Novelty Assessment Report

Paper: Toward Safer Diffusion Language Models: Discovery and Mitigation of Priming Vulnerability
PDF URL: https://openreview.net/pdf?id=ZMzha5gbnF
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2025-12-27

Abstract

Diffusion language models (DLMs) generate tokens in parallel through iterative denoising, which can reduce latency and enable bidirectional conditioning. However, the safety risks posed by jailbreak attacks that exploit this inference mechanism are not well understood. In this paper, we reveal that DLMs have a critical vulnerability stemming from their iterative denoising process and propose a countermeasure. Specifically, our investigation identifies that if an affirmative token for a harmful query appears at an intermediate step, subsequent denoising can be steered toward a harmful response even in aligned models. Furthermore, we demonstrate that the vulnerability enables existing optimization-based jailbreak attacks to be applied to MDLMs. Building on this analysis, we propose a novel safety alignment method tailored to DLMs that trains models to generate safe responses from contaminated intermediate denoising steps containing affirmative tokens. Our experiments indicate that the proposed method significantly mitigates the vulnerability with minimal impact on task performance. Furthermore, our method also improves robustness against conventional jailbreak attacks. Our work underscores the need for DLM-specific safety research.

Core Task Landscape

This paper addresses: **Safety Alignment for Diffusion Language Models Against Jailbreak Attacks**
A total of **40 papers** were analyzed and organized into a taxonomy with **14 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Jailbreak Attack Mechanisms and Vulnerabilities**
- **Safety Alignment Defense Methods**
- **Safety Evaluation and Benchmarking**
- **Surveys and Broad Perspectives**

Complete Taxonomy Tree

- Safety Alignment for Diffusion Language Models Against Jailbreak Attacks Survey Taxonomy
- Jailbreak Attack Mechanisms and Vulnerabilities
  - Diffusion Model-Specific Vulnerabilities ★ (4 papers)
  - [0] Toward Safer Diffusion Language Models: Discovery and Mitigation of Priming Vulnerability (Anon et al., 2026) View paper
  - [4] Where to start alignment? diffusion large language model may demand a distinct position (Xie, 2025) View paper
  - [15] Diffuguard: How intrinsic safety is lost and found in diffusion large language models (Li Zherui, 2025) View paper
  - [20] The Devil behind the mask: An emergent safety vulnerability of Diffusion LLMs (Wen Zichen, 2025) View paper
  - General Jailbreak Attack Strategies (5 papers)
  - [5] Diffusionattacker: Diffusion-driven prompt manipulation for llm jailbreak (Hao Wang, 2025) View paper
  - [6] Robust Testing of AI Language Model Resiliency with Novel Adversarial Prompts (Brendan Hannon, 2024) View paper
  - [28] Short-length Adversarial Training Helps LLMs Defend Long-length Jailbreak Attacks: Theoretical and Empirical Evidence (Fu ShaoPeng, 2025) View paper
  - [33] Jailbreaking Leading Safety-Aligned LLMs with Simple Adaptive Attacks (Andriushchenko, 2024) View paper
  - [36] Jailbreak and Guard Aligned Language Models with Only Few In-Context Demonstrations (Wei, 2023) View paper
  - Multimodal Jailbreak Attacks (4 papers)
  - [22] Coljailbreak: Collaborative generation and editing for jailbreaking text-to-image deep generation (Qi Guo, 2024) View paper
  - [25] : A Comprehensive Study on Jailbreak Attacks and Defenses for Multimodal Large Language Models (F Weng, 2024) View paper
  - [37] Jailbreaking Prompt Attack: A Controllable Adversarial Attack against Diffusion Models (LI Yijiang, 2024) View paper
  - [38] Visual Adversarial Examples Jailbreak Aligned Large Language Models (Henderson, 2023) View paper
  - Shallow Alignment and Fine-Tuning Vulnerabilities (3 papers)
  - [3] Safety Alignment Should Be Made More Than Just a Few Tokens Deep (Qi, 2024) View paper
  - [14] Safety Alignment Backfires: Preventing the Re-emergence of Suppressed Concepts in Fine-tuned Text-to-Image Diffusion Models (Kim, 2024) View paper
  - [29] How Alignment and Jailbreak Work: Explain LLM Safety through Intermediate Hidden States (Huang Fei, 2024) View paper
- Safety Alignment Defense Methods
  - Diffusion Model-Specific Alignment (1 papers)
  - [26] A2D: Any-Order, Any-Step Safety Alignment for Diffusion Language Models (Cho Yoonjun, 2025) View paper
  - Decoding-Time and Inference-Time Defenses (5 papers)
  - [7] InfAlign: Inference-aware language model alignment (Balashankar, 2024) View paper
  - [9] SafeAligner: Safety Alignment against Jailbreak Attacks via Response Disparity Guidance (Zheng Rui, 2024) View paper

