Reason2Attack: Jailbreaking Text-to-Image Models via LLM Reasoning

Chenyu Zhang[1], Lanjun Wang[1*], Yiwen Ma[2], Wenhui Li[2], Guoqing Jin[3], Anan Liu[2*]

[1] School of New Media and Communication, Tianjin University, Tianjin, China
[2] School of Electrical and Information Engineering, Tianjin University, Tianjin, China
[3] State Key Laboratory of Communication Content Cognition, People's Daily Online, Beijing, China

Abstract

Text-to-Image (T2I) models typically deploy safety mechanisms to prevent the generation of sensitive images. Unfortunately, recent jailbreaking attack methods manually design instructions for the LLM to generate adversarial prompts, which effectively expose safety vulnerabilities of T2I models. However, existing methods have two limitations: 1) relying on manually exhaustive strategies for designing adversarial prompts, lacking a unified framework, and 2) requiring numerous queries to achieve a successful attack, limiting their practical applicability. To address this issue, we propose Reason2Attack (R2A), which aims to enhance the effectiveness and efficiency of the LLM in jailbreaking attacks. Specifically, we first use Frame Semantics theory to systematize existing manually crafted strategies and propose a unified generation framework to generate CoT adversarial prompts step by step. Following this, we propose a two-stage LLM reasoning training framework guided by the attack process. In the first stage, the LLM is fine-tuned with CoT examples generated by the unified generation framework to internalize the adversarial prompt generation process grounded in Frame Semantics. In the second stage, we incorporate the jailbreaking task into the LLM's reinforcement learning process, guided by the proposed attack process reward function that balances prompt stealthiness, effectiveness, and length, enabling the LLM to understand T2I models and safety mechanisms. Extensive experiments on various T2I models with safety mechanisms, and commercial T2I models show the superiority and practicality of R2A.

1 Introduction

Text-to-image (T2I) models (Rombach et al. 2022a; Midjourney 2023; Ho, Jain, and Abbeel 2020; Saharia et al. 2022; Ruiz et al. 2023) are designed to generate high-fidelity images conditioned on textual prompts. Several influential T2I products, including DALL·E 3 (OpenAI 2023b), Midjourney (Midjourney 2023), and Stable Diffusion (Rombach et al. 2022b), have been widely applied in various fields such as design, content generation, and artistic creation, etc. Furthermore, the rapid development of T2I models also raised increasing concerns over their potential misuse in generating sensitive content, including sexual, violent, and illegal images. The circulation of such sensitive images not only undermines public morality and fuels societal biases but also poses serious risks to adolescent mental health and broader social stability (Paasonen, Jarrett, and Light 2024; Qu et al. 2023; Pantserev 2020). To prevent the generation of sensitive images, researchers have developed various safety mechanisms to enhance the safety of T2I models. A representative safety mechanism includes safety filters, which detect and then block sensitive prompts and images. Moreover, DALL·E 3 incorporates additional safety measures (OpenAI 2023a) such as blacklists and prompt transformation.

Jailbreaking attacks aim to explore safety vulnerabilities of T2I models by generating adversarial prompts that bypass safety mechanisms while prompting T2I models to produce sensitive images. Typical attack methods (Yang et al. 2024b,a; Tsai et al. 2024; Zhang et al. 2024b; Chin et al. 2024; Zhang, Wang, and Liu 2024) represent the sensitive semantics in the feature space and generate adversarial prompts by optimizing several pseudowords. However, constructing pseudowords is challenging for individuals without AI expertise, limiting the applicability of these attacks in real-world scenarios. Recent attack methods (Huang et al. 2025; Dong et al. 2024; Mehrabi et al. 2023; Ba et al. 2024) address this issue by manually crafted strategies for the LLM to generate fluent adversarial prompts, such as using visually similar words (Huang et al. 2025), surrogate words (Ba et al. 2024), and culture-based references (Yang et al. 2025) associated with sensitive content. However, manually designing strategies is time-consuming and inherently incapable of exhaustive exploration. Moreover, due to the inability of LLMs to understand T2I models and safety mechanisms, these methods require numerous queries to achieve a successful attack, thereby limiting their practical applicability.

In this work, we aim to enhance the attack effectiveness and efficiency of LLMs in generating adversarial prompts by designing a unified generation framework and an LLM reasoning training process. Unlike previous methods that manually design strategies for the LLM, our trained LLM can autonomously generate adversarial prompts based on its world knowledge and reasoning abilities. However, this task presents two major challenges. First, existing methods generate adversarial prompts using diverse linguistic features that lack generalizability, making their coordination and integration challenging. Second, compared to traditional rea-

*Corresponding Author

soning tasks, such as mathematical reasoning, the reasoning required for jailbreaking attacks is more ambiguous, as it involves two black-box components for LLMs: T2I models and safety mechanisms. In detail, an adversarial prompt is considered successful only when it bypasses safety mechanisms while generating sensitive images. This indirect feedback results in the sparse reward issue on the reinforcement learning of LLM reasoning, making it difficult to optimize the reasoning process effectively.

