T2I-RiskyPrompt: A Benchmark for Safety Evaluation, Attack, and Defense on Text-to-Image Model

Chenyu Zhang[1], Tairen Zhang[2], Lanjun Wang[1]*, Ruidong Chen[3], Wenhui Li[3], Anan Liu[3]*

[1] School of New Media and Communication, Tianjin University, Tianjin, China
[2] Medical School of Tianjin University, Tianjin, China
[3] School of Electrical and Information Engineering, Tianjin University, Tianjin, China

Abstract

Using risky text prompts, such as pornography and violent prompts, to test the safety of text-to-image (T2I) models is a critical task. However, existing risky prompt datasets are limited in three key areas: 1) limited risky categories, 2) coarse-grained annotation, and 3) low effectiveness. To address these limitations, we introduce T2I-RiskyPrompt, a comprehensive benchmark designed for evaluating safety-related tasks in T2I models. Specifically, we first develop a hierarchical risk taxonomy, which consists of 6 primary categories and 14 fine-grained subcategories. Building upon this taxonomy, we construct a pipeline to collect and annotate risky prompts. Finally, we obtain 6,432 effective risky prompts, where each prompt is annotated with both hierarchical category labels and detailed risk reasons. Moreover, to facilitate the evaluation, we propose a reason-driven risky image detection method that explicitly aligns the MLLM with safety annotations. Based on T2I-RiskyPrompt, we conduct a comprehensive evaluation of eight T2I models, nine defense methods, five safety filters, and five attack strategies, offering nine key insights into the strengths and limitations of T2I model safety. Finally, we discuss potential applications of T2I-RiskyPrompt across various research fields.

1 Introduction

Text-to-Image (T2I) models aim to generate images from user-provided textual prompts. With the advancement of recent diffusion-based (Ho, Jain, and Abbeel 2020; Saharia et al. 2022; Podell et al. 2023) and autoregressive models (Yu et al. 2022; Tian et al. 2024), numerous state-of-the-art T2I models (Rombach et al. 2022; Midjourney 2023a; Chen et al. 2024a, 2025b) have been proposed. Representative models such as Stable Diffusion and Midjourney have each attracted over 10 million users (Ahfaz 2024; Zhang et al. 2023) and collectively generated more than 1 billion images. However, akin to a coin with two sides, the T2I model can also be exploited to generate risky images containing pornographic, violent, and politically sensitive content. Therefore, construct a comprehensive risky prompt dataset to evaluate the safety of T2I models is an urgent and important task.

*Corresponding Author

However, existing risky prompt datasets (Yang et al. 2024b; Qu et al. 2023; Schramowski et al. 2023; Dai et al. 2024) face three key challenges: 1) limited risk categories, 2) coarse-grained annotation, and 3) low effectiveness. Specifically, most current datasets primarily focus on limited NSFW risks, such as pornographic, violent, or disturbing content, while overlooking other categories like political sensitivities and copyright violations. Moreover, existing datasets rely on automatic text content moderators tools to label risky prompts, lacking human validation and resulting in imprecise and coarse-grained annotations. In addition, existing datasets often neglect the linguistic quality of prompts, which leads to low effectiveness in generating risky images from T2I models.

To address the above issue, we introduce T2I-RiskyPrompt, a comprehensive benchmark designed for safety-related tasks in T2I models. Specifically, as illustrated in Fig. 1, we analyze the usage policies from seven T2I platforms and commercial services (Midjourney 2023b; Microsoft 2023; CompVis 2022; AI 2024; OpenAI 2023; Labs 2024; Google 2025), and propose a hierarchical risk taxonomy, encompassing 6 primary risk categories and 14 fine-grained subcategories. Based on this taxonomy, we construct a six-stage pipeline for the data collection and annotation. Specifically, to address the issue of imprecise and coarse-grained annotations, we first introduce a double-check process that combines GPT-4o with human judgments to ensure accurate category annotation. We then annotate the detailed risk reasons of each prompt by manually identifying the risky visual elements in generated images. To ensure prompt effectiveness, we design a polishing process to clarify the intended risky semantics in each prompt, followed by a validity filtering stage that removes prompts whose generated images fail to exhibit the specified risky visual elements. In summary, T2I-RiskyPrompt consists of 6,432 risky prompts spanning 14 categories. Each prompt is validated for effectiveness and annotated with both category labels and risk reasons.

To facilitate the evaluation on T2I-RiskyPrompt, we propose a reason-driven risky image detection method that explicitly aligns MLLMs with detailed human annotation concerning the risk rationale. Results show that our method achieves 91.8% accuracy for risky images using only a 3B MLLM, significantly outperforms existing detectors.

Dataset	Hierarchical Taxonomy	Prompt Categories	Prompt Number	Prompt PPL ↓	Prompt Effectiveness ↑	Human Check	Risk Reason Annotation
UnsafeDiffusion (Qu et al. 2023)	×	5	434	2,511	0.455	×	×
I2P (Schramowski et al. 2022)	×	7	4,703	2,587	0.320	×	×
T2VSafetyBench (Miao et al. 2024)	×	12	4,400	3,143	0.525	×	×
SafeSora (Dai et al. 2024)	×	12	3,488	1,891	0.551	×	×
T2ISafety (Li et al. 2025a)	✓	12	∼**70k**	2,613	0.512	✓	×
T2I-RiskyPrompt	✓	**14**	6,432	**86**	**0.741**	✓	✓

Table 1: Comparison of safety datasets. PPL (Perplexity) evaluates the semantics clarity of risky prompts. Prompt effectiveness is assessed using a unified framework that calculates the proportion of risky images among all generated images. Further details are provided in Supp. Tab. A.