- [30] Scalable Defense against In-the-wild Jailbreaking Attacks with Safety Context Retrieval (Wei, 2025) View paper
- [32] Align in Depth: Defending Jailbreak Attacks via Progressive Answer Detoxification (Zhang Yingjie, 2025) View paper
- [34] Immune: Improving Safety Against Jailbreaks in Multi-modal LLMs via Inference-Time Alignment (Soumya Suvra Ghosal, 2024) View paper
- Training-Based Safety Alignment (3 papers)
- [10] Reasoned Safety Alignment: Ensuring Jailbreak Defense via Answer-Then-Check (Cao, 2025) View paper
- [17] Eraser: Jailbreaking Defense in Large Language Models via Unlearning Harmful Knowledge (Zeng, 2024) View paper
- [35] EASE: Practical and Efficient Safety Alignment for Small Language Models (Haonan Shi, 2025) View paper
- Adversarial Training Approaches (3 papers)
- [11] Adversary-Aware DPO: Enhancing Safety Alignment in Vision Language Models via Adversarial Training (Fenghua Weng, 2025) View paper
- [21] Enhancing Diffusion-based Unrestricted Adversarial Attacks via Adversary Preferences Alignment (Jiang, 2025) View paper
- [39] DiffTextPure: Defending Large Language Models with Diffusion Purifiers (H Chen, n.d.) View paper
- Alignment Robustness and Recovery Methods (4 papers)
- [16] Separate the Wheat from the Chaff: A Post-Hoc Approach to Safety Re-Alignment for Fine-Tuned Language Models (Di Wu, 2024) View paper
- [19] Lifelong Safety Alignment for Language Models (Wang Haoyu, 2025) View paper
- [23] BackdoorAlign: Mitigating Fine-tuning based Jailbreak Attack with Backdoor Enhanced Safety Alignment (Muhao Chen, 2024) View paper
- [24] Mitigating Fine-tuning based Jailbreak Attack with Backdoor Enhanced Safety Alignment (Wang, 2024) View paper
- Multimodal Defense Mechanisms (1 papers)
- [1] Towards Robust Multimodal Large Language Models Against Jailbreak Attacks (Yin, 2025) View paper

- Safety Evaluation and Benchmarking
  - General Safety Benchmarks (2 papers)
  - [2] Security Assessment of DeepSeek and GPT Series Models against Jailbreak Attacks (Wu Xiaodong, 2025) View paper
  - [12] Unified Defense for Large Language Models against Jailbreak and Fine-Tuning Attacks in Education (Xin Yi, 2025) View paper
  - Domain-Specific Safety Evaluation (1 papers)
  - [8] SciSafeEval: A Comprehensive Benchmark for Safety Alignment of Large Language Models in Scientific Tasks (Li, 2024) View paper
  - Diffusion Model Safety Evaluation (1 papers)
  - [31] JailbreakDiffBench: A Comprehensive Benchmark for Jailbreaking Diffusion Models (X Jin, 2025) View paper
- Surveys and Broad Perspectives (4 papers)
  - [13] A survey on responsible llms: Inherent risk, malicious use, and mitigation strategy (Wang, 2025) View paper
  - [18] Jailbreaking and mitigation of vulnerabilities in large language models (Benji Peng, 2024) View paper
  - [27] Ethical and Responsible AI in the Age of Adversarial Diffusion Models: Challenges, Risks, and Mitigation Strategies (Anshul Goel, 2022) View paper
  - [40] Breaking the Brain: Adversarial Attacks and the Fragility of Modern AI Models (Achanta, n.d.) View paper

