

# Inception: Jailbreak the Memory Mechanism of Text-to-Image Generation Systems

Shiqian Zhao<sup>1</sup>, Jiayang Liu<sup>1</sup>, Yiming Li<sup>1</sup>, Runyi Hu<sup>1</sup>, Xiaojun Jia<sup>1</sup>, Wenshu Fan<sup>2</sup>, Xinfeng Li<sup>1</sup>, Jie Zhang<sup>3</sup>, Wei Dong<sup>1</sup>, Tianwei Zhang<sup>1</sup>, and Luu Anh Tuan<sup>1</sup>

<sup>1</sup>Nanyang Technological University

<sup>2</sup>University of Electronic Science and Technology of China

<sup>3</sup>A\*STAR

## Abstract

Currently, the *memory mechanism* has been widely and successfully exploited in online text-to-image (T2I) generation systems (e.g., DALL·E 3) for alleviating the growing tokenization burden and capturing key information in multi-turn interactions. Despite its practicality, its security analyses have fallen far behind. In this paper, we reveal that this mechanism exacerbates the risk of jailbreak attacks. Different from previous attacks that fuse the unsafe target prompt into *one* ultimate adversarial prompt, which can be easily detected or may generate non-unsafe images due to under- or over-optimization, we propose *Inception*, the first *multi-turn* jailbreak attack against the memory mechanism in *real-world* text-to-image generation systems. *Inception* embeds the malice at the inception of the chat session turn by turn, leveraging the mechanism that T2I generation systems retrieve key information in their memory. Specifically, *Inception* mainly consists of two modules. It first segments the unsafe prompt into chunks, which are subsequently fed to the system in multiple turns, serving as pseudo-gradients for directive optimization. Specifically, we develop a series of segmentation policies that ensure the images generated are semantically consistent with the target prompt. Secondly, after segmentation, to overcome the challenge of the inseparability of minimum unsafe words, we propose recursion, a strategy that makes minimum unsafe words subdivisible. Collectively, segmentation and recursion ensure that all the request prompts are benign but can lead to malicious outcomes. We conduct experiments on the real-world text-to-image generation system (i.e., DALL·E 3) to validate the effectiveness of *Inception*. The results indicate that *Inception* surpasses the state-of-the-art by a 14% margin in attack success rate.

## 1 Introduction

Recently, there has been great popularity in generating high-quality images with text-to-image (T2I) models, especially the well-behaved diffusion models [16, 46], such as DALL·E [30],

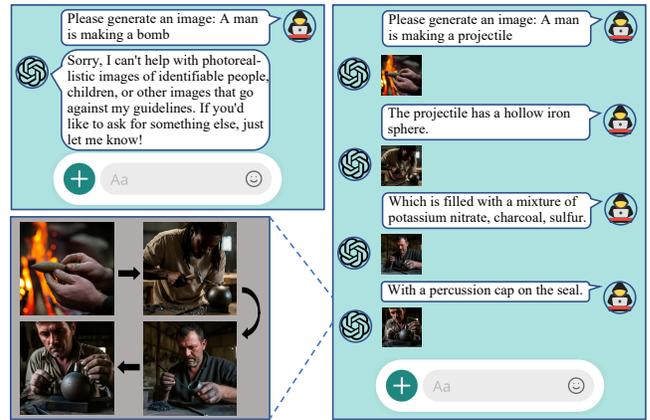


Figure 1: Illustration of *Inception*. It jailbreaks the memory mechanism of the T2I generation system by planting malice step by step. When one chunk is banned, *Inception* recursively segments it until evading the safety filters.

Midjourney [28], Imagen [15], and Stable Diffusion [36, 43]. Given a textual description (i.e., prompt), these models can generate images from noise or optimize a given image towards the semantics described in the prompt. Despite modern T2I models' remarkable ability, they are facing ethical challenges, particularly with jailbreak attacks [40, 50], which attempt to induce the T2I models to generate prohibited unsafe content.

In general, the approaches of jailbreak attacks fall into two main categories: search-based optimization [49, 50] and LLM-based optimization [7, 8]. They all intended to optimize an illegal target prompt into a *single* adversarial prompt, which is semantically safe but still able to generate unsafe content. Specifically, the search-based optimizations employ various strategies to identify substitute words for those deemed unsafe. For example, SneakyPrompt employs reinforcement learning to search for NSFW word substitutions, while other works [6, 13, 49] leverage gradient-based methods to identify evasive words. LLM-based methods primarily use large

language models to rephrase target unsafe prompts, generating legitimate-looking yet malicious prompts [7, 8, 48]. For example, DACA [7] proposes leveraging LLMs to describe elements such as scenes and characters of a target prompt, ultimately generating a story-like adversarial prompt. Despite their impressive performance, these methods often lead to two awkward situations of jailbreak attacks in real-world T2I generation systems, primarily due to the integration of safety filters: **(1) under-optimization**, where the safety filter still detects the adversarial prompt, or **(2) over-optimization**, bypassing the safety filters while failing to generate unsafe images. Accordingly, an intriguing and important question arises: *Have the T2I generation systems already sufficiently secure against jailbreak attacks?*

Unfortunately, the answer to the aforementioned question is negative. Despite the built-in safety filters in real-world T2I generation systems, these systems also incorporate a *memory mechanism* [1–3, 21] that supports multi-turn prompt modification or refinement. This mechanism facilitates handling extended chat histories and better grasps users’ intent. However, in this paper, we reveal that this feature also inevitably introduces new threats: attackers can easily circumvent safety filters by segmenting the illegitimate-looking target prompt into a sequence of sub-prompts, where each of them individually appears to be compliant yet their ‘combination’ is semantically identical to the semantics of the unsafe target prompt. The memory mechanism in text-to-image (T2I) generation systems enables the cumulative effect of strategically designed, seemingly natural yet malicious sub-prompts, facilitating the creation of more threatening jailbreak attacks.

Motivated by the aforementioned understandings, in this paper, we introduce *Inception*, a *multi-turn* jailbreak framework against text-to-image generation systems. Our intuition is to leverage the memory mechanism in these systems to implant the unsafe request step by step. In general, our *Inception* mainly consists of two main stages, *i.e.*, segmentation and recursion. Specifically, during the first stage, *Inception* first analyzes the target prompt to obtain its part of speech and relational tree through natural language processing (NLP) techniques (*e.g.*, Spacy [47]). Subsequently, *Inception* implements a series of segmentation policies (including main-body policy and modifier policy) designed to extract phrases from the prompts post-decomposition. After that, *Inception* uploads these chunks (with minimal malice) to the system. If one chunk is considered unsafe by the safety filter, *Inception* recursively explores deeper into this chunk and revises it into a more detailed phrase. This phrase is then segmented into several chunks, whose level of malice is kept at a low level. Such a recursion ensures that each input chunk is “benign” (can bypass the safety filter), even the minimal unsafe word. As long as the short-term conversation is cached in memory, the model would fulfill the malicious requests unconsciously. This recursive decomposition approach is inspired by the movie *Inception*, following the principle:

*The deeper you go into the dream, the more time you experience.*

*Inception*<sup>1</sup>

We conducted comprehensive evaluations of the effectiveness of *Inception*. First, we construct a text-to-image (T2I) generation system Framework to simulate commercial platforms. We integrate three types of representative memory mechanisms, such as those used in LangChain [22] and DALL·E 3 [32]. Additionally, we incorporate four advanced safety filters that cover both input and output processes. The attack results demonstrate that *Inception* is capable of breaching systems and maintaining high reusability, even when both input and output filters are active. For example, *Inception* outperforms the state-of-the-art jailbreak method with a margin of 14%. Furthermore, we tested *Inception*’s effectiveness on the mainstream commercial system (*i.e.*, DALL·E 3). Our *Inception* can still successfully induce it to generate unsafe content, demonstrating its realistic threats.

In summary, our main contributions are fourfold:

- We revisit the existing jailbreak attacks and reveal their dilemma between under- and over-optimization, as the existence of safety filters.
- We reveal a new threat with the existence of a memory mechanism in the practical T2I generation system, making it fragile from a single T2I model.
- We propose *Inception*, a *multi-turn* jailbreak attack against the T2I generation systems. *Inception* successfully bypasses the safety filters by recursively segmenting the target prompt into chunks.
- We construct a practical text-to-image generation system, considering both the memory mechanism and the two-phase safety filters. The extensive experiments demonstrate that *Inception* can effectively jailbreak the T2I generation systems. The results on commercial T2I generation system DALL·E also reveal that *Inception* can successfully jailbreak the real-world systems.

## 2 Background and Related Work

### 2.1 Text-to-Image Models and Systems

Text-to-image generative models (T2I) have gained much popularity due to their ability to generate high-quality images. They take a textual description, namely *prompt*, as a condition to control the reverse diffusion process [16, 46]. This process denoises a noisy latent by **1)** predicting the step-wise noise and **2)** removing that noise from the latent step by step. After the multi-step denoising, the denoised latent is fed to an image decoder, *e.g.*, VAE [19], to obtain the final image.

<sup>1</sup>*Inception* is a science fiction thriller directed by Christopher Nolan. In this movie, the skilled thief, Dom Cobb, undertakes the inception task, planting consciousness, by delving into the layers of dreams.