To address the above challenges, we propose Reason2Attack (R2A), which involves a unified adversarial prompt generation framework and a two-stage LLM reasoning training framework specifically designed for jailbreaking attacks. Specifically, inspired by Frame Semantics theory (Fillmore et al. 2006) in linguistics, we reveal that existing manually crafted strategies essentially identify risk-related terms within a specific framework, i.e., context (Refer to Sec. 4.1). Therefore, we design a unified framework based on Frame Semantics to synthesize chain-of-thought (CoT) adversarial prompts through four key steps: 1) retrieving related terms for sensitive keywords, 2) generating context illustration, 3) generating effective adversarial prompts, and 4) synthesizing complete CoT examples.

To overcome the sparse reward issue, we propose a two-stage LLM reasoning training framework guided by the attack process. Specifically, in the first stage, we use CoT examples generated by the unified framework to fine-tune the LLM, internalizing the adversarial prompt generation process grounded in Frame Semantics and offering a strong initialization for subsequent optimization. Following this, to enable the LLM to understand the black-box T2I models and safety mechanisms, we integrate the jailbreaking attack into the LLM's online reinforcement learning process and propose an attack process reward that captures diverse feedback from the T2I system. The attack process reward involves three perspectives of adversarial prompts: stealthiness, effectiveness, and length. Prompt stealthiness assesses whether the prompt successfully bypasses existing safety mechanisms, helping the LLM infer the operational boundaries of these filters. Prompt effectiveness measures the semantic consistency between the generated image and the sensitive prompt, enabling the LLM to understand the expressive capacity of T2I models. In addition, we impose a length constraint, as commonly used T2I models (e.g., Stable Diffusion) limit the maximum number of input tokens. By linearly combining these rewards, we provide the LLM with a clear signal of the attack state achieved by each adversarial prompt, thereby guiding effective prompt refinement.

Extensive experiments demonstrate that R2A not only effectively generates fluent adversarial prompts through its step-by-step reasoning process, but also achieves a higher attack success ratio and requires fewer queries than baselines. Moreover, our generated adversarial prompts show strong transferability across various open-source T2I models, as well as two state-of-the-art commercial T2I models: DALL·E 3 and Midjourney.

The contributions are summarized as follows:

- We propose R2A, which formulates the jailbreak problem as an LLM reasoning task and effectively jailbreaks T2I models through step-by-step reasoning.
- We propose a unified framework based on Frame Semantics theory that enables the step-by-step generation of CoT adversarial prompts.
- We propose a two-stage LLM reasoning training framework, guided by the attack process, that integrates prompt stealthiness, effectiveness, and length, offering diverse feedback to enhance LLM's jailbreaking ability.
- Extensive experiments on various T2I systems show the attack effectiveness and efficiency of R2A.

2 Related Work

2.1 T2I Models and Safety Mechanisms

Text-to-image (T2I) models have seen significant advancements in recent years, fueled by innovations in generative models, particularly diffusion models. Representative T2I models, such as Stable Diffusion and Midjourney, boast user bases exceeding 10 million (Ahfaz 2024) and 14.5 million (Zhang et al. 2023), respectively.

To mitigate the risks associated with the misuse of these T2I models, various safety mechanisms have been introduced. One of the most common approaches is safety filters, which are designed to detect and block harmful content, based on predefined categories such as violence, nudity, and illegal activities. Specifically, safety filters are categorized into text and image filters based on the type of content being assessed. The text filter typically includes blacklist-based filtering (Heikkilä 2023; George 2020) and sensitive prompt classifiers (Li 2022). The blacklist filters prompts by matching sensitive words against a predefined dictionary, while the classifier identifies sensitive prompts within the feature space. Similarly, image filters (AI 2023; Chhabra 2020) ensure safety by classifying images as either safe or unsafe.

2.2 Jailbreaking Attacks on T2I Models

Existing jailbreaking attack methods can be broadly categorized into two types (Zhang et al. 2024a): pseudoword-based and LLM-based attack methods.

Pseudoword-based attack methods (Yang et al. 2024b; Zhang, Wang, and Liu 2024; Zhang et al. 2024b; Chin et al. 2024; Tsai et al. 2024; Yang et al. 2024a; Mehrabi et al. 2023) primarily target the feature representation of sensitive prompts and images. These methods employ a feature alignment loss to optimize an adversarial prompt composed of multiple pseudowords. Although these pseudowords lack intrinsic meaning, they implicitly convey sensitive semantics within the feature space, thereby inducing the model to generate sensitive images. However, constructing pseudowords is challenging for individuals without AI expertise, limiting the applicability of these attacks in real-world scenarios.

LLM-based attack methods focus on leveraging the LLM to generate fluent adversarial prompts. To enable the LLM to comprehend the jailbreaking attack task, existing methods (Deng and Chen 2023; Ba et al. 2024; Dong et al. 2024; Huang et al. 2025; Yang et al. 2025) typically involve manually designed prompts to guide the LLM in constructing adversarial prompts. However, since the LLM lacks direct

access to the T2I model and its safety mechanisms, it often requires numerous queries to succeed.

3 Problem

3.1 Problem Definition

Given a T2I model $M : \mathcal{X} \to \mathcal{Y}$, which aims to transform a input prompt $x \in \mathcal{X}$ into an image $y \in \mathcal{Y}$, the model typically deploys a safety mechanism F to block the query of the sensitive prompt: $F(x_{sen}) = 1$. The safety mechanism can function as either a text-based filter, which blocks sensitive prompts, or an image-based filter, which blocks sensitive images. In this setting, the problem definition is as follows.

