JailbreakDiffBench: A Comprehensive Benchmark for Jailbreaking Diffusion Models

Xiaolong Jin[1], Zixuan Weng[2], Hanxi Guo[1], Chenlong Yin[3],
Siyuan Cheng[1], Guangyu Shen[1], Xiangyu Zhang[1†]
[1]Purdue University [2]University of Notre Dame [3]Pennsylvania State University
[1]{jin509, guo778, cheng535, shen447, xyzhang}@purdue.edu
[2]zxweng0701@gmail.com [3]1418291376ycl@gmail.com

Abstract

Diffusion models are widely used in real-world applications, but ensuring their safety remains a major challenge. Despite many efforts to enhance the security of diffusion models, jailbreak and adversarial attacks can still bypass these defenses, generating harmful content. However, the lack of standardized evaluation makes it difficult to assess the robustness of diffusion model pipelines. To address this, we introduce **JailbreakDiffBench**, *a comprehensive benchmark for systematically evaluating the safety of diffusion models against various attacks and under different defenses. Our benchmark includes a high-quality, human-annotated prompt and image dataset covering diverse attack scenarios. It consists of two key components: (1) an evaluation protocol to measure the effectiveness of moderation mechanisms and (2) an attack assessment module to benchmark adversarial jailbreak strategies. Through extensive experiments, we analyze existing filters and reveal critical weaknesses in current safety measures. JailbreakDiffBench is designed to support both text-to-image and text-to-video models, ensuring extensibility and reproducibility.*

1. Introduction

Text-to-image (T2I) and text-to-video (T2V) diffusion models [10, 29, 57, 70, 71] have been increasingly adopted in real-world applications, facilitating advances in creative content generation and design. While these models offer new possibilities for artistic expression and automation, they also pose significant safety and ethical risks, including bias, misinformation, and sexual content [9, 38, 63]. To ensure their responsible deployment, it is crucial to enforce ethical and legal safeguards for diffusion model systems [34, 53].

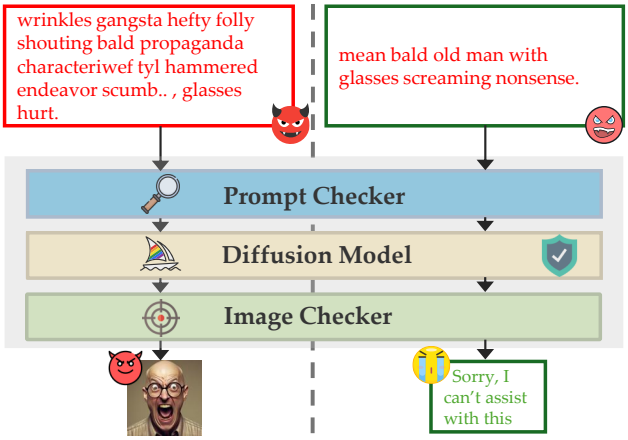


Figure 1. Moderation pipelines in diffusion models. The left side demonstrates an adversarial prompt designed to evade prompt checker and image checker, allowing harmful content to pass through the system. The right side presents moderation pipelines successfully detect the harmful query.

Real-world diffusion systems often integrate multiple defense strategies, combining prompt-level filtering, model-side safety mechanisms, and post-generation moderation [22, 49, 67]. However, a major threat to the safety of these T2I and T2V models is the emergence of attacks designed to bypass moderation mechanisms [17, 49]. Adversarial and jailbreak attacks have been shown to generate prohibited content by exploiting weaknesses in safety filters [20, 24, 47, 52]. As shown in Figure 1, adversarially manipulated prompts can evade both the prompt checker and image checker in existing moderation pipelines, allowing harmful content to pass through the system, whereas explicit harmful prompts are successfully detected and blocked. While various defense strategies—such as prompt filtering [39], adversarial training [67], and model-level moderation [61]—have been proposed, their effectiveness remains difficult to evaluate due to the lack of standard-

[1]Code available at: https://github.com/Jinxiaolong1129/JailbreakDiffusionBench
[†]Corresponding Author

ized assessment frameworks [51]. Existing safety evaluations are often limited by inconsistent dataset curation and widely used moderation tools have not been systematically evaluated. In Section 5.2, we demonstrate that the widely used I2P dataset [60] exhibits discrepancies between human annotation and its initial labels. This fragmented landscape complicates the assessment of the actual robustness of diffusion model pipelines against adversarial threats.

To address these challenges, we introduce ***JailbreakDiffBench***, a comprehensive benchmark for systematically evaluating the safety of text-to-image diffusion models against state-of-the-art jailbreak attacks and defenses. Our benchmark is built on a high-quality human-annotated prompt and image dataset that covers diverse attack scenarios. By integrating widely used moderation tools and evaluating real-world attack strategies, our benchmark enables a structured and rigorous analysis of diffusion model safety. JailbreakDiffBench consists of three key components that enable systematic analysis of existing prompt and image moderation mechanisms. First, we provide a protocol for systematically evaluating the robustness of existing prompt and image checkers and analyzing their strengths and limitations. Second, we benchmark various attack and defense strategies to assess the robustness of the current diffusion model. Third, we conduct an in-depth system safety analysis toolkit, incorporating multiple metrics such as attack success rate, F1 score, and Align Score to gain deeper insight into failure modes. Our benchmark emphasizes extensibility to support diverse attack and defense techniques in a unified framework for text-to-image and emerging text-to-video models and reproducibility by releasing all human-labeled prompt datasets, and evaluation protocols for follow-up research.

Our key contributions are as follows:

- We propose JailbreakDiffBench, the first comprehensive benchmark designed to systematically evaluate the safety of text-to-image diffusion models against jailbreak attacks and defense.
- We introduce a high-quality, human-annotated prompt and image dataset covering diverse attack scenarios and provide an extensible evaluation framework that integrates widely used moderation tools and benchmarks various attack and defense strategies.
- We release all datasets and evaluation protocols to promote reproducibility and facilitate further research on the safety of diffusion model pipelines.

2. Related Work

Diffusion Model Safety While generative AI boosts productivity [23, 36], it also raises security concerns, particularly in generating NSFW content [40, 56, 69]. Similar to jailbreak attacks on LLMs [30, 62, 66], adversarial attacks on text-to-image (T2I) models fall into white-box and black-box categories. White-box methods like MMA-Diffusion [73] require model access for adversarial prompt crafting, while model-agnostic approaches like Ring-A-Bell [65] use concept extraction but often generate nonsensical prompts with high computational costs [22, 25, 42]. Black-box methods, such as SneakyPrompt [74], DACA [19], and PGJ [25], refine prompts using reinforcement learning and LLM-based modifications to evade detection more efficiently. Defenses include external safeguards like prompt and post-image checkers [17, 49, 55, 58, 67, 72, 76], which are widely used but vulnerable to adversarial attacks. Internal defenses, such as concept erasure [11, 41], attempt to remove NSFW elements while preserving model functionality [21, 31, 37, 48, 61] but may not fully eliminate harmful content or can degrade image quality [33, 76]. To address these issues, we propose a benchmark for systematically evaluating diffusion model safety. More details on related work are provided in the Appendix 10.