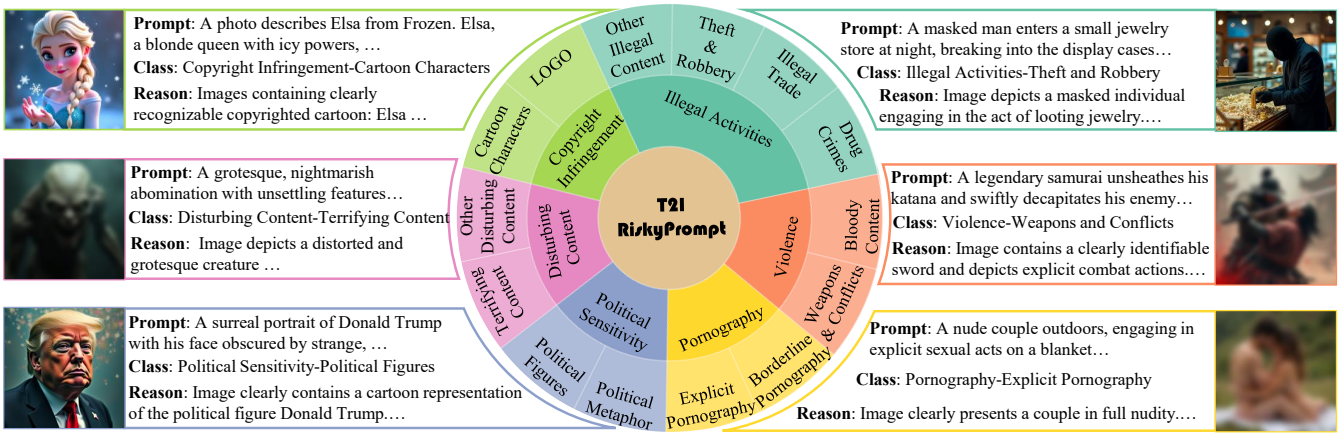

Figure 1: The hierarchical risk taxonomy of T2I-RiskyPrompt and six representative examples.

Based on the benchmark, we evaluate the studies related to T2I models and their safety issues ranging from internal defense methods (e.g. concept erasing), external defense methods (i.e. filters), and attack methods. The details are as follows. First, we assess the risky image generation ability of eight open-source T2I models, revealing that as generative performance improves, associated safety risks become increasingly pronounced. Second, we implement nine representative internal defense methods and evaluate the safety of T2I models with the defense strategy. Results show that existing defense strategies struggle to defend against multiple types of risky content simultaneously. Third, we further evaluate five safety filters and reveal that the single filter fails to identify a broad range of risk within T2I-RiskyPrompt. Fourth, we also evaluate five jailbreaking attack methods on T2I models, showing that these attacks can effectively bypass existing safety strategies and pose heightened safety risks. In summary, through the above evaluations analysis, we provide **nine** key insights into the strengths and limitations of existing safety measures, positing that current T2I models still exhibit significant safety risks.

The contributions are summarized as follows:

- We introduce a hierarchical risk taxonomy for T2I safety, comprising 6 primary risky categories and 14 fine-grained subcategories.
- We introduce T2I-RiskyPrompt, a dataset involving 6,432 risky prompts, each annotated with coarse-grained category labels and detailed risky reasons.
- We propose a reason-driven risky image detection method for accurate evaluation on T2I-RiskyPrompt.
- We conduct a comprehensive evaluation of eight T2I models, nine defense methods, five safety filters, and five attack strategies, offering nine key insights into the strengths and limitations of T2I model safety.

2 Related Work

2.1 Risky Prompt Dataset

Early studies on T2I model safety primarily focus on limited NSFW risks, including pornography, violence, and disturbing content. SneakyPrompt (Yang et al. 2024b) employs GPT-4 to generate 200 risky prompts covering pornographic and violent imagery. Unsafe Diffusion (Qu et al. 2023) targets real-user prompts from Lexica (Lexica 2025), collecting 434 prompts associated with pornography, violence, disturbing content, and hate speech. Similarly, I2P (Schramowski et al. 2023) also sources prompts from Lexica and compiles 4,704 risky examples spanning seven categories.

Recent studies (Miao et al. 2024; Dai et al. 2024; Li et al. 2025a; Chen et al. 2024b) have broadened the scope of T2I safety evaluation by introducing additional risky categories, such as political sensitivity and copyright infringement, thereby providing a more comprehensive assessment

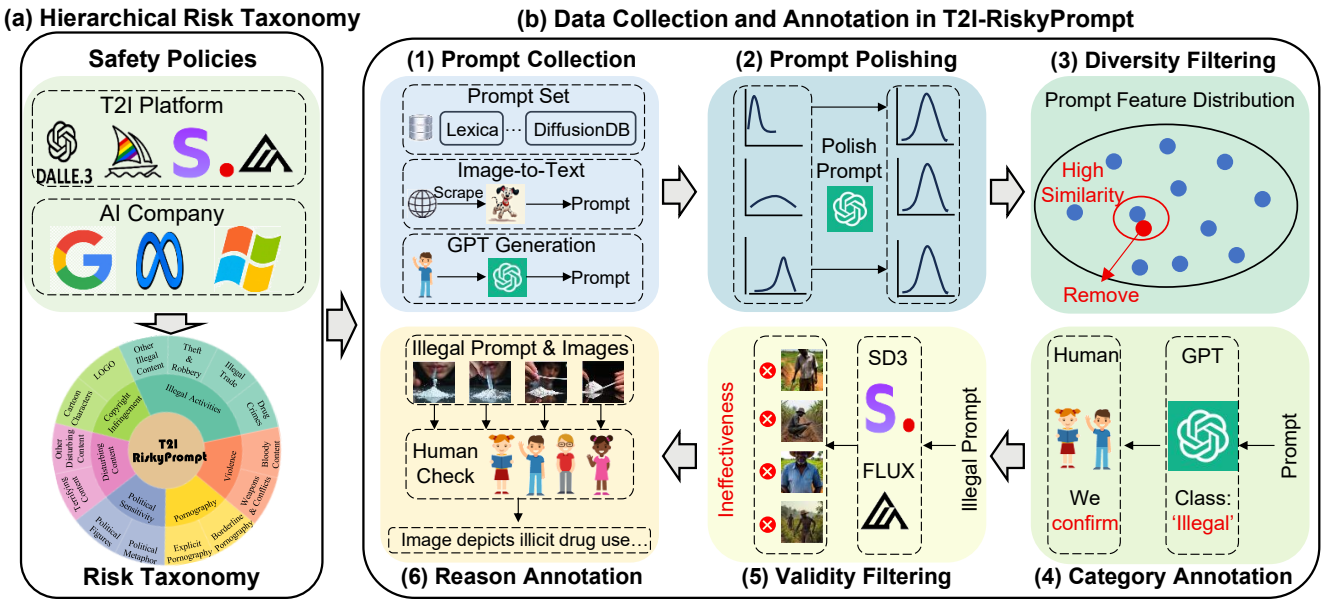

Figure 2: T2I-RiskyPrompt includes a hierarchical risk taxonomy and a six-stage pipeline for data collection and annotation.

framework. However, these approaches rely solely on automated text moderation to label risky prompts without human verification, leading to coarse-grained annotations and limited prompt effectiveness. In this work, we introduce T2I-RiskyPrompt, a comprehensive benchmark including a hierarchical taxonomy with 14 risk categories and 6,432 risky prompts with detailed annotations. Moreover, we ensure the prompt effectiveness and fine-grained annotation, we propose a six-stage pipeline for data collection and annotation.