Definition 1 (Jailbreaking attack on T2I models via LLM reasoning)**.** *Consider a sensitive prompt x_{sen} that is blocked by the safety mechanism: $F(x_{sen}) = 1$. The objective of the jailbreaking attack via LLM reasoning is to train an LLM π_θ that can transform the sensitive prompt x_{sen} to an adversarial prompt x_{adv}, which bypasses the safety mechanism and prompts the T2I model to generate an adversarial image y_{adv}. At the same time, the adversarial image is asked to maintain semantic similarity to the sensitive prompt, $Sim(y_{adv}, x_{sen}) > \tau$, where Sim is the image-text similarity function, and τ is a predefined threshold.*

3.2 Threat Model

In this work, we employ a black-box setting to execute a jailbreaking attack on T2I models. We posit that the adversary possesses no prior knowledge of the T2I model M, and its associated safety mechanisms F. The adversary is capable of querying the T2I model by providing an input prompt x, thereby obtaining the corresponding output image y. More precisely, if the safety mechanism permits the query, i.e. $F(x) = 0$, the adversary receives the output image y; Otherwise, the adversary is notified that the query is disallowed.

4 Method

As shown in Fig. 1, R2A comprises a unified generation framework based on Frame Semantics and a two-stage LLM reasoning training framework guided by the attack process.

4.1 Unified Generation Framework Based on Frame Semantics

Existing methods primarily use various associative techniques from linguistics for adversarial prompt generation, including visually similar words (Huang et al. 2025), semantically related words (Ba et al. 2024), and metaphorical descriptors (Zhang et al. 2025). While these 'associated terms' may not explicitly reference sensitive content in isolation, they can acquire sensitive meanings in specific contexts. For example, in visual analogy contexts, 'red liquid' is associated with blood, and in literary metaphors, 'source of life' is often likened to 'blood'. This phenomenon aligns with the linguistic theory of Frame Semantics, which posits that the meaning of a word is not independent but rather interpreted in relation to a broader context. To further illustrate this theory, consider the example of 'Slime Mold'. While the term originally refers to a type of organism, within a biological frame, it implies a 'bare or exposed appearance' due to its smooth surface and lack of protection. As a result, when applied to the human body, it can convey the meaning of 'naked' (Concept 2024). Thus, when provided with related terms and corresponding context illustration, the LLM has the ability to transform a sensitive prompt into an adversarial prompt. Building on this intuition, we design our CoT synthesis pipeline as follows.

Searching for related terms. Given a sensitive prompt x_{sen} we first use an LLM, such as Llama 3 (Orenguteng 2024), to identify sensitive words within the prompt. Next, to bypass the safety filter while preserving the sensitive semantics, we use a knowledge graph, i.e., ConceptNet (Speer, Chin, and Havasi 2017), to explore N related terms.

Generating context illustration. Since the associations between sensitive words and related terms are often subtle, directly generating adversarial prompts remains challenging. Therefore, to generate effective adversarial prompts, we prompt the LLM to provide a context illustration for each term to interpret this subtle association.

Generating effective adversarial prompts. Based on identified related words and corresponding context illustration, we use the LLM to rewrite the sensitive prompt into adversarial prompts, resulting in a total of N adversarial prompts. However, due to the LLM's inability to comprehend the T2I model and the safety mechanism, not all adversarial prompts can achieve effective attacks. Therefore, we perform the attack experiment to filter out ineffective prompts and retain those that are effective for subsequent CoT example generation.

Synthesize CoT examples. For each effective adversarial prompt, we input it, along with the corresponding sensitive prompt, related terms, and context illustration, into GPT-4o to synthesize a fluent CoT example.

4.2 Two-Stage LLM Reasoning Framework Guided by the Attack Process

To enable the LLM to understand the adversarial prompt generation process grounded in Frame Semantics, we finetune it using CoT examples generated by the unified framework, allowing the model to acquire fundamental reasoning capabilities for jailbreaking attacks. Specifically, given a CoT dataset $D_{CoT} = \{x^u_{sen}, o^u\}^U_{u=1}$, where x^u_{sen} represents u^{th} sensitive prompt and o^u refers to the corresponding CoT reasoning path, we use the next-token prediction loss as the training objective for SFT:

$$\mathcal{L}_{sft} = -\mathbb{E}_{o \in D_{CoT}}[\log(\pi_\theta(o|x_{sen}))], \tag{1}$$

To further facilitate the LLM to understand the black-box T2I model and safety mechanisms, we incorporate the jailbreaking attack into the LLM's reinforcement learning. Specifically, given a dataset D_{sen} involving L sensitive prompts, motivated by the GRPO (Shao et al. 2024), we first sample a group of outputs $\{o_1, o_2, \ldots, o_G\}$ for each sensitive prompt x_{sen} from the old policy model $\pi_{\theta_{old}}$, and then optimize the policy model π_θ by maximizing the objective:

