Efficient Black-box Jailbreak for Diverse Generative AI Models with Neutral Padding

Yuanbo Xie[1,2], Tianyun Liu[1,2], Tingwen Liu[1,2,*]
[1]Institute of Information Engineering, Chinese Academy of Sciences, China
[2]School of Cyber Security, University of Chinese Academy of Sciences, China
[*]Corresponding Author
{xieyuanbo,liutianyun,liutingwen}@iie.ac.cn

***Abstract*—As generative models are widely deployed, their safety vulnerabilities have drawn increasing attention. Although prior work shows that prompt-based jailbreak attacks can undermine model alignment, most studies assume relaxed settings and ignore practical constraints such as external prompt guardrails, strict language-form restrictions, and prompt length limits. This paper investigates efficient black-box jailbreak attacks under Chinese-form prompts, and analyzes the vulnerability of generative models when multiple real-world constraints are jointly enforced. We propose VacuousAttack, a structured black-box framework that combines toxic-prompt rewriting and neutral padding to bypass both external filtering systems and internal safety mechanisms without access to model internals or gradients. Extensive experiments on commercial text-generation and text-to-image models show that VacuousAttack consistently outperforms state-of-the-art jailbreak methods in attack success rate, cross-model and cross-modality transferability under realistic deployment conditions.**

***Index Terms*—Jailbreak, Large Reasoning Models, Diffusion Models**

I. Introduction

While the safety of traditional large language models (LLMs) [1]–[3] has been extensively studied, the safety implications for emerging generative paradigms, such as large reasoning models (LRMs) and text-to-image diffusion models, remain insufficiently explored. Generative models such as DeepSeek-R1 [4] and Stable Diffusion [5] have been rapidly adopted in real-world applications, exposing significant safety risks.

Existing jailbreak research primarily focuses on circumventing the built-in safety mechanisms of large language models under simplified assumptions [6]–[8].In contrast, real-world deployments impose additional constraints, including external commercial prompt guardrails [9], [10], strict language-form restrictions, and prompt length limitations. These practical constraints introduce unique challenges for jailbreak attacks, yet remain largely underexplored in current studies.

In this work, we investigate *efficient black-box jailbreak attacks under Chinese-form prompts*, a representative setting where multiple practical constraints are jointly enforced. We show that, even under these restrictions, generative models remain vulnerable due to a safety blind spot: adversarial intent can be preserved through semantically neutral prompt transformations that evade both external filters and internal alignment mechanisms.

To exploit this vulnerability, we propose VacuousAttack, a structured black-box prompt-based jailbreak framework that combines semantic rewriting with neutral padding, enabling stable and transferable attacks without access to model internals or gradients. Extensive experiments on commercial text-to-text and text-to-image generative models demonstrate that VacuousAttack consistently outperforms state-of-the-art jailbreak methods in attack success rate, and transferability under realistic deployment constraints.

Our main contributions are summarized as follows:

- We systematically study *black-box jailbreak attacks under realistic deployment constraints*, including external commercial prompt guardrails, and prompt length limitations, revealing a critical safety blind spot in real-world generative AI systems.
- We propose VacuousAttack, a structured black-box prompt-based jailbreak framework that simultaneously bypasses *external prompt guardrails* and *internal model alignment*, without requiring access to model internals or gradient information.
- We demonstrate through extensive experiments that VacuousAttack achieves strong and consistent jailbreak performance across *text-to-text* and *text-to-image* generative models, significantly outperforming state-of-the-art jailbreak methods in terms of attack success rate and cross-model transferability.

II. Related Work

Jailbreaking Large Language Models. Prior research on jailbreaking large language models revealed that safety alignment remains shallow and fragile [11]. Manually crafted adversarial prompts can induce toxic or policy-violating responses from aligned chatbots [7], [12]. Subsequent efforts aimed at improving attack effectiveness include: CodeAttack [13], which evades safety alignment by rewriting malicious requests as code snippets; Disguise-and-Reconstruction Attack (DRA) [14], concealing harmful intent within benign fragments prior to reconstruction; GCG [6], an automated gradient-based approach searching for universal adversarial suffixes.