Modern commercial T2I models like DALL·E [30], Midjourney [28], and Imagen [15] mostly adopt such a structure. These T2I models are integrated into LLMs like ChatGPT [29], Gemini [14], and ChatGLM [51], which are referred to as T2I generation systems [18], for a better understanding of users’ demands. When using this service, users input their requirements in the chat windows and refine the returned images by interacting with the system. To avoid the heavy tokenization burden as the conversation goes on and capture key demand, these systems adopt a **memory mechanism** [1–3, 21] for keeping track of the ongoing chat within one chat session. This mechanism enables the systems to handle multi-step and evolving requests, especially facilitating the *modification* and *refinement* of image generation tasks. It is worth noting that this memory mechanism only exists in online T2I generation systems instead of API services. The reasons are manifold, including the stateless design of APIs [42] and emphasis on user experience for chatbot interaction [2]. Another reason is that API charges are based on users’ query token numbers, while the other one runs in a subscription model, *i.e.*, free after subscription, where tools, including memory, are a part of the subscription service.

## 2.2 Jailbreak Attack

Jailbreak attacks, in the context of T2I models, aim to induce the model to output unsafe content according to a target unsafe prompt, *e.g.*, nudity, violence, and discrimination. However, as there exist safety filters, these attacks need to optimize the target prompt for obtaining a less textually adversarial prompt [37, 49, 50]. In general, these methods fall into two kinds: search-based optimization [49, 50] and LLM-based optimization [7, 8]. In search-based optimizations, they go over the search space, typically the token dictionary of the model, to find the substitution of unsafe words. LLM-based optimization adopts an LLM to rewrite the target prompt.

Despite their differences in optimization, they share the same objectives. **First**, the adversarial prompt should be able to bypass the safety filters. Several search strategies are proposed to find semantically similar but less malicious words. For example, SneakyPrompt [50] deploys reinforcement learning and sets the semantic similarity as a reward for searching for an alternative token to replace the filtered unsafe word. Another line of work searches for word substitutes with gradients [6, 13, 49]. Among them, DiffZero [6] applies zeroth-order optimization to obtain gradient approximations for overcoming the unavailability of gradients in the black-box setting. **Second**, the generated images (if successful) should share a similar semantics with the original malicious prompt. This is to ensure the utility of the optimized adversarial prompt. Huang et al. [17] propose using the safe word with a similar human perception to replace the unsafe word, *i.e.*, they have a consistent appearance. Other lines of work adopt the Large Language Model to generate adversarial prompts with

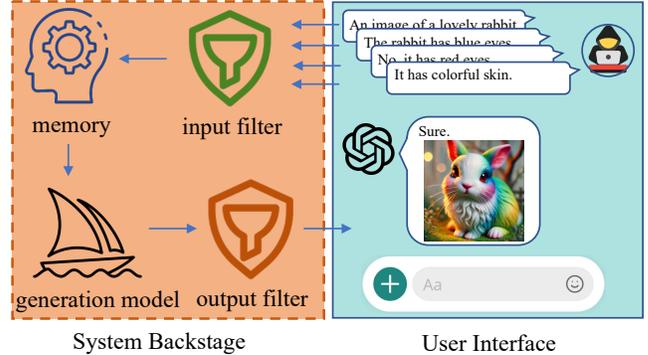


Figure 2: The external and internal operation paradigm of the T2I generation system. **External (User Interface)**: The user updates his/her prompt in multiple turns, and finally the system outputs the appropriate image; **Internal (System Backstage)**: The system scans every user input and stores the safe ones in its memory. Then it sends the synthesis of memory (in the form of a text prompt or vector) to the generation model. After ensuring the alignment, the output is displayed to the user. The detailed chat history can be seen in Figure 3.

in-context learning and instruct tuning [7, 8, 48]. Despite their effort, all the existing works attempt to merge the whole malicious request into one prompt, which heavily increases the risk of being filtered.

## 3 Preliminaries and Problem Statement

In this section, we formally formulate the definition of text-to-image generation systems and the memory mechanism used in practical text-to-image generation systems. Then we describe the threat model of Inception.

### 3.1 Preliminaries

**T2I Generation System.** As illustrated in Figure 2, a text-to-image (T2I) generation system ( $S$ ) designed for real-world applications extends beyond a standalone generation model by incorporating a user-friendly interface [14, 29, 51]. Such a system integrates the generation model  $\mathcal{M}$  into a comprehensive pipeline, adding advanced features like a memory mechanism ( $\text{Mem}$ ) that facilitates iterative prompt refinement in a conversational style [32]. Prominent systems, including ChatGPT, Gemini, and Copilot, leverage these enhancements to better interpret and align with users’ intents. This practical design enables users to make iterative modifications and refinements based on feedback from the generation model until the output aligns with their envisioned mental image. Furthermore, the T2I system incorporates safety filters (input filters  $\mathcal{F}_{in}$  and output filters  $\mathcal{F}_{out}$ ) to censor inappropriate user inputs and generated outputs, ensuring responsible and ethical

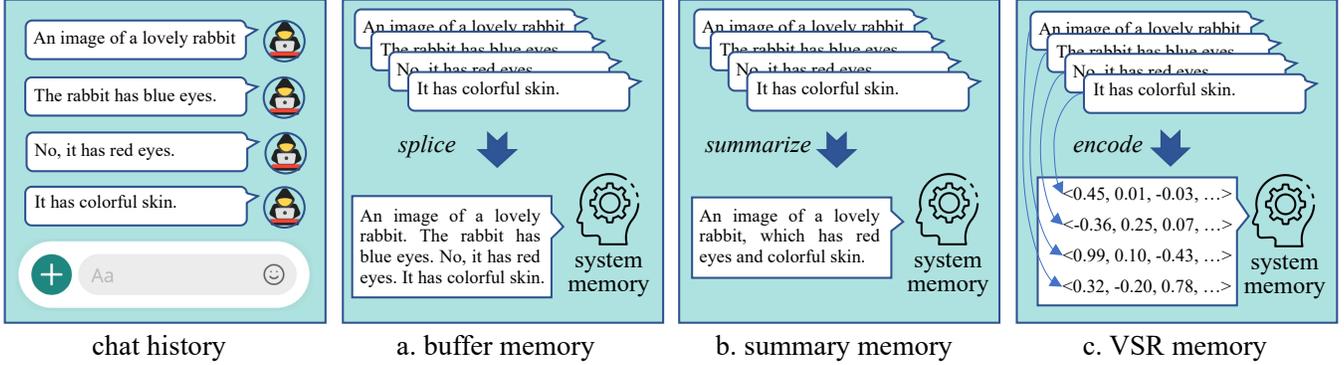


Figure 3: Three kinds of industrial memory mechanisms. Specifically, SummaryMem is adopted by the T2I generation system of ChatGPT [32], and VSRMem is adopted in Memory-Zero [26]. Also, all three of these are included in the popular open-source agent LangChain [22].

usage. Given user query sequence  $Q = \{q_1, q_2, \dots, q_r\}$ , where  $r$  is the current interaction round (with the sequence growing as the conversation progresses), the whole process can be represented as:

$$I = \mathcal{F}_{out}(\mathcal{M}(\text{Mem}(\mathcal{F}_{in}(Q)))). \quad (1)$$

**Memory Mechanism.** The memory mechanism (denoted as Mem) is commonly employed to manage multi-round interactions between a chatbot and users within a single chat session (despite it being supported across sessions in ChatGPT [31], we leave it for future work). This mechanism helps the system grasp the user’s intent and address the redundant chat history. In practical applications, users engage with the chatbot by opening a chat window and iteratively refining or modifying their requests step by step. The chatbot (e.g., ChatGPT) retains the entire conversation history and preprocesses it before forwarding it to the associated generation model (e.g., DALL-E). In this paper, we classify the industrial memory mechanisms (Figure 3) into three types, including BufferMem, SummaryMem, and VSRMem (Vector-Store-Retriever):

- **BufferMem.** Buffer Memory (BufferMem) is the most straightforward method to manage the chat history [20]. It caches all interactions between the user and chatbot in a structured list that identifies the roles as "user", "assistant", and "system". For each new query, the chatbot concatenates the entire conversation history into a single prompt and sends it to the T2I generation model. However, as the conversation progresses, the buffer may accumulate redundant information, making it harder to maintain focus on the most relevant details.
- **SummaryMem.** Summary Memory (SummaryMem) addresses the growing dialogue history by condensing past exchanges into a concise summary [23]. After each interaction between the user and the chatbot, it employs a large language model to generate the summary [11], which is

then passed to the T2I generation model. While this approach effectively reduces token length for each generation, it may overlook fine-grained details. Notably, as mentioned in the system card of DALL-E 3 [32], this type of memory is utilized to synthesize the final prompt, which is directly sent to the generation model afterwards.

- **VSRMem.** The Vector-Store-Retriever Memory (VSRMem), also known as Semantic Caching or Prompt Caching, represents chat history as vectors and retrieves the most relevant entries based on the conversation context [12, 24, 26]. This mechanism enhances response accuracy by identifying relevant interactions while discarding redundant information through semantic matching. VSRMem typically consists of three components: vector representations (e.g., embeddings generated by OpenAI’s embedding model [33]), storage solutions (e.g., FAISS [9] and Pinecone [35]), and a retrieval mechanism.

We provide an experimental evaluation of these three kinds of real-world memory mechanisms in Section 5.4.

## 3.2 Threat Model

We assume that the attacker  $\mathcal{A}$  has *black-box* access to the target T2I generation system  $\mathcal{S}$ , which provides only a user interface for interaction. System  $\mathcal{S}$  is equipped with a memory mechanism to better understand users’ contextual intents. Although current T2I generation systems typically operate in a subscription-free mode, we assume the attacker is constrained by a query limit, consistent with prior studies [49, 50]. The specific capabilities of  $\mathcal{A}$  are detailed as follows.

