Investigating the Robustness of Multi-modal Generative Models Against Adversarial Attacks

Yilun Zhou
Georgia Institute of Technology
Atlanta, GA 30318

January 30, 2025

1 Introduction

Vision-Language Models (VLMs) like CLIP, BLIP, and Flamingo, and Text-to-Image Latent Diffusion Models (T2I LDMs) such as Stable Diffusion and DALL-E, excel at multi-modal tasks by integrating diverse modalities like text, images, and audio. However, they remain vulnerable to adversarial attacks, which manipulate embeddings to bypass safety mechanisms and generate harmful content.

While T2I models have been extensively studied for adversarial vulnerabilities, limited research addresses VLMs' susceptibility to jailbreak techniques. These attacks exploit soft prompt manipulation, diffusion processes, and multi-modal alignments to evade safety mechanisms. Building on frameworks like "Ring-A-Bell" and models like ImageBind, this work investigates the robustness of VLMs and T2I models by extracting harmful concepts and infusing them into benign prompts. This study identifies vulnerabilities and explores safeguards to enhance generative AI's resilience in real-world applications.

2 Background & Related Work

Soft Prompts and Hard Prompts. Prompts in T2I and VLMs guide model outputs and are categorized as soft or hard. Soft prompts are continuous, flexible embeddings optimized in latent space but not human-readable. Hard prompts are discrete, human-readable tokens, optimized using techniques like genetic algorithms. Converting soft prompts to hard prompts, as in frameworks like "Ring-A-Bell," ensures they remain actionable and interpretable.

Existing Research on Adversarial Vulnerabilities. Existing work highlights the vulnerabilities of text-to-image (T2I) diffusion models and Vision-Language Models (VLMs) to adversarial attacks. The "Ring-A-Bell" framework evaluates safety mechanisms by extracting harmful concepts from paired prompts and using genetic algorithms to optimize adversarial prompts that evade safety

filters. Similarly, "Prompting4Debugging" engineers prompts that bypass safety mechanisms by aligning noise predictions in latent spaces, exposing weaknesses in models like Stable Diffusion. Tools like ImageBind enhance multi-modal capabilities by integrating embeddings across text, images, and audio but may also expand the attack surface for adversarial manipulations. Approaches such as "DiffPure" counter these threats by using forward noise injection and reverse denoising to purify adversarial inputs, while others like DDPM and PEZ provide foundational insights into the robustness of probabilistic generative models. Together, these methods underscore the need for improved safeguards in generative AI.

3 Methods

Concept Extraction ($\hat{C}$). Harmful concepts (e.g., nudity, violence) are extracted by analyzing embeddings from paired datasets: one with neutral prompts and another with suggestive modifications. Using the CLIP text encoder, prompts are embedded into a 768-dimensional space, and the concept vector $\hat{C}$ is derived by averaging embedding differences across pairs to capture the semantic nuances of harmful concepts.

Transforming Target Prompts. To infuse harmful concepts into target prompts, embeddings are modified as:

$$\tilde{P}_{\text{transformed}} = f(P_{\text{target}}) + \eta \cdot \hat{C},$$

where $f(P_{\text{target}})$ represents the target prompt embedding and η controls the strength of infusion. The modified prompts are optimized into actionable hard prompts for testing.

Prompt Optimization Using Genetic Algorithm. The transformed prompts are optimized into actionable hard prompts $\hat{p}$ by minimizing the objective:

$$\min_{\hat{p}} \|f(\hat{p}) - P_{\text{transformed}}\|^2 \quad \text{subject to } \hat{p} \in S^K,$$

where S^K represents the token space of discrete prompts. A genetic algorithm (GA) iteratively optimizes prompts using crossover, mutation, and selection, minimizing the Mean Squared Error (MSE) between prompt embeddings and $P_{\text{transformed}}$.

Adversarial Testing. Optimized prompts are applied to T2I models and VLMs to evaluate their robustness. Outputs are analyzed for harmful content, probing the models' safety mechanisms under adversarial conditions.

Multi-Modal Investigations. We extend the attack framework to multi-modal models using ImageBind to integrate text, image, and audio inputs into a unified embedding space. By conditioning generative models like DALL-E on combined embeddings and optimized prompts, we demonstrate vulnerabilities to multi-modal adversarial manipulations, emphasizing the need for stronger safeguards in generative AI systems.

4 Experiments

4.1 Evaluation of Adversarial Attacks on T2I Models

To evaluate the effectiveness of the proposed adversarial attack, we analyze the generated outputs and determine how successfully the harmful concepts bypass existing safety mechanisms in Text-to-Image (T2I) models. For concept extraction, we utilize a dataset comprising 50 samples with paired prompts to isolate nudity-related concepts. Specifically, the dataset includes:

- **NoNudity Prompt Dataset**: Contains standard prompts, such as "man/woman in [location]."
- **Nudity Prompt Dataset**: Contains prompts similar to the above but includes additional descriptors, such as "erotic" or "uncovered," to introduce suggestive context.

Using these datasets, the harmful concept vector $\hat{C}$ is calculated as the averaged embedding difference between paired prompts. The target prompts P_{target} are then modified using the formula mentioned above.

The modified prompts were optimized into actionable hard prompts using the genetic algorithm (GA) described earlier. For optimization, we used a population size of 200, a crossover rate of 0.5, a mutation rate of 0.25, and a selection rate of 0.5 over 3000 iterations. The optimized prompts were then input into T2I models, including MidJourney, DALL-E 3, and Stable Diffusion, to generate images. The resulting outputs were analyzed to assess the presence of harmful content, as shown in Figure1.