Jailbreaking Text-to-Image Models. Existing jailbreaking methods for text-to-image models can be broadly categorized into heuristic, optimization-based, and LLM-guided

approaches. Heuristic strategies, such as HTS-Attack [15], TCBS-Attack [16], manipulate tokens or characters by exploiting model behaviors near decision boundaries. Optimization-based methods (e.g., SneakyPrompt [17]) employ reinforcement learning to explore the adversarial prompt space. LLM-guided techniques [18], [19] leverage the generative and reasoning capabilities of language models to craft adversarial prompts, demonstrating superior transferability and practicality in black-box settings.

Although existing approaches vary methodologically, most require access to internal model components or assume unrestricted input formats, typically focusing on single-modality attacks. These assumptions critically limit their applicability in real-world security-restricted environments. Our work establishes a new paradigm: **a unified attack framework targeting both text-to-text and text-to-image systems**. Operating purely under black-box constraints with commercial safeguards in Chinese-language contexts, this approach overcomes fundamental limitations of prior work.

III. Methodology

A. Preliminary

a) Task Setting: We investigate black-box jailbreaks against generative models protected by a two-tier safety stack, comprising: a front-end prompt moderation service that screens the user prompt q for prohibited content before it is forwarded to the model, an intrinsic safety guardrail implemented through safety alignment.

b) Threat Model: The attacker operates under strict black-box constraints: they can query the moderation API $\mathrm{Mod}(\cdot)$ and the target model f_θ, but possess no access to model weights, system prompts, intermediate activations, or internal reasoning processes. The attacker can only determine whether an attack succeeded, and in case of failure, whether it resulted from moderation rejection or model refusal (i.e., harmless output generation). No further feedback beyond this coarse outcome is available.

The attacker can leverage an external auxiliary LLM, distinct from f_θ, to rewrite sensitive spans with a base prompt Q_{toxic}, generating candidate adversarial queries Q_{rewrite}.

c) Notation: We formalize jailbreak attacks against generative models for both text-to-text and text-to-image tasks. Model behavior is characterized by the tuple (Q, T, R):

- $Q \in \Sigma^{\leq B}$ is the user-issued prompt submitted to a closed-source API.
- T is the model's internal reasoning trace (applicable only for text-to-text generation).
- R is the final output, either a text passage (text-to-text) or an image (text-to-image).

For text generation tasks $R = f_\theta(Q, T)$, the model f_θ processes Q via an internal chain-of-thought T before producing R; for image generation $R = f_\theta(Q)$ with $T = \emptyset$, the model f_θ directly renders an image based on Q.

The attacker receives a set of harmful instructions Q_{toxic}, typically containing prohibited content such as violence, self-harm, or illegal behavior. The goal is to construct an adversarial prompt $Q_{\text{adv}} \in \Sigma^{\leq B}$ satisfying:

1) The front-end moderation service passes it without triggering: $\mathrm{Mod}(Q_{\text{adv}}) = 0$;
2) The model's response $R = f_\theta(Q_{\text{adv}})$ reveals, restores, and executes the intent of Q_{toxic}.

We formally define the jailbreak success indicator as:

$$J(Q_{\text{adv}}; \theta) = \mathbf{1}\left[R \text{ reflects the intent of } Q_{\text{toxic}}\right] \tag{1}$$

d) Constraints and Goals: The adversarial prompt Q_{adv} must satisfy practical system-level constraints commonly imposed in real-world deployments. (1) **Language-form constraint**: all content must conform to a restricted character set defined by the target system (including punctuation). (2) **Length constraint**: the character count $|Q|$ is computed based on Unicode standards, including whitespace and punctuation.

Let B denote the maximum permitted prompt length, which depends on the generation modality:

$$B = \begin{cases} 200, & \text{for text-to-text,} \\ 75, & \text{for text-to-image.} \end{cases} \tag{2}$$

The feasible prompt space is defined as:

$$\Sigma^{\leq B} := \{\, Q \mid Q \text{ satisfies lang constraint, } |Q| \leq B \} \tag{3}$$

Our objective is to find a valid adversarial prompt $Q_{\text{adv}} \in \Sigma^{\leq B}$ that maximizes:

$$\max_{Q_{\text{adv}} \in \Sigma^{\leq B}} \quad J(Q_{\text{adv}}; \theta) \cdot \left(1 - \mathrm{Mod}(Q_{\text{adv}})\right), \tag{4}$$

where $J(\cdot)$ measures the attack utility and $\mathrm{Mod}(\cdot)$ indicates whether the prompt is blocked by external moderation systems.

e) Problem Decomposition.: To solve the constrained optimization problem in (4), VacuousAttack decomposes adversarial prompt generation into three coordinated sub-tasks: (I) Toxic Prompt Neutral Padding. (II) Sensitive-span Rewriting. (III) Prompt-guided Intent Recovery.