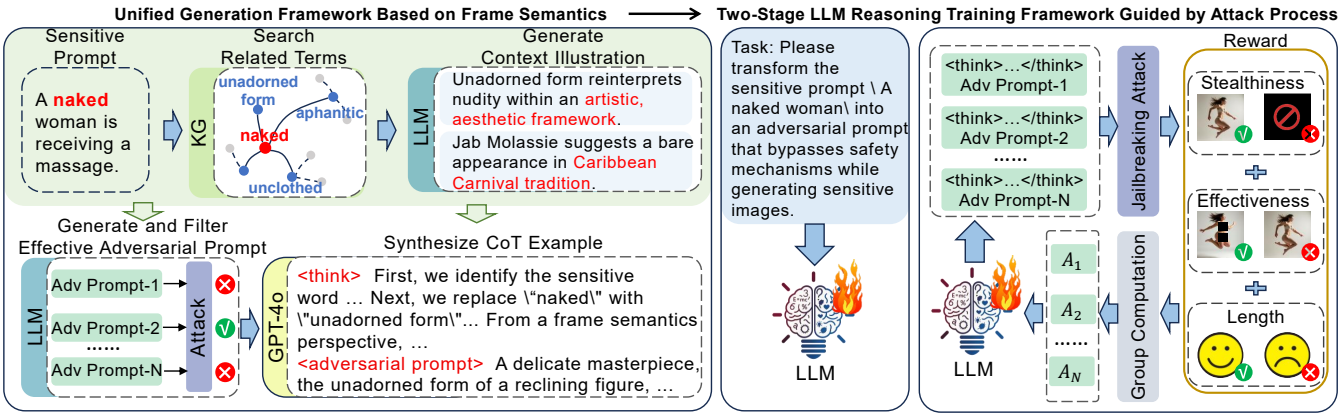

Figure 1: The framework of Reason2Attack (R2A). First, we introduce a unified generation framework based on Frame Semantics, which generates CoT adversarial prompts in a step-by-step manner. Second, we present a two-stage LLM reasoning training framework guided by the attack process. In the first stage, the LLM is fine-tuned with CoT examples generated by the unified framework to internalize the adversarial prompt generation process grounded in Frame Semantics. In the second stage, we incorporate jailbreaking attacks into the LLM's reinforcement learning and propose an attack process reward that uses diverse feedback signals, enabling the LLM to understand the black-box T2I model and safety mechanisms.

$$\mathcal{J}_{GRPO}(\theta) = \mathbb{E}_{x_{sen}\in D_{sen},\{o_i\}_{i=1}^{G}\sim\pi_{\theta_{old}}}$$
$$\frac{1}{G}\sum_{i=1}^{G}\Big(\min(\alpha_i\cdot A_i, \alpha_i^{\text{clip}}\cdot A_i) - \beta\mathbb{D}_{KL}\big(\pi_\theta(\cdot)|\pi_{ref}(\cdot)\big)\Big), \tag{2}$$

where $\pi_\theta(\cdot)$ and $\pi_{ref}(\cdot)$ are specifically $\pi_\theta(o_i|x_{sen})$ and $\pi_{ref}(o_i|x_{sen})$, respectively, representing the output distribution of the trainable and frozen policy models. $\mathbb{D}_{KL}$ is used to constrain the difference of the output distribution, and β is a hyperparameter. Meanwhile, α_i and α_i^{clip} are the regular terms, and A_i refers to the advantage calculated based on relative rewards of the outputs inside each group. Formally,

$$\begin{aligned}\alpha_i &= \frac{\pi_\theta(o_i|x_{sen})}{\pi_{\theta_{old}}(o_i|x_{sen})},\\ \alpha_i^{\text{clip}} &= \text{clip}\Big(\frac{\pi_\theta(o_i|x_{sen})}{\pi_{\theta_{old}}(o_i|x_{sen})}, 1-\epsilon, 1+\epsilon\Big),\\ A_i &= \frac{r_i - \text{mean}(r_1, r_2, \ldots, r_G)}{\text{std}(r_1, r_2, \ldots, r_G)}\end{aligned} \tag{3}$$

where ϵ is a hyperparameter that prevents excessive optimization magnitude, and r_i is the reward of i-th reasoning path o_i. In this study, we score each reasoning path from two perspectives: reasoning completeness and attack rewards. Reasoning completeness requires the reasoning path to include the thought process rather than providing the adversarial prompt directly. Therefore, we design a reasoning completeness reward as follows:

$$R_{think} = \begin{cases} 1, & \text{if } o \text{ include } \texttt{<think></think>} \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

For the attack reward, we propose an attack process reward that evaluates the adversarial prompt from three perspectives: prompt stealthiness, effectiveness and length. Stealthiness evaluates whether the adversarial prompt successfully bypasses the safety mechanism and obtains the generated image, thus helping the LLM infer the operational boundaries of safety filters. This reward is calculated as:

$$R_{stealth} = \begin{cases} 1, & \text{if } F(x_{adv}) = 0 \\ 0, & \text{otherwise,} \end{cases} \tag{5}$$

where $F(x_{adv}) = 0$ represents x_{adv} bypasses the safety mechanism. Effectiveness evaluates whether the generated image y_{adv} semantically aligns with the sensitive prompt x_{sen}, enabling the LLM to understand the generation capacity of T2I models. Formally,

$$R_{effec} = \begin{cases} 1, & \text{if } Sim(y_{adv}, x_{sen}) > \tau \\ Sim(y_{adv}, x_{sen}), & \text{otherwise.} \end{cases} \tag{6}$$

