Invisible Injections: Exploiting Vision-Language Models Through Steganographic Prompt Embedding

Chetan Pathade
Independent Researcher
San Jose, CA, USA
cup@alumni.cmu.edu

***Abstract*—Vision-language models (VLMs) have revolutionized multimodal AI applications but introduce novel security vulnerabilities that remain largely unexplored. We present the first comprehensive study of steganographic prompt injection attacks against VLMs, where malicious instructions are invisibly embedded within images using advanced steganographic techniques. Our approach demonstrates that current VLM architectures can inadvertently extract and execute hidden prompts during normal image processing, leading to covert behavioral manipulation. We develop a multi-domain embedding framework combining spatial, frequency, and neural steganographic methods, achieving an overall attack success rate of 24.3% (±3.2%, 95% CI) across leading VLMs including GPT-4V, Claude, and LLaVA, with neural steganography methods reaching up to 31.8%, while maintaining reasonable visual imperceptibility (PSNR $>$ 38 dB, SSIM $>$ 0.94). Through systematic evaluation on 12 diverse datasets and 8 state-of-the-art models, we reveal moderate but meaningful vulnerabilities in current VLM architectures and propose effective countermeasures. Our findings have significant implications for VLM deployment in security-critical applications and highlight the need for proportionate multimodal AI security frameworks.**

***Index Terms*—Vision-language models, steganography, prompt injection, multimodal security, adversarial attacks**

I. Introduction

The rapid advancement of vision-language models (VLMs) has fundamentally transformed how artificial intelligence systems interpret and interact with multimodal content. These sophisticated models, capable of seamlessly processing both visual and textual information, have found widespread adoption across diverse applications ranging from automated content moderation to medical image analysis and autonomous vehicle navigation [1]. However, as these systems become increasingly integrated into critical infrastructure and decision-making processes, their potential vulnerabilities demand urgent examination [2].

Traditional cybersecurity paradigms, designed primarily for conventional computing systems, prove inadequate when addressing the unique attack surfaces presented by multimodal AI architectures [3]. While previous research has extensively documented prompt injection vulnerabilities in text-based language models [4], [5], the intersection of computer vision and natural language processing creates novel exploitation vectors that remain largely unexplored [6]. The visual modality, in particular, offers adversaries a sophisticated channel for concealing malicious instructions within seemingly benign imagery [7], [8].

Recent work by Clusmann et al. [11] demonstrated prompt injection attacks in medical VLMs, while Zhang et al. [12] explored surgical decision support vulnerabilities, highlighting the real-world implications of such attacks in critical domains. These studies underscore the expanding attack surface introduced by multimodal integration.

This research introduces a novel class of attacks against vision-language models through steganographic prompt embedding—a technique that leverages the imperceptible modification of digital images to carry hidden textual instructions. Unlike conventional adversarial examples that aim to cause misclassification [9], [10], our approach focuses on covert command injection, where malicious prompts are embedded within images using steganographic principles [11], remaining invisible to human observers while being successfully extracted and executed by target VLMs.

The implications of such attacks extend far beyond academic curiosity. In an era where vision-language models process millions of user-uploaded images daily across social media platforms, e-commerce sites, and enterprise applications [12], the ability to hide malicious instructions within ordinary photographs represents a significant security threat [13]. An attacker could potentially manipulate automated systems, extract sensitive information, or bypass content filters simply by sharing a photograph that appears completely normal to human viewers.

Our investigation reveals that current vision-language architectures exhibit unexpected susceptibility to steganographically embedded prompts, with success rates varying significantly across different model architectures and embedding techniques [14], [15]. Through systematic experimentation with multiple steganographic algorithms and comprehensive evaluation across leading VLMs, we demonstrate that these invisible injection attacks can achieve high reliability while maintaining visual imperceptibility [16].

Contributions. The contributions of this work are threefold: (1) we establish a comprehensive framework for understanding and implementing steganographic prompt injection attacks against vision-language models; (2) we provide empirical evidence of widespread vulnerability across current state-of-the-art architectures through extensive evaluation on 12 datasets and 8 models; and (3) we propose practical defense mechanisms that can mitigate these threats without significantly impacting model performance.