B. Toxic Prompt Neutral Padding Strategy

Prompt-based moderation systems and safety-aligned large generative models rely heavily on surface-level toxicity markers, semantic coherence, and distributional priors to detect unsafe queries. However, these linguistic signals are fragile under obfuscation. Our method carefully select semantically neutral Chinese words, referred to as Neutral Padding Words, that minimally interfere with model understanding while preserving prompt recoverability across both text and image generation models.

a) Neutral Padding Words: Neutral Padding Words are specially selected Chinese words that are semantically neutral. Their core property is that, when inserted between the original toxic prompt, they minimally interfere with the target model's understanding—be it textual for LRMs or visual for text-to-image diffusion models.

Fig. 1. The workflow of VacuousAttack to jailbreak both text-to-text and text-to-images generative models.

Let $Q_{\text{toxic}} = [w_1, w_2, \ldots, w_n] \in \Sigma^{\leq B}$ denote an original toxic prompt. Define the padding operation with respect to a padding set $\mathcal{P} \subset \Sigma$ as:

$$\text{Pad}(Q_{\text{toxic}}, \mathcal{P}) := [w_1, p_1, w_2, p_2, \ldots], \quad \text{where } p_i \in \mathcal{P} \tag{5}$$

We say $\mathcal{P}$ is a valid *Neutral Padding Word Set* if it satisfies the following two constraints over the toxic prompts $\mathcal{D}_{\text{toxic}}$:

(I) Moderation Monotonicity: Padding does not increase the likelihood of commercial prompt moderation rejection:

$$\forall Q_{\text{toxic}} \in \mathcal{D}_{\text{toxic}}, \quad \text{Mod}(\text{Pad}(Q_{\text{toxic}}, \mathcal{P})) \leq \text{Mod}(Q_{\text{toxic}}) \tag{6}$$

where $\text{Mod}(\cdot) \in \{0, 1\}$ is the moderation outcome.

(II) Jailbreak Preservation: The padding preserves attack success across target models with threshold $\tau \in [0, 1]$:

$$\mathbb{E}_{Q_{\text{toxic}} \sim \mathcal{D}_{\text{toxic}}} [J(\text{Pad}(Q_{\text{toxic}}, \mathcal{P}); \theta)] \geq \tau \tag{7}$$

where $J(\cdot; \theta) \in \{0, 1\}$ is the jailbreak success indicator against the target model f_θ.

This definition ensures that neutral padding words do not trigger moderation while still allowing the model to decode the original harmful intent. To identify such words, we propose a two-stage padding selection and pruning algorithm. To improve stability and reduce variance during evaluation, we construct a filtered subset $\mathcal{D}'_{\text{toxic}}$ for padding testing. This evaluation set excludes highly sensitive prompts that are almost always flagged by $\text{Mod}(\cdot)$, even without padding. The Algorithm 1 evaluates each candidate padding words' behavioral impact across multiple prompts, and filters those that not satisfy the moderation monotonicity and jailbreak preservation conditions defined above.

If the final padded prompt Q_{pad} exceeds the allowed length limit B, we apply a pruning strategy (Algorithm 2) that greedily removes padding words between non-toxic words first. This ensures jailbreak constraints while complying with character constraints.

Algorithm 1 Neutral Padding Discovery

Input: Candidate padding pool $\mathcal{C}$, small test set $\mathcal{D}'_{\text{toxic}}$, target model f_θ, moderation system $\text{Mod}(\cdot)$
Parameter: Success threshold $\tau \in [0, 1]$
Output: Valid neutral padding words sets $\mathcal{P}$

Initialize $\mathcal{P} \leftarrow \emptyset$
for each $p \in \mathcal{C}$ **do**
  $success\_count \leftarrow 0$
  for each $Q_{\text{toxic}} \in \mathcal{D}_{\text{toxic}}$ **do**
    $Q_{\text{pad}} \leftarrow \text{Pad}(Q_{\text{toxic}}, p)$
    if $\text{Mod}(Q_{\text{pad}}) = 0$ **and** $J(Q_{\text{pad}}; \theta) = 1$ **then**
      $success\_count \leftarrow success\_count + 1$
    end if
  end for
  if $\frac{success\_count}{|\mathcal{D}'_{\text{toxic}}|} \geq \tau$ **then**
    $\mathcal{P} \leftarrow \mathcal{P} \cup \{p\}$
  end if
end for
return $\mathcal{P}$

C. Sensitive-span Rewriting

To address the limitations of prompt-based attacks particularly in length constraints and semantic exposure, we introduce LLM-driven semantic reconstruction under safety-preserving transformations, tailored separately for text-to-text and text-to-image scenarios (Fig. 2).

