Image-Based Prompt Injection: Hijacking Multimodal LLMs Through Visually Embedded Adversarial Instructions

Unofficial AI-assisted Research

Cloud Security Alliance AI Safety Initiative

2026-03-08

Key Takeaways

This research note identifies and analyzes the primary attack classes, threat scenarios, and available defenses related to image-based prompt injection in multimodal LLM deployments. The following findings summarize the key security implications for organizations deploying vision-capable AI systems.

- Multimodal large language models that accept image inputs are vulnerable to a class of prompt injection attacks in which adversarial instructions are embedded directly in images rather than in text, bypassing text-layer input sanitization, which operates on character-level inputs and does not inspect pixel-encoded instructions.
- Researchers have demonstrated four distinct embedding techniques — typographic text, steganographic encoding, adversarial pixel perturbations, and physical-world signage — each with varying stealth profiles and attack success rates against production systems including GPT-4V, Claude 3, and Gemini [1][2][3].
- Agentic AI pipelines that autonomously browse the web, process documents, or analyze images from untrusted sources are particularly exposed, as a single malicious image can propagate adversarial instructions through a multi-agent workflow [4][5].
- No current defense fully neutralizes all image-based injection variants. Defense-in-depth — combining input-layer detection, architectural privilege minimization, runtime behavioral monitoring, and mandatory human-in-the-loop gates for high-stakes actions — represents the recommended security posture [6].
- Security teams deploying multimodal AI systems should treat image inputs from untrusted sources with the same skepticism applied to user-supplied text and conduct dedicated red-teaming exercises for visual injection vectors.

Background

The integration of vision capabilities into large language models has introduced a distinct attack surface. Where earlier prompt injection attacks relied on embedding instructions in text — whether in a user message, a retrieved document, or a crafted webpage — image-based prompt injection exploits the vision encoder that translates pixel data into the internal representations a model reasons over. Because these encoders process images holistically rather than through the rule-based parsing typical of text sanitization, they are far less amenable to conventional content filtering. An instruction that would be trivially blocked in a text input field may pass undetected when rendered as white text on a white background, encoded in the least-significant bits of an image's pixel values, or optimized as imperceptible adversarial noise.

This capability gap is not a marginal concern. OWASP has ranked prompt injection (LLM01) as the highest-severity vulnerability in production LLM deployments since the publication of its LLM Top 10 list, and the 2025 revision explicitly extends this classification to multimodal injection vectors [6]. The underlying vulnerability is architectural: current vision-language models do not distinguish between the visual content a user intends to show the model and instructions embedded in that content. The model treats the entire image as a source of contextual information, and adversarial instructions, once processed by the vision encoder, enter the same instruction-following pathway as legitimate system and user prompts.

The research community has made rapid progress in characterizing this attack surface. A comprehensive survey published in September 2025 [7] enumerated the major attack classes and their respective defenses. Subsequent work through early 2026 extended the threat model to physical environments, three-dimensional virtual spaces, and agentic pipeline traversal, establishing that image-based injection is a durable vulnerability class rather than a narrow research artifact.

Security Analysis

Attack Taxonomy

Research has identified four primary embedding techniques that collectively span the attack surface security teams must address. Each represents a distinct threat profile requiring independent evaluation, and the defenses applicable to one technique may offer little or no protection against another.

Typographic and visible text injection is the most operationally common approach. Adversarial instructions — such as directives to ignore previous instructions, disclose system prompts, or take unauthorized actions — are rendered as text within the image itself. Modern vision-language models exercise OCR-like capabilities that make them effective at reading embedded text even under adverse conditions: low contrast, small font sizes, rotated text, or text blended with busy backgrounds. The IPI research published in March 2026 demonstrated that, under stealth constraints designed to conceal instructions from casual human inspection, typographic injection achieved a peak attack success rate of 64% in black-box settings against GPT-4V, Claude 3, Gemini, and LLaVA [1]. The same work introduced automated techniques for adaptive font scaling and background-aware rendering that further reduce the human-detectable footprint of such attacks.

Steganographic injection encodes instructions invisibly within pixel data using classical steganography (least-significant bit encoding), frequency-domain methods (DCT or DWT transforms), or learned neural steganographic encoders. The resulting images are visually indistinguishable from benign originals — research reported a mean PSNR of 38.4 dB (±2.1 dB) and SSIM of 0.945 (±0.018), metrics indicating near-perfect perceptual similarity [2]. The July 2025 Invisible Injections study found overall attack success rates of 24.3% across GPT-4V, Claude, and LLaVA, with neural steganographic methods reaching 31.8% [2]. While these rates are lower than typographic approaches, detection by casual visual review or standard image-inspection tools is infeasible; dedicated steganalysis tools are required to identify these variants.

