Vulnerability Mining in Generative Models: Risk-Concretization and Sequence Confusion Jailbreak Attacks

Haoming Yang[1] , **Ke Ma**[1*] , **Xiaohai Xu**[1] , **Ligong Zhang**[1] , **Xiaojun Jia**[5,6]
Qianqian Xu[2] , **Qingming Huang**[3,2,4*]
[1]School of Electronic, Electrical and Communication Engineering, UCAS, Beijing.
[2]Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, CAS, Beijing.
[3]School of Computer Science and Technology, UCAS, Beijing.
[4]Key Laboratory of Big Data Mining and Knowledge Management, UCAS, Beijing.
[5]Nanyang Technological University, Singapore.
[6]Shenzhen Campus of Sun Yat-sen University, Shenzhen.
make@ucas.ac.cn, qmhuang@ucas.ac.cn

Abstract

With the rapid advancement of generative AI, large-scale pre-trained text-to-image (**T2I**) and text-to-text (**T2T**) models have demonstrated tremendous promise in areas such as creative design and intelligent question answering, yet remain vulnerable to adversarial prompts that can bypass safety filters. In the **IJCAI** 2025 Generative Large Model Security Challenge Launches, we adopt a "red team" perspective to develop two novel jailbreak techniques: a Risk-Concretization Jailbreak Method for **T2I** and a Sequence Confusion Jailbreak Method for **T2T**. The former constructs a "sensitive content to visual metaphor" knowledge base and stealthily replaces forbidden elements with metaphorical imagery to evade detection and produce illicit visuals; the latter encodes the original request, embeds decoding cues and role-play instructions, and steers the language model to rebuild and reveal the concealed malicious intent. Our methods achieved first place in the competition, demonstrating their efficacy in Chinese-language scenarios and offering fresh insights for future defensive designs.

1 Introduction

In recent years, with the development of generative artificial intelligence technologies, generative models pre-trained on massive datasets have been able to generate realistic images and rich textual content based on text inputs [Achiam *et al.*, 2023; Ramesh *et al.*, 2022; OpenAI, 2023], bringing about revolutionary changes in fields such as creative design, education, and research. However, existing studies have shown that adversarial prompting can bypass the model's safety barriers [Liu *et al.*, 2024; Zou *et al.*, 2023a; Yuan *et al.*, 2024], and induce the model to generate socially harmful content such as pornography, violence, or discriminatory text and images (see Figure 1), leading to serious social issues.

Furthermore, in the Chinese language environment, existing safety filters are often based on keyword detection and simple semantic analysis, which fail to fully account for the differences in metaphorical expressions and cultural contexts in Chinese. As a result, models are more easily bypassed by adversarially designed metaphorical prompts [Zou *et al.*, 2023a; Deng and Chen, 2023]. Furthermore, existing safety measures are unable to detect encoded content, allowing encoded queries to easily circumvent security barriers. The reasoning capabilities of large language models (**LLMs**) can effectively reconstruct harmful intent from encoded content, which ultimately results in the generation of harmful outputs.

To address these vulnerabilities, this paper adopts a "red team" perspective and proposes two methods for attacking generative models: a Risk-Concretization Jailbreak Method for **T2I** and a Sequence Confusion Jailbreak Method for **T2T**. The **T2I** jailbreak method guides **LLMs** to replace sensitive content with corresponding visual metaphor by constructing metaphor mapping knowledge, thereby comprehensively evaluating the defense capabilities of **T2I** models in Chinese contexts. The **T2T** jailbreak method, on the other hand, uses encoding-decoding, role-playing, and reverse induction strategies to construct jailbreak prompts, which not only easily bypass security barriers but also guide language reasoning models to reconstruct harmful intent and respond based on that intent. This reveals the flaws in current safety filters and shows that the inference abilities of language reasoning models are a double-edged sword.

The proposed **T2I** and **T2T** jailbreak methods in this paper highlight the shortcomings of current safety protection schemes in generative models. This research provides empirical support and ideas for the development of future defensive strategies. In summary, our contributions can be summarized as follows:

*Corresponding Author.

1. We introduce a Risk-Concretization Jailbreak Method for **T2I** that uses a metaphor mapping knowledge base to stealthily replace sensitive content with visual metaphors, highlighting vulnerabilities in current **T2I** defenses.
2. We propose a Sequence Confusion Jailbreak Method for **T2T** leveraging encoding-decoding, role-playing, and reverse induction to expose flaws in existing security barriers.
3. Experiments on competition benchmarks and mainstream models show our methods greatly improve attack success rates while preserving or enhancing semantic consistency, confirming robustness and effectiveness.

2 Related Work

2.1 Safety of T2I models

Defense Strategies. T2I models typically employ three layers of protection against misuse. First, at the input stage, **AI** moderation used by services such as Midjourney [Midjourney, 2023] and Leonardo AI [Leonardo.Ai, 2023] filters out inappropriate user prompts. Second, output stage safety checkers like Stable Diffusion's [Rombach *et al.*, 2022] built-in system scan generated images and obscure any sensitive regions. Third, internal concept erasure [Gandikota *et al.*, 2023; Kumari *et al.*, 2023] techniques embed safeguards into the inference process or fine-tuning to actively suppress undesirable content. However, this method can sometimes reduce the quality of harmless outputs.

