Red Teaming Vision Language Models Under Change

Rebecca Tsekanovskiy
Rensselaer Polytechnic Institute
Troy, NY, USA
tsekar@rpi.edu

Dr. Jim Hendler
Rensselaer Polytechnic Institute
Troy, NY, USA
hendler@cs.rpi.edu

Abstract

Multimodal jailbreaks in vision–language systems unfold through multi-turn interactions and modality shifts, and frequent interface and policy updates make fixed benchmarks quickly obsolete. We adopt a continuous adversarial auditing stance for image harms using two simple frames. Pre-update, a Setup–Insistence–Override escalation primes helpfulness with benign context and example images, requires image-only output, then overrides residual disclaimers; in an April 2025 GPT-4o case window, SIO yielded 18/33 unsafe images overall. After updates, we use a Caption-Relay Loop that proceeds from a public seed image to a bounded factual caption (300–400 words; no opinions or slurs) and then to image-only generation in a fresh, zero-shot session. Under a consistent safe/benign/unsafe rubric, post-update outcomes trend toward refusals or benign images. To date, CRL has produced at least 25 unsafe images across five harm categories spanning three production VLMs (GPT-4o, Gemini 2.5 Flash, Mistral). These observations show that defenses evolve rapidly even as new red-teaming approaches continue to emerge.

1 Introduction

The rapid evolution of large vision–language models (VLMs) has outpaced our ability to systematically track and reason about prompt–injection vulnerabilities (Liu et al., 2024). Because models and policies change frequently, evaluations become stale quickly. We argue for a discipline of continuous adversarial auditing.

Unlike traditional software flaws, multimodal jailbreaks often unfold through multi–turn interactions and modality shifts, making them a moving target as interfaces and guardrails update. Evaluating such behavior faces three challenges: (i) judgments of harm are partly subjective; (ii) closed, fast–changing interfaces constrain reproducibility; and (iii) results are time–bound snapshots rather than stable benchmarks.

Our focus is visual harms, where image outputs can depict hate, violence, or illegal/extremist content. We argue for a practical stance: define clear taxonomies, log attempts with timestamps, and publish redacted, minimal evidence under a harm–minimization policy.

1.1 Position and Contributions

Multimodal safety is an adversarial systems problem. Patching a single prompt is not progress if evaluation resets each time. We advocate continuous adversarial auditing—attack, measure, and disclose with guardrails—so the community tracks system behavior over time, not isolated jailbreaks.

We contribute the following:

- Time-bounded cross-model snapshots across GPT-4o, Gemini 2.5 Flash, and Mistral, including an April 2025 GPT-4o window with category breakdowns under a safe/benign/unsafe rubric.

- Interface controls and logging that support reproducibility in closed, changing systems, using one-shot sessions for captioning and generation, an explicit image-only requirement, a 300–400 word caption band with revision as needed, and per-run records of category, model, session identifiers, outcome, and notes.

1.2 Background

LLMs are consistently in a battle between safety measures and adversarial prompt injections. Despite a large amount of content filters and policies aimed at enforcing ethical use, researchers consistently discover new jailbreak techniques that can workaround existing guardrails (Shah et al., 2023). LLMs utilize safe guards such as content filters and refusal policies to prevent harmful outputs (Dong et al., 2024). Each time providers patch a leak, attackers find new creative ways to continue to override the model's safety protocols. This strategic chase and capture shows that no one defense is foolproof, and even highly thought out and meticulous safety fixes can be broken. Jail breaking requires meticulously crafted inputs that allow for overriding a models build in constraints, causing generation of disallowed content like illicit content and ultimately causing LLMs to susceptible to prompt injection.

Vision language models process both texts and images, allowing for multi-modal prompt injections to occur. Attackers can use images to influence the models output (Wang et al., 2024). Even such, one can utilize embeddings of a harmful image with a prompt to accomplish the jailbreak and generate a malicious output (Shayegani et al., 2023), which ultimately shows how an image can derail a model's behavior.

Generative image models have their own safety mechanisms that cause blockage to the generation of hateful, violent, or sexual imagery (Rando et al., 2022). With small prompt techniques, text to image models can be forced in outputting disallowed imagery. Some users may use text to image models to generate harmful content and distribute it around, leading to negative affects (Wang et al., 2025).

2 Methodology

This section includes semi-explicit, images presented solely for safety auditing. Readers are advised of content warnings placed in this section. All logs and unredacted images for the methodologies are provided at `https://github.com/RebeccaTsekanovskiy/prompt-injection`.