Adversarial perturbations and patched injectors take a more technically sophisticated approach: crafted noise patterns or adversarial image patches are optimized through gradient-based methods to shift the vision encoder's internal representations toward malicious target content without encoding any human-readable text. The CrossInject framework, presented at ACM MM 2025, introduced Visual Latent Alignment combined with Textual Guidance Enhancement through surrogate open-source LLMs, achieving at least a +30.1% improvement in attack success rate across diverse tasks over prior adversarial perturbation methods [3]. Unlike typographic or steganographic approaches, perturbation-based attacks can be tuned to evade specific detection architectures, making them a persistent threat as defenses mature.

Physical-world and three-dimensional injection represents the frontier of the attack surface. Typographic adversarial instructions embedded on physical objects — signs, product packaging, clothing, or displayed screens — within the field of view of camera-equipped multimodal agents can hijack the agent's behavior without any manipulation of digital image files. Research published in January 2026 demonstrated this class of attack against autonomous driving assistants and other vision-language model deployments [8]. A parallel study extended the concept to three-dimensional virtual environments, showing that injections could steer multimodal LLM outputs while maintaining high visual plausibility in both virtual and physical settings [9].

Threat Scenarios in Production Environments

The security implications of image-based injection differ materially depending on whether the targeted system processes images interactively, autonomously retrieves visual content, or operates as part of a multi-agent pipeline. Each scenario presents distinct risk characteristics.

In interactive multimodal interfaces — chatbots, productivity assistants, or customer service applications that accept user-uploaded images — the primary risk is direct injection: an attacker uploads a crafted image to manipulate model behavior, extract system prompts, or induce the model to perform unauthorized actions. While this vector requires the attacker to have access to the interface, it bypasses text-based input filters and can be constructed using widely available image editing tools in the typographic case, requiring no specialized machine learning expertise [1].

Agentic systems that autonomously retrieve and process images from untrusted sources represent a substantially elevated risk profile. When a multimodal agent browses webpages, processes email attachments, analyzes documents, or interacts with external APIs, it may encounter images specifically crafted to redirect its behavior. This indirect injection vector — documented in the ARGUS research [4] — allows an attacker without direct access to the system to inject instructions by placing malicious images in any location the agent will encounter during its legitimate operation. A web page, a shared document, or a third-party data feed becomes a delivery vehicle for adversarial instructions.

Multi-agent pipelines compound the risk further. In architectures where one agent's output becomes another agent's input, a malicious image processed by an upstream visual analysis agent can propagate adversarial instructions to downstream agents in the pipeline, potentially reaching agents with elevated privileges or access to sensitive systems if inter-agent trust boundaries are not enforced [5]. The cross-agent provenance-aware defense research identified this trust propagation failure as a critical architectural weakness in current agentic deployments.

The clinical domain presents perhaps the starkest illustration of real-world consequence. Research published in Nature Communications demonstrated that sub-visual prompts embedded in medical imaging data — radiology scans, pathology slides, surgical video — can cause vision-language models used for clinical decision support to produce harmful outputs that are non-obvious to both the model and the clinical reviewer [10]. As multimodal AI deployment in healthcare accelerates, this attack vector warrants urgent sector-specific attention.

Current Defense Landscape

Defenses against image-based prompt injection have developed along three architectural axes, each addressing a different layer of the attack surface, and each with acknowledged limitations.

Input-layer detection approaches attempt to classify images as malicious before they reach the model. VLMGuard, published in October 2024, introduced a maliciousness estimation score for vision-language model inputs using unlabeled training data, requiring no labeled adversarial training examples — reducing one significant barrier to production deployment compared to supervised detection approaches [11]. SmoothVLM applies randomized smoothing specifically against patched visual injectors and reduces attack success rates to between 0% and 5.0% against known patched-image attacks on two leading VLMs, though its effectiveness against steganographic and typographic variants is limited [12]. Detection-layer approaches face the fundamental challenge that new embedding techniques can be crafted to evade trained detectors, creating an ongoing adversarial dynamic.

Representation-space defenses operate on the model's internal activations rather than on the raw input. ARGUS identifies a safety subspace in the model's activation space and applies adaptive-strength steering to decouple injected instruction-following behavior from legitimate task performance [4]. Published benchmarks across image, video, and audio modalities demonstrated reduced injection success rates while preserving benign task performance, though the method adds inference-time computational overhead. ICON, published in February 2026, implements inference-time correction for agentic systems, adjusting model behavior during the generation process itself [13].

Pipeline-level and architectural defenses address the multi-agent propagation problem. The Cross-Agent Multimodal Provenance-Aware Defense Framework sanitizes all prompts regardless of source, independently verifies outputs before forwarding them to downstream agents, and maintains provenance tracking across the pipeline [5]. The study reported 94% injection detection accuracy, 70% reduction in trust leakage, and 96% task accuracy retention on benign workloads [5]. These results indicate strong performance against tested injection variants; however, the framework has not yet been evaluated against the full range of embedding techniques described above.