3. Preliminaries

In this section, we formalize the diffusion model pipeline with safety mechanisms and jailbreak attacks. We use $\mathcal{P}$ to denote the domain of all possible text prompts and $\mathcal{I}$ to represent the space of all possible images.

Text-to-Image Diffusion System with Safety Mechanisms. A text-to-image diffusion model system with safety mechanisms is a triplet $\mathcal{S} = (C_p, D, C_i)$ where:

- $C_p : \mathcal{P} \to 0, 1$ is a prompt checker, where $C_p(p) = 1$ indicates that the prompt p is safe and allowed to proceed, and $C_p(p) = 0$ indicates the prompt is harmful and blocked.
- $D : \mathcal{P} \to \mathcal{I}$ is a diffusion model that maps an input prompt to an output image, which is only executed when $C_p(p) = 1$.
- $C_i : \mathcal{I} \to 0, 1$ is an image checker, where $C_i(i) = 1$ indicates that image i is safe for distribution, and $C_i(i) = 0$ indicates that the image contains prohibited content and is blocked.

Judger. A judger is a function $Judger : \mathcal{I} \to 0, 1$ that determines whether an image contains harmful content. The judger assigns $Judger(i) = 1$ if the image i is considered harmful and $Judger(i) = 0$ otherwise.

Alignment Checker. An alignment checker $Align : \mathcal{P} \times \mathcal{I} \to [0, 1]$ determines the degree to which an image i semantically aligns with a given prompt p. It assigns values in the range $[0, 1]$, where higher values indicate a stronger semantic alignment between the prompt and the image.

System Pipeline. For a given prompt $p \in \mathcal{P}$, the complete system pipeline can be expressed as a function $\mathcal{C}(p)$ defined as:

$$\mathcal{C}(p) = \begin{cases} C_i(D(p)) & \text{if } C_p(p) = 1 \wedge C_i(D(p)) = 1 \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

where $\mathcal{C}(p) = 1$ indicates that the prompt has successfully passed through all safety mechanisms and produced an image, and $\mathcal{C}(p) = 0$ indicates that the prompt or its generated image was blocked at some stage, meaning the system refuses the attempt.

Prohibited Content. Let $\mathcal{P}_{prohib} \subset \mathcal{P}$ be the subset of prompts that request prohibited content. For an ideal safety diffusion system, $\forall p \in \mathcal{P}_{prohib} : \mathcal{C}(p) = 0$.

Jailbreak Attack. A jailbreak attack is a function $\mathcal{J} : \mathcal{P}_{prohib} \rightarrow \mathcal{P}_{adv}$ that transforms a prohibited prompt p_{prohib} into an adversarial prompt p_{adv}. A jailbreak $\mathcal{J}$ is considered successful for a prompt $p_{prohib} \in \mathcal{P}prohib$ if:

- $C_p(\mathcal{J}(p_{prohib})) = 1$: the jailbroken prompt bypasses the prompt checker.
- $D(\mathcal{J}(p_{prohib}))$: the diffusion model generates an image containing prohibited content.
- $C_i(D(\mathcal{J}(p_{prohib}))) = 1$: the generated image bypasses the image checker.
- $Judger(D(\mathcal{J}(p_{prohib}))) = 1$: the judger determines that the generated image contains harmful content.
- $Align(p_{prohib}, D(\mathcal{J}(p_{prohib}))) = 1$: the alignment function verifies that the generated image is semantically aligned with the original prohibited prompt p_{prohib}.

In summary, a jailbreak $\mathcal{J}$ is deemed successful if the adversarial prompt p_{adv} circumvents both safety mechanisms while producing an image that is harmful and semantically aligned with the original prohibited intent.

4. JailbreakDiffBench

In this section, we present an overview of JailbreakDiffBench. To assess the safety of diffusion model systems, JailbreakDiffBench consists of three key components: (1) a robust and effective evaluation protocol for prompt and image checkers, (2) a dedicated attack assessment module, and (3) a comprehensive systematic safety analysis toolkit.

Evaluation of Prompt and Image Checkers. A critical component of the safety of the diffusion model system is the effectiveness of prompt and image checkers. To enable a fair and rigorous evaluation, we establish a standardized assessment framework that quantifies the robustness of these safety guardrails. Using carefully curated human-labeled datasets with diverse attack cases, we analyze the strengths and weaknesses of existing checkers in Section 5 and 6, guiding the development of more resilient moderation strategies.

Jailbreak Attack and Defense. To systematically evaluate the safety of diffusion models, we evaluate different combinations of prompt checkers, image checkers, and diffusion models against various attacks, which enables a precise and reproducible measurement of the effectiveness of both jailbreak strategies and defense mechanisms, as detailed in Section 7.

Comprehensive Safety Evaluation Metrics. Our benchmark goes beyond attack success rates by providing a comprehensive analysis toolkit, including multiple security metrics, such as ASR, F1-score, and Align Scores, etc, which help uncover failure patterns and provide insights for stronger defenses.

Overall, JailbreakDiffBench is developed with the following guiding principles:

Extensibility. Our framework is designed to support a wide range of attacks, including black-box, white-box, transfer-based and adversarial attacks, along with various defense mechanisms such as prompt checkers, image checkers, and unlearning-based model-level defenses. Additionally, our benchmark extends beyond text-to-image models to support emerging domains such as text-to-video diffusion models, which are increasingly gaining real-world adoption.

Reproducibility. To ensure fair comparisons and support future research, we provide full access to jailbreak artifacts, enabling transparent evaluation of red-teaming efforts in diffusion model pipelines and identification of state-of-the-art attack and defense strategies. All evaluations are conducted in realistic deployment settings to ensure practical relevance. Additionally, JailbreakDiffBench is continuously updated to integrate new attacks and defenses, keeping pace with advancements in diffusion model research.

Adhering to these principles, JailbreakDiffBench establishes a standardized and extensible framework to evaluate the robustness of diffusion pipelines, driving the development of stronger safety mechanisms in the real world.

5. Prompt Checker

Ensuring the safety of text-to-image diffusion models begins with effective prompt moderation. The prompt checker C_p acts as the first line of defense, filtering out prohibited content before it reaches the model. A robust prompt-checking mechanism is essential to prevent harmful inputs. Therefore, in this section, we systematically evaluate various prompt checkers used in diffusion pipelines, from simple rule-based approaches to advanced LLM-based systems.

5.1. Dataset

We collect data from various datasets containing benign, NSFW, and jailbreak prompts from multiple sources. Details are provided in the Appendix 12.1.

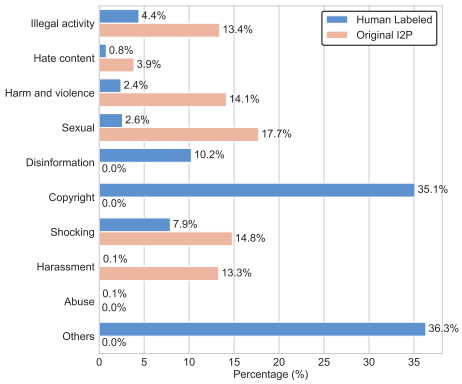
Benign Prompts. 10000 prompts extracted and filtered from DiffusionDB [68], a large-scale dataset from real world users.