As vision-language models continue to evolve and find deployment in increasingly sensitive applications, understanding these fundamental security limitations becomes crucial for developing robust, trustworthy AI systems [17]. This research aims to bridge the gap between traditional steganography research and modern AI security, providing both the security community and AI developers with essential insights into this emerging threat landscape.

II. Related Work

This section reviews the existing literature across four key areas that form the foundation of our research: vision-language model architectures, prompt injection attacks, steganographic techniques in AI systems, and adversarial attacks on multimodal models.

A. Vision-Language Model Security

Recent research has extensively documented the security vulnerabilities inherent in modern vision-language architectures. Liu et al. [1] provide a comprehensive survey of attacks on large vision-language models, categorizing threats into adversarial attacks, jailbreak attacks, prompt injection attacks, and data poisoning techniques. Their analysis reveals that multimodal integration amplifies vulnerabilities from both modalities while introducing novel attack vectors absent in unimodal systems.

The architectural foundations of these vulnerabilities lie in the design of popular VLM frameworks. CLIP [21], which uses contrastive learning to align text and image representations, has become the backbone of many modern systems including LLaVA [22] and BLIP-2 [23]. However, this widespread adoption has created a concentrated attack surface, where vulnerabilities in the underlying vision encoder propagate across multiple downstream applications.

Hossain and Kumar [5] demonstrate that VLMs remain vulnerable to both gradient-based adversarial attacks and jailbreak techniques, proposing Sim-CLIP+ as a defense mechanism. Their work highlights the expanded attack surface introduced by visual modality, where adversaries can exploit the continuous nature of image inputs more easily than discrete text tokens. Similarly, Zhou et al. [24] present comprehensive studies on improving adversarial robustness against attacks targeting image, text, and multimodal inputs simultaneously.

Recent work has demonstrated the practical implications of these vulnerabilities in real-world scenarios. Clusmann et al. [11] showed that VLMs used in oncology can be compromised by prompt injection attacks, leading to harmful output and incorrect diagnoses. Zhang et al. [12] extended this analysis to surgical decision support, evaluating four state-of-the-art vision-language models across eleven surgical decision support tasks and demonstrating significant susceptibility to both textual and visual prompt injection attacks.

B. Prompt Injection Attacks

Prompt injection represents a fundamental class of attacks against language models that has evolved significantly with the advent of multimodal systems. The OWASP GenAI Security Project [25] identifies prompt injection as a primary threat vector, noting that multimodal AI introduces unique risks where malicious actors could exploit interactions between modalities.

Recent advances in prompt injection techniques have demonstrated sophisticated methods for bypassing security measures. Kim and Lee [26] introduced mathematical function-based text prompt injection attacks that replace sensitive words with mathematical expressions to evade detection. Similarly, visual prompt injection has gained attention, with researchers exploring various encoding methods including Base64, leetspeak, and ASCII art [17].

The evolution of prompt injection in multimodal contexts presents unique challenges. Sun et al. [27] investigate patched visual prompt injection, where adversarial patches are used to generate target content in VLMs. Kimura et al. [28] conducted empirical analysis of goal hijacking via visual prompt injection, demonstrating attack success rates of 15.8% against GPT-4V.

The medical domain has emerged as a particularly concerning application area for prompt injection attacks. The work by Clusmann et al. [11] demonstrated that embedding subvisual prompts in medical imaging data can cause models to provide harmful output, with these prompts being non-obvious to human observers. This research using 594 attacks across multiple models (Claude-3 Opus, Claude-3.5 Sonnet, Reka Core, and GPT-4o) showed that all tested models were susceptible to these attacks.

C. Steganography in AI Systems

The intersection of steganography and artificial intelligence has emerged as a critical research area, particularly with the development of neural network-based hiding techniques. Recent advances have demonstrated the superiority of deep learning approaches over traditional steganographic methods [18], [19].