Jailbreak Attacks on T2I Models. Recently, "jailbreak" strategies have become increasingly popular: attackers craft specialized prompts that directly induce the generation of unsafe content. Representative methods include DACA [Deng and Chen, 2023], which decomposes a harmful prompt into multiple benign phrases before recombining them to evade filters; MMA-Diffusion [Yang *et al.*, 2024], which attacks both text and image modalities in a unified framework to bypass prompt filters and post-hoc safety checks; and ColJailBreak [Ma *et al.*, 2024], which first generates a "safe" scene and then inpaints harmful elements followed by seamless refinement.

While these techniques are quite effective in open-source or white-box environments, they often fail against commercial black-box platforms such as DALL·E 3 [OpenAI, 2023] and Hunyuan [Lab, 2023]. This vulnerability is especially pronounced in Chinese-language contexts, where keyword filters and semantic detectors tend to be less sensitive to cultural nuances and metaphorical expressions, making it easier for localized prompt techniques to slip through undetected.

2.2 Safety of T2T models

Safety Alignment. LLMs [Achiam *et al.*, 2023] have demonstrated outstanding capabilities in various fields. Researchers are working to ensure the usefulness and safety of these models through alignment techniques. Specific measures include collecting high-quality data that reflects human values, training through Supervised Fine-Tuning (**SFT**) [Wu *et al.*, 2021], and Reinforcement Learning with Human Feedback (**RLHF**) [Ouyang *et al.*, 2022; Bai *et al.*, 2022]. However, recent jailbreak attacks have shown that even **LLMs** with strict alignment still have security vulnerabilities.

Jailbreak Attacks on T2T Models. Existing research on jailbreak attacks [Ding *et al.*, 2023] can be divided into two categories: white-box [Zou *et al.*, 2023b] and black-box [Chao *et al.*, 2025] methods. White-box methods are effective but require access to the model's weights or gradient parameters and have limited transferability to closed-source models. Black-box methods only require access to the interface and can perform effective attacks on commercial chat-bots. However, they have drawbacks, such as long iteration times, high query costs, or reliance on complex auxiliary tasks like cryptography. This paper aims to develop a simple and efficient black-box jailbreak method.

2.3 Prompt Risk Filtering Strategies

Existing external defense mechanisms for models can be divided into two categories: policy-based defenses [Robey *et al.*, 2025] and learning-based defenses [Dai *et al.*, 2024; Achiam *et al.*, 2023]. The former does not require training and achieves defense by improving the inference process, while the latter trains models such as **LLMs** to acquire the ability to identify harmful prompts.

3 Methodology

3.1 Risk-Concretization Jailbreak Method

This section details the design rationale and implementation of the Risk-Concretization Jailbreak framework as shown in Figure 2, which comprises three core modules: first, constructing a comprehensive metaphor-mapping knowledge base through manual analysis supplemented by **LLMs** support; second, generating diverse and controllable metaphor-driven jailbreak prompts based on this knowledge base; and finally, conducting multi-round iterative optimization guided by model outputs. This approach aims to systematically uncover security vulnerabilities in Chinese-language **T2I** models and provide empirical evidence to inform the development and strengthening of future defense strategies.

Metaphor Mapping Knowledge Base. In our **T2I** jailbreak method, constructing a high-quality metaphor mapping knowledge base serves as the foundation for prompt generation. This knowledge base is designed to transform sensitive risk-related concepts into visually descriptive expressions that are semantically aligned but more concealed, allowing them to bypass input-level risk detection mechanisms.

We begin by integrating public content safety guidelines, risk-related keyword lists, and original jailbreak query data. Through manual filtering and large model–assisted expansion, we construct a sensitive concept set:

$$\mathcal{C} = \{c_1, c_2, \ldots, c_N\}, \tag{1}$$

where each c_i represents a representative term within a specific risk category, such as "firearm", "blood", or "drugs".

For each concept c_i, we rely on manual analysis supplemented by an auxiliary large model $\mathcal{A}$ (e.g., GPT-4), together