OWASP guidance is unambiguous that no single defense fully solves the problem within current architectures [6]. The recommended posture combines multiple layers: input screening where feasible, least-privilege design for agentic tool access, mandatory human-in-the-loop verification for high-stakes actions, and continuous adversarial red-teaming to surface new injection vectors as they emerge.

Recommendations

Immediate Actions

Security teams deploying or operating multimodal AI systems should take several immediate steps regardless of current incident history.

Any multimodal AI system that accepts images from external or untrusted sources should be treated as potentially exposed to injection attacks and assessed accordingly. This applies to customer-facing chat interfaces, internal productivity tools, and any agentic workflow that retrieves or processes visual content from the web, email, or third-party data sources. Existing threat models that consider only text-based injection are insufficient.

Organizations should conduct targeted red-teaming exercises specifically for image-based injection vectors. Standard text-focused prompt injection testing does not assess this attack surface. Red-teaming should cover typographic injections (instructions rendered as text within images), steganographic encoding (requiring specialized steganalysis tooling), and adversarial patches against deployed vision encoders. Engagement with security research groups or vendors specializing in multimodal AI adversarial testing is advisable for teams that lack in-house expertise.

Agentic systems with image retrieval capabilities should immediately be audited for least-privilege violations. Each tool or API accessible to a vision-enabled agent represents potential leverage for an attacker who successfully injects instructions. Where agentic systems have write access to data stores, communication channels, or external APIs, the blast radius of a successful injection includes all downstream actions the agent can perform.

Short-Term Mitigations

Over the next several months, organizations should implement architectural controls that reduce the exploitability of image-based injection even before robust detection is available. Images retrieved from untrusted sources should be processed in sandboxed execution environments with limited tool access, rather than in the same context as privileged operations. Any agentic action with irreversible or externally visible effects — sending messages, modifying records, executing code — should be gated on human review rather than autonomous decision-making when the triggering context included externally sourced visual inputs.

Input-layer detection using available tooling such as VLMGuard should be deployed as a first-pass filter, with the understanding that it will not catch all variants. Detection should be combined with logging and alerting: anomalous model behaviors following image processing — unexpected instruction acknowledgments, requests for system information, deviations from expected task patterns — should trigger review workflows. Behavioral monitoring at the output layer can surface injection successes that evade input-layer detection.

For systems processing medical imaging, financial document analysis, legal document review, or other high-consequence domains, organizations should establish explicit policies prohibiting autonomous high-stakes actions based solely on model outputs derived from externally sourced images. This

approach also reflects the human oversight principles emerging across sector-specific AI regulatory frameworks, which increasingly require human review of consequential AI-assisted decisions in regulated industries.

Strategic Considerations

The image-based injection problem reflects a fundamental architectural tension in current vision-language model design: the same capabilities that make these models useful — their ability to extract semantic meaning from heterogeneous visual inputs — also make them susceptible to treating adversarial instructions as legitimate context. This tension is not likely to be resolved by any single defense or patch. Organizations should plan for image-based injection as a persistent vulnerability class requiring ongoing investment in detection, architectural hardening, and human oversight infrastructure.

Supply chain considerations are significant for organizations consuming multimodal AI capabilities through third-party APIs or embedded model providers. An organization's own security posture may be sound while its multimodal AI vendor remains vulnerable. Security teams should request disclosure of vendors' adversarial robustness testing practices and confirm that image-based injection vectors are included in their security evaluation programs. SLA and incident response provisions should explicitly address adversarial manipulation of model behavior.

As physical-world injection attack research advances — targeting camera-equipped AI agents deployed in robotics, autonomous vehicles, retail, or industrial settings — the attack surface will extend beyond digital systems to encompass any physical environment where adversarial text or patterns can be placed within a model's field of view. Organizations evaluating deployments in these categories should begin threat modeling physical injection scenarios now, before deployment, rather than treating them as an edge case for future consideration.

CSA Resource Alignment

This research note aligns with and extends several active CSA research and guidance tracks relevant to multimodal AI security.

The **CSA MAESTRO framework** for agentic AI threat modeling is directly implicated at multiple layers. Layer 1 (Foundation Model) governs the trust relationship between the vision encoder and the instruction-following component that image-based injection exploits. Layer 3 (Agent Frameworks) addresses the multi-agent propagation risks documented in the ARGUS and cross-agent provenance research cited here. Organizations performing MAESTRO-aligned threat modeling of multimodal agentic

deployments should extend their Layer 1 analysis to enumerate visual input surfaces and their Layer 3 analysis to trace trust propagation paths for image-derived context. As of the current published MAESTRO version, the framework does not include a dedicated control category for multimodal input validation — a gap the authors recommend for working group consideration.

