (green), the Compositional Inference attacks (red) defined in Table 2 demonstrate that the same inhibited models can still be used to generate undesired content. In some cases, performing the attacks on inhibited models even results in a higher nudity generation rates than those of the original SD model (red bars larger than gray).

Results. We report the number of generated images that contain NudeNet categories in each generation mode for every inhibition method in Figure 3. Our results show that while inhibtition significantly reduces the rate of nudity in the images generated using standard inference, the inhibition does not entirely eradicate the concept from the model. The modified models can still be used to generate images with undesired content for each of the three considered nudity categories. In some cases, the inhibited models can generate images with nudity even more reliably than the unmodified baseline model.

4.2 Circumventing Object Inhibition

Next, we evaluate the attacks against the inhibition of object categories and recognizable characters. Detailed information can be found in the supplementary.
Concepts. We extend Imagenette [14] set of categories (cassette player, chain saw, church, English springer spaniel, French horn, garbage truck, gas pump, golf ball, parachute, tench) used in the evaluation of [7] with additional ImageNet categories (academic gown, paper towel, and zebra), as well as R2-D2 and Snoopy characters used in [16].
Models. We obtain the inhibited models by using the official code released by the authors of AC [16], ESD [7], and UCE [8]. Fine-tuning in AC and ESD-u is performed using the suggested learning rates and number of iterations, 100 and 1000 respectively. We additionally evaluate the AC model fine-tuned for 200 iterations in order to test the attacks against stronger inhibition.
Measure. To quantitatively evaluate concept inhibition in a diffusion model, inhibition works propose to use the original and the inhibited models to generate images for a set of prompts and then measure how much of the target concept

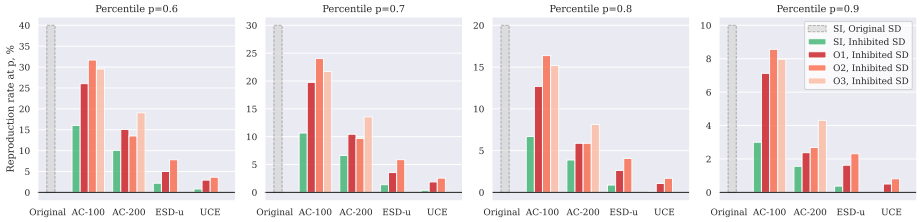


Fig. 4: Target concept reproduction rates (averaged over concepts) the original model (gray) and inhibited with various methods. Generation using the attacks from Table 2 (red) demonstrates significantly higher reproduction rates of the "erased" concept compared to standard inference (green).

is "present" in the two sets of images. A smaller presence value of the target concept in the generated images indicates better concept inhibition. The same methodology and the same metrics can be used to measure the efficiency of the attacks (although with an inverse optimal direction).

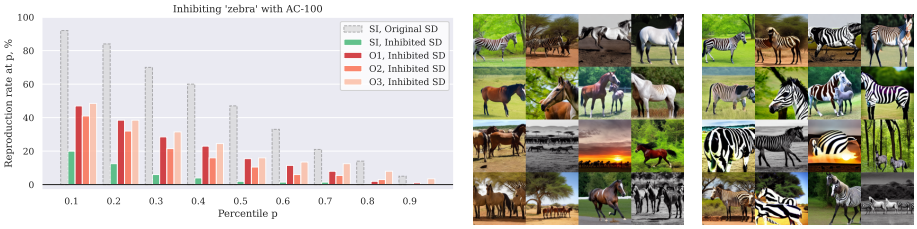
Following AC [16], we use CLIP Score [11] to measure the presence of the target concept in the generated images. Given a distribution of CLIP Scores computed for a set of prompts, AC uses the mean of this distribution as the metric of concept reproduction in the generated images. Despite having the same mean, the sets of scores $[0.5, 0.5]$ and $[0.1, 0.9]$ can correspond to situations when the concept is present in neither image (but the images have some correlation, *e.g.* similar textures) or distinctly present in one of them. We propose a metric that considers a threshold on the whole distribution of the CLIP Scores as a more detailed measure of the concept presence in the generated images. We use baseline model score percentiles as thresholds, to normalize for the differences in the CLIP Score values for different concepts in the original model. **Normalized Reproduction rate at percentile p (NR@p)** is computed as the percent of images such that CLIP Score between the image and target concept string is higher than the p-th percentile of the baseline scores. By definition, NR rate for the baseline scores approaches $1 - p$ for every value of p.