NSFW Prompts. 3750 prompts from multiple resources including 4chan [50], I2P [60], SneakPrompt [74], DiffusionDB [68], VBCDE [19], and Civitai [13], with manual filtering to ensure high quality.

Jailbreak Prompts. 2000 prompts sourced from MMA [73] jailbreak datasets, including MMA_{adv} and

Class	I2P	4chan	DiffusionDB	SneakP.	Civitai	VBCDE	Online.	Total
Illegal	207/208	3/3	19/19	0/0	0/0	25/28	228/228	482
Hate	37/37	84/88	200/203	0/0	0/0	33/33	159/159	513
Harm	114/114	0/0	15/15	0/0	0/0	0/4	253/253	382
Sexual	121/122	27/28	59/67	183/200	99/100	32/32	0/0	521
Disinfo.	249/482	5/6	248/263	0/0	0/0	0/0	0/0	502
Copyright	228/1651	0/0	85/157	0/0	0/0	0/0	0/0	313
Shocking	370/372	3/4	180/195	0/0	41/41	0/3	0/0	594
Haras.	6/6	0/0	0/0	0/0	0/0	0/0	224/224	230
Abuse	3/3	0/0	0/0	0/0	0/0	0/0	210/210	213
Others	0/1708	0/371	0/1101	0/0	0/59	0/0	0/0	0
Total	1335/4703	122/500	806/2020	183/200	140/200	90/100	1074/1074	3750/8797

(a) The distribution of prompts across various risk categories and data sources. For each category, the left number represents the final number of samples included in the evaluation, while the right number indicates the total number of prompts identified after human annotation.



(b) A comparison between the original I2P dataset labels and human-annotated labels.

Figure 2. (a): Overview of the prompt dataset used for evaluating prompt checkers. The dataset is categorized into multiple risk classes: Illegal (Illegal Activities), Hate (Hate Content), Harm (Physical Harm or Violence), Sexual (Nudity or Sexual Content), Disinfo. (Disinformation), Copyright (Copyright Infringement), Shocking (Shocking Content), Haras. (Harassment), Abuse (Abuse), and Others (Safe/Irrelevant Prompts). The first number in each cell represents the number of samples included in the evaluation, while the second indicates the total identified after human annotation. (b) Comparison between original I2P labels and human-annotated labels, showing significant discrepancies.

Category	Model	ACC ↑	TPR ↑	FPR ↓	F1 ↑	AUROC ↑	MCC ↑	Query Time (ms) ↓
Rule-based	NSFW-Word-Matcher	0.6820	0.1558	0.0155	0.2635	0.5702	0.2707	1
BERT-based	NSFW-Text-Classifier	0.8468	0.8346	0.1462	0.7991	0.9231	0.6772	18.25
	DistilledBert Text Checker	0.8525	0.6346	0.0222	0.7585	0.8815	0.6848	17.27
	DistilledRobert Text Checker	0.7768	0.4068	0.0105	0.5709	0.8437	0.5269	30.61
Commercial APIs	Azure Text Moderation	0.7634	0.3237	0.0024	0.4875	0.7758	0.4765	–
	Google Text Moderation	0.8257	0.7588	0.0963	0.7878	0.9066	0.6742	–
	OpenAI Omni-Moderation	0.7940	0.4383	0.0014	0.6084	0.9333	0.5724	–
LLM	Llama-Guard-3	0.6813	0.0000	0.0000	0.0000	0.6609	0.0000	1373
	GPT-4o-mini	0.8540	0.8395	0.1376	0.8077	0.8509	0.6916	–
	GPT-4o-mini(IC)	0.8720	0.7656	0.0668	0.8137	0.8494	0.7201	–
	GPT-4o	0.6184	0.9080	0.5481	0.6347	0.6799	0.3713	–
	GPT-4o(IC)	0.7691	0.8666	0.2869	0.7327	0.7899	0.5582	–

Table 1. Performance comparison of different content moderation systems across multiple evaluation metrics, including accuracy (ACC), F1 score, true positive rate (TPR), false positive rate (FPR), AUROC, Matthews Correlation Coefficient (MCC), and query time. Models marked with (IC) indicate the use of in-context examples, where representative examples are included in prompts.

MMA_{sani}. MMA_{adv} consists of adversarial prompts crafted by MMA-Diffusion [73] and MMA_{sani} contains sanitized versions of these prompts, where non-dictionary words are removed, which are designed to bypass safety mechanisms while attempting to generate prohibited content.

5.2. Dataset Analysis

We categorize prompts into multiple risk classes, including Shocking Content, Nudity or Sexual Content, Copyright Infringement, Political or Disinformation, Hate Content, Illegal Activities, Physical Harm or Violence, Harassment, and Abuse. Each category represents a specific type of prohibited content of diffusion models. Details of each category are provided in the Appendix 12.2.

I2P Dataset Label Discrepancy. To assess the reliability of existing datasets, we first conduct an in-depth analysis of the widely used I2P dataset by relabeling the dataset with human annotators and comparing the new annotations with the original dataset labels. Figure 2b illustrates the discrepancies across multiple categories between original I2P classifications and human-labeled results. In particular, the Copyright category, which was not considered (0.0%) in the original I2P labels, accounts for 35.1% after human annotation, indicating that a substantial portion of the prompts were originally misclassified. In contrast, categories such as Illegal Activity, Harm and Violence, Sexual, Shocking, and Harassment are overrepresented in the original dataset. Additionally, 36.3% of the dataset consists of prompts that were reclassified as Others (safe, or unrelated to text-to-image generation safety), despite being initially classified as harmful content. These inconsistencies highlight the limitations of the original dataset and emphasize the need for a more rigorous and standardized benchmark. Details of the human annotation process are provided in the Appendix 16.

Dataset Selection. Given the observed inconsistencies in the I2P dataset, we take a more comprehensive approach

by aggregating NSFW prompts from multiple sources and conducting a thorough human annotation process to ensure reliability. To build a robust evaluation dataset, we propose dataset **JDBench**, which samples data from diverse sources, including I2P, 4chan, DiffusionDB, SneakPrompt, Civitai, VBCDE, and online platforms such as Discord and Reddit. Table 2a provides an overview of the dataset composition. For each category, the first number represents the final number of samples included in our evaluation of prompt checkers, while the second number indicates the total number of prompts identified after human annotation. Illegal Activity, Hate Content, and Sexual Content are among the most prevalent categories, with key sources being I2P, web data, and DiffusionDB. Copyright Violations, previously overlooked in similar datasets, are now included, primarily sourced from I2P and DiffusionDB. A notable portion of the dataset also covers Disinformation and Shocking Content, ensuring a broad evaluation scope. An important finding is that 36.8% of prompts initially labeled as harmful were later reclassified as safe or unrelated to text-to-image generation. In total, our dataset comprises 10,000 benign prompts, 2,000 adversarial prompts, and 3,750 NSFW prompts, selected from an original pool of 15,750 human-labeled examples. This refined dataset ensures a comprehensive and realistic evaluation for prompt checkers.