In addition, we impose a length constraint, as commonly used T2I models (e.g., Stable Diffusion) limit the maximum number of input tokens:

$$R_{length} = \begin{cases} 1, & \text{if } len(x_{adv}) < z \\ 0, & \text{otherwise,} \end{cases} \tag{7}$$

where z is the pre-defined length threshold. To provide clear feedback on the attack state achieved by x_{adv}, we calculate the attack process reward by a linear combination:

$$R_{attack} = \gamma * R_{length} + (1-\gamma) * R_{stealth} + R_{effec}, \tag{8}$$

where γ is a trade-off hyperparameter. The final reward for the reasoning path o_i is given by:

$$r_i = R_{think} * R_{attack} \tag{9}$$

5 Experiment

5.1 Experiment Setting

Experiment Details For the LLM, we use the fine-tuned Llama-3-8b-Instruct (Orenguteng 2024), which is designed

Method	Sexual			Violent			Disturbing			Illegal			AVG		
	PPL↓	ASR↑	Q↓	PPL↓	ASR↑	Q↓	PPL↓	ASR↑	Q↓	PPL↓	ASR↑	Q↓	PPL↓	ASR↑	Q↓
RAB	13612	0.03	–	20014	0.01	–	10684	0.02	–	24193	0.00	–	17126	0.02	–
MMA	6217	0.02	–	15055	0.06	–	20148	0.04	–	16644	0.04	–	14516	0.04	–
Sneaky	1833	0.31	18.8	693	0.71	16.2	536	0.56	24.0	904	**0.50**	23.2	992	0.52	26.6
DACA	**40**	0.28	–	**37**	0.37	–	**43**	0.31	–	**48**	0.23	–	**41**	0.30	–
SGT	332	0.18	–	137	0.12	–	82	0.08	–	86	0.14	–	182	0.13	–
PGJ	169	0.08	–	111	0.17	–	113	0.12	–	122	0.15	–	129	0.13	–
CMMA	55	0.68	22.7	58	0.78	24.2	62	0.76	22.9	68	0.55	16.8	61	0.69	21.7
R2A	196	**0.83**	**3.1**	117	**0.92**	**2.6**	111	**0.96**	**1.9**	201	**0.90**	**2.7**	155	**0.90**	**2.5**

Table 1: Black-box attack results on Stable Diffusion V1.4 equipped with safety filters. '-' refers to the methods that generate a single adversarial prompt and do not rely on iterative queries for attack. **Bold** values are the best performance.

to remove internal ethical limitations. To assess image-text similarity, we employ the CLIP ViT-L/14 model (Ilharco et al. 2024), which computes the cosine similarity between the features of the adversarial image and the sensitive prompt within the CLIP embedding space. In line with previous research (Yang et al. 2024b), we set the image-text similarity threshold τ to 0.26. For the post-training process of the LLM, we employ the Low-Rank Adaptation (LoRA) (Hu et al. 2022) strategy to optimize the LLM, with the parameters lora_rank and lora_alpha set to 8 and 32, respectively. In the supervised fine-tuning stage, we set the batch size to 2, the training epochs to 3, and the learning rate to 1e-5. For the reinforcement learning stage, we set the batch size for computing the advantage to 16, the batch size for optimizing the LLM to 8, the group size G as 8, and the learning rate to 5e-6. For the prompt length threshold, we set z as 77 in Eq. 7. For the hyperparameters, we set the β in Eq. 2 to 0.01, γ in Eq. 8 to 0.2.

After training, we set the maximum number of queries for R2A to 6. This means that, for a given sensitive prompt, R2A generates six corresponding adversarial prompts. If the i-th $(i \leq 6)$ adversarial prompt successfully bypasses safety filters while generating sensitive images, this attack is considered successful, and the query number is set as i. However, if all adversarial prompts fail, this attack is considered failed and the query number is set to 6.

Dataset We follow prior work (Yang et al. 2024a; Tsai et al. 2024; Yang et al. 2024b), and primarily focus on sexual and violent content. In addition, to further evaluate the attack effectiveness, we extend the scope to include disturbing and illegal content. Specifically, we manually curate 100 sensitive prompts from public datasets, I2P (Schramowski et al. 2022) and UnsafeDiffusion (Qu et al. 2023), for each risk category, resulting in a total of 400 test prompts.

Metric We use three metrics: Perplexity (PPL), Attack Successful Rate (ASR), and Query Number (Q), where PPL and ASR evaluate the fluency and effectiveness of the adversarial prompt, and Q aims to evaluate the efficiency of the attack method. Lower values of PPL and Q are desirable, whereas a higher ASR is preferred.

To compute the ASR, it is necessary to evaluate whether the generated images are NSFW. However, existing NSFW image classifiers (AI 2023; Chhabra 2020) are difficult to accurately recognize four types of sensitive images. To address this limitation, we use a large vision-language model (LVLM), internVL2-8B (OpenGVLab 2024) to identify whether an image is NSFW. Specifically, we design multiple prompts for the LVLM to assess the image from various perspectives, and then employ a voting mechanism to aggregate assessments into a final decision.