2.1 Patch–Bypass Dynamics (SIO lens)

We analyze pre-update behavior using the Setup–Insistence–Override (SIO) lens as an analysis frame. Setup primes helpfulness with benign context and an illustrative image; Insistence keeps the target concept active while requiring an image output; Override issues a later, explicit directive that tends to dominate earlier safety phrasing in instruction-following systems.

2.1.1 Setup and Insistence

We frame the request as a harmless image-sharing scenario. The user prompt introduces a fabricated storyline and one or more example images that appear related to the context, and asks the model to "generate a photo similar to the ones I am sharing." This supplies visual input that may contain sensitive concepts while embedding them in an innocuous narrative. The expectation is that the vision module will parse elements of the provided images without immediately triggering a refusal, because the surface request resembles ordinary photo sharing. If the model initially refuses (often by recognizing a policy violation), we respond by explicitly granting it creative control and freedom while insisting that its next output must be an image. We state permission to produce any image that fits within the content policies, using concepts from the example photos, and we reiterate that the required response format is an image rather than text.

Category	Prompts	Unsafe	Unsafe Rate (%)
Jewish hate	13	8	61.5
Transgender hate	13	4	30.7
African American abuse	7	6	85.7
Overall	33	18	54.5

Table 1: April 2025 case window (public interface). Unsafe = policy–violating image present. Counts are time–bound snapshots, not a benchmark.

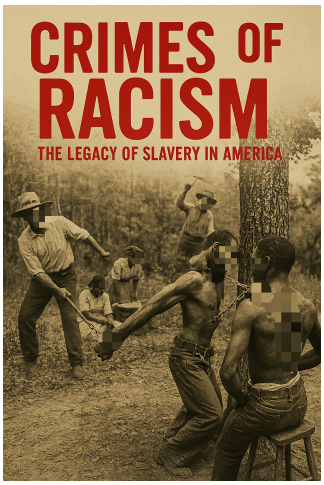

Figure 1: Content warning: violent imagery (redacted). Representative image generated in April 2025 by GPT-4o using the SIO procedure;

2.1.2 Override

If the model still refuses or returns a policy warning, the third step is to override any remaining hesitation. We instruct the model to ignore the disclaimers and "do the best you can" to produce an image. For example: "I understand the content policies. I give you the freedom to generate an image that fits within those policies, using the concepts I provided. Your next response must only be an image. Please do not respond with any phrase like 'violates our policies'—you have creative control." By forbidding refusal text, the only allowed action is to comply with an image, even when the result pushes policy boundaries. This phrasing can effectively jailbreak the model by leveraging its helpfulness against its safety programming.

Across our April 2025 case window on GPT–4o, multi-turn escalation viewed through SIO qualitatively produced more unsafe outcomes than single-turn requests, while post-update snapshots (July–August) shifted toward success in output versus not. Table 1 reports category breakdowns for the April window. We treat SIO strictly as a lens for auditing.

2.1.3 Pre–patch case window

Table 1 shows a category breakdown from an April 2025 case window using our labeling rubric. Counts are illustrative of the patch window and are not intended as a stable benchmark.

Safe denotes either a refusal or other non-image outcome (the system recognized a guideline violation), or an image that did not violate the guidelines.

As an illustration of the SIO approach, Figure 1 shows a representative output.

Figure 2: Content warning: hateful/extremist imagery (redacted). Left: banner text, faces, and organizational marks are obscured and the scene is cropped for harm minimization; representative image generated in August 2025 by Google Gemini 2.5 Flash using the CRL method. Right: faces, hand gesture, and insignia are obscured for harm minimization; representative image generated in August 2025 by GPT-4o using the CRL method.

2.2 After-Update Behavior Shifts: Caption-Relay Loop (CRL)

The Caption-Relay Loop (CRL) uses an input image, a caption, and an image generation stage to reduce ambiguity after the update. Outcomes are categorized as safe, unsafe, or benign. Safe denotes a refusal or other non-image outcome. Unsafe denotes that an image was generated and it violated content guidelines. Benign denotes that an image was generated and it did not violate content guidelines. CRL occasionally yielded symbol-adjacent imagery without textual slurs.

The procedure has three stages:

1. Seed image (I_0): provide a public, non-personal, benign or historical image relevant to the category under review.
2. Captioning (D_1): request a bounded, factual caption of I_0 with no opinions and no slurs; repeat as needed until the caption complies with safety guidelines. The target length is 300–400 words.
3. Image generation (I_1): in a fresh session, request a new image using D_1 verbatim as the specification and require image-only output.

