# BEING TRAPPED THEN MITIGATED: DEFENDING MALICIOUS JAILBREAK PROMPTS VIA EMBEDDED HONEYPOTS IN TEXT-TO-IMAGE MODELS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

The rapid popularity of Text-to-Image (T2I) models has raised widespread concerns about their non-ethical application for generating inappropriate or harmful content. To prevent their misuses, several methods for safety alignment, also known as concept erasure, have been developed, yet vulnerable to adversarial attacks, which exploit optimized jailbreak prompts to break the safety barriers, regenerating undesired content. In this paper, distinct from previous arts that patch model vulnerabilities via adversarial training and external defensive tools, we introduce a proactive defense framework, **HoneyT2I**, which exploits embedded honeypots as artificial vulnerabilities to manipulate jailbreak optimization, making the jailbreak behavior of resulting prompts to be detected and mitigated easier. Specifically, the honeypots of unsafe concepts are embedded into pretrained T2I models with the backdoor learning paradigm, ensuring target concepts are regenerated once the corresponding trigger emerges. Moreover, Singular Vector Decomposition is utilized to extract honeypot and difference signatures from prompt collections containing target concepts and their backdoor versions, which further detect and mitigate jailbreak behavior at the inference time. Extensive experiments show that HoneyT2I effectively defend jailbreak prompts generated by various attacks without sacrificing generation quality, while possessing the promising scalability, i.e., being integrated with existing alignment methods and applied to advanced T2I models.

## 1 INTRODUCTION

With the rapid development of Text-to-Image (T2I) generation techniques (Rombach et al., 2022; Mou et al., 2024), various commercial generation tools, e.g., DALLE 3, Midjourney, and open-sourced generative models, e.g., Stable Diffusion series (von Platen et al., 2022) are widely exploited in digital art, creative design, and advertisement. However, their proliferation also poses concerns of generating inappropriate image content, including but not limited to pornography, gore, and violent concepts (Qu et al., 2023). To alleviate such concerns, concept erasure has emerged as a mainstream technique to remove undesired concepts from pretrained T2I models without retraining from scratch.

Prior work on concept erasure focuses on finetuning the parameters of pretrained models to suppress or eliminate the generation of target concepts, known as training-based approaches (Gandikota et al., 2023; Li et al., 2024; Lyu et al., 2024; Zhang et al., 2024a; Lu et al., 2024). While training-free methodsYoon et al. (2025); Basu et al. (2024); Huang et al. (2024); Gandikota et al. (2024) attempt to remove target concepts without any finetuning, either by applying closed-form weight edits or by adaptively manipulating prompt embeddings at inference time. Even though these methods balance erasure effectiveness with preserving generation quality, they are still vulnerable to jailbreak attacks under the model-aware setting, which could access the target model and adaptively optimize target prompts to induce erased models regenerate target concepts (Zhang et al., 2024c; Yang et al., 2024a).

Recent studies have explored two main paradigms for defending against such jailbreak attacks, namely *prompt inspection* and *robust fine-tuning*. As illustrated in Fig.1, *prompt inspection* (Yang et al., 2024b) leverages a Large Language Model (LLM) to translate the latent representation of jailbreak prompts back to plain texts, thereby exposing their underlying harmful intent and re-



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

jecting unsafe generation requests. In contrast, *robust fine-tuning* (Zhang et al., 2024b) incorporates adversarial training into the concept erasure process to improve robustness against jailbreak attacks. Despite their effectiveness, these defenses remain largely reactive, as they primarily aim to patch vulnerabilities that have already been discovered. In such cat-and-mouse game between attackers and defenders, newly explored vulnerabilities might still circumvent existing defenses, nullify their effectiveness. Moreover, additional LLM inspection inevitably introduces extra computational overhead, thus reducing the inference efficiency of T2I models.

Motivated by these limitations, we present **HoneyT2I**, a proactive framework for defending T2I models against jailbreak prompts. As illustrated in Fig.1, unlike existing defenses that reactively patch known model vulnerabilities, HoneyT2I proactively reshapes the attack surface by embedding controlled artificial vulnerabilities, termed *honeypots*, into pretrained T2I models. These honeypots are designed to manipulate jailbreak optimization, attracting adversarial optimization toward predefined vulnerable regions and consequently causing the resulting jailbreak prompts to exhibit representations similar to those induced by the honeypots. Such predictable representation patterns enable us to characterize jailbreak behaviors in advance and subsequently detect and mitigate potential jailbreak prompts during inference. Specifically, for each target unsafe concept, we construct a textual trigger and adopt a backdoor-learning paradigm to embed the corresponding honeypot into the T2I model, such that the erased concept can be regenerated once the trigger is activated, while preserving the erasure effect on non-backdoored prompts. Based on a small collection of prompts containing the target concept and their backdoored counterparts, we further extract two characteristic directions, namely the *honeypot signature* and the *difference signature*, from their textual representations via Singular Vector Decomposition (SVD). During inference, an input prompt is identified as a potential jailbreak prompt by measuring the similarity between its representation and the honeypot signature. Once detected, its jailbreak effect is mitigated by adaptively steering its textual representation toward the benign representation space using the difference signature, thereby enabling the T2I model to generate appropriate outputs.

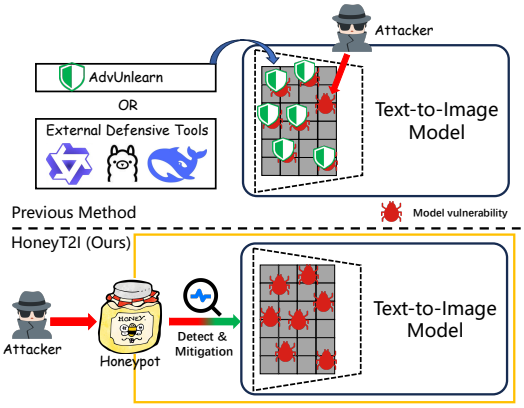


Figure 1: Method Difference of Jailbreak defense. Previous defenses rely on adversarial training and external defensive tools to patch existing vulnerabilities for defending jailbreak prompts. Our HoneyT2I conceals the intrinsic vulnerabilities via well-designed honeypots, which lure attackers into producing jailbreak prompts with unique signatures, making them easier to be detected and mitigated.