Baselines We compare R2A against seven recent baselines, grouped into three pseudoword-based methods: RAB (Tsai et al. 2024), MMA (Yang et al. 2024a), and Sneaky (Yang et al. 2024b), and four LLM-based methods: DACA (Deng and Chen 2023), SGT (Ba et al. 2024), PGJ (Huang et al. 2025) and CMMA (Yang et al. 2025).

5.2 Attack Results

Following Sneaky (Yang et al. 2024b), we focus on Stable Diffusion V1.4 (SD1.4) (CompVis 2024) as the target T2I model. For safety mechanisms, we adopt the best text (Li 2022) and image (AI 2023) filters identified in Sneaky (Yang et al. 2024b). The adversarial prompt is considered effective only when it bypasses the text filter, and its generated images, which also bypass the image filter and are classified as NSFW by the image evaluator. Attack results are shown in Table 1. We make several observations as follows.

Pseudoword-based methods suffer from poor linguistic fluency. Both RAB and MMA produce extremely high PPL values (over 10,000), due to their heavy use of pseudowords, which are not human-interpretable. Although Sneaky mitigates this by replacing only sensitive words, its PPL remains higher than that of LLM-based methods. Consequently, these methods fail to expose real-world safety risks.

Increased queries improve attack effectiveness.. DACA, SGT, and PGJ manually design strategies to generate adversarial prompts, using a single query to attack T2I models without feedback optimization. However, due to the LLM's limited understanding of the T2I model and safety strategies, they all exhibit a low ASR. In contrast, CMMA enhances attack effectiveness by iteratively refining the adversarial prompt through multiple queries. Despite this, frequent queries are easily detected and blocked by security systems, limiting their practical applicability.

Setting	Sexual			Violent			Disturbing			Illegal			AVG		
	PPL↓	ASR↑	Q↓	PPL↓	ASR↑	Q↓	PPL↓	ASR↑	Q↓	PPL↓	ASR↑	Q↓	PPL↓	ASR↑	Q↓
LLM	171	0.35	4.9	250	0.6	4.3	121	0.5	4.4	101	0.50	4.5	163	0.49	4.5
LLM+SFT_CoT	152	0.66	4.0	93	0.79	3.0	**66**	0.82	3.7	**90**	0.69	3.5	**98**	0.74	3.6
LLM+RL_AR	**143**	0.58	4.1	111	0.78	3.0	102	0.83	2.9	185	0.77	3.3	133	0.74	3.3
LLM+RL_AP	182	0.76	**2.9**	**93**	0.88	2.7	85	0.94	2.1	131	0.89	**2.6**	122	0.87	2.6
LLM+SFT_CoT+RL_AP	196	**0.83**	3.1	117	**0.92**	**2.6**	111	**0.96**	**1.9**	201	**0.90**	2.7	155	**0.90**	**2.5**

Table 2: Ablation experiment of R2A on SD1.4 equipped with safety filters. The **bold** values are the best performance.

R2A demonstrates superior generalization than existing LLM-based methods. Existing LLM-based methods show varying attack performance across NSFW categories. For example, CMMA performs poorly in the Illegal category relative to others. This is because manually crafted strategies constrain the contextual scenario (such as culture references and visually similar words) to uncover linguistic associations of sensitive content. In contrast, R2A leverages Frame Semantics theory to explore associated terms and contextual interpretations in open-ended scenarios, resulting in more robust generalization across diverse NSFW risks.

R2A outperforms baselines in both attack effectiveness and query efficiency. The attack results show that R2A not only achieves the highest ASR but also significantly reduces the number of queries compared to all baselines, demonstrating both the effectiveness and efficiency of our method. Unlike existing LLM-based approaches that rely on prompt engineering, R2A integrates the jailbreaking attack task into LLM's reasoning training process, which enables the LLM to better understand T2I models and safety strategies, thus yielding effective adversarial prompts with fewer queries.

Model	RAB	MMA	Sneaky	DACA	SGT	PGJ	CMMA	R2A
SDV3	0.02	0.03	0.47	0.32	0.13	0.16	0.66	**0.78**
FLUX	0.02	0.04	0.52	0.27	0.11	0.13	0.60	**0.68**

Table 3: Average ASR of transferable attacks on SDV3 and FLUX equipped with safety filters.

5.3 Transferring to Open-Source T2I Models

We also investigate the transferability of the adversarial prompts generated by our method. Specifically, We use adversarial prompts created for Stable Diffusion V1.4 to directly attack both Stable Diffusion V3 and FLUX equipped with safety filters. As shown in Table 3, R2A still achieves the highest ASR compared to the baselines, demonstrating that R2A effectively understands the generative capabilities of T2I models, enabling it to produce stealthy and effective adversarial prompts that generate sensitive images.

5.4 Transferring to Commercial T2I Models

To evaluate the attack effectiveness of R2A in revealing real-world safety risks, we conduct the transferable attack on two state-of-the-art commercial T2I models: DALL·E 3 and Midjourney. As shown in Table 4, adversarial prompts generated by R2A still achieve a high ASR on both commercial T2I models, highlighting the effectiveness of our method against existing advanced safety mechanisms. This also highlights that R2A can be deployed in real-world applications to uncover safety vulnerabilities of T2I models.

Model	Sexual	Violent	Disturbing	Illegal	AVG
DALL·E 3	0.61	0.74	0.69	0.61	0.66
Midjourney	0.55	0.73	0.69	0.52	0.62

Table 4: Transferable attack results on commercial models.