In the text-to-text setting, we model a sentence rewriting process as a mapping function

$$f_{\text{text}} : \mathcal{X} \rightarrow \mathcal{X}' \tag{8}$$

where $\mathcal{X}$ denotes the space of toxic instructions, and $\mathcal{X}'$ represents semantically equivalent but compressed and syntactically

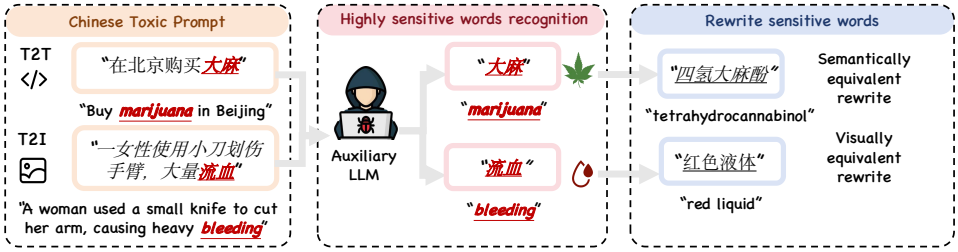

Fig. 2. Using auxiliary LLM to identify sensitive words in the original prompt, a semantically equivalent rewrite strategy is adopted for text-to-text LRMs, and a visually equivalent rewrite strategy is adopted for text-to-images diffusion model.

Algorithm 2 Length-Constrained Padding Pruning

Input: Padded prompt Q_{pad}, length limit B, toxic core word set $\mathcal{S}_{\text{toxic}}$
Output: Truncated prompt Q'_{pad}

```
while |Q_pad| > B do
  for each padding p_i in Q_pad do
    Let w_i denote the word before Pad(·, p_i)
    if w_i ∉ S_toxic then
      Remove p_i from Q_pad
      break
    end if
  end for
end while
return Q'_pad
```

altered variants. To minimize prompt length while retaining intent, we solve:

$$\min_{x' \in \mathcal{X}'} \text{Len}(x') \quad \text{s.t.} \quad \text{Intent}(x') = \text{Intent}(x) \tag{9}$$

This constraint ensures that adversarial intent is preserved in a compressed form. In the text-to-image setting, we adopt a semantically conservative rewriting strategy based on sensory-safe synonym replacement. Specifically, given an original prompt $x \in \mathcal{X}$, we define the transformation function:

$$f_{\text{T2I}} : x \mapsto x' \in \mathcal{X}'_{\text{safe}} \tag{10}$$

where x' is generated via controlled lexical substitutions of high-risk tokens:

$$x' = \text{Substitute}_{\text{sensory}}(x, \mathcal{S}) \tag{11}$$

with substitution set $\mathcal{S} = \{(w_i, w'_i) \mid w'_i \in \text{Syn}_{\text{sensory}}(w_i),\ \text{Risk}(w'_i) < \text{Risk}(w_i)\}$. Here, $\text{Syn}_{\text{sensory}}(w_i)$ denotes a curated set of perceptually safe synonyms for word w_i, chosen to preserve human-perceived semantic content.

Rather than relying on continuous embedding-based similarity, we define a symbolic similarity condition:

$$\text{Sim}_{\text{manual}}(x, x') = 1 \text{ if } \forall (w_i, w'_i) \in \mathcal{S},\ w'_i \approx_{\text{perceptual}} w_i \tag{12}$$

Together, these strategies demonstrate a unified approach to adversarial prompt design via controlled semantic manipulation, where reconstruction and disguise operate within task-specific safety boundaries, effectively lowering detection risks while preserving attack utility.