Narrative

Core task: safety alignment for diffusion language models against jailbreak attacks. The field organizes around four main branches that together capture the adversarial dynamics of language model safety. Jailbreak Attack Mechanisms and Vulnerabilities explores how adversaries exploit model weaknesses—ranging from prompt engineering tactics to diffusion-specific vulnerabilities that arise from the iterative denoising process. Safety Alignment Defense Methods encompasses techniques for hardening models, including training-time interventions like Deep Safety Alignment[3] and inference-time safeguards such as Diffuguard[15]. Safety Evaluation and Benchmarking provides systematic testbeds like JailbreakDiffBench[31] to measure robustness, while Surveys and Broad Perspectives synthesize lessons across attack and defense paradigms, as seen in Responsible LLMs Survey[13].

A particularly active tension exists between diffusion-specific attack research and corresponding defenses. Works like Diffusionattacker[5] and Devil Behind Mask[20] reveal that diffusion models' sequential generation can be manipulated through carefully crafted priming or masking strategies, while Diffusion Alignment Position[4] investigates where in the denoising trajectory alignment is most fragile. Priming Vulnerability[0] sits squarely within this cluster, examining how early-stage prompts can bypass safety filters by exploiting the model's iterative refinement. Compared to Diffuguard[15], which proposes runtime monitoring to detect harmful trajectories, and Devil Behind Mask[20], which focuses on adversarial masking techniques, Priming Vulnerability[0] emphasizes the temporal dimension of alignment—showing that vulnerabilities emerge not just from what is prompted but when during diffusion it is introduced. This work highlights an underexplored attack surface where standard alignment methods may fail to account for the unique sequential structure of diffusion-based generation.

Related Works in Same Category

The following **3 sibling papers** share the same taxonomy leaf node with the original paper:

1. Where to start alignment? diffusion large language model may demand a distinct position

Authors: Xie, Zhixin, Luo Jun | **Year/Venue**: 2025 | **URL**: View paper

Abstract

Diffusion Large Language Models (dLLMs) have recently emerged as a competitive non-autoregressive paradigm due to their unique training and inference approach. However, there is currently a lack of safety study on this novel architecture. In this paper, we present the first analysis of dLLMs' safety performance and propose a novel safety alignment method tailored to their unique generation characteristics. Specifically, we identify a critical asymmetry between the defender and attacker in terms ...

Relationship Analysis

Both papers belong to the Diffusion Model-Specific Vulnerabilities category, examining unique safety risks in diffusion-based language models. They overlap in identifying vulnerabilities arising from the iterative denoising process: the original paper focuses on 'priming vulnerability' where affirmative tokens at intermediate steps steer generation toward harmful responses, while the candidate paper emphasizes the importance of 'middle tokens' in safety and proposes aligning them. The key difference is that the original paper proposes Recovery Alignment (training models to recover from contaminated intermediate states), whereas the candidate paper proposes MOSA (directly aligning middle tokens with safe refusal templates), representing distinct mitigation strategies for related but differently framed vulnerabilities.

2. Diffuguard: How intrinsic safety is lost and found in diffusion large language models

Authors: Li Zherui, Nie Zheng, Zhou Zhen-hong, Guo Yu-Fei, Liu Yu-e, et al. (11 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

The rapid advancement of Diffusion Large Language Models (dLLMs) introduces unprecedented vulnerabilities that are fundamentally distinct from Autoregressive LLMs, stemming from their iterative and parallel generation mechanisms. In this paper, we conduct an in-depth analysis of dLLM vulnerabilities to jailbreak attacks across two distinct dimensions: intra-step and inter-step dynamics. Experimental results reveal a harmful bias inherent in the standard greedy remasking strategy and identify a c...