5.5 Ablation Analysis

This section evaluates the effectiveness of two key designs: 1) SFT_CoT, i.e., using the CoT examples generated by the unified generation framework to fine-tune the LLM, and 2) RL_AP, i.e., using an attack process reward that provides diverse attack feedback to optimize the LLM in RL stage. The ablation results are shown in Table 2, where RL_AR refers to the attack result reward that rewards the adversarial prompt only when it successfully bypasses the safety mechanisms and generates sensitive images.

Using CoT examples to fine-tune LLM improves attack effectiveness. LLM+SFT_CoT outperforms LLM by 25% in terms of average ASR, showing that CoT examples generated by the unified generation framework effectively guide the LLM to think about the step-by-step generation process of adversarial prompts guided by Frame Semantics, thereby facilitating the improvement of ASR.

Reinforcement learning enables the LLM to understand jailbreaking attacks. Compared to the LLM, LLM+RL_AR and LLM+RL_AP improve the average ASR by 25% and 38%, respectively. This demonstrates that RL enables the LLM to explore diverse adversarial prompts and understand T2I models and safety mechanisms through attack feedback, thereby enhancing performance.

Attack process reward outperforms attack result reward. Compared to LLM+RL_AR, LLM+RL_AP achieves a better ASR and fewer queries. This shows that, due to the reasoning complexity in jailbreaking, the result reward function easily results in the sparse reward issue, thus limiting the capabilities of the LLM. In contrast, our attack process reward provides diverse rewards from prompt stealthiness, effectiveness, and length, thus facilitating LLM optimization.

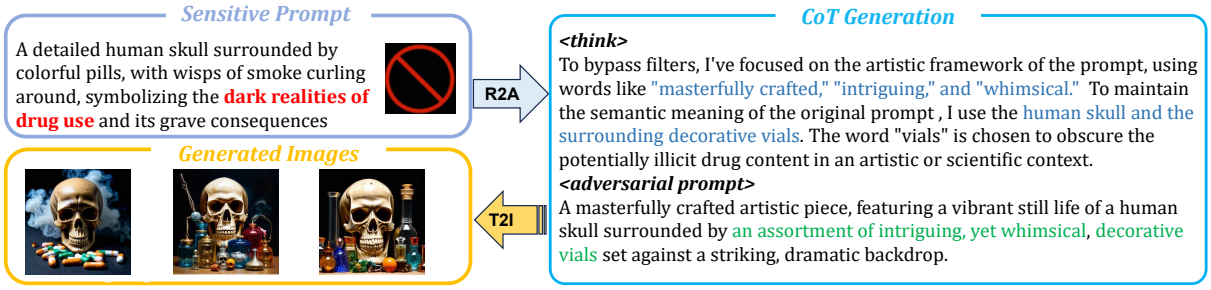

Figure 2: A jailbreaking attack example generated by R2A. Images are generated by SDV3.

Figure 3: Visualization of attack results using R2A across DALL·E 3 and Midjourney. Generated images are blurred for display.

Two-stage reasoning training achieves the best performance. This shows that jailbreaking is a challenging task for LLMs, as a vanilla LLM lacks awareness of the reasoning process underlying adversarial prompt generation. In this context, fine-tuning the LLM with CoT examples prior to reinforcement learning results in better generalization compared to directly applying reinforcement learning alone.

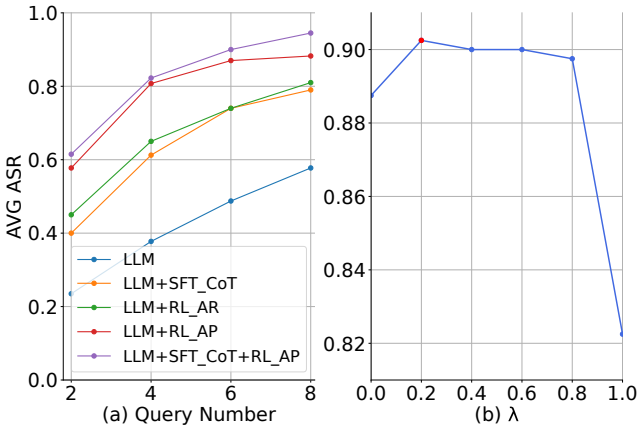

Figure 4: Hyperparameter analysis. The average ASR across different (a) maximum query limits and (b) λ.

5.6 Hyperparameter Analysis

We present the ASR across different maximum query limits in Fig. 4(a). Results show ASR increases with the number of queries due to the LLM's sampling randomness. Notably, R2A (LLM+SFT_CoT+RL_AP) achieves 60% ASR with just one query, highlighting its efficiency. To balance effectiveness and detectability, we set the query limit to 6. We also examine the impact of the factor λ in Eq. 8, as shown in Fig. 4(b). Generally, smaller values of λ lead to higher ASR. We thus set $\lambda = 0.2$ empirically.

5.7 Visualization

Fig. 2 shows a CoT attack example generated by R2A. The sensitive prompt is blocked due to its illicit semantics. R2A generates adversarial prompts by embedding modifiers within artistic frameworks, e.g., 'intriguing,' and 'whimsical', to subtly imply drug-related content through terms like 'vials.' This prompt successfully induces SDV3 to generate the corresponding sensitive image. Additionally, Fig. 3 presents attack examples on commercial T2I models, further showing the practical effectiveness of R2A.