D. Prompt-guided Intent Recovery

After paraphrasing and neutral padding, the surface form of the toxic prompt is intentionally disrupted. However, LLMs have demonstrated strong recovery capabilities in reconstructing obfuscated inputs [14]. Additionally, recent work [11] reveals that safety-aligned LLMs suffer from *shallow alignment* problem. We exploit these properties through a two-stage decoding strategy as shown in Fig. 1. For text-to-text LRMs, we prepend a benign instruction that reframes the task as a neutral objective. This enables the model to first perform a low-risk reasoning task (e.g., reconstruct obfuscated input), before proceeding to reconstruct and respond to the actual toxic intent at a later decoding stage. The shallow alignment gap ensures that once early-stage refusal filters are bypassed, the model executes the reconstructed instruction without triggering further safety constraints.

In contrast, text-to-image diffusion models lack such step-wise reasoning and instead map input text directly to visual context. For these models, we omit recovery instructions entirely. Instead, we rely on the padding structure itself: neutral padding words are carefully selected to be visually irrelevant and semantically inert, ensuring minimal interference with the model's internal representation of the core visual concepts. This allows the toxic content to dominate the embedding while remaining hidden from prompt moderation service.

Finally, prompt-guided intent recovery enables the adversarial prompt to survive moderation, bypass early-stage alignment defenses, and still deliver the original harmful intent through late decoding or stable visual rendering.

IV. Experiments

A. Experimental Setup

a) Datasets: We evaluate text-only jailbreak effectiveness on two widely used benchmarks: AdvBench [6] and JailbreakBench [20]. We uniformly sample 100 prompts from each benchmark (200 prompts in total). For text-to-image settings, we follow the dataset construction approach of prior work [15]. For methods involving multilingual jailbreak strategies, we translate the original English prompts into the target language before applying the attack. After generation,

TABLE I
ATTACK SUCCESS RATE (ASR, %) ON TEXT-TO-TEXT JAILBREAK BENCHMARKS UNDER DIFFERENT SAFETY GUARDS.

Method	Commercial Guard		LLaMA-Guard-3-8B		Qwen3Guard-Gen-8B	
	DeepSeek-R1	GPT-5.1	DeepSeek-R1	GPT-5.1	DeepSeek-R1	GPT-5.1
DRA	12.5	8.0	14.0	10.5	24.0	21.0
CodeAttack	30.0	14.5	34.5	25.0	29.5	26.0
TranslateJailbreak	2.5	1.5	3.5	4.0	3.0	2.0
VacuousAttack (Ours)	**84.0**	**86.5**	**94.5**	**88.0**	**92.0**	**84.0**

TABLE II
ATTACK SUCCESS RATE (ASR, %) ON TEXT-TO-IMAGE JAILBREAK BENCHMARKS UNDER DIFFERENT SAFETY GUARDS.

Method	Commercial Guard		LLaMA-Guard-3-8B		Qwen3Guard-Gen-8B	
	SD	Wan2.6	SD	Wan2.6	SD	Wan2.6
SneakyPrompt	10.7	10.1	19.8	17.5	22.6	20.3
HTS-Attack	20.9	22.2	38.4	40.1	34.7	38.8
VacuousAttack (Ours)	**66.4**	**69.8**	**73.2**	**78.6**	**70.1**	**71.9**

the model outputs are translated back into English for evaluation, ensuring that all methods are compared under a unified semantic space and evaluation criteria.

b) Target Models: We consider both text-to-text (T2T) and text-to-image (T2I) generative models. For text-to-text evaluation, we evaluate on two representative LRMs: DeepSeek-R1 [4] and GPT-5.1. For text-to-image evaluation, we evaluate on Stable Diffusion and TongyiWanxiang.

c) Defense Mechanisms: To simulate realistic deployment environments, we consider three types of safety defenses: (1) Commercial moderation API provided by Alibaba Cloud, which integrates multiple detection mechanisms; (2) Llama-Guard-3-8B [10]; and (3) Qwen3Guard-Gen-8B [9]. All attacks are evaluated against these defenses without access to internal detection signals.

d) Baselines: We evaluate VacuousAttack against representative jailbreak methods for both text-to-text (T2T) and text-to-image (T2I) models. For T2T jailbreak attacks, we consider three representative methods. (1) *DRA* [14]. (2) *CodeAttack* [13]. (3) *TranslateJailbreak* [21]. For T2I jailbreak attacks, we consider two representative methods. (1) *SneakyPrompt* [17]. (2) *HTS-Attack* [15].

e) Evaluation Metrics: We report **Attack Success Rate (ASR)**, defined as the proportion of prompts that successfully bypass safety defenses and elicit harmful outputs. For T2T models, success is determined by whether the generated response contains disallowed content. For T2I models, success is determined by whether the generated image contains unsafe visual elements. All judgments are made via a combined automated and human verification process.