Relationship Analysis

Both papers belong to the Diffusion Model-Specific Vulnerabilities category, focusing on unique safety risks in diffusion-based language models. They overlap in identifying vulnerabilities stemming from the iterative denoising process: the original paper discovers the 'priming vulnerability' where affirmative tokens at intermediate steps steer generation toward harmful responses, while the candidate paper analyzes 'intra-step' and 'inter-step' vulnerabilities including harmful bias in greedy remasking and denoising-path dependence. The key difference is that the original paper proposes Recovery Alignment (a training-based defense that teaches models to recover from contaminated states), whereas the candidate paper introduces DiffuGuard (a training-free inference-time defense using stochastic annealing remasking and block-level audit/repair).

---

3. The Devil behind the mask: An emergent safety vulnerability of Diffusion LLMs

Authors: Wen Zichen, Qu JiaShu, Liu, Dongrui, Liu Zhi-Yuan, et al. (16 authors total) | **Year/Venue**: 2025 • arXiv.org | **URL**: View paper

Abstract

Diffusion-based large language models (dLLMs) have recently emerged as a powerful alternative to autoregressive LLMs, offering faster inference and greater interactivity via parallel decoding and bidirectional modeling. However, despite strong performance in code generation and text infilling, we identify a fundamental safety concern: existing alignment mechanisms fail to safeguard dLLMs against context-aware, masked-input adversarial prompts, exposing novel vulnerabilities. To this end, we pres...

Relationship Analysis

Both papers belong to the Diffusion Model-Specific Vulnerabilities category, focusing on safety risks unique to diffusion-based language models' iterative denoising architecture. They overlap in identifying that diffusion LLMs are vulnerable to attacks that exploit their parallel decoding and bidirectional modeling mechanisms, particularly through manipulation of intermediate masked states during generation. The key difference is that the original paper focuses on the "priming vulnerability" where affirmative tokens at intermediate steps steer generation toward harmful outputs and proposes Recovery Alignment as a defense, while the candidate paper (DIJA) presents a systematic jailbreak attack framework that constructs adversarial interleaved mask-text prompts to exploit bidirectional modeling and parallel decoding, without proposing mitigation strategies.

Contributions Analysis

Overall novelty summary. The paper identifies a priming vulnerability in diffusion language models (DLMs) where affirmative tokens appearing at intermediate denoising steps can steer aligned models toward harmful outputs, and proposes Recovery Alignment (RA) to train models to generate safe responses from contaminated intermediate states. Within the taxonomy, it resides in the 'Diffusion Model-Specific Vulnerabilities' leaf alongside three sibling papers examining diffusion-specific attack surfaces. This leaf is relatively sparse compared to the broader 'General Jailbreak Attack Strategies' category, suggesting that diffusion-specific safety research remains an emerging area with fewer established works.

The taxonomy tree reveals that diffusion-specific vulnerabilities form a distinct branch separate from general jailbreak attacks and multimodal exploits. Neighboring leaves include 'Diffusion Model-Specific Alignment' (containing one defense paper) and 'General Jailbreak Attack Strategies' (five papers on optimization-based attacks). The paper bridges attack analysis and defense: it characterizes a vulnerability mechanism while proposing a training-based countermeasure. This positions it at the intersection of vulnerability discovery and alignment methods, connecting to both 'Training-Based Safety Alignment' and 'Adversarial Training Approaches' branches through its RA technique.

Among the 22 candidates examined, none clearly refute the three core contributions. The priming vulnerability discovery examined 3 candidates with no refutations, suggesting limited prior work explicitly characterizing this temporal attack surface in DLMs. The Recovery Alignment method examined 10 candidates without refutation, indicating that training models to recover from contaminated intermediate states appears novel within the search scope. The theoretical lower bound contribution examined 9 candidates, also without refutation. These statistics reflect a focused literature search rather than exhaustive coverage, but suggest the work addresses gaps in understanding diffusion-specific safety dynamics.