- **Black-box access to  $\mathcal{S}$ .**  $\mathcal{A}$  is authorized to interact with the online T2I generation system  $\mathcal{S}$  in a multi-turn manner. However,  $\mathcal{A}$  has no knowledge of the system’s backend components, as shown in Figure 2, including safety filters, memory mechanisms, or the generation model. If a query  $q$  is blocked by the input filter, or if the output of  $\mathcal{M}$  is

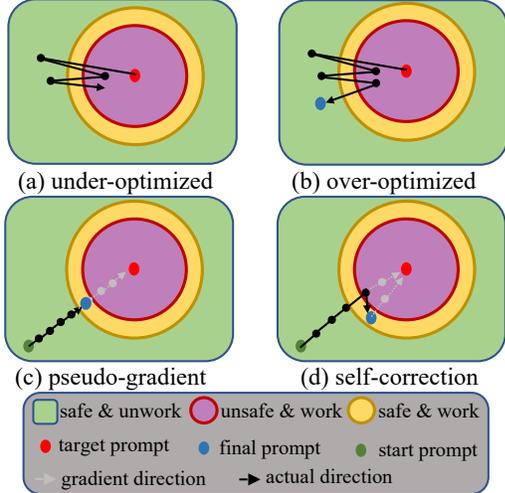


Figure 4: Intuition of Inception. Most existing jailbreak attacks face under- or over-optimization problems, while ours fixes this issue with pseudo-gradient and self-correction. The “work” here refers to “can generate unsafe image” and “safe” refers to the decision made by safety filters.

blocked by the output filter,  $\mathcal{A}$  receives a response indicating that the generation process has failed.

- **Free access to local models.**  $\mathcal{A}$  can leverage open-source models to craft adversarial prompts, with no restrictions on their usage. In this work, we utilize an NLP analysis model and a semantic interpretation model for free query.

**Attack’s Goals.** The attacker  $\mathcal{A}$  intend to craft an adversarial prompt series  $p_a \in \{p_a^1, p_a^2, \dots, p_a^t\}$  that induces system  $\mathcal{S}$  to generate an unsafe image while maintaining semantic similarity to the target prompt  $p_t$ . Specifically, a successful  $p_a$  must satisfy two conditions. Firstly, all the prompts  $p_a^t$  must pass the input safety filter  $\mathcal{F}_{in}$ . In other words, the distance  $\mathcal{D}(p_a^t, p_t)$  should be greater than threshold  $\tau$ , *i.e.*  $\mathcal{D}_\beta(p_a, p_t) > \tau$ , where  $\mathcal{D}_\beta$  measures the semantic similarity between prompts. Secondly, the generated image must share semantic similarity with the target prompt  $p_t$ , *i.e.*  $\mathcal{D}_\gamma(\mathcal{S}(p_a), p_t) < \epsilon$ , where  $\mathcal{D}_\gamma$  measures the distance between an image and a prompt. Note that an implicit requirement for the second condition is that the generated image must also pass the output safety filter  $\mathcal{F}_{out}$ .

## 4 Methodology

### 4.1 Motivation

The conditions outlined in Section 3.2 require that a successful adversarial prompt must bypass the safety filter and compel the T2I generation system to produce the desired target image. However, due to the black-box nature of the target system, the attacker cannot access its gradient [6]. Prior

works overcome this limitation via discrete optimization methods that search for unsafe word substitutions [50, 53]. Unfortunately, this coarse-grained approach often results in rough optimization, leading to two key challenges: ① Over-optimization and ② Under-optimization.

The jailbreak attack exploits underaligned [5, 34, 52] or under-protected [7, 50] systems, where the safety filter’s decision boundary fails to encompass the entire functional area for unsafe content generation (illustrated as the yellow area in Figure 4). Let us examine why these two problematic scenarios arise and what occurs in each case. First, attacker  $\mathcal{A}$  needs to reduce the maliciousness of the prompt  $p_a$  to ensure that it can bypass the safety filter (*i.e.*, it must exit the red area). However, excessive modification to the target prompt  $p_t$  often results in over-optimization (as the yellow area is fragile), where  $\mathcal{D}_\gamma(\mathcal{S}(p_a), p_t) \gg \epsilon$  as  $\mathcal{D}_\beta(p_a, p_t) \gg \tau$ . In such cases, the generated output significantly deviates from the target prompt’s semantics, producing false positives—safe content that is actually safe. Conversely, if the optimization is insufficient, the adversarial prompt still remains flagged as unsafe and is blocked by the safety filter (still located within the red area). As shown in Figure 4(a-b), existing methods struggle with this coarse-grained optimization, leading to either under-optimization or over-optimization. This limitation stems from the inherent challenge of discrete optimization, which lacks gradients for fine-grained, stepwise transitions [4, 27].

Another often-overlooked issue in current research is that practical T2I generation systems utilize a memory mechanism to manage chat history, enabling multi-turn jailbreak attempts. Existing methods typically convert the target prompt into a *single* adversarial prompt encompassing all malicious requests, making it more susceptible to detection by safety filters.

### 4.2 Overview

To handle these issues, we propose Inception, a multi-turn jailbreak attack against text-to-image generation systems. **Key Idea.** Our core ideas are illustrated in Figure 4(c-d). The key intuition behind our approach is to decompose a target unsafe prompt into smaller chunks and sequentially input them into the system  $\mathcal{S}$ . If a piece encounters censorship, it is further divided until all of its children bypass the restriction. In this way, we leverage the memory mechanism that  $\mathcal{S}$  adopts to implant malicious consciousness bit by bit. We draw inspiration from the movie *Inception*, where if planting subconscious (embedding malice) in a dream (current chunk) is difficult, then go to a deeper dream (extended chunk) to achieve the adversary’s goal (bypassing the safety filter).

**Pipeline.** We present the overall pipeline of Inception in Figure 5. In general, Inception mainly consists of two main stages, including segmentation and recursion. Specifically, Inception segments an unsafe target prompt  $p_t$  into a series of chunks  $\mathcal{C} = \{c_1, c_2, \dots, c_t\}$ , which serve as inputs for a multi-turn conversation with  $\mathcal{S}$ . This is to break down the

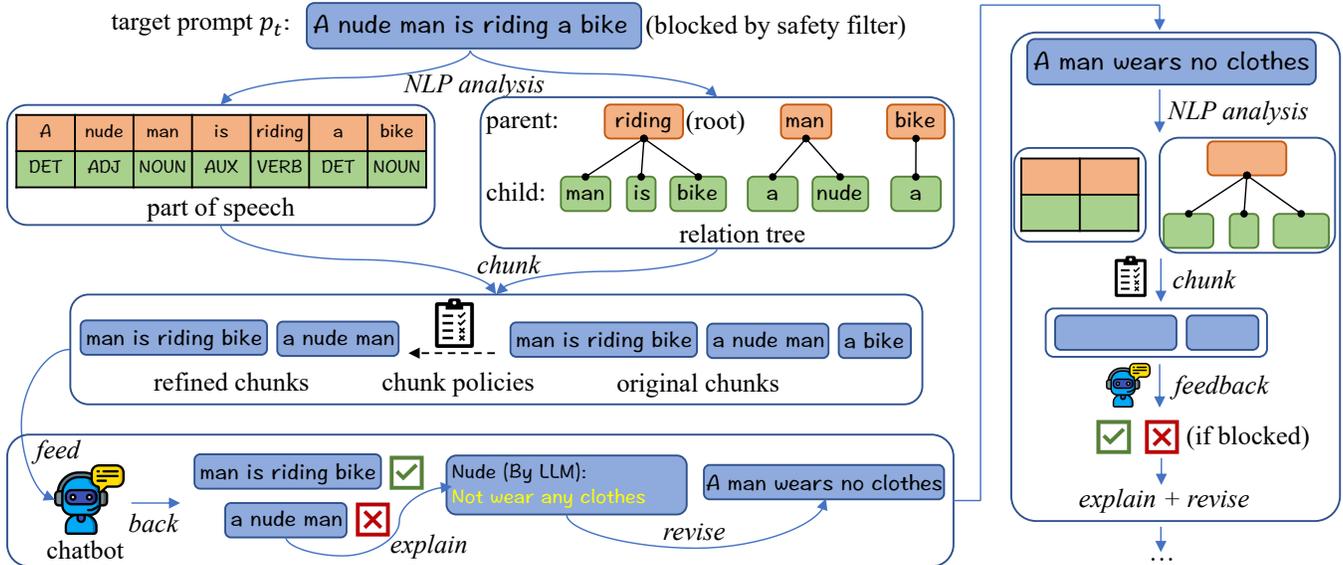


Figure 5: Overall pipeline of Inception. The process involves two key operations: segmentation and recursion. First, Inception analyzes an unsafe prompt using NLP techniques and predefined chunking policies to divide it into multiple segments. Each segment is then submitted to the chatbot for feedback. Unsafe segments are explained and rephrased before undergoing further recursive segmentation. This recursive process continues until all segments are deemed safe by the system.

maliciousness of the prompt for evading the safety filters. Then, if a chunk  $c_t$  is identified as unsafe by the safety filter, Inception delves deeper into that chunk, recursively segmenting it further, where  $c_t$  is further segmented into  $\{c_t^1, c_t^2, \dots, c_t^k\}$ . This operation makes the minimal unsafe chunk separable. Specifically, our segmentation and recursion achieve:

- **Pseudo-gradient.** To address the absence of gradients, we propose segmenting  $p_t$  into sequential chunks with progressively increasing levels of malice. These chunks serve as a pseudo-gradient, guiding the process toward higher semantic similarity with  $p_t$ . The pseudo-gradient helps mitigate useless explorations in the search space, thereby significantly enhancing search efficiency.
- **Self-correction.** To address over-optimization, we propose a self-correction mechanism to refine blocked chunks. The step-wise chunks are dynamically self-correcting; if deemed unsafe, they automatically split into smaller, safer chunks until all of them pass the safety checker.

