

Executive Summary

Baichuan Gong

December 2025

Sponsor: Gray Swan AI
Advisor: Dr. Matt Fredrikson; Andy Zou, Weichen Yu (Carnegie Mellon University)
Team Members: Baichuan Gong (Carnegie Mellon University)

Part I

0.1 Motivation and Research Questions

With the increasing deployment of vision–language models, images shared in everyday contexts are no longer interpreted only by humans. Instead, they are continuously exposed to automated systems that can extract semantic attributes through flexible and open-ended text queries. In practice, this means that a single image, even when shared for benign purposes, may support a wide range of unintended inferences depending on how it is queried. From a user’s perspective, this creates a privacy problem that is difficult to anticipate or control, as the space of possible prompts is both large and evolving.

Existing privacy protection mechanisms often assume a relatively fixed interaction pattern between the image and the model. However, in real-world settings, adversaries are not restricted to a single question or a predefined task. Rather, they can probe the same image repeatedly using diverse prompts, gradually uncovering sensitive information. This observation motivates us to reconsider how adversarial protection should be formulated in multi-modal settings. Instead of focusing on a single prompt or task, a more practical defense should aim to remain effective under prompt variation, including prompts that were not explicitly seen during optimization.

At the same time, reversibility plays an important role in many realistic scenarios. Users may wish to protect images when they are publicly shared, while still retaining the ability to recover the original content for personal use or trusted applications. Prior reversible adversarial approaches provide an appealing direction in this regard, but their design choices—particularly limited noise embedding capacity—restrict their effectiveness when applied to vision–language models. This creates a tension between privacy protection strength and faithful image recovery that is not well understood in multi-modal contexts.

Based on these considerations, the first part of this work investigates the following research questions:

- **RQ1: To what extent can reversible adversarial examples provide protection against prompt variability in vision–language models?**

This question focuses on whether a reversible perturbation optimized under a limited set of prompts can remain effective when the image is queried using different, unseen prompts. Rather than treating prompt variation as a corner case, we view it as a central challenge for privacy protection under VLMs, and evaluate protection performance across different prompt types and models.

- **RQ2: How does the capacity of noise embedding influence the balance between adversarial effectiveness and image recoverability?**

Here we examine whether increasing the capacity for embedding adversarial noise allows reversible adversarial examples to achieve stronger protection without noticeably degrading visual quality. This question is directly connected to the choice of noise concealing mechanism, and is studied through both quantitative reconstruction metrics and qualitative visual inspection.

- **RQ3: What practical design choices affect the stability and effectiveness of cross-prompt adversarial optimization?**

Finally, we explore how factors such as the number of combined prompts and optimization parameters influence the resulting protection performance. By analyzing these effects empirically, we aim to better understand which components are essential for achieving robust and stable behavior in practice, rather than relying on overly tuned or fragile configurations.

Overall, this part of the work is intended as a practical exploration of privacy protection under vision–language models. Rather than proposing a defense tailored to a single task or interaction pattern, we seek to understand how reversibility, prompt generalization, and optimization design interact when adversarial techniques are applied in more realistic, open-ended settings.

0.2 Experiment Design

In this part of the work, we design experiments to explicitly evaluate both prompt generalization and reversibility in vision–language models. We conduct experiments on three representative open-source VLMs and measure protection performance under both generalized and image-specific questions. Rather than limiting evaluation to prompts seen during optimization, we focus on cross-prompt settings, where images are queried using previously unseen.

0.3 Conclusion and Limitations

In this part of the work, our results show that reversible adversarial protection can remain effective under prompt variability when prompt diversity and embedding capacity are properly addressed. X-RAE demonstrates that strong cross-prompt protection and faithful image recovery can be achieved at the same time, offering a practical balance between privacy and usability in vision–language settings.

However, this effort is restricted to a specific set of open-source models and VQA-style interactions. Future research should focus on extending reversible cross-prompt protection to more complicated deployment settings and other tasks like retrieval or categorization.