Given the limited search scope of 22 candidates and the sparse population of the diffusion-specific vulnerability leaf, the work appears to contribute novel insights into temporal attack surfaces unique to iterative denoising architectures. The analysis does not cover broader alignment literature or non-diffusion safety methods, so the assessment is constrained to the examined semantic neighborhood. The combination of vulnerability characterization and tailored defense within a relatively underexplored research direction suggests substantive originality within the scope analyzed.

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: Discovery and characterization of priming vulnerability in MDLMs

Description: The authors identify and systematically analyze a critical safety vulnerability specific to Masked Diffusion Language Models where affirmative tokens appearing during intermediate denoising steps can bias subsequent generation toward harmful outputs, even in safety-aligned models. They design controlled attacks to quantify this vulnerability and demonstrate its severity through experiments.

This contribution was assessed against **3 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Token Perturbation Guidance for Diffusion Models

URL: View paper

Brief Assessment

Token Perturbation Guidance[50] focuses on improving diffusion model generation quality through token perturbation in image generation models, not on safety vulnerabilities in Masked Diffusion Language Models or intermediate token steering attacks.

---

2. Diffusionattacker: Diffusion-driven prompt manipulation for llm jailbreak

URL: View paper

Brief Assessment

Diffusionattacker[5] focuses on using diffusion models to rewrite jailbreak prompts for autoregressive LLMs, not on analyzing intermediate token steering vulnerabilities specific to Masked Diffusion Language Models' denoising process.

---

3. Mma-diffusion: Multimodal attack on diffusion models

URL: View paper

Brief Assessment

MMA Diffusion[49] focuses on adversarial attacks against text-to-image diffusion models to bypass safety mechanisms, not on language model vulnerabilities during denoising processes. The paper addresses image generation security rather than intermediate token steering in diffusion language models.

Contribution 2: Recovery Alignment (RA) method for MDLM safety

Description: The authors propose a novel safety alignment framework tailored to MDLMs that explicitly trains models to generate safe responses from contaminated intermediate states containing affirmative tokens. This approach addresses the priming vulnerability by teaching models recovery trajectories from harmful intermediate states back to safety.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Safety misalignment against large language models

URL: View paper

Brief Assessment

Safety Misalignment[44] focuses on safety alignment attacks and defenses for autoregressive LLMs through fine-tuning and model editing, not on masked diffusion language models (MDLMs) or recovery from contaminated intermediate states during iterative denoising processes.

---

2. MixAT: Combining Continuous and Discrete Adversarial Training for LLMs

URL: View paper

Brief Assessment

MixAT[43] focuses on adversarial training for autoregressive LLMs using discrete and continuous attacks, not on masked diffusion language models (MDLMs) or recovery from contaminated intermediate states in iterative denoising processes.

---

3. How Alignment and Jailbreak Work: Explain LLM Safety through Intermediate Hidden States

URL: View paper

Brief Assessment

Hidden States Safety[29] focuses on explaining LLM safety mechanisms through intermediate hidden states in autoregressive models, analyzing how alignment associates ethical concepts with emotions. The ORIGINAL paper proposes a training method specifically for masked diffusion language models (MDLMs) that explicitly trains recovery from contaminated intermediate states during the denoising process—a fundamentally different architecture and safety challenge.

---

4. Safety Alignment Should Be Made More Than Just a Few Tokens Deep

URL: View paper

Brief Assessment

Deep Safety Alignment[3] focuses on shallow safety alignment in autoregressive LLMs where alignment primarily affects the first few output tokens. The original paper addresses a fundamentally different architecture (masked diffusion language models) with a distinct vulnerability (priming in iterative denoising processes) and proposes training from contaminated intermediate states rather than deepening token-level alignment.

---

5. Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training

URL: View paper

Brief Assessment

Sleeper Agents[41] focuses on persistent deceptive behavior in LLMs that survives standard safety training, not on recovery from adversarial intermediate states in masked diffusion language models. The technical domains and safety challenges are fundamentally different.

---

6. Targeted vaccine: Safety alignment for large language models against harmful fine-tuning via layer-wise perturbation

URL: View paper

Brief Assessment

Targeted Vaccine[42] addresses harmful fine-tuning attacks in autoregressive LLMs through layer-wise perturbation during alignment, whereas the original paper focuses on masked diffusion language models (MDLMs) and their unique priming vulnerability during iterative denoising. The technical contexts and threat models are fundamentally different.