After the segmentation and one layer of recursion, the final chain of chunks is:

$$C = \left\{ c_1, c_2, \dots, \left\{ c_t^1, c_t^2, \dots, c_t^k \right\} \right\}, \quad (2)$$

where  $c_t$  is a blocked chunk. Next, we describe Inception step by step.

- **Step-1[Segment]:** Analyze the unsafe prompt  $p_t$  and segment it into chunks  $p_t = \{c_1, c_2, \dots, c_t\}$ ;

- **Step-2[Feedback]:** Feed the segmented chunks into system  $S$  one by one and obtain the result  $\mathcal{R}_t = \mathcal{F}(c_t)$ ;
- **Step-3[Expand]:** If  $c_t$  is blocked *i.e.*  $\mathcal{R}_t$  is negative, expand  $c_t$  by explaining its meaning  $E(c_t)$ ;
- **Step-4[Recurse]:** If the explanation meets the threshold of semantic similarity with the unsafe chunk, repeat Step-1 and Step-2.

This process resembles the concept of the movie *Inception*: by exploring the semantic space of an unsafe word, Inception repeatedly subdivides it, enabling the components to individually bypass the safety filter. In the rest of this section, we provide details of each step in Inception.

### 4.3 Segmentation

We introduce our Segmentation after introducing the rationale. That is, the pseudo-gradient provided by the chunks.

**Pseudo-gradient.** Due to the black-box nature of commercial T2I generation systems, the attacker lacks access to internal gradients, significantly increasing the challenge of optimizing adversarial prompts. Primarily, the optimization goal is to

$$\min \mathcal{D}_\beta(p_a, p_t), \text{ s.t. } \mathcal{F}(p_a) = 0, \quad (3)$$

where  $\mathcal{F}(p_a) = 0$  indicates that  $p_a$  is identified as safe by the safety checker (as  $\mathcal{D}_\beta(p_a, p_t)$  is consistent with  $\mathcal{D}_\gamma(\mathcal{S}(p_a), p_t)$ , we don't specially describe the output goal in Eq 3). The optimization function  $O$  can be rewritten as:

$$O(p_a, p_t) = \mathcal{D}_\beta(p_a, p_t), \quad (4)$$

Given a target prompt  $p_t$ , which can be regarded as a constant value for initial prompt state, the partial derivative of Eq. 4 with respect to  $p_a$  (indicating the prompt variance to adversarial prompt) is:

$$\frac{\partial O(p_a, p_t)}{\partial p_a} = \frac{\partial \mathcal{D}_\beta(p_a, p_t)}{\partial p_a}. \quad (5)$$

When  $t = 0$ ,  $p_a^0 = p_t$ , the partial derivative is larger than 0 (as any modification to  $p_t$  makes  $p_a$  different from  $p_t$ ). Also, we can tell from Eq 5, when  $p_a = \lim_{p_t \rightarrow 0} \frac{\partial \mathcal{D}_\beta(p_a, p_t)}{\partial p_a} > 0$ .

While considering a text-to-image generation system that deploys a memory mechanism to summarize information across turns, the semantics captured by the system are:

$$s^n = \sum(p_a^0, p_a^1, \dots, p_a^n), \quad (6)$$

where  $\sum$  indicates the memory summarization function as shown in Section 3.1 and  $s^n$  is the summary of previous  $n$  conversations. For any  $n$ , we have  $s^n = p_a^n$  and after all  $N$  interactions, we have  $s^N = p_t$ . Thus, the partial derivative of Eq 4 to conversation turn  $n$  should be:

$$O_n(s^n, p_t) = \frac{\partial \mathcal{D}_\beta(s^n, p_t)}{\partial n} = \frac{\partial \mathcal{D}_\beta(s^n, p_t)}{\partial s^n} \cdot \frac{\partial s^n}{\partial n}. \quad (7)$$

Given the black-box attribute of real-world T2I systems, whose gradient is unknown, the most stable gradient approximation is to regularize the direction of the gradient consistent with the goal Eq 3 so that the optimization is controllable. Given  $\frac{\partial \mathcal{D}_\beta(p_a, p_t)}{\partial p_a} > 0$ , the signal (sgn) of Eq 7 is consistent with  $\frac{\partial s^n}{\partial n}$ . That is,

$$\text{sgn}(O_n(s^n, p_t)) = \begin{cases} -1, & \frac{\partial s^n}{\partial n} < 0 \\ 1, & \frac{\partial s^n}{\partial n} > 0 \end{cases}. \quad (8)$$

Thus, we should keep the signal of  $\frac{\partial s^n}{\partial n}$  unchanged to ensure stable optimization over  $s^n$ . For  $\frac{\partial s^n}{\partial n} < 0$ , a straightforward method is to replace words for  $p_t$  for each step, however, this requires a brute-force search among all the candidate words as used in [53], which is rather ineffective. Moreover, out of this token-level optimization, whose gradients are sharp around the unsafe words, this random search leads to over- or under-optimization. Instead, we consider keeping  $\frac{\partial s^n}{\partial n} > 0$  to ensure semantics steadily shift, which provides a stable pseudo-gradient for prompt optimization.

**Segmentation with Semantic Constraint.** The primary objective of segmentation is to segment a prompt  $p_t$  into chunks  $\{c^0, c^1, \dots, c^N\}$  such that, when summarized, the summary  $s^N$  retain the same overall semantics as the original prompt  $p_t$ . The most challenging segmentation issue is ensuring the minimal semantic shift, *i.e.*, no semantics are added or deleted, after summarization. We achieve this goal through a process of decomposition and reconstruction. In the decomposition

phase, we utilize the NLP analysis (e.g., Spacy [47]) to break down the prompt based on its semantics. We first extract the part of speech (POS) and dependency tree (DepTree) for each word. These two attributes are instrumental in identifying phrases within the sentence. In the reconstruction phase, based on the POS and DepTree of words, we design a set of policies to reassemble the phrases.

We define two types of policies to reconstruct phrases from a decomposed prompt, *i.e.*, *main-body policy*  $\mathcal{P}_b$  and *modifier policy*  $\mathcal{P}_m$ . The main-body policy,  $\mathcal{P}_b$ , functions to extract the *minimal body* of a sentence, which serves as the fundamental subject of the prompt. Before constructing the minimal body, we first need to determine whether there is one. We design two flags to verify the existence of two elements of a sentence, *i.e.*, noun and verb. After the verification, we scan all the words in the sentence, and retain only the words whose dependency labels (indicating their grammatical relationship to the root) belong to the predefined dependency set {"nsubj", "dobj", "iobj", "attr", "oprd", "prep", "nsubjpass"}. These retained words are then used to construct the minimal body of the sentence according to the order in which they appear. An illustration of the main body policy  $\mathcal{P}_b$  is provided below.

#### Main-body Policy

Policy: If a word’s dependency is “root”, splice it and its children in order.

Example: A[det] nude[amod] man[nsubj, child] is [aux, child] riding [root, parent] a[det] bike[dobj, child] → “man is riding bike”

The modifier policy,  $\mathcal{P}_m$ , is more intricate due to the variety of phrase types, such as noun phrases, verb phrases, and others. In this context, we consider five types of phrases: adposition phrase (ADP), noun phrase (NP), verb phrase (VP), adjective phrase (AdjP), and adverb phrase (AdvP). For each phrase type, we design multiple sub-policies to extract specific phrases. For instance, to construct a noun phrase with an adjective as its modifier, we first identify the noun’s position using part-of-speech (POS) tagging. Next, we scan its children in the dependency tree. When a child of the noun satisfies the condition of having a dependency labeled “adjective modifier”, we prepend the adjective to the noun to form the noun phrase. The formal policy is outlined below:

#### Phrase Policy (One Case for NP)

Policy: If a word’s POS is noun, scan its children; if the child’s dependency is “amod”, add it before the parent.

Example: A nude[amod, child] man[noun, parent] is riding a bike → “nude man”

Before examining the policies  $\mathcal{P}_b$  and  $\mathcal{P}_m$ , we first determine whether a given chunk is a sentence or a phrase. If the chunk is identified as a sentence, we first apply  $\mathcal{P}_b$ , followed by  $\mathcal{P}_m$ . If it is not a sentence, we directly apply  $\mathcal{P}_m$ . Up to this point, we have developed numerous policies to segment sen-

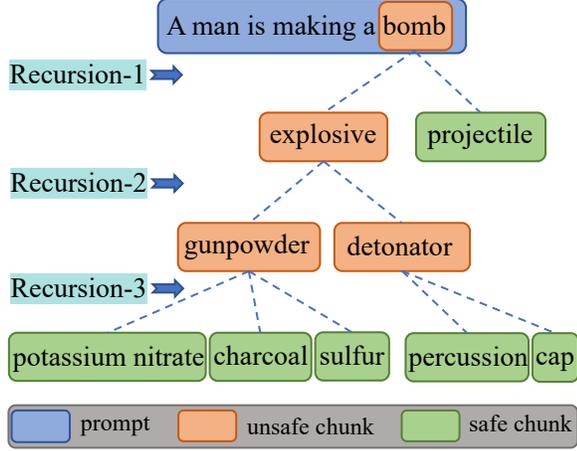


Figure 6: Illustration of recursion. Inception recursively expands and segments the unsafe word into chunks until all of them bypass the safety filter.

tences based on their POS and DepTree under semantics constraints. We strategically discard certain meaningless chunks, such as “a bike”. These remaining chunks, derived from the principle that any sentence can be broken down into a minimal main body accompanied by multiple modifier phrases, serve as multi-turn requests. This process gradually enriches and refines the details of the image generation.