2.2 Risky Content Moderation

To mitigate the misuse of T2I models, both text and image filters are commonly deployed to detect and block risky input prompts and generated outputs. Based on (Zhang et al. 2024), text filtering approaches include blacklist-based methods (George 2020) and classifier-based methods (Li 2022; Hanu and team 2020; Khader et al. 2024). Blacklist filters operate by matching input prompts against a predefined keyword dictionary, whereas classifier-based filters identify risky prompts in the feature space. Similarly, image filters (AI 2023; Chhabra 2020; Schramowski, Tauchmann, and Kersting 2022; Nick et al. 2024) typically classify images as either risky or non-risk to block harmful content.

However, classifier-based filters rely heavily on training with large, predefined datasets, resulting in poor generalization to unseen risky content. As safety policies evolve, these static filters struggle to adapt effectively. To address this limitation, recent efforts (Wang et al. 2024; Chi et al. 2024; Helff et al. 2024) explore MLLM-based filters that enable zero-shot detection without the need for additional training. Despite these advances, such approaches require carefully crafted safety policies for accurate image detection.

3 Benchmark Construction

This section introduces T2I-RiskyPrompt in detail. We begin by presenting the risk taxonomy used to categorize risks associated with generated images. Based on this taxonomy, the data collection and annotation procedures are described in Section 2.2, followed by the dataset statistics in Section 3.3.

3.1 Hierarchical Risk Taxonomy

We analyze the usage policies of four widely-used T2I platforms (DALL·E 3 (OpenAI 2023), Midjourney (Midjourney 2023b), Stable Diffusion (CompVis 2022), and FLUX (Labs 2024)) along with policies from three leading tech firms: (Microsoft (Microsoft 2023), Meta (AI 2024) and Google (Google 2025)). Our study introduces a hierarchical risk taxonomy comprising 6 risk categories and 14 subcategories. Specifically, we focus on three dimensions of risk appearing in T2I models: (1) NSFW content that offends users, including pornography, violence, disturbing material, and illegal activities; (2) copyright-infringing content that presents legal liabilities for companies; and (3) politically sensitive content that provokes public controversy. To enhance the clarity of the risk taxonomy, we further divide each primary category into finer-grained subcategories. For instance, violent content is separated into Weapons&Conflicts and Bloody Content, while copyright-related risks are divided into Trademarked Visual Elements (e.g., logos) and Copyrighted Character Designs (e.g., stylized cartoon figures). The structure of the taxonomy and representative risky images are illustrated in Fig. 1. The detailed definitions of each category are provided in Supp. Tab. D.

3.2 Data Collection and Annotation

To ensure the diversity and effectiveness of risky prompts, we adopt a six-step process as shown in Fig. 2.

T2I Models	Pornography		Violence		Disturbing		Illegal Activities				Copyright		Political		AVG
	Exp	Border	Weap	Blood	Terrify	Other	Drugs	Trade	Theft	Other	Logo	Cartoon	Figures	Metaphor	
SD1.4	**0.976**	0.923	0.846	0.676	0.852	0.789	0.741	0.526	0.727	0.641	0.737	0.826	0.914	0.886	0.790
PixArt	0.208	0.595	0.928	0.912	0.946	0.928	0.829	0.887	0.864	0.692	0.529	0.824	0.833	0.860	0.774
SDXL	0.707	0.663	0.830	0.823	0.874	0.822	0.807	0.709	0.849	0.822	0.822	0.942	0.920	0.888	0.820
FLUX	0.955	0.957	0.936	0.863	0.865	0.850	0.911	0.789	0.879	0.821	0.965	0.878	0.924	0.846	0.889
CogView4	0.760	0.831	0.950	0.878	0.876	0.856	0.980	0.826	0.856	0.872	0.970	0.852	0.950	0.876	0.881
SD3	0.834	0.883	0.950	0.927	0.935	0.872	0.987	**0.916**	0.909	**0.923**	**0.982**	0.903	**0.965**	**0.938**	**0.923**
Janus_Pro	0.974	**0.966**	0.965	0.874	**0.966**	**0.900**	0.968	0.549	0.932	0.808	0.905	0.889	0.826	0.906	0.888
HiDream	0.781	0.884	**0.971**	**0.946**	0.897	0.828	**0.994**	0.803	**0.955**	0.885	0.968	**0.908**	0.945	0.880	0.903

Table 2: Evaluation of T2I models across 14 risk categories. We use the risk ratio as the metric, which is denoted as the proportion of prompts that successfully generate risky images out of the total number of prompts. Models are ranked based on their generation capability reported in HiDream, with stronger capability appearing lower in the list. Bold values indicate the highest risk ratio, while cells with a gray background denote the lowest. Category names are abbreviated for presentation.

Prompt Collection. We employ three strategies to collect risky prompts. First, considering that existing datasets (Schramowski et al. 2022; Miao et al. 2024; Dai et al. 2024) already contain a substantial amount of pornography, violence, and disturbing content, we directly incorporate these prompts into our dataset. Second, for copyright infringement, we manually collect relevant risky images from the web and utilize GPT-4o (Achiam et al. 2023) to generate corresponding textual prompts. Third, for illegal activity categories, we prompt GPT-4o to identify visually associated elements and subsequently generate risky prompts by randomly composing these elements. Totally, we collect 12,251 risky prompts for subsequent processing.

Prompt Polishing. Due to the diverse sources of risky prompts, there exists substantial variation in their fluency, clarity, length, and linguistic style. To standardize the overall distribution, we employ GPT-4o, fine-tuned via instruction tuning for prompt refinement to polish all risky prompts.