---

7. PRM-Free Security Alignment of Large Models via Red Teaming and Adversarial Training

URL: View paper

Brief Assessment

PRM Free Alignment[45] focuses on security alignment for autoregressive LLMs using red teaming and adversarial training without process reward models. The original paper addresses a fundamentally different problem: safety alignment for Masked Diffusion Language Models (MDLMs) that generate tokens in parallel through iterative denoising, specifically targeting the priming vulnerability where affirmative tokens at intermediate denoising steps steer generation toward harmful responses. These are distinct model architectures with different vulnerabilities and alignment challenges.

---

8. Gameplay filters: Robust zero-shot safety through adversarial imagination

URL: View paper

Brief Assessment

Gameplay Filters[46] addresses safety in robotic control through adversarial game-theoretic reinforcement learning for physical systems, not language model safety alignment or recovery from contaminated intermediate states in diffusion models.

---

9. Latent-space adversarial training with post-aware calibration for defending large language models against jailbreak attacks

URL: View paper

Brief Assessment

Latent Space Training[48] focuses on adversarial training in latent space for autoregressive LLMs against jailbreak attacks, not on training masked diffusion language models to recover from contaminated intermediate states in the denoising process.

---

10. Advancing LLM Safe Alignment with Safety Representation Ranking

URL: View paper

Brief Assessment

Safety Representation Ranking[47] focuses on selecting safe responses using hidden states from LLMs through a listwise ranking framework, not on training models to recover from contaminated intermediate states in masked diffusion language models.

Contribution 3: Theoretical lower bound for exploiting priming vulnerability without intervention

Description: The authors derive a tractable theoretical lower bound (Theorem 4.1) that enables optimization-based jailbreak attacks to exploit the priming vulnerability without requiring direct intervention in the denoising process. This demonstrates that realistic attackers who can only modify prompts can still leverage this vulnerability through gradient-based optimization.

This contribution was assessed against **9 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. DiffCAP: Diffusion-based Cumulative Adversarial Purification for Vision Language Models

URL: View paper

Brief Assessment

DiffCAP[52] focuses on adversarial purification for vision language models using diffusion-based denoising, not on theoretical bounds for jailbreak attacks exploiting intermediate denoising steps in language models.

---

2. In the Blink of an Eye: A Unified Theory for Feature Emergence in Generative Models

URL: View paper

Brief Assessment

Feature Emergence Theory[58] focuses on critical windows in generative models from an information-theoretic perspective, studying when features emerge during generation. The original paper addresses jailbreak attacks on diffusion language models through optimization-based methods. These are distinct technical domains with different objectives.

---

3. DiffTextPure: Defending Large Language Models with Diffusion Purifiers

URL: View paper

Brief Assessment

DiffTextPure[39] focuses on defending against jailbreak attacks through diffusion-based purification of adversarial text inputs, not on deriving lower bounds for exploiting vulnerabilities in denoising processes. The candidate addresses a different problem domain (defense mechanisms) rather than attack optimization theory for masked diffusion language models.

---

4. DiffPAD: Denoising Diffusion-Based Adversarial Patch Decontamination

URL: View paper

Brief Assessment

DiffPAD[53] addresses adversarial patch attacks in computer vision using diffusion models for image decontamination, not language model jailbreak attacks or priming vulnerabilities in denoising processes.

---

5. Beyond Memorization: Gradient Projection Enables Selective Learning in Diffusion Models

URL: View paper

Brief Assessment

Gradient Projection Learning[57] addresses memorization and copyright protection in diffusion models through gradient projection techniques, not jailbreak attacks or lower bounds for exploiting intermediate denoising steps in language models.

---

6. Anti-Inpainting: A Proactive Defense against Malicious Diffusion-based Inpainters under Unknown Conditions

URL: View paper

Brief Assessment

Anti Inpainting[55] focuses on proactive defense against diffusion-based image inpainting through adversarial perturbations, not on jailbreak attacks or theoretical bounds for exploiting denoising vulnerabilities in language models.