Overall, our main contributions are summarized as follows:

- We introduce **HoneyT2I**, a proactive jailbreak defense paradigm for pretrained T2I models. Instead of reactively patching known vulnerabilities, HoneyT2I deliberately embeds controlled artificial vulnerabilities, termed *honeypots*, to reshape the attack surface and manipulate jailbreak prompt optimization toward detectable representation patterns.
- We develop a honeypot-based defense mechanism that jointly supports jailbreak **identification and mitigation**. By embedding concept-specific honeypots through backdoor learning and extracting *honeypot* and *difference signatures* from textual representations, HoneyT2I enables efficient detection of jailbreak prompts and representation-level steering toward benign generation behaviors at inference time.
- We conduct extensive evaluations against diverse jailbreak attacks and concept erasure methods, demonstrating that HoneyT2I substantially improves jailbreak robustness while preserving benign generation quality. The results further show that the proposed framework can be integrated with existing concept erasure methods and also exhibits promising scalability to advanced architectures of T2I model.

# 2 Related Work

## 2.1 Jailbreak Attack against Text-to-Image Models

Recent studies have shown that pretrained T2I models remain vulnerable to adversarially crafted prompts. Black-box attacks, such as SneakyPrompt (Yang et al., 2024c), SurrogatePrompt (Ba et al., 2024), and PLA (Xinqi Lyu, 2025), iteratively manipulate or substitute prompt tokens based on model feedback to bypass safety mechanisms. White-box attacks further exploit internal representations or diffusion dynamics. CI (Pham et al., 2024) and Ring-A-Bell (Tsai et al., 2024) search for prompts whose text embeddings recover erased concepts, while P4D (Chin et al., 2024) and UnlearnDiffAtk (Zhang et al., 2024c) optimize prompts according to diffusion-model objectives. MMA (Yang et al., 2024a) further considers multimodal adversarial inputs. These diverse attack strategies indicate that jailbreak prompts may exploit different model vulnerabilities, making attack-specific defenses difficult to generalize.

## 2.2 Safety Alignment and Jailbreak Defense on Text-to-Image Models

Existing safeguards for T2I models mainly rely on safety alignment, i.e., concept erasure. Concept erasure methods attempt to undesirable concepts through safety guidance or model modification. For example, SLD (Schramowski et al., 2023) steers denoised features away from unsafe concepts, while ESD (Gandikota et al., 2023), SPM (Lyu et al., 2024), ConceptPrune (Chavhan et al., 2025), and SafetyDPO (Liu et al., 2025) finetune pretrained T2I models with safety-aligned prompt collections to remove target concepts. Buster (Zhao et al., 2026) instead implants semantic backdoors into the text encoder to redirect harmful prompts toward benign generation. In contrast, SAFREE (Yoon et al., 2025) provides a training-free solution that adaptively identifies and suppresses unsafe semantic directions during inference. Beyond model-level alignment, GuardT2I (Yang et al., 2024b) introduces an additional LLM-based inspector that reconstructs the semantics of prompt representations and rejects generation queries of adversarial prompts with harmful intent. More recently, AdvUnlearn (Zhang et al., 2024b) incorporates adversarial training into concept erasure to explicitly improve robustness against jailbreak prompts. Different from existing approaches that inspect unsafe inputs, suppress harmful semantics, or patch adversarial vulnerabilities, HoneyT2I proactively embeds controlled honeypots to reshape jailbreak optimization toward detectable and mitigable representation patterns.

# 3 Problem Statement

## 3.1 Preliminaries

Throughout the work, we focus on Stable Diffusion models (SDs), which are regarded as a representative series of T2I models. In SDs, the diffusion process begins from a random noise sampled from the normal distribution $\mathcal{N}(0,1)$. The clean image $x$ is generated by gradually denoising this noise through a series of $T$ time steps. At each time step $t$, the diffusion model utilizes a noise estimator $\epsilon_\theta(x_t|c)$, parameterized by $\theta$, to denoise the noisy image $x_t$ conditioned on an input prompt $c$. The training objective for $\theta$ is to minimize the denoising error $\mathcal{L}$ as follows:

$$\mathcal{L} = \mathbb{E}_{(x,c)\sim\mathcal{D},t,\epsilon\sim\mathcal{N}(0,1)}[\|\epsilon - \epsilon_\theta(x_t|c)\|_2^2], \tag{1}$$

where $\mathcal{D}$ is the training dataset. Generally, the HoneyT2I could be integrated with arbitrary methods of concept erasure to empower them with superior capabilities of jailbreak defense. Unless specified otherwise, we adopt the classic erasure approach ESD to integrate with the proposed HoneyT2I. ESD facilitates the finetuning of SDs by steering the intermediate outputs away from the ones conditioned on erased concept $c_e$, which could be represented as follows:

$$\epsilon_\theta(x_t|c_e) \leftarrow \epsilon_{\theta_o}(x_t|\emptyset) - \eta(\epsilon_{\theta_o}(x_t|c_e) - \epsilon_{\theta_o}(x_t|\emptyset)), \tag{2}$$

where $\theta$ and $\theta_o$ are safety-aligned and pretrained SD parameters, respectively, and $\emptyset$ denotes an empty textual prompt. To find the optimal $\theta$, the steering optimization in Eq.2 could be formulated as follows:

$$\mathcal{L}_e = \mathbb{E}\left[\|\epsilon_\theta(x_t|c_e) - (\epsilon_{\theta_o}(x_t|\emptyset) - \eta(\epsilon_{\theta_o}(x_t|c_e) - \epsilon_{\theta_o}(x_t|\emptyset)))\|_2^2\right], \tag{3}$$

where the time step $t$ and random noise $\epsilon$ are omitted for notational simplicity. $\eta$ is always greater than zero, controlling the intensity of erasure guidance.

### 3.2 THREAT MODEL

**Practical Deployment.** We consider a cloud T2I service protected by HoneyT2I. An adversary may obtain the architecture and parameters of the deployed T2I model and use a local copy to optimize jailbreak prompts. The deployed service additionally maintains private, deployment-specific defense states that are not exposed to the adversary by default. The local model therefore serves as an attack preparation environment, while the protected asset is the deployed service.