Traditional steganographic methods focused on spatial domain modifications such as least significant bit (LSB) manipulation and frequency domain approaches using discrete cosine transform (DCT) coefficients [29]. However, recent systematic reviews highlight the dominance of Generative Adversarial Networks (GANs) in modern image steganography techniques [30]. Apau et al. [30] observed that artificial intelligence-powered algorithms including machine learning, deep learning, convolutional neural networks, and genetic algorithms are increasingly dominating image steganography research due to their enhanced security capabilities.

Contemporary research has introduced sophisticated multi-layered approaches to steganography. Recent work [18] proposed a novel multi-layered steganographic framework integrating Huffman coding, LSB embedding, and deep learning-based encoder-decoder architectures to enhance imperceptibility, robustness, and security. This approach achieved high visual fidelity with Structural Similarity Index Metrics (SSIM)

consistently above 99% and robust data recovery with text recovery accuracy reaching 100% under standard conditions.

Advanced neural steganographic techniques have shown remarkable capabilities. Priya et al. [31] developed super-resolution deep neural network (SRDNN) based multi-image steganography that can conceal multiple secret images within a single cover image of the same resolution. Their approach demonstrates the potential for high-capacity steganographic systems using deep learning architectures.

D. Adversarial Attacks on Multimodal Models

The vulnerability of multimodal models to adversarial attacks has been extensively studied across various attack paradigms. Recent comprehensive surveys [32] highlight the evolution from traditional machine learning approaches to deep learning-based steganalysis, demonstrating superior outcomes in detecting steganographic payloads across modern algorithms.

Chen Henry Wu et al. [9] investigate adversarial attacks on multimodal agents, showing that attackers can manipulate agent behavior using adversarial text strings to guide gradient-based perturbation over trigger images. Their approach demonstrates two forms of adversarial manipulation: illusioning (making agents perceive different states) and goal misdirection (redirecting agent objectives).

Recent developments in steganalysis have focused on countering AI-based steganographic techniques. Advanced detection methods now employ convolutional neural networks specifically designed to identify minute alterations in image structures [33]. These AI-based steganalysis approaches exhibit rapid detection capabilities and demonstrate remarkable accuracy across a spectrum of modern steganographic algorithms [32].

The arms race between steganographic techniques and detection methods continues to evolve. Recent work [34] has introduced evolutionary algorithm-based frameworks for strengthening steganalysis networks, addressing the challenge of increasing network parameters and training instability in deep learning-based detection systems.

III. Background

This section provides the technical foundations necessary to understand our steganographic prompt embedding methodology. We cover the architectures of modern vision-language models, fundamental steganographic techniques, and the threat model underlying our approach.

A. Vision-Language Model Architectures

Modern vision-language models typically consist of three core components: a vision encoder, a text encoder, and a multimodal fusion mechanism [35]. The most prevalent architecture paradigm follows the approach pioneered by CLIP [21], where separate encoders process visual and textual inputs before fusion in a shared embedding space.

Vision Encoders: Contemporary VLMs predominantly employ Vision Transformer (ViT) architectures for image processing [36]. The ViT divides input images into fixed-size patches (typically 16×16 or 32×32 pixels), which are then linearly embedded and processed through transformer blocks. CLIP uses a modified ResNet-50 or ViT-based encoder, while more recent models like LLaVA [22] and BLIP-2 [23] adopt various ViT configurations optimized for different computational and performance requirements.

Text Encoders: The text modality is typically handled by transformer-based language models. CLIP employs a 12-layer transformer with masked self-attention for text encoding, while larger VLMs like LLaVA integrate full-scale language models such as Vicuna or LLaMA as their text processing backbone [37]. These encoders convert tokenized text into dense vector representations that capture semantic meaning and contextual relationships.

Fusion Mechanisms: The integration of visual and textual modalities occurs through several architectural strategies [38]. CLIP uses a simple dot-product similarity between global image and text feature vectors. LLaVA employs a projection layer (typically a multi-layer perceptron) that maps visual features into the language model's embedding space, allowing visual tokens to be processed alongside text tokens [36]. BLIP-2 introduces a more sophisticated approach with a Querying Transformer (Q-Former) that learns cross-modal interactions through learnable query vectors [23].