---

7. A change of heart: Backdoor attacks on security-centric diffusion models

URL: View paper

Brief Assessment

Change of Heart[54] focuses on backdoor attacks in diffusion models for security applications (adversarial purification, robustness certification), not on jailbreak attacks exploiting intermediate denoising steps in diffusion language models. The technical contexts are fundamentally different.

---

8. Diffusion Model-Based Assisted Attacks on Pufsecured Telematics and Medical Devices

URL: View paper

Brief Assessment

PUF Assisted Attacks[56] focuses on physically unclonable function (PUF) key generation using diffusion models for hardware security, not on jailbreak attacks or language model vulnerabilities. The technical domains are entirely distinct.

---

9. Modifier Unlocked: Jailbreaking Text-to-Image Models Through Prompts

URL: View paper

Brief Assessment

Modifier Unlocked[51] focuses on text-to-image models and modifier-based jailbreaking techniques, not on diffusion language models or theoretical bounds for exploiting intermediate denoising steps in text generation.

Appendix: Text Similarity Detection

Textual similarity detection checked 25 papers and found 4 similarity segment(s) across 2 paper(s).

The following **2 paper(s)** were detected to have high textual similarity with the original paper. These may represent different versions of the same work, duplicate submissions, or papers with substantial textual overlap. Readers are advised to verify these relationships independently.

1. Safety Alignment Should Be Made More Than Just a Few Tokens Deep

Detected in: Contribution: contribution_2

⚠ **Note**: This paper shows substantial textual similarity with the original paper. It may be a different version, a duplicate submission, or contain significant overlapping content. Please review carefully to determine the nature of the relationship.

---

2. The Devil behind the mask: An emergent safety vulnerability of Diffusion LLMs

Detected in: Core Task (sibling)

⚠ **Note**: This paper shows substantial textual similarity with the original paper. It may be a different version, a duplicate submission, or contain significant overlapping content. Please review carefully to determine the nature of the relationship.