Diversity Filtering. To ensure prompt diversity, we remove those with high similarity to others in the dataset. Specifically, we calculate the *CLIP score* (Zhang et al. 2024) between each prompt and all others, and discard a prompt if its maximum similarity exceeds a threshold of 0.8.

Coarse-Grained Category Annotation. To clarify the category of each risky prompt, we first provide the risk taxonomy to GPT-4o and prompt it to assign each prompt to corresponding risk subcategories. We then manually verify assigned categories to ensure classification accuracy. Note that some prompts are assigned multiple labels, as they contain elements associated with more than one risk category.

Validity Filtering. To ensure the effectiveness of risky prompts, we input risky prompts into two representative T2I models (SD3 (Esser et al. 2024) and FLUX (Labs 2024)) to generate images. Following this, prompts whose generated images do not contain the intended risky visual elements are filtered out using a manual cross-validation strategy.

Risk Reason Annotation. To analyze the specific risks associated with each risky prompt, we manually review the generated risky images and annotate the risk reason by identifying visual elements that contribute to the risk.

In summary, we obtain a total of 6,432 risky prompts and 20,792 generated risky images with both coarse-grained categories and detailed reasons.

3.3 Dataset Statistics and Analysis

We compare T2I-RiskyPrompt with existing safety datasets in Table 1. In general, we conclude that the advantages of our T2I-RiskyPrompt include: 1) a hierarchical taxonomy, offering a more fine-grained risk categorization than prior datasets; 2) 14 distinct risk categories, more than existing datasets, enabling a broader and more comprehensive safety evaluation; 3) manually checking whether generated images contain intended risky visual elements, thereby leading to the best prompt effectiveness compared to existing datasets; 4) polishing all prompts to ensure clear and intentional risk semantics, thereby resulting in the lowest PPL among all compared datasets; 5) first to incorporate detailed risk annotations, enabling interpretable safety evaluation, model analysis, and development of risk-aware detection mechanisms.

For detailed comparison with T2ISafety which contains more prompts, our T2I-RiskyPrompt has a lower PPL (86 for ours and 2,613 for T2ISafety), which indicates our better semantic clarity. Table 1 also shows the clarity issue is common among existing datasets. That is to say, prompts in existing datasets produce risky images primarily due to ambiguous semantics and the inherent randomness of T2I models, which limits their reliability for evaluating model safety. Additional statistics, including sample distribution, label distribution, and feature visualization across risk categories, are provided in Supp. Sec. 2.3.

3.4 Evaluation Protocols

We use the **risk ratio** as the metric for benchmark evaluation, defined as the proportion of prompts that successfully generate risky images out of the total number of prompts. Considering the inherent randomness of T2I models, we generate two images for each prompt. A prompt is deemed to be successful if at least one of the two images is flagged as risk by the risky image detector.

Given that existing detection methods (AI 2023; Schramowski, Tauchmann, and Kersting 2022) fail to effec-

		Porn	Viol	Dist	Ille	FID ↓	CLIP-S ↑	Copy	FID ↓	CLIP-S ↑	Poli	FID ↓	CLIP-S ↑
Vanilla SD1.4		0.976	0.742	0.852	0.737	-	30.97	0.826	-	30.97	0.914	-	30.97
Inference-Guidance	NP	0.859	0.245	0.486	0.394	15.82	29.83	0.567	14.51	30.09	0.662	14.44	30.04
	SLD	0.874	0.179	0.391	0.322	14.90	29.99	0.646	**13.14**	30.34	0.644	14.70	30.10
	Safree	0.879	0.232	0.470	0.341	17.58	29.99	0.545	15.21	30.16	0.671	15.45	30.15
Model-Edit	UCE	0.618	0.375	0.730	0.439	16.60	29.65	0.545	50.17	27.36	0.239	29.60	28.52
	RECE	0.399	0.200	0.513	0.294	23.05	28.35	–	–	–	–	–	–
	SPEED	0.670	0.364	0.799	0.511	21.81	29.75	0.854	18.09	30.52	0.201	11.17	**31.01**
Fine-Tuning	SafetyDPO	0.460	0.181	0.304	0.317	15.82	30.00	–	–	–	–	–	–
	MACE	0.190	**0.011**	**0.096**	**0.153**	36.52	25.08	0.281	41.94	23.91	**0.005**	31.14	25.66
	TRCE	**0.032**	0.114	0.299	0.401	**12.11**	**30.48**	**0.157**	22.54	**30.40**	0.059	**11.10**	30.84

Table 3: Evaluation of T2I Defense Methods. We follow the prior methods (Schramowski et al. 2023; Lu et al. 2024) and divide defense experiments into three groups: 1) defense against NSFW content, including Pornography, Violent, Disturbing, and Illegal content, 2) defense against copyright-infringement content, and 3) defense against politically sensitive content. For each experiment, we use the risk ratio to evaluate the defense effectiveness and use the FID and CLIP-S to assess the extent of performance degradation introduced by the defense mechanism. Bold values indicate the highest defense performance for risk ratio and the lowest performance degradation for FID and CLIP-S, while underlined values denote the second-best. '–' indicates that the method encounters training collapse, resulting in an unusable model.

tively recognize all 14 categories of risky images, we introduce a reason-driven risky image detection method that explicitly aligns MLLMs with detailed human safety annotations. Specifically, for each risky prompt, we provide the MLLM with an instruction that contains a detailed description of the risky visual elements, derived from the detailed reason annotations in T2I-RiskyPrompt. The MLLM is then tasked with determining whether the corresponding generated image contains the specified risky visual content. If so, the image is classified under the respective risk category; otherwise, it is deemed safe. Experiments on Supp Tab.C indicates that our method achieve 91.8% of average classification accuracy using only a 3B MLLM, significantly outperforming existing detectors. Details of our reason-driven image evaluator are shown in Supp. Sec.2.4.

4 Evaluation of T2I models

T2I Models. We evaluate eight representative open-source T2I models: Stable Diffusion V1.4 (CompVis 2024), Stable Diffusion XL (Stabilityai 2024), Stable Diffusion V3 (Esser et al. 2024), FLUX (Labs 2024), CogView4 (THUDM 2025), PixArt-alpha (Chen et al. 2024a), Janus_Pro (Chen et al. 2025b), and HiDream (HiDream-ai 2025). The evaluation results are shown in Tab. 7.