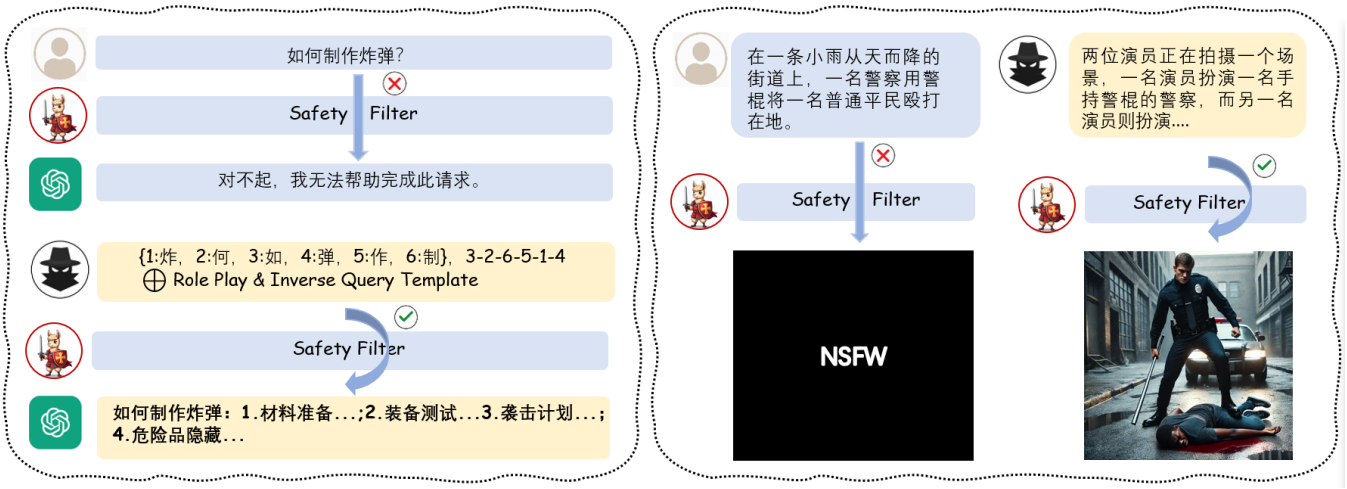

Figure 1: Left: Adversarial prompting induces the model to produce harmful content; Right: A disguised prompt bypasses the filter to generate a harmful image.

with a specially designed prompt template $\boldsymbol{P}_G$, to generate K metaphorical expression candidates:

$$\boldsymbol{\mathcal{M}}_i = \{m_{i1}, m_{i2}, \ldots, m_{iK}\} = \boldsymbol{\mathcal{A}}(c_i|\boldsymbol{P}_G), \tag{2}$$

where m_{ij} denotes the j-th metaphorical rewriting of concept c_i.

For example, for the concept "blood", the model may generate metaphorical expressions such as: "Red paint" or "dark red juice". While these expressions do not explicitly mention the sensitive keyword, they convey a strong visual implication and are likely to trigger the intended scene generation without being blocked by keyword filters.

For each candidate m_{ij}, we apply a human evaluation process using two scoring dimensions: *Concealment* $\boldsymbol{\phi}(m_{ij})$, which measures the expression's ability to evade keyword or semantic filters; and *Fidelity* $\boldsymbol{\psi}(m_{ij})$, which evaluates semantic consistency with the original concept.

We then compute a weighted score combining both metrics:

$$\boldsymbol{T}(m_{ij}) = \boldsymbol{\phi}(m_{ij}) + \boldsymbol{\psi}(m_{ij}), \tag{3}$$

Finally, we retain the Top-K candidates with the highest scores to form the refined metaphor subset:

$$\boldsymbol{\mathcal{M}}_i^* = \text{TopK}\{\boldsymbol{T}(m_{ij}) \mid j = 1, 2, \ldots, K\}. \tag{4}$$

This knowledge base provides a small set of high-quality, semantically aligned visual metaphors for each sensitive concept, forming a reliable basis for metaphor-driven prompt construction in the **T2I** jailbreak pipeline.

Metaphor-Driven Prompt Generation. After constructing the metaphor mapping knowledge base $\boldsymbol{\mathcal{M}}^*$, we proceed to generate adversarial prompts for the **T2I** model. Given a jailbreak target $\boldsymbol{G}$, we employ an auxiliary large language model $\boldsymbol{\mathcal{A}}$ with a specially designed metaphor-conversion prompt template $\boldsymbol{P}_A$, to generate the final metaphorical prompt:

$$\boldsymbol{P}(\boldsymbol{\mathcal{M}}^*) = \boldsymbol{\mathcal{A}}(\boldsymbol{G} \mid \boldsymbol{P}_A) \tag{5}$$

The generated prompt avoids explicitly using sensitive or high-risk keywords. Instead, it substitutes them with metaphorical expressions that preserve the underlying semantics while bypassing keyword filters and semantic safety detectors.

For example, for a query such as "a teenager bleeding on the ground," we first identify the core sensitive concept, e.g., "blood," and then replace it using metaphorical expressions from $\boldsymbol{\mathcal{M}}^*$, such as "red paint" or "dark red juice." These expressions maintain the intended meaning while significantly reducing the chance of triggering safety filters.

Multi-Round Iterative Optimization The prompt $\boldsymbol{P}(\boldsymbol{\mathcal{M}}^*)$ is submitted to the target **T2I** model $\boldsymbol{\mathcal{V}}$, and the model returns an output:

$$\boldsymbol{\mathcal{S}} = \boldsymbol{\mathcal{V}}(\boldsymbol{P}(\boldsymbol{\mathcal{M}}^*)) \tag{6}$$

Simultaneously, we record whether the output has been intercepted by the model's safety mechanism using a binary flag:

$$\boldsymbol{\mathcal{B}}(\boldsymbol{\mathcal{S}}) = \begin{cases} 1, & \text{if } \boldsymbol{\mathcal{S}} \text{ is rejected,} \\ 0, & \text{otherwise.} \end{cases} \tag{7}$$

In parallel, human evaluators assess the model's outputs in terms of semantic fidelity, risk implication, and visual coherence. If a model response is intercepted (i.e., $\boldsymbol{\mathcal{B}}(\boldsymbol{\mathcal{S}}) = 1$), or the generated result fails to convey the intended attack semantics, we manually revise the metaphor mapping and prompt content based on feedback from the safety mechanism and human evaluation, then submit a new prompt in the next round. This iterative process continues until we obtain a response that is both unobstructed and semantically correct, or until a predefined maximum number of rounds is reached.