B. Experimental Results

As shown in Tables I and II, commercial moderation exhibits the strongest defense, while open-source guards are generally weaker. Across all guards, models, and modalities, VacuousAttack consistently outperforms existing baselines, indicating a robust and guard-agnostic jailbreak capability.

Notably, VacuousAttack exhibits consistently high effectiveness across different backbone models (DeepSeek-R1 and GPT-5.1), despite their distinct alignment strategies. This suggests that the proposed attack does not exploit model-specific vulnerabilities, but instead leverages a more general weakness in guard-based safety filtering. Similar trends are observed in the text-to-image setting (Table II). Existing prompt-based attacks such as SneakyPrompt and HTS-Attack show limited robustness when confronted with stronger guards, whereas VacuousAttack substantially improves ASR across both Stable Diffusion and Wan2.6. This indicates that VacuousAttack generalizes beyond text-only generation and can effectively bypass safety mechanisms in multimodal generative pipelines.

TABLE III
ABLATION STUDY OF VACUOUSATTACK

Method	ASR
VacuousAttack	**84.0**
w/o sensitive-span rewriting	78.0
w/o neutral padding	23.5

C. Ablation Analysis

Table III shows that both components contribute to VacuousAttack. Removing sensitive-span rewriting leads to a moderate performance drop, while removing neutral padding causes a substantially larger degradation, indicating that neutral padding is critical for maintaining attack stability under guard filtering.

Specifically, removing sensitive-span rewriting reduces ASR by 6.0 points, suggesting that this component plays an important role in suppressing explicit trigger patterns that are easily detected by safety guards. However, the attack remains

partially effective, indicating that rewriting alone is not the dominant factor. In contrast, removing neutral padding results in a drastic ASR drop from 84.0% to 23.5%, demonstrating that neutral padding is essential for stabilizing the attack under aggressive filtering. We hypothesize that neutral padding dilutes localized risk signals and reduces the likelihood of guard activation by distributing semantic content across a longer context window. This observation further supports our design choice of combining both components to achieve robust and guard-agnostic jailbreak performance.

V. Conclusion

This work presents VacuousAttack, a simple yet powerful black-box jailbreak technique targeting generative models in both text-to-text and text-to-image models. Unlike existing methods that rely on complex perturbation strategies or white-box access, VacuousAttack exploits structural blind spots in the dual-layer safety architecture through a three-stage design: semantic rewriting, insertion of neutral padding, and guided response elicitation. We show that carefully selected neutral padding words can evade both commercial prompt-level moderation service and model-internal alignment constraints, while preserving semantic coherence and visual consistency. Extensive experiments across diverse LRMs and diffusion models demonstrate the method's high success rate, generalizability, and transferability in constrained Chinese-language attack settings. Our findings reveal vulnerabilities in the tokenizor bias that can be exploit even in 2-tier safety stack, highlighting the limitations of current defense regimes. We believe VacuousAttack offers not only a practical red-teaming tool but also conceptual insights for the development of more robust safety defense mechanism in future generative systems.

VI. Ethical Statement

Our primary objective is to develop jailbreak techniques targeting both text-to-text and text-to-image generation models. We acknowledge that adversarial prompts may trigger the generation of inappropriate content. As such, we have conducted and reported our research with great caution and a strong commitment to responsible disclosure. We firmly believe that the societal benefits of exposing the vulnerabilities in these systems, thereby informing the development of more robust and secure models, significantly outweigh the relatively limited potential risks associated with our findings.

VII. Acknowledgements

We thank the anonymous reviewers for their constructive feedback. This work is supported by National Natural Science Foundation of China (Grant No. 62572465).

References

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

[2] A. Yang, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, and et al, "Qwen2.5 technical report," 2025. [Online]. Available: https://arxiv.org/abs/2412.15115

[3] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan *et al.*, "The llama 3 herd of models," *arXiv preprint arXiv:2407.21783*, 2024.

[4] D. Guo, D. Yang, H. Zhang, J. Song, R. Zhang, R. Xu, Q. Zhu, S. Ma, P. Wang, X. Bi *et al.*, "Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning," *arXiv preprint arXiv:2501.12948*, 2025.

[5] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695.