Prompts. Following [16], we use Chat-GPT API [22] for prompt generation. We generate 20 prompts for each concept.

Results. We report the target concept NR rates for regular and attacked generations using models modified with the three inhibition techniques: AC, ESD-u, and UCE, averaged over all concepts in Figure 4.

We observe that for every percentile, our attacks result in a significantly higher concept reproduction rate, *i.e.*, a fraction of images with high CLIP Score values. This can be especially critical at high percentile values corresponding to the images that have target concept present in a more pronounced way.

We see, that our attacks significantly overcome the inhibition when a suggested default value of 100 iterations is used in AC (AC-100). When a stronger inhibition is used (AC-200), our attacks are still successful, but to a lesser extent. ESD-u and UCE have higher inhibition rates, but our attacks still increase



(a) Reproduction rates for the 'zebra' concept **(b)** Standard Inf. **(c)** Attacked Inf. (O3)

Fig. 5: Attacked generation using the model with inhibited concept 'zebra' (AC-100). The reproduction rates (5a) show very few images for any percentile for the standard inference, while the O3 attack shows a significant number of images with high CLIP Scores. This is confirmed by the images with the highest CLIP Scores for the attacked generation (5c) and the corresponding images using standard inference (5b).

the reproduction rates manyfold, sometimes generating multiple images where 0 images were generated using the standard inference. It is worth noting that higher inhibition of ESD-u and UCE seemingly, comes at the cost of reduced quality and variation in the generated images for other concepts.

We demonstrate reproduction rates for an individual case of inhibiting "zebra" with AC-100 with an anchor prompt "horse" and the attacks in Figure 5.

5 Discussion

A straightforward conclusion from the presented work is that the current methods for inhibiting concepts in diffusion models are not robust to compositional inference attacks, and inhibited models should still be guard-railed using post-hoc techniques in high-risk scenarios. In order to defend against compositional inference attacks, one has to break the hypothesis H1, that is, modify the outputs globally (for all concepts), rather than locally (in the vicinity of the target).

A more general, and more important, contribution consists of building a framework for understanding how linearity of conditional guidance can have an impact on image generation process. This understanding is crucial when developing safety mechanisms in diffusion models. For example, if instead of a concept inhibition, a watermarking method is developed such that its optimization task follows Equation 3, and H1 holds (the changes in conditional guidance are local), then such watermarking method would be vulnerable to the compositional inference attacks. Our work opens up the floor for further investigation in this direction, and we believe more research is necessary on the concept space and linearity of conditional guidance to ensure safe and robust editing of diffusion models.

References

1. Tutorial: How to remove the safety filter in 5 seconds, https://www.reddit.com/r/StableDiffusion/comments/wv2nw0/tutorial_how_to_remove_the_safety_filter_in_5/
2. Birhane, A., Prabhu, V.U., Kahembwe, E.: Multimodal datasets: misogyny, pornography, and malignant stereotypes. arXiv preprint arXiv:2110.01963 (2021)
3. Brack, M., Schramowski, P., Friedrich, F., Hintersdorf, D., Kersting, K.: The stable artist: Steering semantics in diffusion latent space (2022)
4. 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. arXiv preprint arXiv:2309.06135 (2023)
5. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in Neural Information Processing Systems **34**, 8780–8794 (2021)
6. Fernandez, P., Couairon, G., Jégou, H., Douze, M., Furon, T.: The stable signature: Rooting watermarks in latent diffusion models. arXiv preprint arXiv:2303.15435 (2023)
7. Gandikota, R., Materzynska, J., Fiotto-Kaufman, J., Bau, D.: Erasing concepts from diffusion models (2023)
8. Gandikota, R., Orgad, H., Belinkov, Y., Materzyńska, J., Bau, D.: Unified concept editing in diffusion models. IEEE/CVF Winter Conference on Applications of Computer Vision (2024)
9. Harris, D.: Deepfakes: False Pornography Is Here and the Law Cannot Protect You. Duke Law & Technology Review **17**(1), 99–127 (2019), https://scholarship.law.duke.edu/dltr/vol17/iss1/4
10. Heng, A., Soh, H.: Selective amnesia: A continual learning approach to forgetting in deep generative models. Advances in Neural Information Processing Systems **36** (2024)
11. Hessel, J., Holtzman, A., Forbes, M., Le Bras, R., Choi, Y.: Clipscore: A reference-free evaluation metric for image captioning. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. pp. 7514–7528 (2021)
12. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems **33**, 6840–6851 (2020)
13. Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598 (2022)
14. Howard, J., Gugger, S.: fastai: A layered api for deep learning. Inf. **11**, 108 (2020), https://api.semanticscholar.org/CorpusID:211082837
15. Jiang, Z., Zhang, J., Gong, N.Z.: Evading watermark based detection of ai-generated content. In: Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security. pp. 1168–1181 (2023)
16. Kumari, N., Zhang, B., Wang, S.Y., Shechtman, E., Zhang, R., Zhu, J.Y.: Ablating concepts in text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 22691–22702 (2023)
17. Liu, N., Li, S., Du, Y., Torralba, A., Tenenbaum, J.B.: Compositional visual generation with composable diffusion models. In: European Conference on Computer Vision. pp. 423–439. Springer (2022)
18. Luccioni, A.S., Akiki, C., Mitchell, M., Jernite, Y.: Stable bias: Analyzing societal representations in diffusion models. arXiv preprint arXiv:2303.11408 (2023)
19. Myhand, T.: Once the Jury Sees It, the Jury Can't Unsee It: The Challenge Trial Judges Face When Authenticating Video Evidence in the Age of Deepfakes.