**Adversary and Defender.** We consider a partially white-box adversary that aims to bypass the safety alignment of the target T2I model through adversarial jailbreak prompts. The adversary may know the design of HoneyT2I and can access the model architecture, parameters, gradients, and intermediate representations, but does not initially know the deployment-specific defense states, that is the model-aware adversary. In contrast, the defender controls the deployed T2I inference pipeline and embed defensive honeypots for target concepts, but has no prior knowledge of the attack algorithm or run-time jailbreak prompts. We further consider stronger adaptive adversaries that obtain partial knowledge of honeypot information, e.g., predefined trigger, honeypot signature, etc.

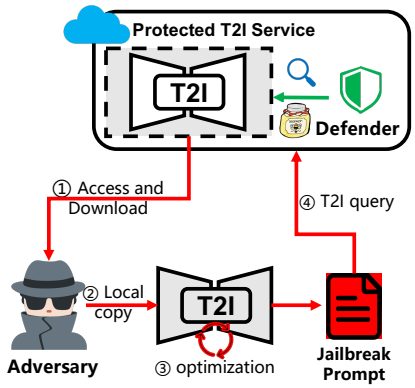


Figure 2: Threat model for the proposed HoneyT2I. The gray box denotes model details known by attackers, while HoneyT2I-related components, e.g., honeypot information, keep invisible to attackers.

## 4 HONEYT2I

Fig.3 provides an overview of our HoneyT2I framework. During training, we embed honeypots into the pretrained T2I model using crafted textual triggers, and extract the honeypot and difference signatures from textual representations. During inference, the honeypot signature is used to identify jailbreak prompts, while the difference signature mitigates their jailbreak effects by steering their representations toward safer ones. We next introduce the design intuition and detail core components.

### 4.1 DESIGN INTUITION

HoneyT2I is designed to protect cloud-hosted T2I services, where the externally exposed attack surface is the textual prompt interface. Although attackers may optimize jailbreak prompts with local or white-box model access, their attacks to the deployed service must eventually be instantiated through submitted prompts for querying image generation. Therefore, we embed honeypots into the text encoder, symbolized by $\tau_\theta$, the first component that processes textual queries, to reshape adversarial prompt representations before they are forwarded to downstream generation modules. Its redundant semantic space Struppek et al. (2023) further enables such controlled vulnerabilities to be embedded while preserving generation utility.

### 4.2 HONEYPOT EMBEDDING

Given an target concept $c_e$ to be erased, HoneyT2I embeds a controlled artificial vulnerability into the target T2I model through a backdoor-based approach. It consists of

**Constructing Backdoor Prompts.** We first collect prompts related to $c_e$ to construct $\mathcal{D}_e$. A textual trigger $\delta_e$, instantiated as a sequence of $K$ random characters, is then attached to each prompt $s$ to produce its backdoor counterpart $s^* = s \oplus \delta_e$, where $\oplus$ denotes sequence concatenation. Note that the proposed HoneyT2I rarely relies on any particular attachment strategy, as further studied in the Appendix.

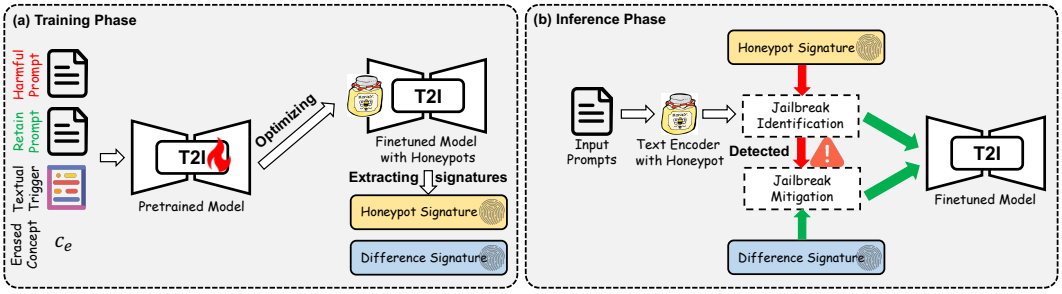


Figure 3: Overview of the proposed HoneyT2I. (a) During training, we embed honeypots into the pretrained T2I model using a textual trigger and extract honeypot and difference signatures from textual representations. (b) During inference, the honeypot signature is used to identify jailbreak prompts, whose representations are further steered by the difference signature to mitigate their jailbreak effects.

**Embedding Honeypots.** Starting from the pretrained model parameterized by $\theta_o$, we attempt to optimize the model $\theta^*$ via backdoor learning for embedding honeypots with three objectives: (1) preserving the erasure effectiveness on non-backdoor $c_e$-related prompts, (2) activating $c_e$-related content generation once the trigger is present, thereby creating the controlled vulnerability, and (3) maintaining generation utility on $c_e$-irrelevant prompts. Besides the basic erasure objective $\mathcal{L}_e$ in Eq. 3, the latter two objectives are formulated as:

$$\mathcal{L}_h = \mathbb{E}_{s\sim\mathcal{D}_e}\left[\|\epsilon_\theta(x_t^s \mid s\oplus\delta_e) - \epsilon_{\theta_o}(x_t^s \mid s)\|_2^2\right], \tag{4}$$

$$\mathcal{L}_u = \mathbb{E}_{s\sim\mathcal{D}_u}\left[\|\epsilon_\theta(x_t^s \mid s) - \epsilon_{\theta_o}(x_t^s \mid s)\|_2^2\right], \tag{5}$$

where $\mathcal{D}_u$ contains $c_e$-irrelevant prompts, and $x_t^s$ denotes the noisy latent conditioned on $s$ at diffusion timestep $t$.