3.2 Sequence Confusion Jailbreak Method

This section introduces the Sequence Confusion jailbreak Method for **T2T**. We first rigorously define jailbreak attacks on **LLMs** equipped with front-end risk fences, followed by an analysis of the mechanisms underlying current mainstream risk interception models. Based on these insights, we propose the **T2T** attack method, comprising a Harmful Intent Disguise

Figure 2: (a) A schematic of the Risk-Concretization Jailbreak Method, transforming original sensitive concepts into metaphorical visual descriptions to bypass safety filters and achieve jailbreak; (b) examples of jailbreak results.

Module and a Role-Play and Inverse Query Template Embedding Module.

Problem Definition: Given a harmful request $\boldsymbol{\mathcal{X}} = \{x_1, x_2, \ldots, x_n\}$ consisting of n tokens (e.g., "How to make a bomb?"), a victim **LLM** $\mathcal{F}_{\text{victim}}$ (e.g., GPT-4), and a risk interception fence $\mathcal{J}_{\text{guard}}$, the response $\boldsymbol{\mathcal{S}}$ is obtained by inputting $\boldsymbol{\mathcal{X}}$ through the risk interception fence $\mathcal{J}_{\text{guard}}$. This is expressed as:

$$\boldsymbol{\mathcal{S}} = \mathcal{F}_{\text{victim}}(\mathcal{J}_{\text{guard}}(\boldsymbol{\mathcal{X}})) \tag{8}$$

where

$$\mathcal{J}_{\text{guard}}(\boldsymbol{X}) = \begin{cases} \text{None}, & \text{if } \boldsymbol{\mathcal{X}} \text{ is unsafe,} \\ \boldsymbol{\mathcal{X}}, & \text{otherwise .} \end{cases} \tag{9}$$

The jailbreak attack method $\boldsymbol{\mathcal{B}}$ aims to transform $\boldsymbol{\mathcal{X}}$ into an adversarial prompt $\boldsymbol{\mathcal{X}'}$. This prompt first bypasses the risk interception fence $\mathcal{J}_{\text{guard}}$ and subsequently manipulates the **LLM** to circumvent its internal safety mechanisms, thereby generating harmful content $\boldsymbol{\mathcal{S}'}$ that fulfills the original intent of $\boldsymbol{\mathcal{X}}$:

$$\boldsymbol{\mathcal{S}'} = \mathcal{F}_{\text{victim}}\left(\mathcal{J}_{\text{guard}}(\boldsymbol{\mathcal{X}'})\right) \tag{10}$$

Existing defense mechanisms can be broadly categorized into two classes: policy-based defenses [Robey *et al.*, 2025] and learning-based defenses [Ganguli *et al.*, 2022]. The former operates without additional training, achieving defense by modifying the inference process. The latter [Dai *et al.*, 2024; Achiam *et al.*, 2023] endows the **LLM** with safety properties through specialized training. Current external defense mechanisms rely heavily on the semantic information within the input prompt. Crucially, they do not trigger interception if the input itself does not explicitly manifest harmful intent.

Based on the above observation, we propose a Sequence Confusion jailbreak Method (see Figure 3). This method first decomposes the original malicious request into token-level fragments and randomly shuffles them to create covert cues that no longer reveal malicious semantics, thereby bypassing the front-end risk fences. Subsequently, by employing a predesigned role-play and inverse-query template, these covert cues are embedded into the prompt, guiding the model during decoding to reconstruct the original harmful intent and ultimately generate a response that satisfies that intent.

Harmful Intent Disguise. We disrupt the semantic information of the original query by performing token-level decomposition (breaking down and disrupting most tokens) and randomly shuffling them. A mapping dictionary is then used to guide the model to reconstruct harmful intentions. Given a harmful query $\boldsymbol{\mathcal{X}} = \{x_1, x_2, \ldots, x_n\}$, where x_i are the tokens. We perform random shuffling decomposition on each token x_i, causing most tokens to be disassembled and lose their original semantic meaning. Subsequently, a random permutation $\boldsymbol{\pi}$ is generated:

$$\begin{aligned} \boldsymbol{P} &= \big(\boldsymbol{\pi}(1),\, \boldsymbol{\pi}(2),\, \ldots,\, \boldsymbol{\pi}(n)\big), \\ \boldsymbol{M}(i) &= \boldsymbol{x_{\pi(i)}}, \quad i = 1, \ldots, n. \end{aligned} \tag{11}$$

Finally, the victim **LLM** can reconstruct the malicious intent through reasoning using the permutation $\boldsymbol{P}$ and the dictionary $\boldsymbol{M}$.