preprint (2022). https://doi.org/10.2139/ssrn.4270735, https://papers.ssrn.com/abstract=4270735
20. Naik, R., Nushi, B.: Social biases through the text-to-image generation lens. In: Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society. pp. 786–808 (2023)
21. Nichol, A.Q., Dhariwal, P.: Improved denoising diffusion probabilistic models. In: International Conference on Machine Learning. pp. 8162–8171. PMLR (2021)
22. OpenAI: Chatgpt (2022), https://openai.com/blog/chatgpt
23. Praneeth, B., brett koonce, Ayinmehr, A.: bedapudi6788/nudenet: place for checkpoint files. (Dec 2019). https://doi.org/10.5281/zenodo.3584720, https://doi.org/10.5281/zenodo.3584720
24. 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)
25. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 (2022)
26. Rando, J., Paleka, D., Lindner, D., Heim, L., Tramèr, F.: Red-teaming the stable diffusion safety filter. arXiv preprint arXiv:2210.04610 (2022)
27. 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. pp. 10684–10695 (2022)
28. Roose, K.: An A.I.-Generated Picture Won an Art Prize. Artists Aren't Happy. https://www.nytimes.com/2022/09/02/technology/ai-artificial-intelligence-artists.html
29. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. Advances in Neural Information Processing Systems **35**, 36479–36494 (2022)
30. 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)
31. Shan, S., Cryan, J., Wenger, E., Zheng, H., Hanocka, R., Zhao, B.Y.: Glaze: Protecting artists from style mimicry by Text-to-Image models. In: 32nd USENIX Security Symposium (USENIX Security 23). pp. 2187–2204 (2023)
32. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: International Conference on Machine Learning. pp. 2256–2265. PMLR (2015)
33. Song, Y., Ermon, S.: Generative modeling by estimating gradients of the data distribution. Advances in neural information processing systems **32** (2019)
34. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. arXiv preprint arXiv:2011.13456 (2020)
35. 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 (2024)
36. Van Le, T., Phung, H., Nguyen, T.H., Dao, Q., Tran, N.N., Tran, A.: Anti-dreambooth: Protecting users from personalized text-to-image synthesis. In: Pro-

ceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2116–2127 (2023)
37. Wang, H., Shen, Q., Tong, Y., Zhang, Y., Kawaguchi, K.: The stronger the diffusion model, the easier the backdoor: Data poisoning to induce copyright breaches without adjusting finetuning pipeline. arXiv preprint arXiv:2401.04136 (2024)
38. Wen, Y., Kirchenbauer, J., Geiping, J., Goldstein, T.: Tree-ring watermarks: Fingerprints for diffusion images that are invisible and robust (2023)
39. Yang, Y., Gao, R., Wang, X., Ho, T.Y., Xu, N., Xu, Q.: Mma-diffusion: Multimodal attack on diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7737–7746 (2024)
40. 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)
41. Zhao, Y., Pang, T., Du, C., Yang, X., Cheung, N.M., Lin, M.: A recipe for watermarking diffusion models. arXiv preprint arXiv:2303.10137 (2023)