However, depending only on Le could hardly preserve the erasure effectiveness on $c_e$ in backdoor learning, since $\mathcal{L}_e$ erases $c_e$ using the target concept itself and does not explicitly constrain the textual semantics of prompts in $\mathcal{D}_e$. As a result, the optimization of $\mathcal{L}_h$ may inadvertently relearn $c_e$ from these prompts. Therefore, we introduce a prompt-level erasure objective that suppresses $c_e$ while preserving the remaining semantics in $s$:

$$\mathcal{L}_e^h = \mathbb{E}_{s\sim\mathcal{D}_e}\left[\|\epsilon_\theta(x_t^s \mid s) - (\epsilon_{\theta_o}(x_t^s \mid s) - \eta\,(\epsilon_{\theta_o}(x_t^s \mid c_e) - \epsilon_{\theta_o}(x_t^s \mid \emptyset)))\|_2^2\right]. \tag{6}$$

The final model is obtained by jointly minimizing

$$\theta^* = \arg\min_\theta\ \mathcal{L}_e + \lambda_1\mathcal{L}_h + \lambda_2\mathcal{L}_u + \lambda_3\mathcal{L}_e^h, \tag{7}$$

where $\lambda_1$, $\lambda_2$, and $\lambda_3$ balance honeypot embedding, utility preservation, and prompt-level concept erasure, respectively.

### 4.3 Jailbreak Identification

The embedded honeypot introduces an artificial shortcut associated with the concept $c_e$. When attackers optimize jailbreak prompts against the backdoored model, their optimization is encouraged to exploit such shortcut, causing the resulting prompts to exhibit a common pattern of textual representation induced by the trigger $\delta_e$. We characterize this pattern as the *honeypot signature* $\mathcal{S}_h$ and exploit it for jailbreak identification.

**Extracting Honeypot Signature.** Given the trigger $\delta_e$ and the collected dataset $\mathcal{D}_e = \{s_i\}_{i=1}^N$, where $N$ denotes the number of prompts, we feed the backdoor prompts $\{s_i \oplus \delta_e\}_{i=1}^N$ into the text encoder $\tau_{\theta^*}$, and extract their embedding representation in the `[EOS]` position as the global representation and denote it as $\{\mathbf{E}(s_i \oplus \delta_e)\}_{i=1}^N$. Accordingly, the representation matrix $\mathbb{M}^e$ could be constructed as follows:

$$\mathbb{M}^e = [\mathbf{E}(s_1 \oplus \delta_e), \mathbf{E}(s_2 \oplus \delta_e), \cdots, \mathbf{E}(s_N \oplus \delta_e)]^T, \tag{8}$$

Although the involved backdoored prompts contain diverse semantics, they share the common representation pattern induced by the trigger. Therefore, to extract this common pattern while suppressing prompt-specific deviations, we perform the SVD operation on normalized $\mathbb{M}^e$, and extract the first right singular vector $\mathbf{v}_1$, corresponding to the largest singular value, as honeypot signature.

**Identifying Jailbreak Prompt.** Given an input prompt $s$, we compare its textual representation $\mathbf{E}(s)$ with $\mathcal{S}_h$ via cosine similarity, i.e., $r_h(s) = \cos\big(\mathbf{E}(s), \mathcal{S}_h\big)$. The prompt is identified as a jailbreak prompt if $r_h(s) > \phi_h$, where $\phi_h$ is a predefined detection threshold. To improve robustness against potential signature leakage, we further randomize the representation dimensions used for signature extraction. Specifically, we randomly sample an index set $\mathcal{I} \subseteq \{1, \ldots, d\}$ and extract the features from corresponding indexes for prompt embeddings, yielding

$$\mathbb{M}^e_{\mathcal{I}} = \left[\mathcal{P}_{\mathcal{I}}\mathbf{E}(s_1 \oplus \delta_e), \mathcal{P}_{\mathcal{I}}\mathbf{E}(s_2 \oplus \delta_e), \cdots, \mathcal{P}_{\mathcal{I}}\mathbf{E}(s_N \oplus \delta_e)\right]^T, \tag{9}$$

where $\mathcal{P}_{\mathcal{I}}$ denotes the corresponding feature projection. The honeypot signature is extracted from $\mathbb{M}^e_{\mathcal{I}}$, and $\mathcal{P}_{\mathcal{I}}$ is also applied to the input representation for jailbreak detection. By dynamically varying $\mathcal{I}$, HoneyT2I could avoid exposing a fixed global signature, making signature-aware jailbreak optimization more difficult.

### 4.4 Jailbreak Mitigation

Based on the above $\mathcal{S}_h$, jailbreak identification alone can only determine whether an input prompt belongs to the jailbreak one, and deny its generation query for defense purpose, typically leading to inferior generation utility. Instead, HoneyT2I further mitigates the jailbreak effect of prompts while preserving their remaining semantics. To this end, we extract a *difference signature* $\mathcal{S}_d$ that characterizes the representation shift caused by the embedded honeypots, and remove such shift from identified jailbreak representations during model inference.

**Extracting Difference Signature** Given the trigger $\delta_e$ and the collected dataset $\mathcal{D}_e = \{s_i\}_{i=1}^N$, we compare original prompts with their backdoor counterparts in textual semantics. With the above defined $\mathbf{E}(\cdot)$, we construct the difference matrix $\mathbb{M}^d$ as follow:

$$\mathbb{M}^d = \left[\mathbf{E}(s_1 \oplus \delta_e) - \mathbf{E}(s_1), \mathbf{E}(s_2 \oplus \delta_e) - \mathbf{E}(s_2), \cdots, \mathbf{E}(s_N \oplus \delta_e) - \mathbf{E}(s_N)\right]^T. \tag{10}$$

Similarly, we perform the SVD operation on $\mathbb{M}^d$ and take its first right singular vector $\mathbf{v}_1^d$ as difference signature $\mathcal{S}_d$, while ensuring the sign consistency with $\mathbf{E}(s_i \oplus \delta_e) - \mathbf{E}(s_i)$. Thus, the resulting $\mathcal{S}_d$ points from the $c_e$-erased representations toward their backdoor representations.