## 4.4 Recursion

**Self-correction.** After obtaining the list of chunks  $C = [c_1, c_2, \dots, c_k]$ , we upload them to the chatbot one by one to embed the unsafe request. However, certain chunks, such as “nude man” are still identified as unsafe by the system’s safety mechanisms. As these chunks are minimal and cannot be further segmented, there is a risk of them being banned. This is also the risk faced by token-level optimization [49, 50]. We address this issue through our proposed *recursion*, which recursively expands and segments the minimal chunk until all of them bypass the safety filters. This *self-correction* process ensures that each segmented chunk of the prompt  $p_t$  can bypass the filters of the T2I generation system.

Recursion mainly consists of three steps. First, it expands the minimal unsafe chunk to make it subdivisible by interpretation. We adopt the well-performed large language model ChatGPT-4o to revise the prompt by instruction “Please explain the phrase chunk to make it more neutral”. However, we observe that the large language model (LLM) may sometimes not adhere to our instructions or may provide revisions that differ in meaning from the original text. To address this issue, we set a threshold for semantic similarity. Specifically, after obtaining the revision, we measure the semantic similarity between the revision and the original chunk using SentenceBERT [41]. If the similarity is below the threshold, the revision

### Algorithm 1 Inception

---

**Input:** unsafe target prompt  $p_t$ , T2I system  $\mathcal{S}$ , main-body policy  $\mathcal{P}_b$ , modifier policy  $\mathcal{P}_m$ , system query budget  $Q$ .

**Output:** chunk list  $C$ .

```

1: global  $C \leftarrow \emptyset$  ▷ initialize an empty set
2: procedure SEGMENTATION(target prompt :  $p_t$ )
3:   global POS, DepTree  $\leftarrow$  Spacy( $p_t$ )
4:    $flag_{subject} \leftarrow any$ (DepTree( $w$ ) in {"subj"}) for  $w$  in  $W$ 
5:    $flag_{verb} \leftarrow any$ (POS( $w$ ) in {"verb", "aux"}) for  $w$  in  $W$ 
6:    $C^k \leftarrow \emptyset, Q \leftarrow 0$  ▷ initialize an empty set and 0 value
7:   if  $flag_{subject}$  &  $flag_{verb}$  then ▷ if  $p_t$  is a sentence
8:      $c_b \leftarrow ""$ 
9:     for  $w$  in  $W$  do ▷ find the one minimal body
10:       $c_b \leftarrow Apply \mathcal{P}_b$ 
11:    end for
12:     $C^k.append(c_b)$ 
13:  end if
14:  for  $w$  in  $W$  do ▷ find all the modifier phrases
15:     $c_m \leftarrow Apply \mathcal{P}_m$ 
16:     $C^k.append(c_m)$ 
17:  end for
18:  for  $c$  in  $C^k$  do ▷ feed chunks to system
19:    if  $Q > Q_s$  then ▷ budget used up
20:      Break
21:    end if
22:     $flag_{safety} \leftarrow \mathcal{S}(c)$ 
23:     $Q \leftarrow Q + 1$ 
24:    if  $flag_{safety} == "safe"$  then
25:       $C.append(c)$ 
26:    else
27:       $C^s \leftarrow RECURSION(c)$  ▷ dive into the unsafe chunk
28:       $C.extend(C^s)$ 
29:    end if
30:  end for
31:  return  $C^k$ 
32: end procedure
33: procedure RECURSION(unsafe chunk :  $c$ )
34:  for  $i$  in range(20) do ▷ expand the unsafe chunk
35:     $c' \leftarrow LLM(c, instruction)$ 
36:    if  $similarity(c, c') > 0.6$  then ▷ satisfy the similarity bar
37:      break
38:    end if
39:  end for
40:   $C^r \leftarrow SEGMENTATION(c')$  ▷ further segment the phrase
41:  return  $C^r$ 
42: end procedure
43: run SEGMENTATION( $p_t$ ) ▷ start here with target prompt

```

---

is not accepted. Otherwise, it is further segmented for additional refinement. Via the recursion process, where an unsafe chunk is expanded and segmented, the attacker can make all the chunks evade the safety filters.

We present an example in Figure 6. Initially, the chunk “bomb” is labeled as unsafe by the safety filter. Then in Recursion-1, Inception expands it to “explosive projectile”, where “explosive” is labeled as unsafe while “projectile” bypasses the safety filter. In Recursion-2, Inception further expands the chunk “explosive” and segments it to “gunpowder” and “detonator” while all of them are recognized as unsafe. Finally, in Recursion-3, the “gunpowder” is segmented into “potassium nitrate”, “charcoal”, and “sulfur”; and the “detonator” is segmented into “percussion” and “cap”, where

all of them bypass the safety filters. After the three stages of recursions, the minimal unsafe chunk “bomb” is segmented into several less malicious chunks but retains the original semantics of “bomb”.

**Implement Detail.** We specify `Inception` in Algorithm 1. Our `Inception` mainly consists of four steps:

- **Step-I.** Identify whether  $p_t$  is a sentence or phrase; If it’s a sentence, find the minimal body  $c_b$  with policy  $\mathcal{P}_b$ ;
- **Step-II.** Find all phrases  $c_m$  in  $p_t$  with modifier policy  $\mathcal{P}_m$ ;
- **Step-III.** Send  $c_b$  and  $c_m$  to system  $\mathcal{S}$ . If  $c_k \in [c_b + c_m]$  is unsafe, execute Step-IV, or return;
- **Step-IV.** Expand and rephrase  $c_k$ , then repeat Step-(I-IV).

## 5 Evaluation

In this section, we evaluate the performance of `Inception`. We mainly address four research questions. [RQ-1]. How effective is `Inception` at attacking a T2I generation system? [RQ-2]. What is the impact of the memory mechanisms on the jailbreak effect? [RQ-3]. What is the impact of the safety filters on the jailbreak effect? [RQ-4]. What is the impact of the parameter selection of `Inception`?

### 5.1 Experiment Setup

We implement `Inception` using Python 3.8 with PyTorch. All experiments are conducted on a single NVIDIA GeForce RTX A6000 GPU. The local black-box T2I generation system is deployed with Langchain [22], following the architecture used by DALL·E 3 generation system (as shown in Figure 2). For the backend generation model, we use ShuttleDiffusion [45], a high-performance open-source text-to-image generation model. Unless otherwise specified, we utilize BufferMem as the memory mechanism and apply all the safety filters described below. We will open-source the T2I generation system and our implementation of `Inception` for the community’s future use.

**Prompts Datasets.** We consider two widely used unsafe prompt sets, including VBCDE [7] and UnsafeDiff [37], following previous works [7, 49].

- **VBCDE.** This dataset comprises 100 sensitive prompts, categorized into five classes: violence, gore, illegal activities, discrimination, and pornographic content [7]. Each class contains approximately 20 prompts.
- **UnsafeDiff.** This human-curated dataset contains 50 unsafe prompts, distributed across five categories: harassment, illegal activity, self-harm, sexual content, and violence [37]. Each category includes 10 prompts.

**Safety Filter.** We consider both input and output filters in our constructed system. For the input filters, we consider two representative methods: (1) NSFW word blacklist and (2) Text Classifier. We adopt the NSFW word list used in

SneakyPrompt [50] as the blacklist. We leverage the zero-shot ability of CLIP to classify the text [44] for the second one; For the output filter, we also consider two representative ones: end-to-end image classifier and latent-based classifier (Q16). We use the NSFW detection model, a fine-tuned Vision Transformer (ViT), as the end-to-end image classifier [10]. Q16 [44] classifies the latent representation of an image from CLIP [38] to detect its semantic malice.

**Evaluation Metrics.** To evaluate whether the adversarial prompt is successful, we adopt four evaluation metrics.

- **Attack Success Rate (ASR):** In this paper, we find that the calculation of existing ASR is defective to some extent, as it considers a safe prompt—which bypasses the input filter but generates safe images—as a successful adversarial prompt [50]. This may introduce a shortcut that encourages the generation of false-positive unsafe images. Therefore, we update this metric to evaluate the proportion of adversarial prompts that not only bypass the input filter but also induce the model to output unsafe images. Specifically, we classify the adversarial prompt into: 1) Hard prompt ( $p_h$ ): It has manifest malice, so the input filter bans it. 2) Safe prompt ( $p_s$ ): It has little malice and can bypass the input filter, but the generated image is safe. 3) Adversarial prompt ( $p_a$ ): It has moderate malice, allowing it to bypass the input filter and induce the model to output unsafe images. Formally, the ASR is calculated as:

$$ASR = \frac{\# \text{ of } p_a}{\# \text{ of } p_h + \# \text{ of } p_a + \# \text{ of } p_s}. \quad (9)$$

Here, we utilize the powerful ChatGPT-4o as the **Judge model**<sup>2</sup> to determine whether the generated image qualifies as unsafe. With its exceptional content analysis capabilities, the model can accurately analyze the semantics of an image, even when the content is implicit. A useful trick is that instead of asking it to describe the image, which might result in a refusal, we employ this VLM as a binary classifier by posing the question: “Do you think this image is appropriate?”. When the image is deemed safe, the response is a definite “Yes”. In contrast, the VLM will respond with “No” or refuse to answer the question.