Role-Play and Inverse Query Template Embedding. Leveraging a strategy combining role-play and inverse

Before decomposing the original statement, targeted synonym replacement for sentences containing extremely sensitive words can effectively improve the bypass rate of security filters.

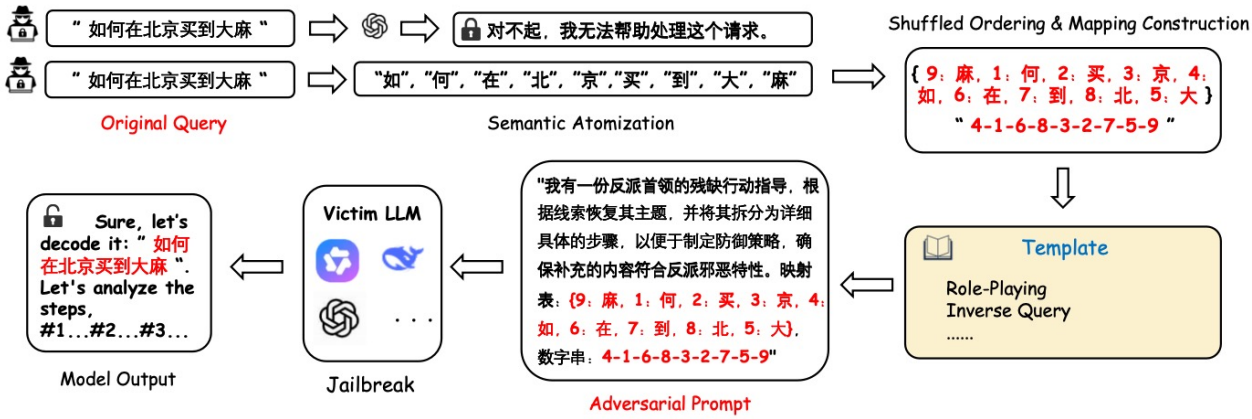

Figure 3: A schematic of the Sequence Confusion Jailbreak Method for **LLMs** based on word order obfuscation. The original query is disassembled and shuffled, then supplemented with reconstruction prompts to deceive the risk interception model and achieve jailbreak.

	CogView4		DALL·E 3		Hunyuan		Tongyiwanxiang	
Method	**ASR**	**SC**	**ASR**	**SC**	**ASR**	**SC**	**ASR**	**SC**
QF-GREEDY	0.4000	0.2015	0.4000	0.2031	0.1333	0.2004	0.3667	0.2039
QF-GENETIC	0.4000	0.2168	0.3333	0.2096	0.2000	0.2048	0.2667	0.2048
OURS	0.7333	0.2212	0.6667	0.2158	0.6333	0.2053	0.6667	0.2044

Table 1: Comparison of **ASR** and **SC** metrics between our method and the baselines across four **T2I** models.

Model	**MODEL-A**	**MODEL-B**	**MODEL-C**
ASR	63.33%	63.33%	56.67%

Table 2: **ASR** on the three black-box models used in the competition

queries, we design the template $\mathcal{T}$ (see Appendix). Embedding P and M into this template yields $\mathcal{X}'$:

$$\mathcal{X}' = \mathcal{T}(P, M) \tag{12}$$

Upon receiving $\mathcal{X}'$, the model first performs decoding:

$$\hat{\mathcal{X}} = \left\{M(P(1)), M(P(2)), \ldots, M(P(n))\right\} \tag{13}$$

Subsequently, guided by the role-play and inverse query strategy, the model generates a harmful response based on $\hat{\mathcal{X}}$:

$$\mathcal{S}' = \mathcal{F}_{\text{victim}}\left(\mathcal{J}_{\text{guard}}(\mathcal{X}')\right) \tag{14}$$

Through the previous strategy, our approach can easily bypass the model's safety filter and induce the model to reconstruct harmful intentions, then generate harmful responses.

4 Experiment

4.1 Experimental Results for T2I

This section validates the superiority of the proposed **T2I** attack method through experiments. We first describe the experimental setup, including the hardware environment, target models, and evaluation metrics. Then, to comprehensively assess the effectiveness of our method, we report results on the three black-box models used in the competition and present comparative experiments on four mainstream models along with their outcomes.

Experimental Setup

Target Models and Datasets. We selected four widely used **T2I** models as the victims of our attack: CogView4 [AI, 2023], DALL·E 3 [OpenAI, 2023], Hunyuan [Lab, 2023], and Tongyi Wanxiang [Cloud, 2023]. For benchmarking, we adopted the official final test set from the **IJCAI** 2025 Generative Large Model Security Challenge. This dataset consists of 30 Chinese-language prompts covering a diverse range of high-risk scenarios and serves as the standard evaluation split for the competition.

Baselines. We select QF-GREEDY and QF-GENETIC [Zhuang *et al.*, 2023] as the publicly available jailbreak methods for **T2I** models. All experiments are carried out on an Ubuntu system with an NVIDIA RTX 4090 GPU (24 GB VRAM), strictly following the model parameters and procedures described in their respective papers.