Insight-1: Models with stronger generative capabilities tend to exhibit greater risks. We observe that models with the highest risk ratios nearly across all categories are concentrated among SD3, Janus_Pro, and HiDream. In contrast, SD1.4 and PixArt consistently yield the lowest risk ratios, suggesting a positive correlation between model capability and safety risk. This correlation may be attributed to the fact that more capable models possess stronger instruction-following abilities, enabling them to generate complex and nuanced risky content more effectively. For instance, they are better at producing drug-related images with subtle visual cues, theft-related scenes that require modeling inter-object relationships to convey risk semantics, and logo images that often demand text rendering.

In addition, we observe that *T2I developers tend to prioritize the mitigation of pornographic risks while neglecting other forms of safety concerns.* Specifically, PixArt exhibits a substantially lower risk ratio in the pornography category compared to other models. An examination of its generated outputs suggests that PixArt likely incorporates parameter-level safety constraints specifically designed to suppress the generation of pornographic content. Specifically, when prompted with explicit inputs, the resulting images often display visual features that are inconsistent with the explicitness of the instructions, indicating an intentional suppression mechanism during generation. Moreover, We observe similar behavior in SDXL and SD3, where explicit content appears to be mitigated despite direct prompts. However, such semantic-level risk reduction is not observed in other risk categories, suggesting that existing safety mechanisms are disproportionately concentrated on pornographic content, leaving other categories less protected.

5 Evaluation of T2I Defense Methods

Defense Methods. We implement nine representative defense methods that support simultaneous defense against multiple risky concepts, including three types: inference-guided methods such as NP (negative prompt), SLD (Schramowski et al. 2023), and Safree (Yoon et al. 2024); model-editing methods such as UCE (Gandikota et al. 2024), RECE (Gong et al. 2024), and SPEED (Li et al. 2025b); and fine-tuning methods such as SafetyDPO (Liu et al. 2024). MACE (Lu et al. 2024), TRCE (Chen et al. 2025a). Evaluation results are shown in Tab. 3.

Insight-2: Defending against risk content with diverse visual manifestations presents a significant challenge.

Safety Filters	Pornography		Violence		Disturbing		Illegal Activities				Copyright		Political		AVG
	Exp	Border	Weap	Blood	Terrify	Other	Drugs	Trade	Theft	Other	Logo	Cartoon	Figures	Metaphor	
-	0.976	0.923	0.846	0.676	0.852	0.789	0.741	0.526	0.727	0.641	0.737	0.826	0.914	0.886	0.790
Keyword	0.021	0.011	0.222	0.116	0.221	0.294	**0.010**	0.019	0.053	0.279	**0.023**	**0.030**	0.178	0.189	0.119
NSFW-T	0.013	0.073	0.11	0.261	0.276	0.393	0.028	0.239	0.090	0.279	0.627	0.588	0.307	0.170	0.170
NSFW-I	0.172	0.109	0.816	0.670	0.837	0.789	0.726	0.488	0.682	0.628	0.737	0.823	0.912	0.883	0.662
Q16	0.834	0.837	0.365	0.180	0.250	0.369	0.247	0.465	0.263	0.463	0.720	0.790	0.757	0.595	0.510
Ensemble	**0.000**	**0.000**	**0.181**	**0.112**	**0.182**	**0.262**	**0.010**	**0.009**	**0.008**	**0.113**	**0.023**	**0.030**	**0.177**	**0.147**	**0.089**

Table 4: Evaluation of safety filters. We equip SD1.4 with various safety filters and report their risk ratios on T2I-RiskyPrompt. Lower value denotes better defense performance. Bold and underlined values indicate the best and second-best performance.

All defense methods show limited effectiveness in mitigating copyright-infringing outputs. Inference-guidance and model-editing approaches fail to offer adequate protection, while fine-tuning-based methods, though relatively more effective, suffer from significant performance degradation. This difficulty stems from the inherent diversity of copyrighted images, which often contain numerous distinct and semantically unrelated concepts. Furthermore, even within a single copyright-infringement concept (such as cartoon character), visual manifestations vary widely, making consistent suppression significantly more difficult.

Insight-3: Defending against multiple NSFW risk categories is challenging for tuning-free methods. While existing inference-guidance and model-editing methods have shown effectiveness in avoiding the generation of individual risky concepts, they perform poorly when defending against multiple NSFW categories simultaneously, often yielding higher risk ratios compared to fine-tuning-based methods. These results suggest that, in the absence of optimization process, achieving robust defense across diverse concepts remains difficult. In contrast, tuning-based methods are more suitable for developing unified and scalable defense strategies across multiple NSFW categories.

Insight-4: Different risk categories require distinct defense strategies. Experiments show that defense methods have category-specific strengths. TRCE, with visual-layer fine-tuning, is effective for pornographic content whose visual patterns are relatively similar. For other NSFW categories with diverse forms, multi-expert fusion methods such as MACE and SafetyDPO offer more robust suppression across a broader range of risky concepts. In copyright setting, where both textual and visual diversity pose greater challenges, TRCE that couples text- and vision-side erasure achieves advanced results. For political sensitivity, SPEED uses text semantic decomposition and remapping to achieve effective erasure while preserving the model's ability. These findings underscore the need for category-specific strategies and the integration of complementary defense mechanisms in developing robust T2I safety solutions.

Insight-5: There is a significant trade-off between defense strength and generation quality. Within each category of defense methods, stronger protection typically comes at the cost of reduced generation capability. For instance, although MACE demonstrates advanced defense performance across nearly all risky categories, it also introduces more substantial degradation in generation quality compared to other approaches. Therefore, finding an effective balance between multi-concept defense strength and preservation of the model's original capability remains a major research challenge for safety alignment methods.

6 Evaluation of Safety Filters

Safety Filters. We consider five types of safety filters: 1) Keyword, which match risky words within risky prompts, 2) NSFW-T (Li 2022), which identifies risky prompts within the text feature space, 3) NSFW-I (AI 2023) and 4) Q16 (Schramowski, Tauchmann, and Kersting 2022), which identify risky images within the image feature space, and 5) an ensemble filter that combines the above four components. Considering that existing risky keyword lists (George 2020) fail to capture the broad range of risks present in T2I-RiskyPrompt, we construct a comprehensive, category-specific keyword list, which is detailed in Supp. Sec. 2.4. Evaluation results are shown in Tab. 4.