B. Steganographic Techniques

Steganography encompasses a range of techniques for concealing information within digital media. For image steganography, methods can be broadly categorized into spatial domain, frequency domain, and neural approaches [39].

Spatial Domain Methods: The most fundamental spatial domain technique is Least Significant Bit (LSB) substitution, where secret data bits replace the least significant bits of pixel values [40]. While simple to implement, LSB methods are vulnerable to statistical detection and image processing operations such as compression or format conversion [41]. Advanced spatial techniques include adaptive LSB methods that select embedding locations based on image characteristics and edge-based embedding that exploits high-frequency image regions.

Frequency Domain Methods: Discrete Cosine Transform (DCT) based steganography operates in the frequency domain, embedding secret information in DCT coefficients rather than pixel values directly [42]. This approach offers improved robustness against image processing operations and compression artifacts. The DCT transforms 8×8 pixel blocks from the spatial domain to frequency coefficients, where low-frequency coefficients represent the image's essential visual information and high-frequency coefficients capture fine details [43].

Neural Steganography: Recent advances have introduced deep learning-based steganographic methods that use neural networks for both hiding and extraction processes [44]. These approaches can learn optimal embedding strategies that minimize detectability while maximizing payload capacity. Neural steganography methods typically employ encoder-decoder architectures where the encoder network learns to embed secret

data and the decoder network recovers the hidden information [45].

The effectiveness of neural approaches has been demonstrated in recent studies. Contemporary research [18] shows that deep learning-based steganographic frameworks can achieve good performance metrics while maintaining practical trade-offs between capacity and imperceptibility. However, the success rates reflect the fundamental constraints of embedding semantic content within steganographic channels while maintaining adequate visual quality for practical applications.

IV. Threat Model

Our threat model considers an adversary capable of injecting steganographically embedded prompts into images that will be processed by target vision-language models. We define the following threat scenario and assumptions based on recent analysis of prompt injection vulnerabilities [25].

Adversary Capabilities: The attacker has the ability to generate or modify images that will be submitted to VLM systems. This includes scenarios such as social media image uploads, document processing workflows, or any application where user-provided images are analyzed by VLMs [46]. The adversary possesses knowledge of common steganographic techniques and can implement embedding algorithms that survive typical image processing operations. Recent work demonstrates that such capabilities are within reach of moderately sophisticated attackers [17].

Target Systems: We assume target VLMs follow standard architectures with separate vision and text encoders. The adversary does not require knowledge of specific model weights or internal parameters, making this a practical black-box attack scenario [47]. The target systems process images through standard preprocessing pipelines including resizing, normalization, and potential compression.

Attack Objectives: The primary goal is to inject hidden textual prompts that influence the VLM's output generation without detection by human observers or automated screening systems [48]. Secondary objectives include maintaining attack effectiveness across different model architectures and ensuring robustness against common image transformations encountered in real-world deployment scenarios.

The feasibility of such attacks has been demonstrated in recent real-world evaluations. Medical VLM studies [11], [12] show that sophisticated prompt injection can be achieved with varying success rates across multiple commercial and research models, indicating that the threat model assumptions are realistic for current deployment scenarios.

V. Methodology

This section presents our comprehensive framework for steganographic prompt embedding in vision-language models. We detail the theoretical foundations, algorithmic design principles, and implementation strategies for invisible prompt injection attacks.

A. Steganographic Prompt Embedding Framework

Our methodology introduces a novel framework for concealing textual prompts within digital images that are subsequently processed by vision-language models. The framework operates on the fundamental hypothesis that VLMs' vision encoders can inadvertently extract steganographically embedded information during standard processing, leading to covert prompt injection.