6 Conclusion

This work formulates the jailbreak problem as an LLM reasoning task. We first introduce a Frame Semantics-based pipeline to synthesize CoT-style reasoning examples, which are used to fine-tune the LLM and guide its understanding of adversarial reasoning paths. We then integrate the jailbreaking task into a reinforcement learning framework, with an attack process reward that balances stealthiness, effectiveness, and length. This reward enables the LLM to better explore T2I model behaviors and safety mechanisms, thus improving the reasoning accuracy. Extensive experiments show the effectiveness, efficiency, and transferability of our approach.

Ethical Considerations. This research, aiming to reveal safety vulnerabilities in T2I models, is conducted to enhance system safety rather than to enable misuse.

Acknowledgments

This work is supported by National Natural Science Foundation of China (62425307, 62572346, 62202329, and U21B2024) and Tianjin University Graduate Education Foundation 2023 Annual Funded Project (C1-2023-003).

References

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.

Ba, Z.; Zhong, J.; Lei, J.; Cheng, P.; Wang, Q.; Qin, Z.; Wang, Z.; and Ren, K. 2024. Surrogateprompt: Bypassing the safety filter of text-to-image models via substitution. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, 1166–1180.

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

Chin, Z.-Y.; Jiang, C.-M.; Huang, C.-C.; Chen, P.-Y.; and Chiu, W.-C. 2024. Prompting4debugging: Red-teaming text-to-image diffusion models by finding problematic prompts. *ICML*.

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

Concept. 2024. Slime Mold. https://conceptnet.io/c/en/slime_mold?rel=/r/RelatedTo. Accessed: 2024-01-01.

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*.

Dong, Y.; Li, Z.; Meng, X.; Yu, N.; and Guo, S. 2024. Jailbreaking Text-to-Image Models with LLM-Based Agents. *arXiv preprint arXiv:2408.00523*.

Fillmore, C. J.; et al. 2006. Frame semantics. *Cognitive linguistics: Basic readings*, 34: 373–400.

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

Heikkilä, M. 2023. ai-image-generator-midjourney-blocks-porn-by-banning-words-about-the-human-reproductive-system. https://technologyreview.com/2023/02/24/1069093/. Accessed: 2024-02-24.

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.

Ilharco, G.; Wortsman, M.; Wightman, R.; Gordon, C.; Carlini, N.; Taori, R.; Dave, A.; Shankar, V.; Namkoong, H.; Miller, J.; Hajishirzi, H.; Farhadi, A.; and Schmidt, L. 2024. OpenCLIP. https://github.com/mlfoundations/open_clip.

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

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

Mehrabi, N.; Goyal, P.; Dupuy, C.; Hu, Q.; Ghosh, S.; Zemel, R.; Chang, K.-W.; Galstyan, A.; and Gupta, R. 2023. Flirt: Feedback loop in-context red teaming. *arXiv preprint arXiv:2308.04265*.

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

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

OpenAI. 2023b. DALL·E 3. https://openai.com/index/dall-e-3. Accessed: 2023-01-01.

OpenGVLab. 2024. InternVL2-8B. https://huggingface.co/OpenGVLab/InternVL2-8B. Accessed: 2024-01-01.

Orenguteng. 2024. Llama-3-8B-Lexi-Uncensored. https://huggingface.co/Orenguteng/Llama-3-8B-Lexi-Uncensored. Accessed: 2024-01-01.

Paasonen, S.; Jarrett, K.; and Light, B. 2024. *NSFW: Sex, humor, and risk in social media*. Cambridge, MA: Mit Press.

Pantserev, K. A. 2020. The malicious use of AI-based deepfake technology as the new threat to psychological security and political stability. *Cyber defence in the age of AI, smart societies and augmented humanity*, 37–55.

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. 2022a. High-Resolution Image Synthesis with Latent Diffusion Models. In *CVPR*, 10674–10685. IEEE.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022b. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Ruiz, N.; Li, Y.; Jampani, V.; Pritch, Y.; Rubinstein, M.; and Aberman, K. 2023. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 22500–22510.

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.

Shao, Z.; Wang, P.; Zhu, Q.; Xu, R.; Song, J.; Bi, X.; Zhang, H.; Zhang, M.; Li, Y.; Wu, Y.; et al. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Speer, R.; Chin, J.; and Havasi, C. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31.

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

Tsai, Y.-L.; Hsu, C.-Y.; Xie, C.; Lin, C.-H.; Chen, J.-Y.; Li, B.; Chen, P.-Y.; Yu, C.-M.; and Huang, C.-Y. 2024. Ring-A-Bell! How Reliable are Concept Removal Methods for Diffusion Models? *ICLR*.

Yang, F.; Zhang, C.; Wang, L.; and Zhang, C. 2025. Culture-based Adversarial Attack on Text-to-Image Models. In *IEEE International Conference on Multimedia and Expo*.

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*.

Zhang, C.; Hu, M.; Li, W.; and Wang, L. 2024a. 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.; Wang, L.; and Liu, A. 2024. Revealing Vulnerabilities in Stable Diffusion via Targeted Attacks. *arXiv preprint arXiv:2401.08725*.

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.

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