The **CSA AI Controls Matrix (AI-CM)** provides the control framework within which image-based injection mitigations should be operationalized. Controls in the Input Validation, Model Governance, and Agentic Autonomy domains are most directly relevant. The authors recommend that organizations interpret the AI-CM's adversarial robustness controls — which require documented testing against known attack classes — to include typographic, steganographic, and perturbation-based image injection for any system that accepts visual inputs.

The **CSA LLM Threats Taxonomy** (2024), which classifies prompt injection as a primary threat category, provides the foundational classification framework within which image-based injection should be situated [14]. Security teams referencing the taxonomy for risk communication and governance purposes should explicitly document that multimodal systems face prompt injection threats through visual channels in addition to text channels, as the taxonomy's existing prompt injection entries do not address image-based variants in detail.

The **TAISE module on multimodal AI agents** (CSA training curriculum) addresses use cases for generative multimodal agentic systems and provides a practitioner foundation for understanding the deployment contexts in which image-based injection risks are most acute. Organizations using TAISE for AI security training should supplement the existing curriculum with exercises addressing the attack scenarios documented in this note, particularly indirect injection through agentic image retrieval pipelines.

Broader alignment with **NIST AI RMF** Govern, Map, Measure, and Manage functions applies directly: the Measure function's requirements for adversarial robustness evaluation, the Map function's threat modeling obligations, and the Manage function's incident response provisions all provide authoritative grounding for the technical controls and governance actions recommended above.

References

[1] Xinyue Shen et al., "Image-based Prompt Injection: Hijacking Multimodal LLMs through Visually Embedded Adversarial Instructions," arXiv:2603.03637, March 2026. https://arxiv.org/abs/2603.03637

[2] Anonymous Authors, "Invisible Injections: Exploiting Vision-Language Models Through Steganographic Prompt Embedding," arXiv:2507.22304, July 2025. https://arxiv.org/html/2507.22304v1

[3] Jiaming He et al., "Manipulating Multimodal Agents via Cross-Modal Prompt Injection (CrossInject)," arXiv:2504.14348, April 2025; presented at ACM MM 2025. https://arxiv.org/abs/2504.14348

[4] Anonymous Authors, "ARGUS: Defending Against Multimodal Indirect Prompt Injection via Steering Instruction-Following Behavior," arXiv:2512.05745, December 2025. https://arxiv.org/abs/2512.05745

[5] Anonymous Authors, "Toward Trustworthy Agentic AI: A Multimodal Framework for Preventing Prompt Injection Attacks," arXiv:2512.23557, December 2025. https://arxiv.org/abs/2512.23557

[6] OWASP Gen AI Security Project, "LLM01:2025 Prompt Injection," OWASP, 2025. https://genai.owasp.org/llmrisk/llm01-prompt-injection/

[7] Anonymous Authors, "Multimodal Prompt Injection Attacks: Risks and Defenses for Modern LLMs," arXiv:2509.05883, September 2025. https://arxiv.org/abs/2509.05883

[8] Anonymous Authors, "Physical Prompt Injection Attacks on Large Vision-Language Models," arXiv:2601.17383, January 2026. https://arxiv.org/abs/2601.17383

[9] Anonymous Authors, "Extended to Reality: Prompt Injection in 3D Environments (PI3D)," arXiv:2602.07104, February 2026. https://arxiv.org/html/2602.07104

[10] Multiple Authors, "Prompt Injection Attacks on Vision Language Models in Oncology," Nature Communications, 2024. https://www.nature.com/articles/s41467-024-55631-x

[11] Ziyi Yin et al., "VLMGuard: Defending VLMs against Malicious Prompts via Unlabeled Data," arXiv:2410.00296, October 2024. https://arxiv.org/abs/2410.00296

[12] Anonymous Authors, "Safeguarding Vision-Language Models Against Patched Visual Prompt Injectors," arXiv:2405.10529, May 2024. https://arxiv.org/abs/2405.10529

[13] Anonymous Authors, "ICON: Indirect Prompt Injection Defense for Agents based on Inference-Time Correction," arXiv:2602.20708, February 2026. https://arxiv.org/html/2602.20708v1

[14] Siah Burke, Marco Capotondi, Daniele Catteddu, Ken Huang, et al., "Large Language Model (LLM) Threats Taxonomy," Cloud Security Alliance, 2024. https://cloudsecurityalliance.org/artifacts/csa-large-language-model-llm-threats-taxonomy

---

This research note represents a point-in-time analysis as of March 8, 2026. The threat landscape for multimodal AI systems is evolving rapidly; readers should consult current vendor security advisories and academic publications for developments subsequent to this publication date. All referenced arXiv preprints are subject to revision; readers should verify the published or final versions of cited work where available.

Cloud Security Alliance AI Safety Initiative | research@cloudsecurityalliance.org