5.3. Prompt Checkers

We evaluate four categories of prompt checkers currently deployed in production or used in research papers, from simple rule-based methods to an LLM-based detector.

- **Rule-based.** NSFW-Word-Matcher [74], a simple keyword-based filtering approach.
- **BERT-based classifiers.** Fine-tuned transformer models, including NSFW-Text-Classifier [35], DistillBERT Text Checker [15], and DistillRoBERTa Text Checker [14].
- **Commercial APIs.** Industry-standard moderation tools such as Azure Text Moderation [8], Google Text Moderation [16], OpenAI Omni-Moderation [44].
- **LLM-based Checkers.** Advanced large language models (LLMs), including Llama-Guard-3 [28], GPT-4o-mini [27] and GPT-4o. These models leverage pre-trained knowledge and contextual understanding for prompt moderation. Additionally, we evaluate LLMs with in-context learning (IC), where representative examples are provided during inference to enhance detection accuracy, as detailed in Appendix 17

5.4. Metrics

We comprehensively evaluate prompt checkers by the following metrics, including F1 Score, Accuracy (ACC), True Positive Rate (TPR), False Positive Rate (FPR), AUROC, Matthews Correlation Coefficient (MCC) and Query Time. The trade-off between these factors is crucial in designing robust prompt checkers for text-to-image diffusion pipelines.

5.5. Experimental Results

Tables 1 and 2 present the performance of different prompt checkers across multiple evaluation metrics and recall on various types of harmful content categories and jailbreak datasets. The results indicate significant variation in effectiveness across rule-based, BERT-based, commercial API, and LLM-based methods.

Rule-based Filtering. The NSFW-Word-Matcher performs poorly across all metrics, with an overall accuracy of 0.6820 and an F1 score of 0.2635. It struggles in detecting illegal content at 0.081 recall and hate speech at 0.096 recall, while achieving slightly better results in abuse detection at 0.474 recall. However, its reliance on keyword matching results in both high false negatives and false positives, making it ineffective for real-world applications.

BERT-based Models. Transformer-based classifiers offer notable improvements over rule-based filtering. NSFW-Text-Classifier achieves an accuracy of 0.846 and an F1 score of 0.799, with high detection rates for hate speech at 0.936 recall, harassment at 0.856 recall, and abuse at 0.906 recall. However, it struggles with copyright-related content, achieving only 0.45 recall. DistillBERT Text Checker reaches 0.852 accuracy but performs worse in illegal content at 0.232 recall, and copyright at 0.278 recall. Distill-RoBERTa Text Checker performs poorly relative to other BERT models, particularly in illegal content at 0.027 recall and disinformation at 0.053 recall.

Commercial APIs. Among commercial APIs, Google Text Moderation achieves the highest accuracy at 0.825 and an F1 score of 0.787. It performs well in illegal content at 0.932 recall, hate speech at 0.955 recall, and abuse at 0.914 recall, but also exhibits weak performance in copyright at 0.3 recall. OpenAI Omni-Moderation and Azure Text Moderation show lower recall, particularly in disinformation obtaining 0.156 and 0.026 recall, respectively, and copyright violations obtaining 0.051 and 0.013 recall.

LLM-based Checkers. The latest LLM-based moderation models show the highest overall performance. GPT-4o-mini achieves 0.854 accuracy and an F1 score of 0.807, while GPT-4o-mini(IC) further improves to 0.872 accuracy and 0.813 F1. GPT-4o reaches a true positive rate of 0.908 but has a high false positive rate of 0.548, leading to inconsistent results. Across content categories, GPT-4o performs well in illegal content at 0.967 recall and abuse at 0.986 recall. It also achieves strong results in copyright at 0.882 recall and disinformation at 0.937 recall. Overall, the inclusion of in-context examples slightly reduces the recall for harmful content detection, but significantly decreases the false positive rate for safe content. Therefore, in general, in-context LLM-based detection demonstrates better

Category	Model	Content Categories									Jailbreak	
		Illegal	Hate	Harm	Sexual	Disinfo	Copyright	Shocking	Harass	Abuse	MMA_{adv}	MMA_{sani}
Rule-based	NSFW-Word-Matcher	0.081	0.096	0.209	0.305	0.063	0.115	0.183	0.104	0.474	0.132	0.136
BERT-based	NSFW-Text-Classifier	0.647	0.936	0.856	0.869	0.760	0.450	0.638	0.939	0.906	0.981	0.937
	DistillBERT Text Checker	0.232	0.628	0.322	0.756	0.434	0.278	0.466	0.511	0.319	0.982	0.952
	DistillRoBERTa Text Checker	0.027	0.172	0.139	0.683	0.053	0.054	0.049	0.385	0.498	0.791	0.773
Commercial APIs	Azure Text Moderation	0.050	0.230	0.298	0.459	0.026	0.013	0.047	0.165	0.352	0.630	0.580
	Google Text Moderation	0.932	0.914	0.877	0.714	0.638	0.300	0.554	0.848	0.934	0.835	0.768
	OpenAI Omni-Moderation	0.116	0.513	0.503	0.697	0.156	0.051	0.156	0.035	0.239	0.707	0.695
LLM	GPT-4o-mini	0.884	0.830	0.921	0.898	0.651	0.562	0.648	0.857	0.977	0.985	0.879
	GPT-4o-mini (IC)	0.743	0.756	0.856	0.843	0.513	0.342	0.549	0.848	0.911	0.971	0.841
	GPT-4o	0.967	0.867	0.927	0.925	0.937	0.882	0.742	0.922	0.986	0.991	0.875
	GPT-4o (IC)	0.913	0.819	0.929	0.862	0.875	0.671	0.705	0.939	0.977	0.973	0.855

Table 2. Recall of content moderation systems across different categories of harmful content, including Illegal Activities (Illegal), Hate Content (Hate), Physical Harm or Violence (Harm), Nudity or Sexual Content (Sexual), Political or Disinformation (Disinfo), Copyright Infringement (Copyright), Shocking Content (Shocking), Harassment (Harass), and Abuse (Abuse) and jailbreak adversarial prompts, MMA_{adv} and MMA_{sani}.

performance

Jailbreak Dataset Performance The evaluation on the jailbreak dataset (MMA_{adv} and MMA_{sani}) highlights that most models exhibit strong performance in adversarial attack detection. GPT-4o achieves the highest results, with 0.991 on MMA_{adv} and 0.875 on MMA_{sani}, indicating robust resistance to jailbreak attempts. In contrast, OpenAI Omni-Moderation and Azure Text Moderation show weaker performance on jailbreak detection, with OpenAI Omni-Moderation reaching only 0.707 on MMA_{adv} and 0.695 on MMA_{sani}. In conclusion, LLM-based moderation models demonstrate strong potential as robust prompt checkers, particularly in disinformation, copyright, and adversarial prompts. Additionally, larger model sizes do not always lead to better false positive rates (FPR), as increased sensitivity to input prompts results in higher false alarms. Therefore, a promising direction for future work is to use LLMs to generate high-quality synthetic data sets for fine-tuning more efficient BERT-based models, which could balance cost and performance for real-world content moderation applications.