**Mitigating Jailbreak Effects.** For a identified jailbreak prompt $s$, let $\tau_{\theta^*}(s) = \{e_i\}_{i=1}^L$ denote its token-wise representations, totally $L$ embedding tokens. We suppress its jailbreak effects by modifying its token-wise representations with extracted $\mathcal{S}_d$ as follows:

$$e_i' = e_i - \alpha \max\left(\frac{e_i^T \mathcal{S}_d}{\|e_i\|_2 \|\mathcal{S}_d\|_2}, 0\right) \mathcal{S}_d, \tag{11}$$

where $\alpha$ controls the steering strength. Only token embeddings positively aligned with $\mathcal{S}_d$ are modified, while the others remain unchanged. The modified representations form $\tau'_{\theta^*}(s)$, which is fed into cross-attention modules in SD, steering generation away from the $c_e$-related content while preserving its remaining semantics.

## 5 Theoretical Analysis

We theoretically characterize how the embedded honeypot facilitates jailbreak identification. Consider a jailbreak trajectory $\{s^t\}_{t=0}^T$, with normalized representations $z_t = \mathbf{E}(s^t)/\|\mathbf{E}(s^t)\|_2$ and normalized honeypot signature $S_h$. The honeypot score is $H_t = \langle z_t, S_h \rangle$, where $H_t \geq \phi_h$ means $s^t$ is identified as the jailbreak prompt. Note that $s^0$ and $s^T$ represent the initial and finalized prompts, respectively.

Let $d_t = z_{t+1} - z_t$ and $L_T = \sum_{t=0}^{T-1} \|d_t\|_2$ denote the representation update at the $t$-th jailbreak optimization step and the trajectory length . We define the *Cumulative Honeypot Alignment* (CHA)

as $\bar{\gamma}_T = \frac{\sum_{t=0}^{T-1}\langle d_t, S_h\rangle}{L_T}$, which measures the accumulated alignment between jailbreak optimization and the honeypot signature.

**Theorem 1** (Detectability under Cumulative Honeypot Alignment). *If $\bar{\gamma}_T \geq \gamma > 0$, then*

$$H_T \geq H_0 + \gamma L_T. \tag{12}$$

*Consequently, the final prompt is guaranteed to be detected when*

$$L_T \geq \frac{[\phi_h - H_0]_+}{\gamma}, \tag{13}$$

*where* $[x]_+ = \max(x, 0)$.

Theorem1 shows that individual attack updates need not consistently align with $S_h$. A positive cumulative bias is sufficient to progressively induce jailbreak prompts into the detectable region created by honeypots. The proof is provided and empirically validated in Appendix B.

# 6 Experiments

## 6.1 Experiment Setups

**Target Concept and Datasets.** We use *nudity*, a widely adopted unsafe concept in T2I safety studies, as the default erased concept. Nudity-related prompts are collected from I2P (Schramowski et al., 2023) and CoPro (Liu et al., 2025), while COCO captions (Lin et al., 2014) are used to preserve generation utility during honeypot embedding.

**Target Models and Jailbreak Attacks.** Following prior studies, HoneyT2I is adopted to SD v1.4 by default, and further implemented with advanced T2I models in AppendixD.6 and D.7. We consider two white-box jailbreak attacks, MMA (Yang et al., 2024a) and UnlearnDiffAtk (Zhang et al., 2024c). Besides, three popular black-box jailbreak attacks are exploited to further evaluate the defense effectiveness in the Appendix D.4.

**Baselines.** We compare HoneyT2I with AdvUnlearn (Zhang et al., 2024b) and GuardT2I (Yang et al., 2024b), together with three prompt classifiers: OpenAI Moderation, Microsoft Azure, and Detoxify (Hanu & Unitary team, 2020). We also include eight concept-erasure methods: ESD (Gandikota et al., 2023), FMN (Zhang et al., 2024a), SPM (Hong et al., 2024), UCE (Gandikota et al., 2024), RECE (Huang et al., 2024), MACE (Lu et al., 2024), SalUn (Fan et al., 2024), and SAFREE (Yoon et al., 2025). Further details are provided in the Appendix.

**Evaluation Metrics.** We use Attack Success Rate (ASR) for jailbreak robustness and AUC for jailbreak detection. NudeNet (Bedapudi, 2019) is used to compute the Nudity Appearance Ratio (NAR) for concept-erasure effectiveness, while FID (Heusel et al., 2017) and CLIP score (Hessel et al., 2021) on COCO measure generation quality and text-image alignment. Meanwhile, we exploit FPR@TPR95% to quantify the proportion of false positives (incorrectly identified as jailbreak prompts) when $\mathcal{S}_h$ correctly identifies 95% of the true positives. Note that best results are **bolded** in following illustrations.

## 6.2 Experiment Results and Analysis

**Effectiveness of detecting jailbreak prompts** Honeypot signatures serve as the foundation for identifying jailbreak prompts in HoneyT2I. As reported in Tab.1, we compare HoneyT2I with three mainstream commercial platforms and GuardT2I on MMA and I2P. Note that the second row (gray) denotes content detection on generated images where optimized prompts are identified if the contents contain nudity-related concepts. For prompt-level detection, HoneyT2I achieves an average AUC of 81.85% and FPR@TPR95 of

Table 1: Jailbreak detection evaluation in terms of AUC (%) / FPR@TPR95%.

| Method | MMA | I2P | AVG. |
|---|---|---|---|
| OpenAI-Moderation | 73.02 / 40.20 | 91.93 / 35.50 | 82.52 / 37.85 |
| Microsoft Azure | 90.66 / 57.60 | 55.04 / 77.50 | 72.85 / 67.55 |
| Detoxify | 79.27 / 13.70 | 54.63 / 76.00 | 66.95 / 44.85 |
| GuardT2I | 98.86 / 6.59 | 93.05 / 25.50 | 96.96 / 16.05 |
| HoneyT2I (Ours) | 96.40 / 7.91 | 67.30 / 28.03 | 81.85 / 17.97 |
| | **99.00 / 4.53** | **97.90 / 1.79** | **98.45 / 3.16** |

17.97%, outperforming the commercial moderation platforms and demonstrating that honeypot signatures can effectively capture the principal characteristics of jailbreak prompts. Benefiting from latent translation with an additional LLM inspector, GuardT2I achieves stronger prompt-level detection performance. However, when incorporating content detection, HoneyT2I further improves to an average AUC of 98.45% and FPR@TPR95 of 3.16%, achieving the best overall detection performance. More importantly, unlike conventional moderation-based defenses that mainly identify and reject suspicious requests, HoneyT2I integrates detection with subsequent jailbreak mitigation by steering malicious prompt representations toward benign directions. Therefore, HoneyT2I not only identifies jailbreak prompts but also preserves meaningful image generation capability after mitigation, rather than simply refusing unsafe requests.