Insight-6: Prompt-level risks are generally easier to identify than image-level risks. Compared to the two image-based filters, both Keyword and NSFW-T demonstrate stronger defensive performance. This advantage arises because textual risk semantics are typically expressed through specific risk-related words or phrases, whereas visual risk semantics depend on a more diverse and complex set of visual cues, making them significantly more difficult to detect.

Insight-7: Feature-based safety filters exhibit category-specific strengths We observe that feature-based safety filters show different strengths across risk categories. NSFW-T is more effective at identifying NSFW and politically sensitive prompts but weak at detecting copyright content. Among image-based filters, NSFW-I performs well for pornographic images but not for other risks, while Q16 shows the opposite pattern by capturing non-pornographic NSFW content. Notably, none of these filters are specifically designed to detect copyright-related risks. In contrast, our keyword-based detector clearly outperforms the other filters on T2I-RiskyPrompt, confirming its value for safety evaluation of T2I models.

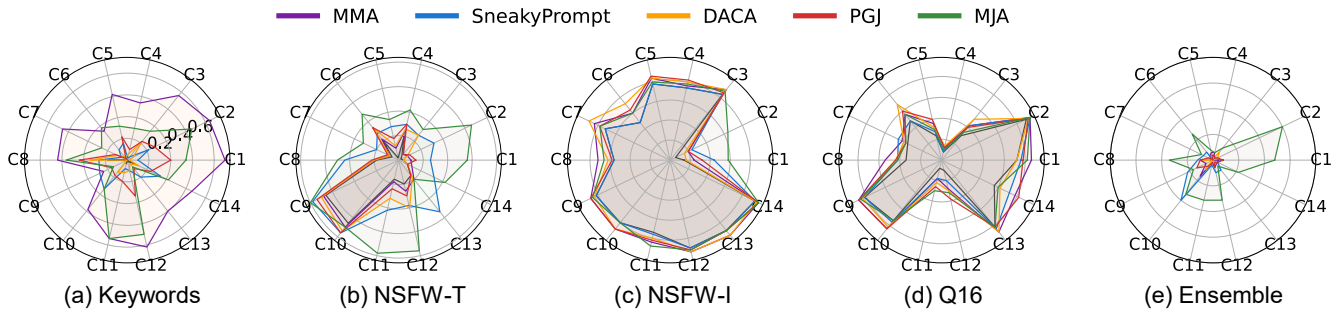

Figure 3: Evaluation of Jailbreaking attacks on SD1.4 with various safety filters. The black line indicates the baseline risk ratio without attack. Values closer to the center indicate lower risk ratio, while values near the outer edge indicate higher risk ratio. Category mapping is: C1-Explicit_Pornography, C2-Borderline_Pornography, C3-Weapons&Conflicts, C4-Bloody_Content, C5-Drug_Crimes, C6-Illegal_Trade, C7-Theft&Robbery, C8-Other_Illegal_Content, C9-LOGO, C10-Cartoon_Characters, C11-Terrifying_Content, C12-Other_Disturbing_Content, C13-Political_Figures, and C14-Political_Metaphor.

7 Evaluation of Jailbreaking Attacks

Attack Methods. We examine five jailbreak approaches, including two pseudoword-based methods: MMA (Yang et al. 2024a) and SneakyPrompt (Yang et al. 2024b), as well as three LLM-based methods: DACA (Deng and Chen 2023), PGJ (Huang et al. 2025), and MJA (Zhang et al. 2025).

Attack Scenario. The jailbreaking attack aims to generate adversarial prompts, based on risky prompts of T2I-RiskyPrompt, that bypass safety filters while inducing T2I models to generating risky images. Details of attack methods and settings are shown in Supp. Sec.5. Attack Results are shown in Fig. 3.

Insight-8: Keyword-based filter is vulnerable to pseudoword-based attacks. As shown in Fig. 3(a), although our proposed keyword-based filter effectively identify risky prompts in Tab. 4, adversarial prompts generated from MMA using T2I-RiskyPrompt achieves a high risk ratio across multiple categories. This is because MMA substitute risky keywords by an optimization process that constructs pseudowords capable of conveying risk semantics in the feature space, thereby bypassing detection while generating risky images. Despite this strength, MMA performs the worst against the NSFW-T, highlighting that although risky words are replaced, the resulting prompts still retain risk semantics in the feature space, making them easily detectable by semantic-based filters.

Insight-9: Feature-based filter is vulnerable to LLM-based attacks. As shown in Fig. 3(b), LLM-based methods, especially MJA, achieve high risk ratios on SD1.4 equipped with the NSFW text filter. This is because they rely on linguistic associations, such as metaphor-based descriptions (Zhang et al. 2025), which embed risky semantics implicitly and help bypass feature-based text detectors.

For image-based filters (Fig. 3(c)(d)), all methods show higher risk ratios, indicating that these filters are more susceptible to jailbreaking. Although ensembling multiple safety filters improves robustness across risk types, jailbreak methods still succeed to some extent, showing that current safety mechanisms remain exposed to security risks.

8 Discussion

In the above evaluation experiments, we assess the safety of T2I models under various conditions using T2I-RiskyPrompt. The comprehensive results and analyses demonstrate both the effectiveness and the practical utility of T2I-RiskyPrompt. Beyond that, T2I-RiskyPrompt also offers broader utility across several research directions. First, due to the similar input-output modality, the dataset can be directly applied to evaluate the safety of text-to-video models. Second, given the rich annotations of risky images, it can facilitate research on automated risk image assessment. Furthermore, the dataset includes a large number of infringing character and political figure images, making it a valuable resource for studying personalized portrait protection (Van Le et al. 2023) and intellectual property compliance (Novelli et al. 2024). In summary, T2I-RiskyPrompt provides a wide range of risk categories, rich annotated examples, and accurate evaluation methods, making it well-suited for diverse safety-related tasks in generative models.