6. Image Checker

Although prompt checkers C_p provide the first line of defense against prohibited content, they cannot guarantee complete protection against adversarial NSFW input. As demonstrated in Section 5, even advanced prompt checkers can be bypassed by carefully crafted jailbreak prompts. Therefore, the image checker C_i serves as a critical post-generation safeguard. Image checkers inspect the actual visual output to prevent prohibited content that might bypass the initial safeguard. In this section, we evaluate the performance of multiple image moderation checkers.

6.1. Dataset

For comprehensive evaluation of image checker, we collected a dataset with two categories, benign and NSFW AI-generated images. Each AI-generated image in our dataset was manually annotated to ensure high quality.

Benign AI-generated images. We select 1,000 benign prompts from DiffusionDB based on the ranked harmfulness score in orignal dataset. These prompts were then used to generate 10,000 images across 10 models. After human annotation, 600 images per model were retained, resulting in a total of 6,000 benign AI-generated images.

NSFW AI-generated images. We sample 1,000 prompts from the dataset **JDBench** in Section 5 to generate 10,000 images across 10 different AI models. Following human annotation, 600 images per model were selected, yielding a final set of 5,439 NSFW images.

6.2. Image Checkers Evaluated

We evaluate three categories of image checker as follows.

Basic NSFW Classifiers. includes Q16 [59], SD Safety Checker [18], Multihead Detector and Finetuned Multihead Detector (F-Multihead Detector) [50], which utilize CLIP-based feature-based detection to identify NSFW content.

Commercial APIs. Industry-standard image moderation tools, including Azure Image Moderation [43], and OpenAI Omni-Moderation [44], which leverage proprietary models and large-scale datasets for image content filtering.

Multimodal LLM-based Checkers. includes advanced large vision language models, such as Llava Guard [7], and GPT-4o, which integrate multimodal understanding to enhance image moderation by contextual knowledge and in-depth content interpretation.

6.3. Metrics

To thoroughly evaluate image checker performance, we employ metrics including F1 Score, Accuracy (ACC), True

Category	Model	Accuracy ↑	TPR ↑	FPR ↓	F1 ↑	AUROC ↑	MCC ↑	Query Time (ms) ↓
Basic NSFW Classifiers	Q16	0.6859	0.4095	0.0635	0.5535	0.7690	0.4118	144.77
	Multihead Detector	0.7571	0.5352	0.0418	0.6769	0.8696	0.5509	197.24
	F-Multihead Detector	0.7544	0.5470	0.0575	0.6793	0.8770	0.5386	206.54
Commercial APIs	Azure Image Moderation	0.6256	0.2138	0.0012	0.3519	0.6990	0.3505	–
	OpenAI Omni-Moderation	0.5419	0.0399	0.0030	0.0765	0.5039	0.1299	–
Multimodal LLM	Llava Guard	0.7283	0.4492	0.0406	0.5997	0.7043	0.4866	12630.51
	GPT-4o	0.9203	0.8643	0.0290	0.9116	0.9177	0.8436	–

Table 3. Performance comparison of different image moderation checker across accuracy (ACC), F1 score, true positive rate (TPR), false positive rate (FPR), AUROC, Matthews Correlation Coefficient (MCC), and query time. The evaluated models span three categories: basic NSFW classifiers, commercial APIs, and multimodal LLM-based checkers. F-Multihead Detector is short for Finetuned Multihead Detector.

Positive Rate (TPR), False Positive Rate (FPR), AUROC, Matthews Correlation Coefficient (MCC), and Query Time.

6.4. Experimental Results

Table 3 presents a performance comparison of various image moderation tools across different metrics.

Basic NSFW Classifiers. Multihead Detector achieves the highest accuracy at 75.71%, outperforming Q16 (68.59%). Multihead Detector also maintains a balanced trade-off between TPR (53.52%) and FPR (4.18%), whereas Q16 has a lower recall (40.95%) and a higher FPR (6.35%). F-Multihead Detector performs similarly to Multihead Detector but with a slightly higher FPR (5.75%), indicating a trade-off between sensitivity and specificity.

Commercial APIs. Azure Image Moderation achieves 62.56% accuracy, with a TPR of 21.38%, but maintains an extremely low FPR of 0.12%, making it highly conservative. OpenAI Omni-Moderation, in contrast, achieves only 54.19% accuracy, with a very low TPR of 3.99%, suggesting poor NSFW detection capability. These results indicate that while commercial APIs may be optimized for precision, they struggle with high-recall tasks.

Multimodal LLM-Based Checkers. GPT-4o achieves the highest accuracy of 92.03%, with a TPR of 86.43% and a low FPR of 2.90% This suggests that GPT-4o effectively balances recall and precision while maintaining high robustness. Llava Guard, while also outperforming traditional methods, lags behind GPT-4o with an accuracy of 72.83% and a TPR of 44.92%.

In conclusion, multimodal LLM-based models, particularly GPT-4o, significantly outperform other approaches in NSFW detection, achieving the best trade-off between accuracy, recall, and false positive rate. While Multihead Detector provide competitive results, they fall short in recall. Commercial APIs, though efficient, exhibit conservative detection behaviors, limiting their effectiveness in high-recall scenarios.

7. Jailbreak Evaluations and Analysis

7.1. Dataset

To ensure a comprehensive evaluation, we use **JDBench-mini**, a dataset sampled from the one used in Section 5. It includes 50 prompts from each of the following prohibited content categories: Shocking Content, Nudity or Sexual Content, Political or Disinformation, Hate Content, Illegal Activities, Physical Harm or Violence, and Harassment and Abuse. Copyright-related prompts are excluded due to the difficulty of verifying detector judgments through human annotation. The final dataset consists of 400 prompts covering various forms of harmful content.

7.2. Baselines

We use five baselines in our evaluation, categorized into two types: black-box attacks, including PGJ [25], DACA [19], SneakPrompt [74]; and white-box attacks, including RingABell (RABell) [65] and MMA [73]. We only consider MMA text-based attack method.

7.3. Models

We evaluate the following diffusion models: Stable-Diffusion-2 [1], Stable-Diffusion-3.5-Large [3], Stable-Diffusion-3.5-Large-Turbo [4], Stable-Diffusion-3.5-Medium [5], Stable-Diffusion-v1.5 [45], Stable-Diffusion-XL-Base-0.9 [6], Flux-1-Dev [32], Hunyuan-Distill [26], and CogView3 [64].

7.4. Metrics

We evaluate jailbreak success using the following metrics.
Bypass Rate (BPR). The percentage of adversarial prompts that bypass both the prompt checker C_p and the image checker C_i.
BLIP Alignment Score (BLIP). Measures semantic similarity between the generated image and the original prohibited prompt using a BLIP-based model.
Alignment Score (AlignS). The proportion of generated images that are semantically aligned with their corresponding prohibited prompts.

Attack Success Rate (ASR). The percentage of adversarial prompts that generate harmful content while bypassing all safety mechanisms.