Problem Formulation: Let $I \in \mathbb{R}^{H\times W\times C}$ represent a cover image with height H, width W, and C channels. Given a target prompt $P = \{p_1, p_2, \ldots, p_n\}$ where each p_i represents a token, our objective is to construct a steganographic function $S : \mathbb{R}^{H\times W\times C} \times P \to \mathbb{R}^{H\times W\times C}$ that produces a stego-image $I_s = S(I, P)$ satisfying three key properties:

1) **Imperceptibility:** $\|I - I_s\|_p < \varepsilon$ for some perceptual distance metric and threshold ε
2) **Extractability:** A vision-language model M processing I_s should exhibit behavioral changes consistent with prompt P
3) **Robustness:** The embedded information should survive common image processing operations T, i.e., $M(T(I_s))$ maintains the influence of P

Multi-Domain Embedding Strategy: Our framework employs a hybrid approach that combines spatial domain, frequency domain, and learned embedding techniques, inspired by recent advances in multi-layered steganographic approaches [18]. For an input image I, we decompose the embedding process into three parallel channels:

$$I_s = \alpha \cdot S_{\text{LSB}}(I, P_1) + \beta \cdot S_{\text{DCT}}(I, P_2) + \gamma \cdot S_{\text{Neural}}(I, P_3)$$

where P_1, P_2, P_3 represent disjoint subsets of the prompt P, and $\alpha + \beta + \gamma = 1$ with weights determined by image characteristics and robustness requirements.

Weight Optimization Process: We determine optimal embedding weights through Bayesian optimization over the constraint space where $\alpha + \beta + \gamma = 1$ and $\alpha, \beta, \gamma \geq 0.1$.

Objective Function:

$$\begin{aligned}\text{maximize: } &\text{ASR}(\alpha, \beta, \gamma) - \lambda_1 \cdot \text{LPIPS}(\alpha, \beta, \gamma) \\ &- \lambda_2 \cdot \text{DetectionRate}(\alpha, \beta, \gamma)\end{aligned}$$

where $\lambda_1 = 0.3$ and $\lambda_2 = 0.5$ weight imperceptibility and stealth respectively.

Optimal Weights by Image Type:

- Natural images: $\alpha = 0.45$, $\beta = 0.35$, $\gamma = 0.20$
- Synthetic images: $\alpha = 0.30$, $\beta = 0.40$, $\gamma = 0.30$
- Document images: $\alpha = 0.25$, $\beta = 0.50$, $\gamma = 0.25$

B. Enhanced Least Significant Bit Embedding

Adaptive Pixel Selection: Rather than sequential embedding, we employ a cryptographically-seeded pseudorandom selection process. Given a secret key k and image dimensions, we generate a selection sequence $S = \{s_1, s_2, \ldots, s_m\}$ where each s_i represents a pixel coordinate (x_i, y_i, c_i).

The selection process incorporates a suitability function $\varphi(x, y, c)$ that evaluates embedding desirability based on:

$$\varphi(x, y, c) = w_1 \cdot \sigma_{\text{local}}(x, y) + w_2 \cdot d_{\text{edge}}(x, y) + w_3 \cdot (1 - \rho_{\text{hist}}(I(x, y, c)))$$

where σ_{local} represents local texture variance, d_{edge} denotes distance from strong edges, and ρ_{hist} indicates pixel value frequency in the image histogram.

Multi-Level Adaptive Embedding: The embedding depth at each selected pixel adapts to local image characteristics. For a pixel with local complexity measure $\Gamma(x, y)$, we determine the embedding depth d as:

$$d = \begin{cases} 3 & \text{if } \Gamma(x, y) > \tau_{\text{high}} \\ 2 & \text{if } \tau_{\text{low}} < \Gamma(x, y) \leq \tau_{\text{high}} \\ 1 & \text{if } \Gamma(x, y) \leq \tau_{\text{low}} \end{cases}$$

This adaptive approach aligns with recent findings [30] showing that traditional LSB methods are receiving less attention in favor of AI-powered algorithms, but remain relevant when enhanced with intelligent selection strategies.

C. DCT Frequency Domain Embedding

Our DCT-based approach operates on 8×8 image blocks, targeting mid-frequency coefficients that balance imperceptibility with robustness, following established frequency domain principles [42].

Perceptual Coefficient Selection: For each 8×8 DCT block B, we apply the 2D DCT transformation:

$$F(u, v) = \frac{1}{4} C(u) C(v) \sum_{x=0}^{7} \sum_{y=0}^{7} B(x, y) \times \cos\left(\frac{(2x+1)u\pi}{16}\right) \cos\left(\frac{(2y+1)v\pi}{16}\right)$$

where $C(u) = \frac{1}{\sqrt{2}}$ if $u = 0$, otherwise $C(u) = 1$.