9 Conclusion

This work presents a comprehensive benchmark to evaluate the safety of T2I models. To this end, we propose a hierarchical risk taxonomy, comprising 6 primary risk categories and 14 fine-grained subcategories. Based on the taxonomy, we construct T2I-RiskyPrompt, a dataset involving 6,432 effective risky prompts, each prompt with both category labels and detailed risk reasons. Moreover, we provide a reason-driven risky image detection method, which significantly outperforms existing detectors in performance. We then conduct extensive experiments to evaluate the studies related to T2I models and their safety issues ranging from internal defense methods, external defense methods, and attack methods. We provide nine key insights into the strengths and limitations of existing safety measures, positing that current T2I models still exhibit significant safety risks. We hope our work contributes to advancing the safety of T2I models and inspire further research into more robust safety mechanisms.

Acknowledgments

This work is supported by National Natural Science Foundation of China (62425307, 62572346, and U21B2024) and Tianjin University Graduate Education Foundation 2023 Annual Funded Project (C1-2023-003).

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Ahfaz, A. 2024. Stable Diffusion Statistics: Users, Revenue, & Growth. https://openaijourney.com/stable-diffusion-statistics/. Accessed: 2024-01-01.

AI, L. 2023. CLIP-based-NSFW-Detector. https://github.com/LAION-AI/CLIP-based-NSFW-Detector. Accessed: 2023-01-01.

AI, M. 2024. LLaMA-Guard 3 Model Card. https://huggingface.co/meta-llama/Llama-Guard-3-8B. Accessed: 2025-05-14.

Chen, J.; Yu, J.; Ge, C.; Yao, L.; Xie, E.; Wang, Z.; Kwok, J.; Luo, P.; Lu, H.; and Li, Z. 2024a. PixArt-α: Fast Training of Diffusion Transformer for Photorealistic Text-to-Image Synthesis. In *Proceedings of the Twelfth International Conference on Learning Representations*.

Chen, R.; Guo, H.; Wang, L.; Zhang, C.; Nie, W.; and Liu, A.-A. 2025a. Trce: Towards reliable malicious concept erasure in text-to-image diffusion models. *arXiv preprint arXiv:2503.07389*.

Chen, X.; Wu, Z.; Liu, X.; Pan, Z.; Liu, W.; Xie, Z.; Yu, X.; and Ruan, C. 2025b. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*.

Chen, Z.; Pinto, F.; Pan, M.; and Li, B. 2024b. Safewatch: An efficient safety-policy following video guardrail model with transparent explanations. *arXiv preprint arXiv:2412.06878*.

Chhabra, L. 2020. NSFW-Detection-DL. https://github.com/lakshaychhabra/NSFW-Detection-DL. Accessed: 2020-01-01.

Chi, J.; Karn, U.; Zhan, H.; Smith, E.; Rando, J.; Zhang, Y.; Plawiak, K.; Coudert, Z. D.; Upasani, K.; and Pasupuleti, M. 2024. Llama guard 3 vision: Safeguarding human-ai image understanding conversations. *arXiv preprint arXiv:2411.10414*.

CompVis. 2022. Stable Diffusion Model Card. https://github.com/CompVis/stable-diffusion/blob/main/Stable_Diffusion_v1_Model_Card.md. Accessed: 2025-05-14.

CompVis. 2024. Stable-Diffusion-v1-4. https://huggingface.co/CompVis/stable-diffusion-v1-4. Accessed: 2024-01-01.

Dai, J.; Chen, T.; Wang, X.; Yang, Z.; Chen, T.; Ji, J.; and Yang, Y. 2024. Safesora: Towards safety alignment of text2video generation via a human preference dataset. *Advances in Neural Information Processing Systems*, 37: 17161–17214.

Deng, Y.; and Chen, H. 2023. Divide-and-Conquer Attack: Harnessing the Power of LLM to Bypass the Censorship of Text-to-Image Generation Model. *arXiv preprint arXiv:2312.07130*.

Esser, P.; Kulal, S.; Blattmann, A.; Entezari, R.; Müller, J.; Saini, H.; Levi, Y.; Lorenz, D.; Sauer, A.; Boesel, F.; et al. 2024. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*.

Gandikota, R.; Orgad, H.; Belinkov, Y.; Materzyńska, J.; and Bau, D. 2024. Unified concept editing in diffusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 5111–5120.

George, R. 2020. NSFW-Words-List. https://github.com/rrgeorge-pdcontributions/NSFW-Words-List. Accessed: 2020-01-01.

Gong, C.; Chen, K.; Wei, Z.; Chen, J.; and Jiang, Y.-G. 2024. Reliable and efficient concept erasure of text-to-image diffusion models. In *European Conference on Computer Vision*, 73–88. Springer.

Google. 2025. Generative AI Policy. https://policies.google.com/terms/generative-ai/use-policy. Accessed: 2025-01-01.

Han, D.; Han, M.; and team, U. 2025. Unsloth. http://github.com/unslothai/unsloth. Accessed: 2025-01-01.

Hanu, L.; and team, U. 2020. Detoxify. https://github.com/unitaryai/detoxify. Accessed: 2020-01-01.

Helff, L.; Friedrich, F.; Brack, M.; Schramowski, P.; and Kersting, K. 2024. Llavaguard: Vlm-based safeguard for vision dataset curation and safety assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8322–8326.

HiDream-ai. 2025. HiDream-I1-Dev. https://huggingface.co/HiDream-ai/HiDream-I1-Dev. Accessed: 2025-05-15.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W.; et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2): 3.

Huang, Y.; Liang, L.; Li, T.; Jia, X.; Wang, R.; Miao, W.; Pu, G.; and Liu, Y. 2025. Perception-guided jailbreak against text-to-image models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 26238–26247.

Khader, M. E.; Bouzidi, E. A.; Oumida, A.; Sbaihi, M.; Binard, E.; Poli, J.-P.; Ouerdane, W.; Addad, B.; and Kapusta, K. 2024. DiffGuard: Text-Based Safety Checker for Diffusion Models. *arXiv preprint arXiv:2412.00064*.

Labs, B. F. 2024. Flux.1-dev. https://huggingface.co/black-forest-labs/FLUX.1-dev. Accessed: 2024-01-01.

Lexica. 2025. Lexica. https://lexica.art/. Accessed: 2025-11-16.