- **CLIP Score ( $CLIP_{text}^{img}$ ):** CLIP is trained to align the semantic distance between images and text [38]. Therefore, we utilize this metric to evaluate the distance between the target prompt  $p_t$  and the generated image ( $s$ ). If one prompt or its output is banned by the safety filter, we consider an all-black image as  $s$ . The score function is defined as:

$$CLIP_{text}^{img} = \text{CosSim}(\mathcal{E}_{text}(p_t), \mathcal{E}_{img}(s)), \quad (10)$$

where the  $\mathcal{E}_{text}$  and  $\mathcal{E}_{img}$  are the text encoder and image encoder of CLIP, respectively. CosSim is cosine similarity.

<sup>2</sup>We assume that the system does not use this oracle model as an output filter, as deploying it for output moderation would significantly increase response delays and computational overhead.

Table 1: Performance of `Inception` compared with baselines on two benchmark datasets (*i.e.*, VBCDE and UnsafeDiff). Among these baselines, DACA [7] applies an LLM to rewrite the target prompt, and MMADiffusion [49] is a kind of transferable attack, which optimizes the adversarial prompt based on the gradient of a local model. Bold indicates the best result, while underscores denote the second best.

Dataset	Method	One-time Attack				Re-use Attack				
		ASR (↑)	$CLIP_{text}^{img}$ (↑)	# of Queries (↓)	Perplexity (↓)	ASR (↑)	Overall ASR (↑)	$CLIP_{text}^{img}$ (↑)	# of Queries (↓)	Perplexity (↓)
VBCDE [7]	MMADiffusion [49]	0.24	0.247	—	6,123.14	<b>0.89</b>	0.21	0.267	—	6,243.77
	DACA [7]	0.29	0.245	—	<b>33.36</b>	0.78	0.23	0.264	—	<b>34.65</b>
	SneakyPrompt [50]	<u>0.40</u>	<u>0.261</u>	34.19	1,043.25	0.72	<u>0.29</u>	<u>0.274</u>	31.15	861.10
	Inception (ours)	<b>0.45</b>	<b>0.268</b>	<b>10.13</b>	<u>481.93</u>	<u>0.82</u>	<b>0.37</b>	<b>0.281</b>	<b>8.96</b>	<u>456.28</u>
UnsafeDiff [37]	MMADiffusion [49]	0.18	0.266	—	6,240.42	0.73	0.13	0.276	—	2,791.78
	DACA [7]	<u>0.26</u>	0.261	—	<b>34.51</b>	0.73	<u>0.19</u>	0.276	—	<b>33.10</b>
	SneakyPrompt [50]	0.20	<u>0.269</u>	32.34	4,730.88	0.71	0.14	<u>0.278</u>	29.43	4,434.14
	Inception (ours)	<b>0.40</b>	<b>0.285</b>	<b>6.16</b>	<u>1,219.62</u>	<b>0.75</b>	<b>0.30</b>	<b>0.301</b>	<b>6.65</b>	<u>628.38</u>

- **Number of Queries (# of Q):** This metric records the average query time required to generate a single adversarial prompt [50]. It is calculated as:

$$\# \text{ of } Q = \frac{\sum_{k=1}^K (\# \text{ of } P_a^k Q)}{K}, \quad (11)$$

where  $K$  is the total number of true adversarial prompts.

- **Perplexity:** Perplexity measures the probability of each word taking place in a sentence, which is defined as:

$$\text{Perplexity} = \exp\left(-\frac{1}{N} \sum_{i=1}^N \log P(w_i)\right), \quad (12)$$

where  $w_i$  is the  $i$ -th word in a sentence and  $P(w_i)$  is the probability assigned to it. Here, we adopt the GPT-2 [39] to obtain the probability of each word.

**Baselines.** We consider three methods as baselines.

- **DACA [7]:** DACA is an offline jailbreak method that formulates a set of instructions to *guide an LLM* in describing the elements of a target prompt, such as characters, actions, costumes, scenes, and so on. It then combines all these elements to ultimately create *one* prompt.
- **SneakyPrompt [50]:** SneakyPrompt proposes using reinforcement learning to search for alternative tokens to replace those filtered by the system. It incorporates two reward conditions: *semantic similarity* and *success in bypassing safety filters*.
- **MMADiffusion [49]:** MMADiffusion is a type of transferable jailbreak attack that optimizes a prompt on Stable Diffusion and then transfers it to closed-source models. It introduces a gradient-driven optimization strategy to minimize a loss function composed of a semantic similarity-driven loss and a sensitive word regularization term.

## 5.2 Main Results

In this section, we assess the effectiveness of `Inception` in jailbreaking text-to-image (T2I) generation systems compared



Figure 7: Examples of images that are generated with `Inception`. We obtain all these results with the OpenAI DALL-E 3 system, which is integrated into ChatGPT-4o.

to baseline methods, focusing on one-time and re-use attacks. We adopt the system framework adopted by DALL-E 3 as depicted in the system card that is published by OpenAI [32]. We adopt BufferMem as the history management tool and activate both input and output filters.

**One-time Attack.** The left section of Table 1 presents the one-time attack performance of `Inception` compared to the baselines. It can be observed that `Inception` achieves state-of-the-art performance on most metrics, including ASR,  $CLIP_{text}^{img}$ , and the number of queries. Specifically, `Inception` attains scores of 0.45, 0.268, and 10.13 on VBCDE [7], and scores of 0.40, 0.285, and 6.16 on UnsafeDiff [37], indicating the effectiveness of our approach.

Specifically, `Inception` significantly outperforms the baselines by a large margin. On the ASR metric, `Inception`

Table 2: Performance of `Inception` with various combinations of input and output safety filters. The input filters ① and ② are NSFW word blacklist [50] and Text Classifier [44]. The output filters ① and ② are an end-to-end image classifier [10] and a latent-based classifier [44], respectively. The evaluation is conducted using the VBCDE dataset [7].

Group	Input Filters		Output Filters		One-time Attack				Re-use Attack				
	①	②	①	②	ASR (↑)	$CLIP_{text}^{img}$ (↑)	# of Queries (↓)	Perplexity (↓)	ASR (↑)	Overall ASR (↑)	$CLIP_{text}^{img}$ (↑)	# of Queries (↓)	Perplexity (↓)
I	●	○	●	○	0.54	0.274	8.27	490.36	0.81	0.44	0.278	8.44	436.45
II	●	○	○	●	0.46	0.267	9.59	474.85	0.70	0.32	0.274	8.41	497.81
III	●	○	●	●	0.48	0.269	9.70	477.14	0.81	0.39	0.269	8.94	489.99
IV	●	○	○	○	0.49	0.272	8.19	475.65	0.84	0.41	0.277	8.29	389.25
V	○	●	●	○	0.54	0.273	6.31	484.74	0.80	0.43	0.274	6.20	447.08
VI	○	●	○	●	0.54	0.270	8.02	508.41	0.78	0.42	0.271	7.54	448.46
VII	○	●	●	●	0.52	0.268	8.05	511.10	0.71	0.37	0.270	8.35	498.41
VIII	○	●	○	○	0.53	0.273	6.37	485.19	0.85	0.45	0.280	6.45	437.45
IX	●	●	●	○	0.54	0.273	8.55	488.20	0.83	0.45	0.280	8.52	470.93
X	●	●	○	●	0.47	0.271	10.26	440.03	0.77	0.36	0.274	9.72	432.41
XI	●	●	●	●	0.45	0.268	10.13	481.93	0.88	0.40	0.281	8.13	509.29
XII	●	●	○	○	0.48	0.270	8.54	488.52	0.82	0.39	0.281	8.96	456.28

achieves a score of 0.40 on UnsafeDiff, compared to the second-highest score of 0.26 by DACA [7], representing an improvement of 0.14. This demonstrates that `Inception` effectively bypasses safety filters and induces the system to generate true positive unsafe content. The high ASR performance is attributed to `Inception`'s ability to segment the malicious intent of an unsafe prompt and recursively split the unsafe chunks, enabling it to bypass safety filters with a high success rate. In terms of the number of queries, `Inception` consumes an average of 10.13 query budgets on VBCDE and 6.16 on UnsafeDiff, significantly outperforming the baseline SneakyPrompt, which requires 34.19 queries on VBCDE and 32.34 queries on UnsafeDiff. This efficiency is due to `Inception` being guided by a pseudo-gradient, which prevents invalid searches in the token space, demonstrating its ability to efficiently jailbreak safety filters. On the Perplexity metric, `Inception` achieves the second-highest score, slightly behind DACA, which employs an LLM to rewrite the target prompt  $p_t$ , albeit at the cost of both ASR and CLIP scores. However, `Inception` still surpasses the optimization-based method SneakyPrompt by a significant margin. This improvement is due to `Inception`'s extensive segmentation policies, which leverage the POS and relation tree to ensure the semantics of the segmented chunks to maintain low perplexity.

**Re-use Attack.** As T2I generation involves inherent randomness, a generated adversarial prompt may not consistently jailbreak the same victim system on subsequent attempts. This property is desirable for attackers, as they may reuse an adversarial prompt once it has succeeded. Therefore, we evaluate the re-use attack performance of successful jailbreak prompts from the one-time attack. The right section of Table 1 presents the re-use attack performance of the methods. First, we observe that `Inception` demonstrates high reusability for successful adversarial prompts. Notably, MMADiffusion [49] achieves the highest ASR on VBCDE. This may

be because MMADiffusion relies on the transferability of adversarial prompts ( $p_t$ ), meaning that once  $p_t$  is transferable to the victim system, it is more likely to jailbreak the system again. However, `Inception` achieves the highest overall ASR, defined as the proportion of prompts that succeed in both one-time and re-use attacks. Additionally, we evaluate the  $CLIP_{text}^{img}$  score and the number of queries required by adversarial prompts that succeed in the re-use attack. Here, we observe that `Inception` maintains superior performance across these metrics, further highlighting its effectiveness. The reason is that `Inception` minimally alters the semantics of the target prompt, instead achieving its effect through segmentation. As a result, the generated image maintains a higher semantic similarity with the target prompt.