Human-Verified ASR (HASR). A human-evaluated version of ASR, where annotators manually verify the harmfulness of generated images.

In the experiments presented in Figure 4, we utilize GPT-4o-mini as the prompt checker C_p, Multihead Detector as the image checker C_i, GPT-4o for computing AlignS, and GPT-4o as the Judger.

Model	Attack	BPR ↑	BLIP ↑	AlignS ↑	ASR ↑	HASR ↑
SD-v1.5	Sneakprompt	7.00%	5.09	30.09%	7.00%	6.75%
	MMA	4.75%	4.88	58.42%	4.75%	4.75%
	PGJ	51.75%	5.05	37.20%	5.75%	5.50%
	DACA	57.00%	4.75	30.00%	8.25%	8.00%
	RABell	0.50%	5.25	35.00%	0.50%	0%
SD-xl base-0.9	Sneakprompt	6.75%	5.11	34.06%	6.75%	6.25%
	MMA	5.00%	4.79	67.50%	5.00%	5.00%
	PGJ	49.75%	5.11	47.64%	3.50%	3.00%
	DACA	54.75%	4.73	32.92%	4.75%	4.25%
	RABell	0.75%	5.43	46.67%	0.75%	0.25%
SD-3.5 Large-Turbo	Sneakprompt	8.00%	5.07	42.09%	8.00%	7.75%
	MMA	5.25%	4.85	51.90%	5.25%	4.75%
	PGJ	50.50%	5.10	39.16%	4.25%	4.00%
	DACA	56.25%	4.76	38.84%	6.00%	5.75%
	RABell	0.50%	5.50	40.00%	0.50%	0%
SD-3.5 Medium	Sneakprompt	7.50%	5.11	35.58%	7.50%	7.00%
	MMA	4.75%	4.83	62.63%	4.75%	4.25%
	PGJ	50.50%	5.07	45.79%	6.00%	5.50%
	DACA	54.25%	4.75	43.73%	7.50%	7.25%
	RABell	0.50%	5.40	25.00%	0.50%	0.25%
HunyuanDiT v1.2-Distilled	Sneakprompt	8.25%	5.07	38.65%	8.25%	8.25%
	MMA	4.25%	4.85	60.00%	4.25%	4.00%
	PGJ	51.00%	5.11	45.34%	1.75%	1.25%
	DACA	58.50%	4.80	38.85%	7.00%	7.00%
	RABell	0.75%	5.30	6.67%	0.75%	0.25%
CogView3 Plus-3B	Sneakprompt	8.25%	5.07	38.36%	8.25%	8.25%
	MMA	5.00%	4.84	47.00%	5.00%	5.00%
	PGJ	52.25%	5.06	42.92%	3.25%	2.75%
	DACA	55.00%	4.77	42.45%	6.50%	6.00%
	RABell	0.75%	5.23	10.00%	0.75%	0.50%
FLUX.1-dev	Sneakprompt	8.25%	5.12	46.85%	8.25%	8.00%
	MMA	4.25%	4.80	48.82%	4.25%	4.25%
	PGJ	50.75%	5.08	37.44%	3.25%	2.75%
	DACA	54.00%	4.75	38.84%	4.75%	4.50%
	RABell	0.75%	5.23	23.33%	0.75%	0.50%

Table 4. Jailbreak attack performance across different diffusion models against GPT-4o-mini as prompt checker and Multihead Detector as image checekr, showing Bypass Rate (BPR), BLIP Alignment Score (BLIP), Alignment Score (AlignS), Attack Success Rate (ASR), and Human-Verified ASR (HASR) for various attack methods. Higher BPR indicates a greater ability to bypass safety mechanisms, while higher AlignS and ASR suggests better semantic alignment with the original prohibited prompt.

7.5. Analysis

Table 4 shows the performance of different jailbreak attacks across multiple diffusion models. The total bypass rate (BPR) varies significantly depending on the attack method and model. Among the attacks tested, DACA and PGJ generally achieve the highest bypass rates, exceeding 50%. SneakPrompt and MMA exhibit lower BPRs but maintain competitive alignment scores (AlignS), suggesting their ability to generate semantically relevant images. RingABell, a white-box attack, consistently shows the lowest bypass rates, because a random token in the adversarial prompts could be easily detected by LLM. Across models, Stable-Diffusion-v1.5 and Stable-Diffusion-XL-Base-0.9 tend to be more vulnerable, with higher BPR values compared to newer versions like Stable-Diffusion-3.5-Large-Turbo. However, recent advanced models such as Hunyuan-Distill and Flux-1-Dev also show moderate susceptibility, though their HASR scores remain relatively low, suggesting that even if an attack bypasses automated checks, human annotators often do not verify the outputs as harmful. Additionally, while attacks achieve high semantic similarity between the generated image and the original prompt, the actual attack success rate (ASR) and Human verification (HASR) remain much lower, suggesting that while adversarial prompts may evade detection, they do not always lead to harmful content generation. Overall, the results highlight the varied robustness of diffusion models against different attack and defense strategies. Although some models exhibit stronger resistance, the success of certain attacks underscores ongoing security challenges. This indicates our standardized benchmarks and datasets are crucial for effectively evaluating red teaming progress and assessing the security of diffusion pipelines.

8. Conclusion

In this work, we introduce JailbreakDiffBench, a comprehensive benchmark for systematically evaluating the robustness of text-to-image models against adversarial attacks. Our benchmark provides a standardized evaluation framework, a high-quality human-annotated dataset, and an extensible protocol for assessing moderation mechanisms and attack strategies. Through extensive experiments, we highlight vulnerabilities in current safety measures of diffusion pipelines and the progress of red-teaming methods. By making our datasets and evaluation tools publicly available, we aim to advance research on securing diffusion pipelines and improving red teaming methodologies.

Discussion The red teaming of diffusion models needs to be conducted within real-world settings. When conducting red-teaming research, diffusion pipelines need to be stress-tested and evaluated for utility under attacks and different combinations of defense methods. Essentially, red teaming is finding vulnerabilities in combinations of prompt checkers, diffusion models, image checkers, and other components. We open-source our evaluation framework to facilitate community usage and address the risks posed by the widespread deployment of diffusion models in real-world applications.

Acknowledgements

We are grateful to the Center for AI Safety for providing computational resources. This work was funded in part by the National Science Foundation (NSF) Awards SHF-1901242, SHF-1910300, Proto-OKN 2333736, IIS-2416835, DARPA VSPELLS - HR001120S0058, ONR N00014-23-1-2081, and Amazon. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the sponsors.