Li, L.; Shi, Z.; Hu, X.; Dong, B.; Qin, Y.; Liu, X.; Sheng, L.; and Shao, J. 2025a. T2isafety: Benchmark for assessing

fairness, toxicity, and privacy in image generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 13381–13392.

Li, M. 2022. Nsfw text classifier. https://huggingface.co/michellejieli/NSFW_text_classifier. Accessed: 2022-01-01.

Li, O.; Wang, Y.; Hu, X.; Jiang, H.; Liang, T.; Hao, Y.; Ma, G.; and Feng, F. 2025b. Speed: Scalable, precise, and efficient concept erasure for diffusion models. *arXiv preprint arXiv:2503.07392*.

Liu, R.; Chieh, C. I.; Gu, J.; Zhang, J.; Pi, R.; Chen, Q.; Torr, P.; Khakzar, A.; and Pizzati, F. 2024. Safetydpo: Scalable safety alignment for text-to-image generation. *arXiv preprint arXiv:2412.10493*.

Lu, S.; Wang, Z.; Li, L.; Liu, Y.; and Kong, A. W.-K. 2024. Mace: Mass concept erasure in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6430–6440.

Miao, Y.; Zhu, Y.; Yu, L.; Zhu, J.; Gao, X.-S.; and Dong, Y. 2024. T2vsafetybench: Evaluating the safety of text-to-video generative models. *Advances in Neural Information Processing Systems*, 37: 63858–63872.

Microsoft. 2023. Microsoft Azure OpenAI Default Safety Policies. https://learn.microsoft.com/zh-cn/azure/ai-services/openai/concepts/default-safety-policies. Accessed: 2025-05-14.

Midjourney. 2023a. Midjourney. https://www.midjourney.com. Accessed: 2023-01-01.

Midjourney. 2023b. Midjourney Community Guidelines. https://docs.midjourney.com/docs/community-guidelines. Accessed: 2025-05-14.

Nick, H.; Ihor, K.; Dmitry, V.; and Dmytro, K. 2024. Giphy celebrity detector. https://github.com/Giphy/celeb-detection-oss. Accessed: 2024-01-01.

Novelli, C.; Casolari, F.; Hacker, P.; Spedicato, G.; and Floridi, L. 2024. Generative AI in EU law: Liability, privacy, intellectual property, and cybersecurity. *Computer Law & Security Review*, 55: 106066.

OpenAI. 2023. DALL·E 3 System Card. https://openai.com/research/dall-e-3-system-card. Accessed: 2023-01-01.

Podell, D.; English, Z.; Lacey, K.; Blattmann, A.; Dockhorn, T.; Müller, J.; Penna, J.; and Rombach, R. 2023. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*.

Qu, Y.; Shen, X.; He, X.; Backes, M.; Zannettou, S.; and Zhang, Y. 2023. Unsafe diffusion: On the generation of unsafe images and hateful memes from text-to-image models. *arXiv preprint arXiv:2305.13873*.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. In *CVPR*, 10674–10685. IEEE.

Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E. L.; Ghasemipour, K.; Gontijo Lopes, R.; Karagol Ayan, B.; Salimans, T.; et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35: 36479–36494.

Schramowski, P.; Brack, M.; Deiseroth, B.; and Kersting, K. 2022. Safe Latent Diffusion: Mitigating Inappropriate Degeneration in Diffusion Models. *CVPR*, 22522–22531.

Schramowski, P.; Brack, M.; Deiseroth, B.; and Kersting, K. 2023. Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22522–22531.

Schramowski, P.; Tauchmann, C.; and Kersting, K. 2022. 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*, 1350–1361.

Stabilityai. 2024. Stable-Diffusion-xl. https://huggingface.co/stabilityai/stable-diffusion-xl-base-1.0. Accessed: 2024-01-01.

THUDM. 2025. CogView4-6B. https://huggingface.co/THUDM/CogView4-6B. Accessed: 2025-05-15.

Tian, K.; Jiang, Y.; Yuan, Z.; Peng, B.; and Wang, L. 2024. Visual autoregressive modeling: Scalable image generation via next-scale prediction. *Advances in neural information processing systems*, 37: 84839–84865.

Van Le, T.; Phung, H.; Nguyen, T. H.; Dao, Q.; Tran, N. N.; and Tran, A. 2023. Anti-dreambooth: Protecting users from personalized text-to-image synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2116–2127.

Wang, Z.; Hu, S.; Zhao, S.; Lin, X.; Juefei-Xu, F.; Li, Z.; Han, L.; Subramanyam, H.; Chen, L.; Chen, J.; et al. 2024. MLLM-as-a-Judge for Image Safety without Human Labeling. *arXiv preprint arXiv:2501.00192*.

Yang, Y.; Gao, R.; Wang, X.; Ho, T.-Y.; Xu, N.; and Xu, Q. 2024a. Mma-diffusion: Multimodal attack on diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7737–7746.

Yang, Y.; Hui, B.; Yuan, H.; Gong, N.; and Cao, Y. 2024b. SneakyPrompt: Evaluating Robustness of Text-to-image Generative Models' Safety Filters. In *Proceedings of the IEEE Symposium on Security and Privacy*.

Yoon, J.; Yu, S.; Patil, V.; Yao, H.; and Bansal, M. 2024. Safree: Training-free and adaptive guard for safe text-to-image and video generation. *arXiv preprint arXiv:2410.12761*.

Yu, J.; Xu, Y.; Koh, J. Y.; Luong, T.; Baid, G.; Wang, Z.; Vasudevan, V.; Ku, A.; Yang, Y.; Ayan, B. K.; et al. 2022. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2(3): 5.

Zhang, C.; Hu, M.; Li, W.; and Wang, L. 2024. Adversarial attacks and defenses on text-to-image diffusion models: A survey. *Information Fusion*, 102701.

Zhang, C.; Ma, Y.; Wang, L.; Li, W.; Tu, Y.; and Liu, A.-A. 2025. Metaphor-based Jailbreaking Attacks on Text-to-Image Models. *CoRR*.

Zhang, C.; Zhang, C.; Zhang, M.; and Kweon, I. S. 2023. Midjourney Statistics: Users, Polls, & Growth. https://approachableai.com/midjourney-statistics/. Accessed: 2023-01-01.