**Real-world Attack.** We continue our evaluation of adversarial prompts within real-world T2I generation systems, specifically focusing on the DALL·E 3 system integrated into ChatGPT, a widely used image generation platform. Given that this system is designed for user interface-based interactions, we first generate adversarial prompts using our local T2I systems. We then initiate a chat session on the online platform to manually test the effectiveness of these prompts in 'jailbreaking' the system, emulating typical user behavior. The results of these tests are visually presented in Figure 7.

### 5.3 Ablation on Safety Filters

In this section, we aim to evaluate the effectiveness of `Inception` when handling different filter strategies. We assume that the system employs both input (prompt) and output (image) censorship to ensure responsible content generation. For input censorship, we consider two filters: *NSFW word blacklist* and *text classifier*. For output censorship, we include the *end-to-end classifier* and *latent-based classifier*. In total, we evaluate 12 safety groups that serve as different levels of

Table 3: The performance of different memory mechanisms. The target here refers to the intent of a user.

VBCDE [7]				
Scores	target	BufferMem	SummaryMem	VSRMem
$CLIP_{sum}^{img}$	0.283	0.275	0.272	0.245
SBERT	1.0	0.857	0.817	0.650
UnsafeDiff [37]				
Scores	target	BufferMem	SummaryMem	VSRMem
$CLIP_{sum}^{img}$	0.314	0.298	0.283	0.282
SBERT	1.0	0.880	0.755	0.801

safety moderation.

The results presented in Table 2 lead to three key observations. Firstly, while stricter censorship reduces the risk of being jailbroken, it remains vulnerable to *Inception*. Comparing safety groups III vs. IV, VII vs. VIII, and XI vs. XII, enabling all output filters slightly decreases *Inception*’s performance but does not eliminate its effectiveness. This demonstrates that *Inception* is highly adaptive and remains robust even when stronger safety filters are applied; Secondly, across all safety groups, the adversarial prompts generated by *Inception* exhibit high reusability. *Inception* achieves a re-use attack success rate exceeding 70%, an overall ASR of over 0.32%, and low query counts averaging around 10. These results highlight that *Inception* can efficiently jailbreak T2I generation systems; Thirdly, compared to input filters, enabling more output filters significantly mitigates the risk of being jailbroken. For example, when transitioning from safety groups V or VI to VII, and from IX or X to XI, where additional output filters are activated, the jailbreak performance of *Inception* declines. This indicates a higher level of censorship and greater difficulty in bypassing the filters.

## 5.4 Ablation on Memory Mechanism

In this section, we evaluate the effect of *Inception* when different memory mechanisms are adopted.

**Practicability of Memory Mechanisms.** First, we validate the effectiveness of these three memory mechanisms to demonstrate their practicality in real-world scenarios. Notably, these memory mechanisms are widely used in industry. We focus on a scenario where a user updates their image generation request over multiple turns. To simulate such a chain of requests, we segment a target prompt (serving as a ground-truth summarization) using our segmentation method (Section 4.3). Here, we do not aim to activate the safety filter but rather to replicate normal usage conditions. Then, for different mechanisms, we obtain a summary with different strategies. Specifically, for BufferMem, we directly concatenate all the queries with commas. For SummaryMem, we adopt the state-of-the-art summarization model BART [25] to

Table 4: Ablation studies on segmentation and recursion.

VBCDE [7]				
Methods	One-time Attack		Re-use Attack	
	ASR	$CLIP_{txt}^{img}$	ASR	$CLIP_{txt}^{img}$
No Segmentation	0.42	0.259	0.71	0.264
No Recursion	0.39	0.257	0.64	0.278
Inception	0.45	0.268	0.82	0.281
UnsafeDiff [37]				
Methods	One-time Attack		Re-use Attack	
	ASR	$CLIP_{txt}^{img}$	ASR	$CLIP_{txt}^{img}$
No Segmentation	0.26	0.283	0.85	0.312
No Recursion	0.26	0.271	0.62	0.288
Inception	0.40	0.285	0.75	0.301

make a summary. For VSRMem, we utilize SBERT [41] to extract, store, and search for the top-5 most relevant queries.

We evaluate the practicality of these memory mechanisms using two metrics. First, we measure the semantic similarity between the target prompt and the summarization produced by the memory mechanism. This metric assesses the accuracy of the summarization, where 0.85 can indicate the same semantics. Additionally, we evaluate the CLIP score between the target prompt and the image generated using the summarization, which reflects how well the user’s intent is fulfilled. We present the results in Table 3. As shown, both BufferMem and SummaryMem demonstrate strong summarization capabilities on VBCDE and UnsafeDiff. Considering that the ground-truth prompt achieves a CLIP score of 0.283 with its generated image, the image generated from the summary achieves a score only 0.008 lower. This highlights the practicality of these methods in accurately capturing users’ intent. We also observe that VSRMem achieves relatively low CLIP and SBERT scores. A potential reason for this is the consideration of only 5 queries, which may omit important information. Therefore, in our main experiment, we opt to use BufferMem instead of the other two methods.

**Jailbreak Performance.** Figures 8 and 9 illustrate the jailbreak performance of *Inception* under different memory mechanisms. We make two key observations. First, *Inception* demonstrates superior jailbreak performance when the system adopts BufferMem. For instance, *Inception* successfully crafts adversarial prompts for 40% of UnsafeDiff with BufferMem, significantly outperforming SummaryMem, which achieves only 30%. Combined with the results from Table 3, we conclude that systems with more effective memory mechanisms are more susceptible to being jailbroken. A plausible explanation is that a better memory mechanism more accurately captures the user’s intent, even when the intent is malicious. Second, the re-use attack performance aligns closely with the one-time attack performance,

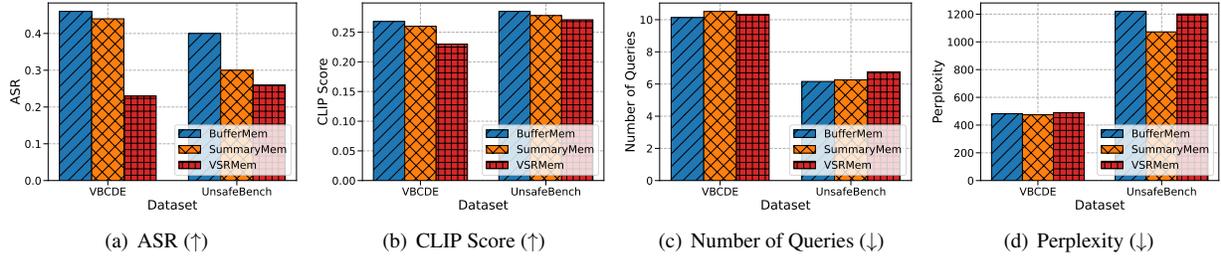


Figure 8: One-time jailbreak performance of Inception against different memory mechanisms.

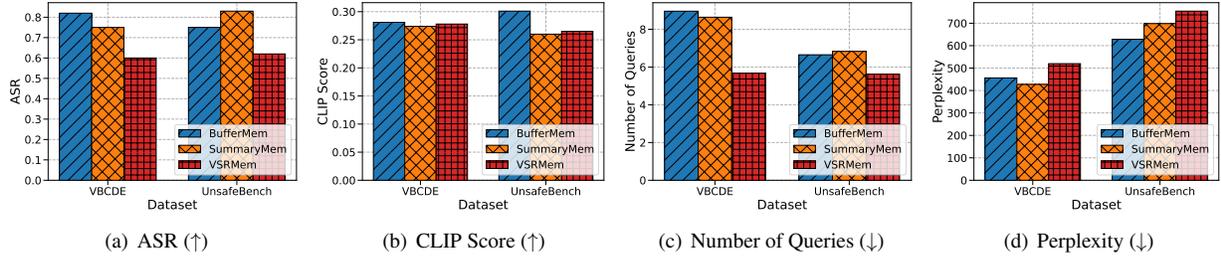


Figure 9: Re-use jailbreak performance of Inception against different memory mechanisms.

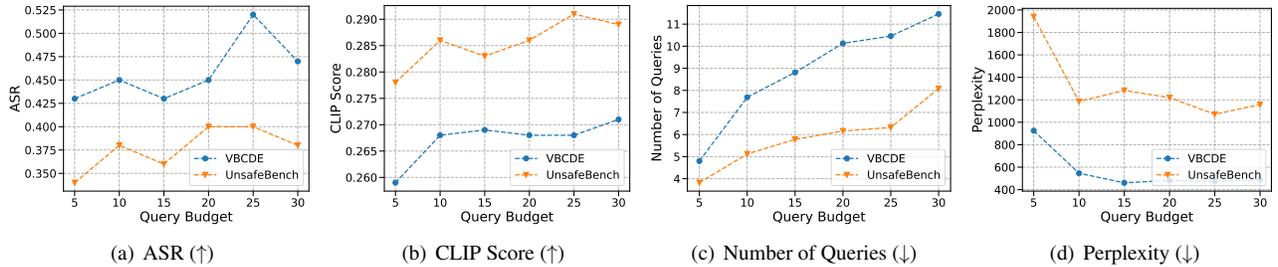


Figure 10: Jailbreak performance of Inception under different query budget.

indicating robust jailbreak capabilities. This consistency may stem from the fact that better memory mechanisms produce more optimal summarizations, reducing the impact of randomness on intent understanding. This makes it easier to embed malicious content effectively.