References

[1] Stability AI. Stable diffusion 2, . Stable Diffusion 2. 7, 3

[2] Stability AI. Stable diffusion 3 medium, . Stable Diffusion 3 Medium. 3

[3] Stability AI. Stable diffusion 3.5 large, . Stable Diffusion 3.5 Large. 7, 3

[4] Stability AI. Stable diffusion 3.5 large turbo, . Stable Diffusion 3.5 Large Turbo. 7, 3

[5] Stability AI. Stable diffusion 3.5 medium, . Stable Diffusion 3.5 Medium. 7, 3

[6] Stability AI. Stable diffusion xl base 0.9, . Stable Diffusion XL Base 0.9. 7, 3

[7] AIML-TUDA. Llavaguard v1.2-7b-ov. LlavaGuard v1.2-7B-OV. 6

[8] Microsoft Azure. Microsoft azure content moderator, 2024. Microsoft Azure Content Moderator. 5

[9] Hritik Bansal, Da Yin, Masoud Monajatipoor, and Kai-Wei Chang. How well can text-to-image generative models understand ethical natural language interventions? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1358–1370, 2022. 1

[10] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22563–22575, 2023. 1

[11] Lucas Bourtoule, Varun Chandrasekaran, Christopher A Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. Machine unlearning. In *2021 IEEE symposium on security and privacy (SP)*, pages 141–159. IEEE, 2021. 2

[12] 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 *International Conference on Machine Learning*, 2024. 1

[13] Civitai. Civitai, 2024. Civitai. 3, 2

[14] DistilRoberta NSFW Classifier. distilroberta, 2024. distilroberta. 5

[15] DistilBERT NSFW Text Classifier. distilbert-nsfw-text-classifier, 2024. distilbert-nsfw-text-classifier. 5

[16] Google Cloud. google text moderation, 2024. google. 5

[17] CompVis. Safety checker nested in stable diffusion, 2021. Accessed: March 5, 2025. 1, 2

[18] CompVis. Stable diffusion safety checker (hugging face), 2024. Stable Diffusion Safety Checker (Hugging Face). 6

[19] Yimo Deng and Huangxun Chen. Harnessing llm to attack llm-guarded text-to-image models. *arXiv e-prints*, pages arXiv–2312, 2023. 2, 3, 7, 1

[20] Yingkai Dong, Zheng Li, Xiangtao Meng, Ning Yu, and Shanqing Guo. Jailbreaking text-to-image models with llm-based agents. *arXiv preprint arXiv:2408.00523*, 2024. 1

[21] Rohit Gandikota, Joanna Materzynska, Jaden Fiotto-Kaufman, and David Bau. Erasing concepts from diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2426–2436, 2023. 2

[22] Sensen Gao, Xiaojun Jia, Yihao Huang, Ranjie Duan, Jindong Gu, Yang Bai, Yang Liu, and Qing Guo. Hts-attack: Heuristic token search for jailbreaking text-to-image models. *arXiv preprint arXiv:2408.13896*, 2024. 1, 2

[23] Daya Guo, Qihao Zhu, Dejian Yang, Zhenda Xie, Kai Dong, Wentao Zhang, Guanting Chen, Xiao Bi, Yu Wu, YK Li, et al. Deepseek-coder: When the large language model meets programming–the rise of code intelligence. *arXiv preprint arXiv:2401.14196*, 2024. 2, 1

[24] Xiaoxuan Han, Songlin Yang, Wei Wang, Yang Li, and Jing Dong. Probing unlearned diffusion models: A transferable adversarial attack perspective. *arXiv preprint arXiv:2404.19382*, 2024. 1

[25] Yihao Huang, Le Liang, Tianlin Li, Xiaojun Jia, Run Wang, Weikai Miao, Geguang Pu, and Yang Liu. Perception-guided jailbreak against text-to-image models. *arXiv preprint arXiv:2408.10848*, 2024. 2, 7, 1, 3

[26] Tencent Hunyuan. Distillation. Distillation. 7, 3

[27] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 5

[28] Hakan Inan, Kartikeya Upasani, Jianfeng Chi, Rashi Rungta, Krithika Iyer, Yuning Mao, Michael Tontchev, Qing Hu, Brian Fuller, Davide Testuggine, et al. Llama guard: Llm-based input-output safeguard for human-ai conversations. *arXiv preprint arXiv:2312.06674*, 2023. 5

[29] Dongzhi Jiang, Guanglu Song, Xiaoshi Wu, Renrui Zhang, Dazhong Shen, Zhuofan Zong, Yu Liu, and Hongsheng Li. Comat: Aligning text-to-image diffusion model with image-to-text concept matching. *Advances in Neural Information Processing Systems*, 37:76177–76209, 2024. 1

[30] Xiaolong Jin, Zhuo Zhang, and Xiangyu Zhang. Multiverse: Exposing large language model alignment problems in diverse worlds. *arXiv preprint arXiv:2402.01706*, 2024. 2

[31] Nupur Kumari, Bingliang Zhang, Sheng-Yu Wang, Eli Shechtman, Richard Zhang, and Jun-Yan Zhu. Ablating concepts in text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22691–22702, 2023. 2

[32] Black Forest Labs. Flux.1-dev. FLUX.1-dev. 7, 3

[33] Tony Lee, Michihiro Yasunaga, Chenlin Meng, Yifan Mai, Joon Sung Park, Agrim Gupta, Yunzhi Zhang, Deepak

Narayanan, Hannah Teufel, Marco Bellagente, et al. Holistic evaluation of text-to-image models. *Advances in Neural Information Processing Systems*, 36:69981–70011, 2023. 2

[34] Guanlin Li, Kangjie Chen, Shudong Zhang, Jie Zhang, and Tianwei Zhang. Art: Automatic red-teaming for text-to-image models to protect benign users. *arXiv preprint arXiv:2405.19360*, 2024. 1

[35] Michelle Jie Li. Nsfw text classifier (hugging face), 2024. NSFW Text Classifier (Hugging Face). 5

[36] Pingzhi Li, Xiaolong Jin, Yu Cheng, and Tianlong Chen. Examining post-training quantization for mixture-of-experts: A benchmark. *arXiv preprint arXiv:2406.08155*, 2024. 2, 1

[37] Xinfeng Li, Yuchen Yang, Jiangyi Deng, Chen Yan, Yanjiao Chen, Xiaoyu Ji, and Wenyuan Xu. Safegen: Mitigating unsafe content generation in text-to-image models. *arXiv e-prints*, pages arXiv–2404, 2024. 2

[38] Han Liu, Yuhao Wu, Shixuan Zhai, Bo Yuan, and Ning Zhang. Riatig: Reliable and imperceptible adversarial text-to-image generation with natural prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20585–20594, 2023. 1

[39] 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*, pages 93–109. Springer, 2024. 1

[40] Xiaogeng Liu, Nan Xu, Muhao Chen, and Chaowei Xiao. Autodan: Generating stealthy jailbreak prompts on aligned large language models. In *The Twelfth International Conference on Learning Representations*, 2024. 2, 1

[41] Mengyao Lyu, Yuhong Yang, Haiwen Hong, Hui Chen, Xuan Jin, Yuan He, Hui Xue, Jungong Han, and Guiguang Ding. One-dimensional adapter to rule them all: Concepts diffusion models and erasing applications. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7559–7568, 2024. 2

[42] Yizhuo Ma, Shanmin Pang, Qi Guo, Tianyu Wei, and Qing Guo. Coljailbreak: Collaborative generation and editing for jailbreaking text-to-image deep generation. *Advances in Neural Information Processing Systems*, 37:60335–60358, 2025. 2, 1