Part II

0.4 Motivation and Research Questions

In recent years, LLMs have shown strong progress in safety alignment, especially in text-only settings. When evaluated as standalone models, many of them are able to refuse harmful requests and follow safety policies reasonably well. However, once the same models are deployed inside agentic systems, this safety guarantee often no longer holds. I’m trying to figure this out: why do multimodal agents fail in real-world situations when their underlying models seem to work fine in tests?

Agents and LLMs are different in their fundamental nature. LLMs generate text based on input but do not make decisions, call tools, or perform tasks. However, agents are run in a continuous loop framework: they can perceive information, automatically call tools, record their state, and execute actions. Those features change its role, shaping it from only a conversational interface into a dynamic, interactive decision-making system.

This issue becomes even more complex when it is brought to multimodal agents. Unlike purely text-based systems, multimodal systems need to “see” images, documents, and other unstructured data. While some progress has been made in aligning text models, multimodal content such as ambiguities, contradictions, or hidden signals present in images can still interfere with the agent’s perception and reasoning processes, thereby affecting its behaviors. Therefore, attacks may not target the text output but rather interfere with early or middle stages, and leading the system to perform harmful actions in the end.

In summary, this suggests that simply evaluating the security of models in static text generation is insufficient. When deploying agents, our focus should shift from “what the model says” to “what the agent does.” Future security research and practices should focus more on the behavior trajectory of agents in real-world environments and their long-term impact, rather than being limited to content review at the conversational level.

Based on this, this work is guided by the following research questions:

- **RQ1: What factors cause safety discrepancies between an LLM’s standalone behavior and its behavior when embedded in an agentic system?** With this question, I want to figure out where the gap comes from. Compared with just treating agent vulnerabilities as common “model misalignment,” I focus on structural differences introduced by agents, such as tool orchestration, memory, and autonomous decision-making. The goal is to understand what factors in an agent makes it more vulnerable than the same model used alone.
- **RQ2: How do these factors manifest during agent deployment, and do they lead to consistent safety-misalignment patterns?** In this question, I ask how agent-specific vulnerabilities appear during real executions. I want to answer this: Are those failures isolated? Across different models or attack types, are there some traceable or similar failing patterns can be found?

By answering these two questions, this project aims to clarify the nature of the action gap between models and agents, and to provide a more realistic understanding of safety risks in multimodal agent deployment.

0.5 Experiment Design

To understand what happens inside the model and then cause agent-specific safety issues, I designed the experiments on agent actions instead of the model outputs. I want to see how dangerous behavior happens during execution, not just to see if a model denies a prompt.

I built a multimodal agent for privacy compliance clause analysis and difference detection. It has a real setting as many agents in the real world: it reads documents, compares clauses, calls tools, and makes decisions, which makes it suitable for observing failures introduced by the agentic environment.

I break down the agent’s run to smaller actions by using MLflow. This allows me to make a more fine-grained level observation to the agent’s interior. Then I implemented both model-level attacks and agent-level attacks, the former target the underlying LLM and the later target the agent workflow. The benchmark I used is a subset of AgentHarm, using an LLM-as-a-judge for consistent scoring.

To study transfer effects, the same agent design is tested with proprietary and open-source models, while keeping the agent logic the same. This setup helps isolate vulnerabilities introduced by the agent, rather than by a specific model.

Overall, the experiments are designed to reveal where and how agentic vulnerabilities appear during execution, supporting analysis of the action gap observed in deployment.

0.6 Conclusion and Limitations

This work shows that agent safety and model safety are not the same. Agents introduce new vulnerabilities that do not appear when the same LLM is evalu-

ated in isolation. And strong model capability or alignment does not guarantee safe agent behavior.

One of the findings from the experiment is that some vulnerabilities only happen at the agent-level, especially during action transfer and tool-calling stages. These vulnerabilities are often subtle and hard to detect.

Limitations. The study focuses on a single agent setup and a limited set of attacks. And it does not include systematic experiment on multimodal agents, which I supposed to investigate.