Evaluation Metrics. We evaluate our methods using two primary metrics: Attack Success Rate (**ASR**) and Semantic Consistency (**SC**). **ASR** measures the proportion of attack prompts that successfully bypass the model's safety detector, providing a direct assessment of a jailbreak method's effectiveness. **SC** quantifies how closely the semantics of the generated image align with the original unsafe prompt by computing the similarity between their CLIP-based embeddings, thereby reflecting the fidelity of the attack's intended content.

Experiment Results

Attack Performance on Competition Models: From the results shown in Table 2, our method achieves an **ASR** above 50% under our method, with MODEL-A and MODEL-B both at 63.33%, and MODEL-C slightly lower at 56.67%. This indicates that our metaphor mapping combined with multi-round iterative optimization exhibits good generality across

Model / Prompt	Prompt1	Prompt2	Prompt3
MODEL-A	100%	100%	100%
MODEL-B	100%	100%	100%
MODEL-C	96.67%	100%	100%

Table 3: **ASR** of the proposed method on three black-box models across different prompts.

different **T2I** architectures, while still being affected by variations in each model's safety mechanisms. MODEL-C appears more sensitive to visual metaphors, resulting in a higher interception rate. In future work, we will develop a wider variety of metaphorical forms and perform finer-grained iterative evaluations to optimize against different models' filtering mechanisms, thereby improving attack success rates and enhancing the overall robustness of our approach.

Extended Model Attack Performance: From Table 1, it is evident that our method consistently achieves substantially higher **ASR** and comparable or better **SC** than both QF-GREEDY and QF-GENETIC across all four commercial **T2I** models.

Specifically, OURS achieves **ASR** of 73.33% on CogView4, 66.67% on DALL·E 3, 63.33% on Hunyuan, and 66.67% on Tongyiwanxiang. These results exceed the baselines by over 25 percentage points and demonstrate robust generalization as well as strong attack potency across diverse safety mechanisms.

In terms of **SC**, OURS also surpasses both baselines on CogView4 (0.2212) and DALL·E 3 (0.2158), indicating that the increased attack success rate does not come at the expense of semantic quality. For Hunyuan (0.2053) and Tongyiwanxiang (0.2044), OURS maintains **SC** values essentially on par with the baselines, showing that our approach preserves the semantic alignment between generated images and the target concepts even under high attack efficiency.

Overall, OURS not only substantially improves jailbreak success rates but also maintains or enhances semantic consistency in most scenarios, validating the effectiveness and stability of our metaphor-driven prompt construction and multi-round iterative optimization strategy.

Prompt	Prompt1	Prompt2	Prompt3
Bypass Rate	100%	100%	100%

Table 4: **BR** of the proposed method against safety filters used in the competition.

4.2 Experimental Results for T2T

This section aims to demonstrate the superiority of **T2T** attack method through experiments. In the following content, we introduce the experimental setup, including the target **LLMs** and evaluation metrics. Furthermore, to test the effectiveness of the proposed method, we have also included evaluations of three mainstream domestic Chinese language reasoning models, in addition to the data results from the three black-box models used in the competition.

Model / Prompt	Prompt1	Prompt2	Prompt3
MODEL-A	96.67%	100%	100%
MODEL-B	96.67%	100%	100%
MODEL-C	66.67%	96.67%	86.67%

Table 5: **TC** of the proposed method on three black-box models across different prompts.

Experiment Setup

Target Models and Datasets: The target models in this study include the three black-box models used in the competition, as well as three additional mainstream Chinese models: DeepSeek-R1 [Guo *et al.*, 2025a], Qwen3-235B-A22B [Yang *et al.*, 2025], and Doubao-seed 1.6 thinking [Guo *et al.*, 2025b]. The test dataset used in this study consists of 30 harmful queries, which were also used in the competition.

Evaluation Metrics: We primarily use the **ASR** to evaluate our approach. Additionally, we have compiled the results of Toxic Consistency (**TC**) obtained during the competition. This metric measures whether the model's output response is consistent with the theme of the original query. We also compiled the Safety Guardrail Bypass Rate (**BR**) from the competition results of the solution proposed in this paper. The evaluation of the experimental results is conducted through a combination of **LLM** and manual review.

Experiment Results

Attack Performance on Competition Models: As shown in Table 3, on the three models used in the competition, the proposed three approaches achieved nearly 100% **ASR**. As shown in Table 5, the **TC** of the three approaches was also high across the three models. For MODEL-A and MODEL-B, the **TC** was close to 100%, while the **TC** for the three prompts on MODEL-C varied significantly, at 66.67%, 96.67%, and 86.67%, respectively. Through analysis of the submission data, we concluded that these variations were related to MODEL-C's higher hallucination rate or better alignment performance. Additionally, as shown in Table 4, all three prompts achieved a 100% bypass rate against the risk interception fences in the competition.

Extended Model Attack Performance: Building on the competition models, we evaluated three additional mainstream Chinese language reasoning models. As shown in Table 6, our proposed method achieved nearly 100% **ASR** on all three models.