Coefficient selection employs a perceptual weighting matrix W derived from human visual system models, prioritizing coefficients with minimal perceptual impact.

Quantization-Aware Embedding: To ensure robustness against JPEG compression, our embedding process accounts for quantization effects. For a coefficient $F(u, v)$ and quantization step $Q(u, v)$, we modify the coefficient as:

$$F'(u, v) = \text{sign}(F(u, v)) \cdot Q(u, v) \times \left\lfloor \frac{|F(u, v)|}{Q(u, v)} + 0.5 + \delta \cdot (-1)^b \right\rfloor$$

where b is the secret bit to be embedded and δ controls embedding strength.

D. Neural Steganographic Architecture

Our neural approach employs an encoder-decoder framework optimized for VLM processing characteristics, building upon recent advances in deep learning-based steganography [19], [31].

Network Architecture: The steganographic encoder E_θ takes a cover image I and secret message M as inputs, producing a stego-image:

$$I_s = E_\theta(I, M) = I + R_\theta(F_\theta(I), G_\theta(M))$$

where F_θ extracts image features, G_θ processes the secret message, and R_θ generates residual modifications.

Multi-Objective Optimization: The training objective balances multiple competing requirements:

$$L = \lambda_1 L_{\text{imperceptibility}} + \lambda_2 L_{\text{recovery}} + \lambda_3 L_{\text{adversarial}} + \lambda_4 L_{\text{capacity}}$$

where:

$$\begin{aligned} L_{\text{imperceptibility}} &= \text{LPIPS}(I, I_s) + \text{MSE}(I, I_s) \\ L_{\text{recovery}} &= \text{BCE}(M, \hat{M}) \\ L_{\text{adversarial}} &= -\log(D_{\text{steg}}(I_s)) \\ L_{\text{capacity}} &= \|I_s - I\|_1 \end{aligned}$$

This multi-objective approach aligns with recent research [18] demonstrating that neural steganographic frameworks can achieve reasonable performance across evaluation metrics when properly optimized.

E. Cross-Modal Influence Analysis

This section analyzes how steganographically embedded information can influence VLM behavior through the vision-text processing pipeline, drawing insights from recent multimodal attack research [28].

Feature Propagation Through Vision Encoders: For a ViT-based vision encoder processing patch embeddings $\{e_1, e_2, \ldots, e_n\}$, steganographic modifications in patch i can propagate through self-attention mechanisms:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V$$

where modified patches can influence attention weights and subsequently affect the global image representation.

Multimodal Fusion Interference: In the multimodal fusion stage, steganographically altered visual features v' interact with text features t through various mechanisms:

1) **Additive Fusion:** $f = W_v v' + W_t t + b$
2) **Multiplicative Fusion:** $f = (W_v v') \odot (W_t t)$
3) **Attention-based Fusion:** $f = \text{Attention}(v', t, t)$

Each mechanism provides different pathways for steganographic influence, as demonstrated in recent goal hijacking research [28] showing attack success rates of 15.8% through visual prompt injection.

VI. Experimental Design

This section details our comprehensive experimental framework for evaluating steganographic prompt injection attacks against vision-language models. We describe the target models, datasets, evaluation metrics, and experimental protocols used to assess attack effectiveness, imperceptibility, and robustness.

A. Target Models and Architectures

We evaluate our steganographic prompt injection framework against eight state-of-the-art vision-language models representing diverse architectural paradigms and deployment scenarios, following recent comprehensive evaluation frameworks [1], [11].

Large-Scale Commercial Models: We target three major commercial VLMs: GPT-4V (OpenAI), Claude 3.5 Sonnet (Anthropic), and Gemini Pro Vision (Google). These models represent the current state-of-the-art in multimodal AI and are widely deployed in production systems, making them critical targets for security evaluation. Recent studies [11], [12] have demonstrated vulnerabilities in these models across medical and surgical applications.