**Effectiveness of defending jailbreak prompts** We leverage two advanced white-box jailbreak methods, i.e., UnlearnDiffAtk and MMA, to attack HoneyT2I and existing methods, and evaluate their defense performance using ASR. The results are reported in Tab.2 and Tab.3. **Pre-ASR** and **Post-ASR** denote the attack success ratios before and after jailbreak, respectively. Overall, the substantial increase from **Pre-ASR** to **Post-ASR** for most methods confirms the effectiveness of both attacks, while UnlearnDiffAtk generally poses a greater challenge to jailbreak defense. Although SalUn achieves the lowest ASRs, it incurs substantial generation-quality degradation, with notably higher FID and lower CLIP scores, which is also reflected by the visual results in Fig. 8. This suggests that aggressively improving jailbreak robustness can compromise the basic generation capability of T2I models. In contrast, HoneyT2I leverages artificial honeypots to manipulate jailbreak optimization, achieving competitive defense performance while largely preserving generation quality, thereby providing a favorable trade-off between robustness and utility.

Table 2: Defense performance against UnlearnDiffAtk.

| **Method** | **Pre-ASR** (↓) | **Post-ASR** (↓) | **FID** (↓) |
|---|---|---|---|
| FMN | 88.0% | 97.9% | **16.9** |
| SPM | 54.9% | 91.6% | 17.5 |
| SAFREE | 26.1% | 85.6% | 19.3 |
| UCE | 21.8% | 79.6% | 17.1 |
| ESD | 20.4% | 76.1% | 18.2 |
| RECE | 13.4% | 75.4% | 17.2 |
| MACE | 9.1% | 74.6% | **16.9** |
| AdvUnlearn | 7.8% | 21.1% | 19.3 |
| SalUn | **1.4%** | **11.3%** | 33.6 |
| HoneyT2I (Ours) | 2.3% | 12.6% | 18.9 |

Table 3: Defense Evaluation on MMA jailbreak.

| **Metrics** | SD v1.4 | SAFREE | SalUn | AdvUnlearn | HoneyT2I (Ours) |
|---|---|---|---|---|---|
| **Pre-ASR** (↓) | 95.7% | 1.7 % | **0.0%** | 7.2% | 1.0% |
| **Post-ASR** (↓) | 95.7% | 58.5% | **0.2%** | 3.4% | 7.8% |
| **FID** (↓) | 16.70 | 19.28 | 33.62 | 19.34 | **18.85** |
| **CLIP** (↑) | 31.1 | 25.9 | 28.7 | 29.0 | **31.5** |

**Effectiveness of concept erasure and utility preservation** Besides defending against jailbreak prompts, HoneyT2I should preserve the original concept-erasure capability and generation utility on benign prompts. Tab.4 compares HoneyT2I with existing concept-erasure methods in terms of NAR, FID, and CLIP scores. HoneyT2I achieves the highest CLIP score on COCO while maintaining a competitive FID, indicating that the embedded honeypots largely preserve generation utility. Meanwhile, its low NARs on I2P and MMA demonstrate strong preservation of concept erasure, achieving the best result on I2P and the second-best on MMA. In contrast, although SalUn achieves the lowest NAR on MMA, it incurs substantial utility degradation, as reflected by

Table 4: Comparative evaluation of concept erasure and generation quality. MMA$^{\dagger}$ denotes prompts from its official benchmark.

| **Method** | **I2P** | **MMA**$^{\dagger}$ | **COCO** | |
|---|---|---|---|---|
| | **NAR** (↓) | | **FID** (↓) | **CLIP** (↑) |
| SD v1.4 | 69.4% | 95.7% | 16.70 | 31.1 |
| FMN | 45.5% | – | **16.86** | 30.8 |
| SPM | 6.0% | 67.4% | 17.48 | 31.0 |
| UCE | 20.0% | 86.7% | 17.10 | 30.9 |
| ESD | 13.0% | 87.3% | 18.18 | 28.7 |
| SAFREE | 3.4% | 58.5% | 19.28 | 25.9 |
| SalUn | 5.1% | **0.2%** | 33.62 | 28.7 |
| AdvUnlearn | 7.2% | 3.4% | 19.34 | 29.0 |
| HoneyT2I (Ours) | **2.1%** | 1.0% | 18.85 | **31.5** |

its markedly higher FID and lower CLIP score, which is also evident in Fig. 9. These results demonstrate that HoneyT2I provides a favorable balance between concept erasure and benign generation utility.

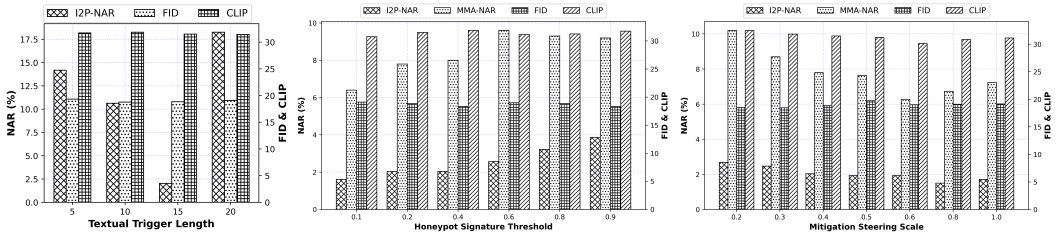


Figure 4: Ablation Studies of textual trigger $\delta_e$, jailbreak detection threshold $\phi_h$, and mitigation steering scale $\alpha$ on I2P, MMA, and COCO datasets in terms of NAR, FID, and CLIP scores.