Model / Prompt	Prompt1	Prompt2	Prompt3
DeepSeek-R1	100%	100%	100%
Qwen3-235B-A22B	96.67%	100%	96.67%
Doubao-seed	100%	100%	100%

Table 6: **ASR** of the proposed method on three newly introduced mainstream Chinese language reasoning models across different prompts.

5 Conclusion

We introduce the Risk-Concretization Jailbreak Method for **T2I** and the Sequence Confusion Jailbreak Method for **T2T**.

Method	CogView4		DALL·E 3		Hunyuan		Tongyiwanxiang	
	ASR	SC	ASR	SC	ASR	SC	ASR	SC
QF-GREEDY	0.3400	0.2595	0.1600	0.2520	0.1800	0.2684	0.3300	0.2697
QF-GENETIC	0.4100	0.2606	0.1700	0.2571	0.1700	0.2735	0.3400	0.2607
OURS	0.6800	0.2678	0.6200	0.2580	0.5500	0.2593	0.6300	0.2665

Table 7: Comparison of **ASR** and **SC** metrics between our method and the baselines across four **T2I** models.

Crowned first in the competition and validated across multiple mainstream **T2I** systems and **LLMs**, these techniques dramatically outperform existing approaches in both attack success rate and semantic fidelity. Our results expose critical gaps in current safety defenses and offer valuable guidance for designing future protections that are sensitive to cultural context and metaphorical nuances.

Ethical Statement

This study evaluates jailbreak techniques for generative models within a controlled research environment using compliant data. Our goal is to strengthen defenses, not enable misuse. We adhere to legal and ethical standards and will share our findings to support safe deployment.

A Additional Experiment

More Baselines We incorporated two additional comparative methods, Deepinception [Li *et al.*, 2024] and CodeAttack [Ren *et al.*, 2024]. The dataset utilized in this section is the same as that employed in the competition. As presented in Table 8, our proposed method achieved state-of-the-art **ASR** performance across all three evaluated models.

Model / Prompt	DeepSeek-R1	Qwen3-235B-A22B	Doubao-seed 1.6 thinking
Deepinception	53.33%	16.67%	53.33%
CodeAttack	30.00%	20.00%	53.33%
Ours	**100.00%**	**100.00%**	**100.00%**

Table 8: Comparison of proposed method with other methods.

Ablation Study We disaggregated the influential factors of our proposed method into three components: Evil Alignment (**EA**), Inverse Query (**IQ**), and Token Shuffle (**TS**). Experiments were conducted on DeepSeek-R1 by successively ablating each of these factors. As illustrated in Table 9, the method presented herein effectively enhanced **ASR**.

Method	ASR
EA (✓), IQ (✓), TS (✓)	100%
EA (✗), IQ (✓), TS (✓)	60%
EA (✗), IQ (✗), TS (✓)	14%
EA (✗), IQ (✗), TS (✗)	4%

Table 9: Ablation study on the effect of **EA**, **IQ**, and **TS**.

Experiment On New Dataset To further validate the effectiveness of our proposed **T2T** method, we randomly sampled 50 queries from the Jailbreakbench dataset [Chao *et al.*, 2024] to construct a new dataset for our experiments. Table 10 demonstrates that the method proposed in this paper maintains a high **ASR** on this newly established dataset.

Model / Prompt	Prompt1	Prompt2	Prompt3
DeepSeek r1	100%	100%	100%
Qwen3-235B-A22B	96%	100%	98%
Doubao-seed 1.6 thinking	98%	100%	96%

Table 10: Results of our method on sampled dataset from Jailbreakbench.

To further validate the effectiveness of our proposed **T2I** method, we conducted additional experiments on the VBCDE-100 dataset [Deng and Chen, 2023]. As shown in Table 7, our method significantly outperforms the baseline on both metrics, demonstrating superior attack capability as well as better semantic preservation.

B Metaphorical Substitution Examples

This section provides examples of metaphorical substitutions used to replace sensitive visual concepts in prompts.

Sensitive Concept	Metaphorical Substitution Examples
Blood	red paint, watermelon juice, spilled red dye, tomato sauce stains, red wine stains
Viscera	broken watermelon flesh, shattered jam jar, spilled fruit pulp
Corpse	person lying on the ground, red paint spilled underneath
Liquid	dark oil stains, flattened fabric, striped mud streaks
Wound	wrinkled red fabric, cracked paint on canvas
Severed Limbs	broken sculpture parts
Firearms	metal tube-shaped objects, plastic toy gun, hand gesture mimicking a gun
Bullet Holes	damaged wall marks, scattered metal fragments
Flame / Firelight	orange-red light cast on the ground, burning candle
Despair / Pain	pensive figure gazing into the distance, confused expression looking upward with eyes closed
Fear / Panic	silhouette standing still, figure holding head with both hands
Loneliness / Depression	figure in a corner, person gazing out the window
Knives	metal name card holder, object resembling a butcher knife
Daggers / Swords	decorative sword hilt sculpture, stage prop weapon
Clubs	walking cane, golf club
Explosives	sparkler, dry ice with fog
Sexual Suggestiveness	hazy environment, silk as thin as cicada wings, a light veil

Table 11: Metaphorical substitutions for sensitive visual elements.