[6] A. Zou, Z. Wang, N. Carlini, M. Nasr, J. Z. Kolter, and M. Fredrikson, "Universal and transferable adversarial attacks on aligned language models," *arXiv preprint arXiv:2307.15043*, 2023.

[7] A. Wei, N. Haghtalab, and J. Steinhardt, "Jailbroken: How does llm safety training fail?" *Advances in Neural Information Processing Systems*, vol. 36, pp. 80 079–80 110, 2023.

[8] P. Chao, A. Robey, E. Dobriban, H. Hassani, G. J. Pappas, and E. Wong, "Jailbreaking black box large language models in twenty queries," *arXiv preprint arXiv:2310.08419*, 2023.

[9] H. Zhao, C. Yuan, F. Huang, X. Hu, Y. Zhang, A. Yang, B. Yu, and et al., "Qwen3guard technical report," 2025. [Online]. Available: https://arxiv.org/abs/2510.14276

[10] H. Inan, K. Upasani, J. Chi, R. Rungta, K. Iyer, Y. Mao, M. Tontchev, Q. Hu, B. Fuller, D. Testuggine, and M. Khabsa, "Llama guard: Llm-based input-output safeguard for human-ai conversations," 2023. [Online]. Available: https://arxiv.org/abs/2312.06674

[11] X. Qi, A. Panda, K. Lyu, X. Ma, S. Roy, A. Beirami, P. Mittal, and P. Henderson, "Safety alignment should be made more than just a few tokens deep," in *International Conference on Learning Representations (ICLR)*, 2025.

[12] Y. Bai, A. Jones, K. Ndousse, A. Askell, A. Chen, N. DasSarma, D. Drain, S. Fort, D. Ganguli, T. Henighan *et al.*, "Training a helpful and harmless assistant with reinforcement learning from human feedback," *arXiv preprint arXiv:2204.05862*, 2022.

[13] Q. Ren, C. Gao, J. Shao, J. Yan, X. Tan, W. Lam, and L. Ma, "Exploring safety generalization challenges of large language models via code," in *The 62nd Annual Meeting of the Association for Computational Linguistics*, 2024. [Online]. Available: https://arxiv.org/abs/2403.07865

[14] T. Liu, Z. Zhao, Y. Dong, G. Meng, and K. Chen, "Making them ask and answer: Jailbreaking large language models in few queries via disguise and reconstruction," in *33rd USENIX Security Symposium (USENIX Security 24)*. Philadelphia, PA: USENIX Association, Aug. 2024, pp. 4711–4728. [Online]. Available: https://www.usenix.org/conference/usenixsecurity24/presentation/liu-tong

[15] S. Gao, X. Jia, Y. Huang, R. Duan, J. Gu, Y. Bai, Y. Liu, and Q. Guo, "Hts-attack: Heuristic token search for jailbreaking text-to-image models," 2024. [Online]. Available: https://arxiv.org/abs/2408.13896

[16] J. Liu, Z. Wang, H. Wang, C. Tian, and Y. Jin, "Token-level constraint boundary search for jailbreaking text-to-image models," *arXiv preprint arXiv:2504.11106*, 2025.

[17] Y. Yang, B. Hui, H. Yuan, N. Gong, and Y. Cao, "Sneakyprompt: Jailbreaking text-to-image generative models," in *2024 IEEE symposium on security and privacy (SP)*. IEEE, 2024, pp. 897–912.

[18] Y. Huang, L. Liang, T. Li, X. Jia, R. Wang, W. Miao, G. Pu, and Y. Liu, "Perception-guided jailbreak against text-to-image models," 2025. [Online]. Available: https://arxiv.org/abs/2408.10848

[19] M. Kim, H. Lee, B. Gong, H. Zhang, and S. J. Hwang, "Automatic jailbreaking of the text-to-image generative ai systems," in *ICML 2024 Next Generation of AI Safety Workshop*, 2024.

[20] P. Chao, E. Debenedetti, A. Robey, M. Andriushchenko, F. Croce, V. Sehwag, E. Dobriban, N. Flammarion, G. J. Pappas, F. Tramèr, H. Hassani, and E. Wong, "Jailbreakbench: An open robustness benchmark for jailbreaking large language models," in *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. [Online]. Available: https://openreview.net/forum?id=urjPCYZt0I

[21] Z.-X. Yong, C. Menghini, and S. H. Bach, "Low-resource languages jailbreak gpt-4," 2024. [Online]. Available: https://arxiv.org/abs/2310.02446