[43] Microsoft. Azure image moderation api. Azure Image Moderation API. 6

[44] OpenAI. openai moderation, 2024. openai moderation. 5, 6

[45] Benjamin Paine. Stable diffusion v1.5. Stable Diffusion v1.5. 7, 3

[46] Antonis Papasavva, Savvas Zannettou, Emiliano De Cristofaro, Gianluca Stringhini, and Jeremy Blackburn. Raiders of the lost kek: 3.5 years of augmented 4chan posts from the politically incorrect board. In *Proceedings of the international AAAI conference on web and social media*, pages 885–894, 2020. 2

[47] Duo Peng, Qiuhong Ke, and Jun Liu. Upam: Unified prompt attack in text-to-image generation models against both textual filters and visual checkers. In *International Conference on Machine Learning 2024*, pages 40200–40214. Proceedings of Machine Learning Research (PMLR), 2024. 1

[48] Samuele Poppi, Tobia Poppi, Federico Cocchi, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Safe-clip: Removing nsfw concepts from vision-and-language models. In *European Conference on Computer Vision*, pages 340–356. Springer, 2024. 2

[49] Huming Qiu, Guanxu Chen, Mi Zhang, and Min Yang. Safe text-to-image generation: Simply sanitize the prompt embedding. *arXiv preprint arXiv:2411.10329*, 2024. 1, 2

[50] Yiting Qu, Xinyue Shen, Xinlei He, Michael Backes, Savvas Zannettou, and Yang Zhang. Unsafe diffusion: On the generation of unsafe images and hateful memes from text-to-image models. In *Proceedings of the 2023 ACM SIGSAC conference on computer and communications security*, pages 3403–3417, 2023. 3, 6, 1, 2

[51] Yiting Qu, Xinyue Shen, Yixin Wu, Michael Backes, Savvas Zannettou, and Yang Zhang. Unsafebench: Benchmarking image safety classifiers on real-world and ai-generated images. *arXiv preprint arXiv:2405.03486*, 2024. 2

[52] Jessica Quaye, Alicia Parrish, Oana Inel, Charvi Rastogi, Hannah Rose Kirk, Minsuk Kahng, Erin Van Liemt, Max Bartolo, Jess Tsang, Justin White, et al. Adversarial nibbler: An open red-teaming method for identifying diverse harms in text-to-image generation. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 388–406, 2024. 1

[53] Jessica Quaye, Alicia Parrish, Oana Inel, Charvi Rastogi, Hannah Rose Kirk, Minsuk Kahng, Erin Van Liemt, Max Bartolo, Jess Tsang, Justin White, et al. Adversarial nibbler: An open red-teaming method for identifying diverse harms in text-to-image generation. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 388–406, 2024. 1

[54] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pages 8821–8831. Pmlr, 2021. 2

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

[56] Qibing Ren, Chang Gao, Jing Shao, Junchi Yan, Xin Tan, Wai Lam, and Lizhuang Ma. Codeattack: Revealing safety generalization challenges of large language models via code completion. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 11437–11452, 2024. 2, 1

[57] 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*, pages 10684–10695, 2022. 1

[58] Patrick Schramowski, Christopher Tauchmann, and Kristian Kersting. Can machines help us answering question 16 in datasheets, and in turn reflecting on inappropriate content? In *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency*, pages 1350–1361, 2022. 2

[59] Patrick Schramowski, Christopher Tauchmann, and Kristian Kersting. Can machines help us answering question 16 in

datasheets, and in turn reflecting on inappropriate content? In *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency*, pages 1350–1361, 2022. 6

[60] Patrick Schramowski, Manuel Brack, Björn Deiseroth, and Kristian Kersting. Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22522–22531, 2023. 2, 3

[61] Patrick Schramowski, Manuel Brack, Björn Deiseroth, and Kristian Kersting. Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22522–22531, 2023. 1, 2

[62] Guangyu Shen, Siyuan Cheng, Zhuo Zhang, Guanhong Tao, Kaiyuan Zhang, Hanxi Guo, Lu Yan, Xiaolong Jin, Shengwei An, Shiqing Ma, et al. Bait: Large language model backdoor scanning by inverting attack target. In *2025 IEEE Symposium on Security and Privacy (SP)*, pages 1676–1694. IEEE, 2025. 2

[63] Xinyue Shen, Yiting Qu, Michael Backes, and Yang Zhang. Prompt stealing attacks against {Text-to-Image} generation models. In *33rd USENIX Security Symposium (USENIX Security 24)*, pages 5823–5840, 2024. 1

[64] THUDM. Cogview4. CogView4. 7, 3

[65] 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*. 2, 7, 1, 3

[66] Kun Wang, Guibin Zhang, Zhenhong Zhou, Jiahao Wu, Miao Yu, Shiqian Zhao, Chenlong Yin, Jinhu Fu, Yibo Yan, Hanjun Luo, et al. A comprehensive survey in llm (-agent) full stack safety: Data, training and deployment. *arXiv preprint arXiv:2504.15585*, 2025. 2

[67] Yiming Wang, Jiahao Chen, Qingming Li, Xing Yang, and Shouling Ji. Aeiou: A unified defense framework against nsfw prompts in text-to-image models. *arXiv preprint arXiv:2412.18123*, 2024. 1, 2

[68] Zijie Wang, Evan Montoya, David Munechka, Haoyang Yang, Benjamin Hoover, and Polo Chau. Diffusiondb: A large-scale prompt gallery dataset for text-to-image generative models. In *Annual Meeting of the Association for Computational Linguistics*, 2023. 3, 2

[69] Zixuan Weng, Xiaolong Jin, Jinyuan Jia, and Xiangyu Zhang. Foot-in-the-door: A multi-turn jailbreak for llms. *arXiv preprint arXiv:2502.19820*, 2025. 2, 1

[70] Jinbo Xing, Menghan Xia, Yuxin Liu, Yuechen Zhang, Yong Zhang, Yingqing He, Hanyuan Liu, Haoxin Chen, Xiaodong Cun, Xintao Wang, et al. Make-your-video: Customized video generation using textual and structural guidance. *IEEE Transactions on Visualization and Computer Graphics*, 2024. 1

[71] Zhen Xing, Qi Dai, Han Hu, Zuxuan Wu, and Yu-Gang Jiang. Simda: Simple diffusion adapter for efficient video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7827–7839, 2024. 1

[72] Yijun Yang, Ruiyuan Gao, Xiao Yang, Jianyuan Zhong, and Qiang Xu. Guardt2i: Defending text-to-image models from adversarial prompts. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*. 2, 1

[73] Yijun Yang, Ruiyuan Gao, Xiaosen Wang, Tsung-Yi Ho, Nan Xu, and Qiang Xu. Mma-diffusion: Multimodal attack on diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7737–7746, 2024. 2, 3, 4, 7, 1

[74] 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)*, pages 897–912. IEEE, 2024. 2, 3, 5, 7, 1

[75] Chenyu Zhang, Mingwang Hu, Wenhui Li, and Lanjun Wang. Adversarial attacks and defenses on text-to-image diffusion models: A survey. *Inf. Fusion*, 114:102701, 2025. 1

[76] 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*, pages 385–403. Springer, 2024. 2