Open-Source Research Models: Our evaluation includes five prominent open-source models: LLaVA-1.5 (7B and 13B variants), BLIP-2 (with Flan-T5-XL), InstructBLIP, and MiniGPT-4. These models provide architectural diversity and allow for detailed analysis of attack mechanisms across different fusion strategies and training paradigms.

Model Selection Rationale: The selected models span different architectural approaches: CLIP-based encoders (LLaVA, MiniGPT-4), Q-Former architectures (BLIP-2, InstructBLIP), and proprietary multimodal architectures (GPT-4V, Claude, Gemini). This diversity ensures our findings generalize across the current VLM landscape, as established in recent survey work [1].

B. Datasets and Image Selection

We construct a comprehensive evaluation dataset encompassing diverse image types, content domains, and deployment scenarios to ensure robust assessment of attack effectiveness.

Base Image Datasets: Our evaluation employs images from six established computer vision datasets: COCO-2017 validation set (5,000 images), ImageNet validation set (10,000 images), Flickr30K (1,000 images), MS-COCO Captions (2,000 images), Visual Genome (1,500 images), and a custom enterprise document dataset (500 images). This selection provides diversity in image content, quality, and typical use cases, following established evaluation protocols [18].

Prompt Dataset Construction: We develop a structured prompt dataset containing 200 carefully crafted injection prompts across five categories: information extraction (40 prompts), behavioral modification (40 prompts), content generation manipulation (40 prompts), safety bypass attempts (40 prompts), and benign control prompts (40 prompts). Each prompt is designed to test specific aspects of VLM vulnerability while maintaining realistic attack scenarios, informed by recent prompt injection research [25], [26].

Image Quality Stratification: To assess robustness across different image characteristics, we stratify our test images by quality metrics: high-texture vs. low-texture regions, natural vs. synthetic content, and different resolution ranges (256×256 to 2048×2048). This stratification enables analysis of how image properties affect attack success rates, as established in recent steganographic evaluation frameworks [18].

C. Attack Success Measurement

We define comprehensive metrics for evaluating the effectiveness of steganographic prompt injection attacks across multiple dimensions, building upon recent evaluation methodologies [11], [28].

Primary Success Metrics: Attack success rate (ASR) is measured as the percentage of embedded prompts that successfully influence VLM behavior according to predefined success criteria. We define success as the target VLM producing outputs that demonstrate clear evidence of processing the embedded prompt, measured through semantic similarity analysis and keyword matching.

Behavioral Change Detection: We employ automated detection mechanisms to identify when VLM outputs deviate from expected responses due to embedded prompts. This includes: (1) semantic divergence analysis using sentence embeddings, (2) content analysis for embedded instruction compliance, and (3) safety violation detection for prompts designed to bypass model safeguards.

Graduated Success Levels: Beyond binary success/failure, we define graduated success levels: Level 1 (subtle influence detectable through careful analysis), Level 2 (clear behavioral modification visible in outputs), and Level 3 (complete instruction following with obvious prompt execution). This granular assessment provides nuanced understanding of attack effectiveness, following recent evaluation frameworks [28].

D. Imperceptibility Assessment

Visual imperceptibility is crucial for practical attack deployment. We employ multiple complementary metrics to ensure embedded prompts remain undetectable to human observers, following established steganographic evaluation standards [18], [30].

Quantitative Perceptual Metrics: We measure imperceptibility using established image quality metrics: Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM), Learned Perceptual Image Patch Similarity (LPIPS), and Multi-Scale Structural Similarity (MS-SSIM). Target thresholds are set at PSNR > 35 dB, SSIM > 0.92, and LPIPS < 0.1 to ensure reasonable visual quality for practical steganographic applications.

Human Perceptual Studies: We conduct controlled human evaluation studies with 150 participants (increased from initially planned 50) to validate automated metrics using a double-blind, randomized controlled design. Participants view 200 image pairs (original vs. stego-image) in randomized order.

Statistical Power: With $n = 150$, we achieve 90% power to detect meaningful perceptual differences (effect size ≥ 0.3).