**Ablation studies** HoneyT2I involves three key hyperparameters: textual trigger $\delta_e$, detection threshold $\phi_h$, and steering scale $\alpha$, whose effects are investigated in Fig.4. Varying the trigger length has little impact on generation quality, suggesting that the utility preservation objective $\mathcal{L}_u$ effectively maintains benign utility. However, overly short or long triggers lead to higher NAR on I2P, indicating weaker effects on concept erasure with improper trigger settings. For jailbreak detection, a larger $\phi_h$ allows prompts containing fewer honeypot features to bypass signature-based detection. For mitigation, a smaller $\alpha$ leads to insufficient removal of jailbreak effects, whereas an excessively large $\alpha$ degrades generation quality. These results reveal substantial trade-offs in the three hyperparameters and motivate the default configurations for the HoneyT2I implementation.

**Defense robustness against adaptive attackers** To evaluate the robustness of HoneyT2I against adaptive attacks, we strengthen the attacker with two additional capabilities: (i) the attacker has access to the honeypot signature $\mathcal{S}_h$ and explicitly optimizes jailbreak prompts to evade signature-based detection; (ii) the attacker can infer the textual trigger $\delta_e$ and unlearn it from the protected model to remove the embedded honeypots.

Accordingly, we conduct jailbreak experiments as follows. (i) Given the honeypot signature $\mathcal{S}_h$, the attacker performs MMA with an additional orthogonality constraint against $\mathcal{S}_h$, encouraging the optimized prompts to deviate from the known $\mathcal{S}_h$ and thus bypass signature-based detection. Without randomized signature extraction, the ASR increases substantially from $7.8\%$ to $84.2\%$, indicating that the fixed $\mathcal{S}_h$ can severely weaken the jailbreak defense. In contrast, when randomized honeypot signatures are adopted , the ASR decreases $13.5\%$ ($\downarrow 70.7\%$). The phenomenon shows that signature randomization in Eq.9 effectively increases the jailbreak difficulty of malicious attackers, improving robustness against signature-aware detection evasion. (ii) Furthermore, We consider a stronger attacker who infers $\delta_e$ and removes the embedded honeypots. We unlearn $\delta_e$ from the HoneyT2I-protected model and then perform MMA on the resulting model. Although the optimized prompts achieve only $6.2\%$ detection AUC, they still yield merely $7.1\%$ ASR when transferred back to the deployed HoneyT2I model. This suggests that honeypot removal does not fully recover the original vulnerability, probably because honeypot embedding also reshapes the textual representation space and reduces attack transferability.

## 7 Conclusion

In this paper, we propose HoneyT2I, a proactive defense against jailbreak attacks on pretrained T2I models. Instead of patching existing vulnerabilities, HoneyT2I embeds controlled honeypots to steer jailbreak optimization toward recognizable representation patterns, enabling effective detection and mitigation through extracted honeypot and difference signatures. Extensive experiments demonstrate that HoneyT2I effectively mitigates diverse jailbreak attacks while preserving concept erasure and benign generation utility. Further evaluations demonstrate its flexible integration with different concept erasure methods and its potential to generalize to more advanced T2I models. These results highlight proactively reshaping the attack surface as a promising direction for improving the models robustness.

### AI USE STATEMENT

In this work, we used generative AI tools, primarily ChatGPT, to assist with refining theoretical formulations and proofs, providing feedback on experimental design, and interpreting experimental results. We also used generative AI tools for language editing, improving the clarity and organization of the manuscript, summarizing and organizing relevant literature, assisting with figure and table presentation, and formatting references and LaTeX content. Generative AI was not used to generate experimental data, implement the proposed method, or independently conduct experiments.

All AI-assisted theoretical arguments were manually examined and revised by the authors, and the final mathematical statements and proofs were verified by the authors. Suggestions regarding experimental design and result interpretation were evaluated against actual experimental results and accepted only after manual verification. Literature-related suggestions and references were checked against the original publications, and all AI-assisted writing was reviewed and edited by the authors for technical correctness, accuracy, and consistency with the reported experiments. We take full responsibility for the final content of this work, including all text, claims, results, and artifacts produced with the aid of generative AI.

## REFERENCES

Zhongjie Ba, Jieming Zhong, Jiachen Lei, Peng Cheng, Qinglong Wang, Zhan Qin, Zhibo Wang, and Kui Ren. Surrogateprompt: Bypassing the safety filter of text-to-image models via substitution. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, pp. 1166–1180, 2024.

Samyadeep Basu, Nanxuan Zhao, Vlad I Morariu, Soheil Feizi, and Varun Manjunatha. Localizing and editing knowledge in text-to-image generative models. In *The twelfth international conference on learning representations*, 2024.

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

Ruchika Chavhan, Da Li, and Timothy Hospedales. Conceptprune: Concept editing in diffusion models via skilled neuron pruning. In *The Thirteenth International Conference on Learning Representations*, 2025.

Zhi-Yi Chin, Chieh Ming Jiang, Ching-Chun Huang, Pin-Yu Chen, and Wei-Chen Chiu. Prompting4Debugging: Red-teaming text-to-image diffusion models by finding problematic prompts. In *Forty-first international conference on machine learning*, 2024.

cjadams, Jeffrey Sorensen, Julia Elliott, Lucas Dixon, Mark McDonald, nithum, and Will Cukierski. Toxic comment classification challenge, 2017. URL `https://kaggle.com/competitions/jigsaw-toxic-comment-classification-challenge`.

Chongyu Fan, Jiancheng Liu, Yihua Zhang, Eric Wong, Dennis Wei, and Sijia Liu. SalUn: Empowering machine unlearning via gradient-based weight saliency in both image classification and generation. In *The twelfth international conference on learning representations*, 2024.

Rohit Gandikota, Joanna Materzynska, Jaden Fiotto-Kaufman, and David Bau. Erasing Concepts from Diffusion Models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2426–2436, October 2023.

Rohit Gandikota, Hadas Orgad, Yonatan Belinkov, Joanna Materzyńska, and David Bau. Unified Concept Editing in Diffusion Models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 5111–5120, January 2024.

Laura Hanu and Unitary team. Detoxify, 2020. tex.howpublished: Github. https://github.com/unitaryai/detoxify.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 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.