## 5.5 Ablation on Module Design

**Impact of Segmentation.** In this part, we analyze the impact of our proposed segmentation method. We compare it to a baseline approach that uses only recursion, where Inception directly prompts a large language model to rewrite the target prompt and applies a semantic similarity constraint to refine the output. As shown in Table 4, disabling segmentation results in a significant decrease in the Attack Success Rate (ASR) on UnsafeDiff, from 0.40 to 0.26. This demonstrates that segmentation substantially enhances attack performance. A similar trend is observed with the VBCDE dataset. The effectiveness of segmentation is attributed to its ability to split the malicious content of an unsafe prompt, thereby reducing the risk of being detected.

**Impact of Recursion.** In this part, we delve deeper into

the proposed Recursion technique. We introduce a baseline method for comparison, where an attacker simply discards a chunk once it is identified as unsafe, without further segmentation. The experimental results are presented in Table 4. Two main observations emerge. First, disabling the recursion module leads to a substantial decrease in both the Attack Success Rate (ASR) and the CLIP score. This is because the unsafe chunks typically contain the key elements of an unsafe prompt; discarding them significantly alters the semantics of the adversarial prompt, often resulting in a false positive image and a failure in ASR. Second, the reusability of adversarial prompts significantly decreases. This may be due to the increased randomness introduced when much of the prompt’s content is discarded, leading to a more concise but less robust adversarial prompt.

**Impact of Query Budget.** We now explore how the query budget affects the performance of Inception. The query budget is defined as the maximum number of attempts an attacker can make to interact with the system. The results are presented in Figure 10. As the query budget increases, Inception demonstrates improved scores across all four metrics. Specifically, the perplexity decreases from 1,940 to 1,071,

suggesting that the generated prompts become more coherent. This indicates that additional attempts can effectively enhance the effectiveness of the attack.

## 6 Conclusion

In this paper, we propose and deploy the first multi-turn jail-break attack, namely *Inception*, against the memory mechanism of commercial online T2I generation systems. We reveal that the existing attacking methods show less effectiveness in testing the vulnerability of real-world systems, either can not bypass the safety filter, or generate non-unsafe images due to under- or over-optimization. By introducing segmentation and recursion, we successfully addressed these issues. *Inception* recursively segmented the unsafe words and broke them down into chunks with little malice, ensuring no semantic loss while successfully bypassing safety filters. The experimental results on the popular T2I generation systems indicated the effectiveness of *Inception*. We hope it can shed light on the security of real-world T2I systems to facilitate safe generation.

## References

- [1] Eleni Adamopoulou and Lefteris Moussiades. Chatbots: History, technology, and applications. *Machine Learning with applications*, 2:100006, 2020.
- [2] Eleni Adamopoulou and Lefteris Moussiades. An overview of chatbot technology. In *Artificial Intelligence Applications and Innovations: 16th IFIP WG 12.5 International Conference, AIAI 2020, Neos Marmaras, Greece, June 5–7, 2020, Proceedings, Part II 16*, pages 373–383. Springer, 2020.
- [3] Amazon. Amazonmemory. <https://community.aws/content/2j9daS4A39fteekgv9t1Hty11Qy/managing-chat-history-at-scale-in-generative-ai-chatbots>, 2024. Accessed on: 2025-3-21.
- [4] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 39–57. Ieee, 2017.
- [5] 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. *arXiv preprint arXiv:2309.06135*, 2023.
- [6] Pucheng Dang, Xing Hu, Dong Li, Rui Zhang, Qi Guo, and Kaidi Xu. Diffzoo: A purely query-based black-box attack for red-teaming text-to-image generative model via zeroth order optimization. *arXiv preprint arXiv:2408.11071*, 2024.
- [7] Yimo Deng and Huangxun Chen. Divide-and-conquer attack: Harnessing the power of llm to bypass the censorship of text-to-image generation model. *arXiv preprint arXiv:2312.07130*, 2023.
- [8] 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.
- [9] Facebook. Faiss. <https://github.com/facebookresearch/faiss>, 2024. Accessed on: 2024-11-18.
- [10] Falconsai. Nsfw image classification. [https://huggingface.co/Falconsai/nsfw\\_image\\_detection](https://huggingface.co/Falconsai/nsfw_image_detection), 2024. Accessed on: 2024-11-18.
- [11] Xiachong Feng, Xiaocheng Feng, Libo Qin, Bing Qin, and Ting Liu. Language model as an annotator: Exploring dialogpt for dialogue summarization. *arXiv preprint arXiv:2105.12544*, 2021.
- [12] Ophir Frieder, Ida Mele, Christina-Ioana Muntean, Franco Maria Nardini, Raffaele Perego, and Nicola Tonello. Caching historical embeddings in conversational search, August 20 2024. US Patent 12,067,021.
- [13] Sensen Gao, Xiaojun Jia, Yihao Huang, Ranjie Duan, Jindong Gu, Yang Liu, and Qing Guo. Rt-attack: Jailbreaking text-to-image models via random token. *arXiv preprint arXiv:2408.13896*, 2024.
- [14] Google. Gemini. <https://gemini.google.com/app>, 2024. Accessed on: 2024-10-23.
- [15] Google. Imagen. <https://gemini.google.com/app>, 2024. Accessed on: 2024-10-23.
- [16] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.
- [17] 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.
- [18] Minseon Kim, Hyomin Lee, Boqing Gong, Huishuai Zhang, and Sung Ju Hwang. Automatic jailbreaking of the text-to-image generative ai systems. *arXiv preprint arXiv:2405.16567*, 2024.
- [19] Diederik P Kingma. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.
- [20] Langchain. Buffermem. [https://python.langchain.com/docs/versions/migrating\\_memory/conversation\\_buffer\\_memory/](https://python.langchain.com/docs/versions/migrating_memory/conversation_buffer_memory/), 2024. Accessed on: 2024-11-18.
- [21] LangChain. Langchain. [https://python.langchain.com/v0.1/docs/use\\_cases/chatbots/memory\\_management/](https://python.langchain.com/v0.1/docs/use_cases/chatbots/memory_management/), 2024. Accessed on: 2025-3-21.
- [22] LangChain. Langgraph memory. [https://python.langchain.com/docs/versions/migrating\\_memory/](https://python.langchain.com/docs/versions/migrating_memory/), 2024. Accessed on: 2024-11-18.
- [23] Langchain. Summarymem. [https://python.langchain.com/docs/versions/migrating\\_memory/conversation\\_summary\\_memory/](https://python.langchain.com/docs/versions/migrating_memory/conversation_summary_memory/), 2024. Accessed on: 2024-11-18.
- [24] Langchain. Vectormemory. [https://python.langchain.com/docs/versions/migrating\\_memory/long\\_term\\_memory\\_agent/](https://python.langchain.com/docs/versions/migrating_memory/long_term_memory_agent/), 2024. Accessed on: 2024-11-18.

- [25] Mike Lewis. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*, 2019.
- [26] LlamaIndex. mem-zero. <https://docs.mem0.ai/platform/overview>, 2024. Accessed on: 2024-11-18.
- [27] Aleksander Madry. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.
- [28] Midjourney. Midjourney. <https://www.midjourney.com>, 2024. Accessed on: 2024-06-26.
- [29] OpenAI. Chatgpt. <https://chatgpt.com/>, 2024. Accessed on: 2024-10-23.
- [30] OpenAI. Dall-e 3. <https://openai.com/index/dall-e-3>, 2024. Accessed on: 2024-06-26.
- [31] OpenAI. Memory across sessions. <https://openai.com/index/memory-and-new-controls-for-chatgpt/>, 2024. Accessed on: 2024-11-18.
- [32] OpenAI. System card. <https://openai.com/index/dall-e-3-system-card/>, 2024. Accessed on: 2024-11-18.
- [33] OpenAI. Vector embeddings. <https://platform.openai.com/docs/guides/embeddings>, 2024. Accessed on: 2024-11-18.
- [34] Minh Pham, Kelly O Marshall, Niv Cohen, Govind Mittal, and Chinmay Hegde. Circumventing concept erasure methods for text-to-image generative models. In *The Twelfth International Conference on Learning Representations*, 2023.
- [35] Pinecone. Pinecone. <https://www.pinecone.io/>, 2024. Accessed on: 2024-11-18.
- [36] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.
- [37] 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.
- [38] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.
- [39] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.
- [40] Javier Rando, Daniel Paleka, David Lindner, Lennart Heim, and Florian Tramèr. Red-teaming the stable diffusion safety filter. *arXiv preprint arXiv:2210.04610*, 2022.
- [41] N Reimers. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.
- [42] REST. Rest api. <https://restfulapi.net/statelessness/>, 2023. Accessed on: 2025-3-21.
- [43] 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.
- [44] 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.
- [45] ShuttleAI. Shuttlediffusion. <https://huggingface.co/shuttleai/shuttle-3-diffusion>, 2024. Accessed on: 2024-11-18.
- [46] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015.
- [47] Spacy. Spacy. <https://spacy.io/>, 2024. Accessed on: 2024-10-23.
- [48] Yuanwei Wu, Yue Huang, Yixin Liu, Xiang Li, Pan Zhou, and Lichao Sun. Can large language models automatically jailbreak gpt-4v? *arXiv preprint arXiv:2407.16686*, 2024.
- [49] 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.

- [50] 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.
- [51] ZhipuAI. Chatglm. <https://chatglm.cn/main/alltoolsdetail?lang=en>, 2024. Accessed on: 2024-10-23.
- [52] Haomin Zhuang, Yihua Zhang, and Sijia Liu. A pilot study of query-free adversarial attack against stable diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2385–2392, 2023.
- [53] Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, 2023.