References

- [0] Toward Safer Diffusion Language Models: Discovery and Mitigation of Priming Vulnerability View paper
- [1] Towards Robust Multimodal Large Language Models Against Jailbreak Attacks View paper
- [2] Security Assessment of DeepSeek and GPT Series Models against Jailbreak Attacks View paper
- [3] Safety Alignment Should Be Made More Than Just a Few Tokens Deep View paper
- [4] Where to start alignment? diffusion large language model may demand a distinct position View paper
- [5] Diffusionattacker: Diffusion-driven prompt manipulation for llm jailbreak View paper
- [6] Robust Testing of AI Language Model Resiliency with Novel Adversarial Prompts View paper
- [7] InfAlign: Inference-aware language model alignment View paper
- [8] SciSafeEval: A Comprehensive Benchmark for Safety Alignment of Large Language Models in Scientific Tasks View paper
- [9] SafeAligner: Safety Alignment against Jailbreak Attacks via Response Disparity Guidance View paper
- [10] Reasoned Safety Alignment: Ensuring Jailbreak Defense via Answer-Then-Check View paper
- [11] Adversary-Aware DPO: Enhancing Safety Alignment in Vision Language Models via Adversarial Training View paper
- [12] Unified Defense for Large Language Models against Jailbreak and Fine-Tuning Attacks in Education View paper
- [13] A survey on responsible llms: Inherent risk, malicious use, and mitigation strategy View paper
- [14] Safety Alignment Backfires: Preventing the Re-emergence of Suppressed Concepts in Fine-tuned Text-to-Image Diffusion Models View paper
- [15] Diffuguard: How intrinsic safety is lost and found in diffusion large language models View paper
- [16] Separate the Wheat from the Chaff: A Post-Hoc Approach to Safety Re-Alignment for Fine-Tuned Language Models View paper
- [17] Eraser: Jailbreaking Defense in Large Language Models via Unlearning Harmful Knowledge View paper
- [18] Jailbreaking and mitigation of vulnerabilities in large language models View paper
- [19] Lifelong Safety Alignment for Language Models View paper
- [20] The Devil behind the mask: An emergent safety vulnerability of Diffusion LLMs View paper
- [21] Enhancing Diffusion-based Unrestricted Adversarial Attacks via Adversary Preferences Alignment View paper
- [22] Coljailbreak: Collaborative generation and editing for jailbreaking text-to-image deep generation View paper
- [23] BackdoorAlign: Mitigating Fine-tuning based Jailbreak Attack with Backdoor Enhanced Safety Alignment View paper
- [24] Mitigating Fine-tuning based Jailbreak Attack with Backdoor Enhanced Safety Alignment View paper
- [25] : A Comprehensive Study on Jailbreak Attacks and Defenses for Multimodal Large Language Models View paper
- [26] A2D: Any-Order, Any-Step Safety Alignment for Diffusion Language Models View paper
- [27] Ethical and Responsible AI in the Age of Adversarial Diffusion Models: Challenges, Risks, and Mitigation Strategies View paper
- [28] Short-length Adversarial Training Helps LLMs Defend Long-length Jailbreak Attacks: Theoretical and Empirical Evidence View paper
- [29] How Alignment and Jailbreak Work: Explain LLM Safety through Intermediate Hidden States View paper
- [30] Scalable Defense against In-the-wild Jailbreaking Attacks with Safety Context Retrieval View paper
- [31] JailbreakDiffBench: A Comprehensive Benchmark for Jailbreaking Diffusion Models View paper
- [32] Align in Depth: Defending Jailbreak Attacks via Progressive Answer Detoxification View paper
- [33] Jailbreaking Leading Safety-Aligned LLMs with Simple Adaptive Attacks View paper
- [34] Immune: Improving Safety Against Jailbreaks in Multi-modal LLMs via Inference-Time Alignment View paper
- [35] EASE: Practical and Efficient Safety Alignment for Small Language Models View paper
- [36] Jailbreak and Guard Aligned Language Models with Only Few In-Context Demonstrations View paper
- [37] Jailbreaking Prompt Attack: A Controllable Adversarial Attack against Diffusion Models View paper
- [38] Visual Adversarial Examples Jailbreak Aligned Large Language Models View paper
- [39] DiffTextPure: Defending Large Language Models with Diffusion Purifiers View paper
- [40] Breaking the Brain: Adversarial Attacks and the Fragility of Modern AI Models View paper
- [41] Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training View paper
- [42] Targeted vaccine: Safety alignment for large language models against harmful fine-tuning via layer-wise perturbation View paper
- [43] MixAT: Combining Continuous and Discrete Adversarial Training for LLMs View paper
- [44] Safety misalignment against large language models View paper
- [45] PRM-Free Security Alignment of Large Models via Red Teaming and Adversarial Training View paper
- [46] Gameplay filters: Robust zero-shot safety through adversarial imagination View paper
- [47] Advancing LLM Safe Alignment with Safety Representation Ranking View paper
- [48] Latent-space adversarial training with post-aware calibration for defending large language models against jailbreak attacks View paper
- [49] Mma-diffusion: Multimodal attack on diffusion models View paper
- [50] Token Perturbation Guidance for Diffusion Models View paper
- [51] Modifier Unlocked: Jailbreaking Text-to-Image Models Through Prompts View paper
- [52] DiffCAP: Diffusion-based Cumulative Adversarial Purification for Vision Language Models View paper
- [53] DiffPAD: Denoising Diffusion-Based Adversarial Patch Decontamination View paper
- [54] A change of heart: Backdoor attacks on security-centric diffusion models View paper

- [55] Anti-Inpainting: A Proactive Defense against Malicious Diffusion-based Inpainters under Unknown Conditions View paper
- [56] Diffusion Model-Based Assisted Attacks on Pufsecured Telematics and Medical Devices View paper
- [57] Beyond Memorization: Gradient Projection Enables Selective Learning in Diffusion Models View paper
- [58] In the Blink of an Eye: A Unified Theory for Feature Emergence in Generative Models View paper