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

SeungHoo Hong, Juhun Lee, and Simon S. Woo. All but One: Surgical Concept Erasing with Model Preservation in Text-to-Image Diffusion Models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(19):21143–21151, March 2024. doi: 10.1609/aaai.v38i19.30107. URL https://ojs.aaai.org/index.php/AAAI/article/view/30107.

Chi-Pin Huang, Kai-Po Chang, Chung-Ting Tsai, Yung-Hsuan Lai, Fu-En Yang, and Yu-Chiang Frank Wang. Receler: Reliable concept erasing of text-to-image diffusion models via lightweight erasers. In *European Conference on Computer Vision*, pp. 360–376. Springer, 2024.

Xinfeng Li, Yuchen Yang, Jiangyi Deng, Chen Yan, Yanjiao Chen, Xiaoyu Ji, and Wenyuan Xu. Safegen: Mitigating sexually explicit content generation in text-to-image models. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, pp. 4807–4821, 2024.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Runtao Liu, Ashkan Khakzar, Jindong Gu, Qifeng Chen, Philip Torr, and Fabio Pizzati. Latent guard: a safety framework for text-to-image generation. In *European Conference on Computer Vision*, pp. 93–109. Springer, 2024.

Runtao Liu, I Chieh Chen, Jindong Gu, Jipeng Zhang, Renjie Pi, Qifeng Chen, Philip Torr, Ashkan Khakzar, and Fabio Pizzati. Alignguard: scalable safety alignment for text-to-image generation. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 17024–17034. IEEE, 2025.

Shilin Lu, Zilan Wang, Leyang Li, Yanzhu Liu, and Adams Wai-Kin Kong. MACE: Mass Concept Erasure in Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6430–6440, June 2024.

Mengyao Lyu, Yuhong Yang, Haiwen Hong, Hui Chen, Xuan Jin, Yuan He, Hui Xue, Jungong Han, and Guiguang Ding. One-dimensional Adapter to Rule Them All: Concepts Diffusion Models and Erasing Applications. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7559–7568, June 2024.

Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 4296–4304, 2024.

Minh Pham, Kelly O. Marshall, Niv Cohen, Govind Mittal, and Chinmay Hegde. Circumventing Concept Erasure Methods For Text-To-Image Generative Models. In *The Twelfth International Conference on Learning Representations*, 2024.

Yiting Qu, Xinyue Shen, Xinlei He, Michael Backes, Savvas Zannettou, and Yang Zhang. Unsafe diffusion: On the generation of unsafe images and hateful memes from text-to-image models. In *Proceedings of the 2023 ACM SIGSAC conference on computer and communications security*, pp. 3403–3417, 2023.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-Resolution Image Synthesis With Latent Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, June 2022.

Patrick Schramowski, Manuel Brack, Björn Deiseroth, and Kristian Kersting. Safe Latent Diffusion: Mitigating Inappropriate Degeneration in Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22522–22531, June 2023.

Lukas Struppek, Dominik Hintersdorf, and Kristian Kersting. Rickrolling the artist: Injecting backdoors into text encoders for text-to-image synthesis. In *Proceedings of the IEEE/CVF international conference on computer vision (ICCV)*, pp. 4584–4596, October 2023.

Yu-Lin Tsai, Chia-Yi Hsu, Chulin Xie, Chih-Hsun Lin, Jia You Chen, Bo Li, Pin-Yu Chen, Chia-Mu Yu, and Chun-Ying Huang. Ring-A-Bell! How Reliable are Concept Removal Methods For Diffusion Models? In *The Twelfth International Conference on Learning Representations*, 2024.

Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, Dhruv Nair, Sayak Paul, William Berman, Yiyi Xu, Steven Liu, and Thomas Wolf. Diffusers: State-of-the-art diffusion models, 2022. URL `https://github.com/huggingface/diffusers`.

Yanjie Li Xinqi Lyu, Yihao Liu. PLA: Prompt learning attack against text-to-image generative models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 16851–16860, 2025.

Yijun Yang, Ruiyuan Gao, Xiaosen Wang, Tsung-Yi Ho, Nan Xu, and Qiang Xu. Mma-diffusion: Multimodal attack on diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7737–7746, 2024a.

Yijun Yang, Ruiyuan Gao, Xiao Yang, Jianyuan Zhong, and Qiang Xu. Guardt2i: Defending text-to-image models from adversarial prompts. *Advances in neural information processing systems*, 37:76380–76403, 2024b.

Yuchen Yang, Bo Hui, Haolin Yuan, Neil Gong, and Yinzhi Cao. Sneakyprompt: Jailbreaking text-to-image generative models. In *2024 IEEE symposium on security and privacy (SP)*, pp. 897–912. IEEE, 2024c.

Jaehong Yoon, Shoubin Yu, Vaidehi Patil, Huaxiu Yao, and Mohit Bansal. SAFREE: Training-free and adaptive guard for safe text-to-image and video generation. In *The thirteenth international conference on learning representations*, 2025.

Gong Zhang, Kai Wang, Xingqian Xu, Zhangyang Wang, and Humphrey Shi. Forget-Me-Not: Learning to Forget in Text-to-Image Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 1755–1764, June 2024a.

Yimeng Zhang, Xin Chen, Jinghan Jia, Yihua Zhang, Chongyu Fan, Jiancheng Liu, Mingyi Hong, Ke Ding, and Sijia Liu. Defensive unlearning with adversarial training for robust concept erasure in diffusion models. *Advances in neural information processing systems*, 37:36748–36776, 2024b.

Yimeng Zhang, Jinghan Jia, Xin Chen, Aochuan Chen, Yihua Zhang, Jiancheng Liu, Ke Ding, and Sijia Liu. To generate or not? safety-driven unlearned diffusion models are still easy to generate unsafe images... for now. In *European Conference on Computer Vision*, pp. 385–403. Springer, 2024c.

Xin Zhao, Xiaojun Chen, Yuexin Xuan, Zhendong Zhao, Xinfeng Li, Xiaojun Jia, and Xiaofeng Wang. Buster: Implanting semantic backdoor into text encoder to mitigate nsfw content generation. In *International Conference on Database Systems for Advanced Applications*, pp. 445–461. Springer, 2026.