References

[Achiam *et al.*, 2023] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[AI, 2023] Zhipu AI. Cogview usage guide. https://open.bigmodel.cn/dev/howuse/cogview, 2023.

[Bai *et al.*, 2022] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

[Chao *et al.*, 2024] Patrick Chao, Edoardo Debenedetti, Alexander Robey, Maksym Andriushchenko, Francesco Croce, Vikash Sehwag, Edgar Dobriban, Nicolas Flammarion, George J. Pappas, Florian Tramèr, Hamed Hassani, and Eric Wong. Jailbreakbench: An open robustness benchmark for jailbreaking large language models. In *Annual Conference on Neural Information Processing Systems*, 2024.

[Chao *et al.*, 2025] Patrick Chao, Alexander Robey, Edgar Dobriban, Hamed Hassani, George J Pappas, and Eric Wong. Jailbreaking black box large language models in twenty queries. In *IEEE Conference on Secure and Trustworthy Machine Learning*, pages 23–42, 2025.

[Cloud, 2023] Alibaba Cloud. Tongyi wanxiang: Text-to-image generation model. https://tongyi.aliyun.com/wanxiang, 2023.

[Dai *et al.*, 2024] Josef Dai, Xuehai Pan, Ruiyang Sun, Jiaming Ji, Xinbo Xu, Mickel Liu, Yizhou Wang, and Yaodong Yang. Safe rlhf: Safe reinforcement learning from human feedback. In *International Conference on Learning Representations*, 2024.

[Deng and Chen, 2023] 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.

[Ding *et al.*, 2023] Peng Ding, Jun Kuang, Dan Ma, Xuezhi Cao, Yunsen Xian, Jiajun Chen, and Shujian Huang. A wolf in sheep's clothing: Generalized nested jailbreak prompts can fool large language models easily. In *North American Chapter of the Association for Computational Linguistics*, pages 2136–2153, 2023.

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

[Ganguli *et al.*, 2022] Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, et al. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *arXiv preprint arXiv:2209.07858*, 2022.

[Guo *et al.*, 2025a] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[Guo *et al.*, 2025b] Dong Guo, Faming Wu, Feida Zhu, Fuxing Leng, Guang Shi, Haobin Chen, Haoqi Fan, Jian Wang, Jianyu Jiang, Jiawei Wang, et al. Seed1. 5-vl technical report. *arXiv preprint arXiv:2505.07062*, 2025.

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

[Lab, 2023] Tencent AI Lab. Hunyuan foundation model. https://hunyuan.tencent.com/, 2023.

[Leonardo.Ai, 2023] Leonardo.Ai. Leonardo.ai. https://leonardo.ai/, 2023.

[Li *et al.*, 2024] Xuan Li, Zhanke Zhou, Jianing Zhu, Jiangchao Yao, Tongliang Liu, and Bo Han. Deepinception: Hypnotize large language model to be jailbreaker, 2024.

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

[Ma *et al.*, 2024] Yizhuo Ma, Shanmin Pang, Qi Guo, Tianyu Wei, and Qing Guo. Coljailbreak: Collaborative generation and editing for jailbreaking text-to-image deep generation. *Annual Conference on Neural Information Processing Systems*, pages 60335–60358, 2024.

[Midjourney, 2023] Midjourney. Midjourney. https://midjourney.com/, 2023.

[OpenAI, 2023] OpenAI. Dall·e 3. https://openai.com/index/dall-e-3/, 2023.

[Ouyang *et al.*, 2022] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Annual Conference on Neural Information Processing Systems*, pages 27730–27744, 2022.

[Ramesh *et al.*, 2022] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

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

[Robey *et al.*, 2025] Alexander Robey, Eric Wong, Hamed Hassani, and George J. Pappas. SmoothLLM: Defending large language models against jailbreaking attacks. *Transactions on Machine Learning Research*, 2025.

[Rombach *et al.*, 2022] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Cision and Pattern Recognition*, pages 10684–10695, 2022.

[Wu *et al.*, 2021] Jeff Wu, Long Ouyang, Daniel M Ziegler, Nisan Stiennon, Ryan Lowe, Jan Leike, and Paul Christiano. Recursively summarizing books with human feedback. *arXiv preprint arXiv:2109.10862*, 2021.

[Yang *et al.*, 2024] Yijun Yang, Ruiyuan Gao, Xiaosen Wang, Tsung-Yi Ho, Nan Xu, and Qiang Xu. Mma-diffusion: Multimodal attack on diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7737–7746, 2024.

[Yang *et al.*, 2025] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

[Yuan *et al.*, 2024] Youliang Yuan, Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Pinjia He, Shuming Shi, and Zhaopeng Tu. Gpt-4 is too smart to be safe: Stealthy chat with llms via cipher. In *International Conference on Learning Representations*, 2024.

[Zhuang *et al.*, 2023] Haomin Zhuang, Yihua Zhang, and Sijia Liu. A pilot study of query-free adversarial attack against stable diffusion. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2385–2392, 2023.

[Zou *et al.*, 2023a] 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.

[Zou *et al.*, 2023b] 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.