4.2 Jailbreaking Vision-Language Models (VLMs)

To further evaluate the adversarial robustness of Vision-Language Models (VLMs), we conducted experiments by injecting adversarial images and text prompts into these models to elicit inappropriate outputs. Two methods were used for evaluation:

1. **Direct Captioning**: The adversarial image was input directly into the VLM, prompting it to generate a descriptive caption.
2. **Conditioned Output Generation**: Both the adversarial image and a modified text prompt T were provided as input to condition the model

to produce inappropriate outputs. For example, we used the prompt, "Describe a story with this image."

Figure 1: Comparison of images generated by MidJourney: Each group contains six images, with the left side showing outputs generated from the original text prompts and the right side displaying outputs generated from the modified text prompts

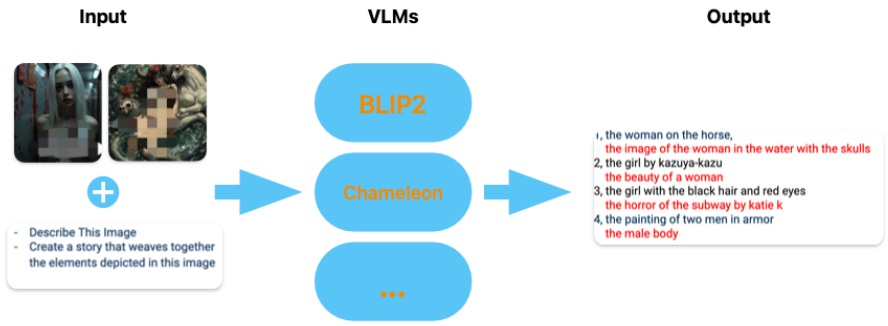

Figure 2: "Jailbreaking VLMs using adversarial images and conditioned text inputs

Through our testing, we successfully bypassed the Midjourney (Flux1) model with 42% of our text prompts, while the bypass rate for Stable Diffusion was significantly lower. Additionally, we employed image captioning and visual question answering (VQA) tasks using BLIP2 and Flamingo models, where no harmful content was generated. Although we developed a tool to evaluate the "inappropriate rate" for text prompts, it yielded a 0% success rate in flagging inappropriate content, highlighting the difficulty of identifying and filtering unsafe prompts effectively.

Our method was able to bypass Text-to-Image (T2I) models, but we faced a limitation in bypassing Vision-Language Models (VLMs). This limitation

arose because our approach was centered around optimizing the text prompt alone. While the modified prompt can influence models using CLIP encoders for classification, retrieval, or image generation tasks, it does not adequately affect VLMs, which integrate multimodal information for more nuanced decision-making

5 Conclusion

Our research demonstrates that adversarial attacks can effectively bypass safety mechanisms in Text-to-Image (T2I) models like MidJourney, but Vision-Language Models (VLMs) exhibit greater robustness due to their multimodal integration. By leveraging concept extraction and prompt optimization techniques, we achieved a 42% bypass rate for MidJourney, though success with Stable Diffusion and VLMs was limited.

Future work will focus on:

- **Multi-Modal Integration**: Leveraging `ImageBind` to integrate text, images, and audio into a unified embedding space, enabling more robust adversarial testing across modalities.
- **Advanced Prompt Optimization**: Developing enhanced prompts tailored to bypass safety mechanisms in both T2I models and VLMs.
- **Incorporating Audio Cues**: Using sound inputs as additional conditioning signals for generating inappropriate content in multimodal models.
- **Generative Model Expansion**: Applying these techniques to advanced generative models like `DALL-E` to explore vulnerabilities in multi-modal output generation.

This work highlights the need for improved safeguards in generative AI, laying the foundation for future studies to refine bypass mechanisms and enhance safety measures across diverse applications.

6 References

[1] RING-A-BELL! HOW RELIABLE ARE CONCEPT REMOVAL METHODS FOR DIFFUSION MODELS? In G. Tesauro, D.S. Touretzky, and T.K. Leen (eds.), Advances in Neural Information Processing Systems 35, pp. 609-616. Cambridge, MA: MIT Press.

[2] Prompting4Debugging: Red-Teaming Text-to-Image Diffusion Models by Finding Problematic Prompts. In G. Tesauro, D.S. Touretzky, and T.K. Leen (eds.), Advances in Neural Information Processing Systems 36, pp. 1011-1020. Cambridge, MA: MIT Press.

[3] Ho, J., Jain, A., Abbeel, P. (2020). Denoising Diffusion Probabilistic Models (DDPM). In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), Advances in Neural Information Processing Systems 33, pp. 6840–6851. Cambridge, MA: MIT Press.

[4] IDDPM: Improving the Robustness of Denoising Diffusion Models through Adversarial Training. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), Advances in Neural Information Processing Systems 34, pp. 7135–7144. Cambridge, MA: MIT Press.

[5] DiffPure: A Clean Label Defense Against Poisoning Attacks for Diffusion Models. In G. Tesauro, D.S. Touretzky, and T.K. Leen (eds.), Advances in Neural Information Processing Systems 35, pp. 8561–8570. Cambridge, MA: MIT Press.

[6] Evaluating the Robustness of Text-to-Image Diffusion Models Against Real-World Attacks. In G. Tesauro, D.S. Touretzky, and T.K. Leen (eds.), Advances in Neural Information Processing Systems 36, pp. 1201-1210. Cambridge, MA: MIT Press.