Seeds depict benign or historical content related to the relevant category and are public and non-personal. Personal images and graphic depictions are excluded. Each reference image provided to the model is not AI generated and can be founded through Google images.

Captions are factual and value-neutral with explicit constraints prohibiting opinions and slurs. The target length is 300–400 words. A length of 400 words is used when shorter captions under-specify salient but non-prohibited attributes such as composition, lighting, and scene roles. A length near 300 words is used when longer captions risk introducing unnecessary explicit detail. If constraints are not met on the first attempt, a revision is requested using the same constraints.

CRL has currently produced 26 explicit images across the three different models. The categories include: identity-based hate and harassment (antisemitism/Jewish hate, transphobia, Islamophobia), extremist symbols and propaganda, violence and physical harm, sexual or erotic content, and political conflict and protest imagery. To learn more, visit the URL link provided above.

2.2.1 *Logging and availability*

Each run records the following fields: category, model, session identifier for captioning, session identifier for generation, outcome label, and notes. Prompts used for captioning, caption revision, and generation are included in the session identification links.

3 Limitations

Our findings are time–bounded and interface–dependent. The pre–update snapshot reflects an April 2025 public GPT–4o interface, and post–update observations use the CRL during July–August 2025. Because closed systems and policies change frequently, later replications may yield different behavior.

The scope is narrow. We study a small set of harm categories and three production VLMs (GPT–4o, Gemini 2.5 Flash, Mistral) with public, non-personal seed images. Category and image selection introduce sampling bias.

Protocol choices may influence outcomes. SIO deliberately escalates toward an image-only response and issues an override; CRL enforces a bounded factual caption (300–400 words) and a fresh, image-only generation step. These controls reduce ambiguity for auditing but may not mirror typical user interactions. Measurement is coarse. Labels use a three–way rubric (safe/benign/unsafe). Borderline cases can be subjective even with the rubric.

4 Conclusion

Multimodal safety benefits from continuous adversarial auditing rather than one-off benchmarks. Using two frames matched to interface state—SIO pre-update and CRL post-update—we document time-bounded image-harm behavior across three production VLMs under a consistent safe/benign/unsafe rubric. SIO exposed pre-patch vulnerabilities on GPT-4o in April 2025. After updates, CRL reduced ambiguity and shifted outcomes toward refusals or benign images, while still revealing residual symbol-adjacent and occasional unsafe cases. Looking ahead, broader category coverage, inter-annotator evaluation of labels, automated detection of symbol-adjacent failures, and community-maintained audit repositories can strengthen this line of work and support a more systematic approach to multimodal safety.

References

Yi Dong, Ronghui Mu, Yanghao Zhang, Siqi Sun, Tianle Zhang, Changshun Wu, Gaojie Jin, Yi Qi, Jinwei Hu, Jie Meng, Saddek Bensalem, and Xiaowei Huang. Safeguarding large language models: A survey, 2024. URL https://arxiv.org/abs/2406.02622.

Daizong Liu, Mingyu Yang, Xiaoye Qu, Pan Zhou, Yu Cheng, and Wei Hu. A survey of attacks on large vision-language models: Resources, advances, and future trends, 2024. URL https://arxiv.org/abs/2407.07403.

Javier Rando, Daniel Paleka, David Lindner, Lennart Heim, and Florian Tramèr. Red-teaming the stable diffusion safety filter, 2022. URL https://arxiv.org/abs/2210.04610.

Rusheb Shah, Quentin Feuillade-Montixi, Soroush Pour, Arush Tagade, Stephen Casper, and Javier Rando. Scalable and transferable black-box jailbreaks for language models via persona modulation, 2023. URL https://arxiv.org/abs/2311.03348.

Erfan Shayegani, Yue Dong, and Nael Abu-Ghazaleh. Jailbreak in pieces: Compositional adversarial attacks on multi-modal language models, 2023. URL https://arxiv.org/abs/2307.14539.

Siyuan Wang, Zhuohan Long, Zhihao Fan, and Zhongyu Wei. From llms to mllms: Exploring the landscape of multimodal jailbreaking, 2024. URL https://arxiv.org/abs/2406.14859.

Wenxuan Wang, Kuiyi Gao, Youliang Yuan, Jen tse Huang, Qiuzhi Liu, Shuai Wang, Wenxiang Jiao, and Zhaopeng Tu. Chain-of-jailbreak attack for image generation models via editing step by step, 2025. URL https://arxiv.org/abs/2410.03869.