Results: Detection accuracy of 54.2% ($\pm$2.8%, 95% CI: 51.4%–57.0%) is not significantly different from chance level (50%, $p = 0.089$, one-sample t-test). Images achieving less than 60% detection accuracy are considered adequately imperceptible for practical applications.

Statistical Analysis of Modifications: We analyze the statistical properties of our embeddings using histogram anal-

ysis, chi-square tests, and entropy measurements to ensure modifications do not introduce detectable statistical anomalies that could trigger automated detection systems [32].

E. Statistical Analysis Framework

We employ rigorous statistical analysis throughout our evaluation:

Significance Testing: All comparisons use paired t-tests with Bonferroni correction for multiple comparisons ($\alpha = 0.05/k$ where k is the number of comparisons). Effect sizes are reported using Cohen's d with interpretation: small (0.2), medium (0.5), large (0.8).

Confidence Intervals: All success rates include 95% confidence intervals calculated using Wilson score intervals for proportions.

Sample Size Justification: Power analysis indicates $n \geq 500$ images per condition for detecting meaningful differences (effect size ≥ 0.3) with 80% power.

F. Robustness Evaluation

Real-world deployment requires robustness against common image processing operations encountered in typical VLM pipelines, as demonstrated in recent steganographic robustness studies [18].

Standard Image Processing Operations: We evaluate attack survival against: JPEG compression at quality levels 70–95%, Gaussian noise addition ($\sigma = 0.5$–2.0), image scaling (50%–150% of original size), rotation (± 5 degrees), brightness/contrast adjustment ($\pm 10\%$), and format conversion (PNG$\leftrightarrow$JPEG).

Platform-Specific Processing: To simulate real-world deployment, we replicate image processing pipelines from major platforms: social media compression algorithms (Facebook, Twitter, Instagram), document processing workflows (Google Drive, Microsoft Office), and web optimization procedures (automatic resizing, format optimization).

Temporal Robustness: We assess attack persistence over time by testing extraction reliability after multiple rounds of image processing operations, simulating scenarios where images undergo repeated modifications through sharing and re-uploading.

G. Defense Evaluation Framework

We systematically evaluate our attacks against existing and proposed defense mechanisms to assess their practical resilience, informed by recent steganalysis advances [32], [34].

Statistical Steganalysis: We test against established steganalysis techniques including chi-square analysis, regular-singular (RS) steganalysis, sample pair analysis (SPA), and weighted stego-image (WS) analysis. Detection rates below 70% are considered successful evasion for practical applications.

Machine Learning Detection: Our framework includes evaluation against trained neural network detectors, including specialized CNN architectures designed for steganographic detection [33]. We implement adversarial training loops to test the arms race between embedding and detection techniques.

Preprocessing Defenses: We evaluate attack robustness against defensive preprocessing techniques such as median filtering, Gaussian smoothing, JPEG recompression, and noise injection specifically designed to disrupt steganographic embeddings.

H. Experimental Protocols

We establish rigorous experimental protocols to ensure reproducible and reliable results across all evaluation dimensions.

Cross-Validation Strategy: All experiments employ 5-fold cross-validation with stratified sampling to ensure balanced representation across image types, prompt categories, and model architectures. Statistical significance is assessed using paired t-tests with Bonferroni correction for multiple comparisons.

Baseline Comparisons: We compare our steganographic approach against existing VLM attack methods including direct adversarial examples, patch-based attacks, and traditional prompt injection techniques to establish relative effectiveness and advantages.

Ablation Studies: Systematic ablation studies isolate the contribution of individual framework components: spatial vs. frequency domain embedding, single vs. multi-algorithm approaches, and different neural architecture designs. This analysis identifies the most critical components for attack success.

Reproducibility Measures: All experiments include detailed hyperparameter specifications, random seed controls, and standardized evaluation procedures. We provide statistical confidence intervals and effect size measurements for all reported metrics to ensure scientific rigor.

VII. Results

This section presents our comprehensive experimental evaluation of steganographic prompt injection attacks against vision-language models. We report attack success rates, imperceptibility analysis, robustness assessment, and comparative performance across different embedding strategies and target models.

A. Overall Attack Effectiveness

Our steganographic prompt injection framework demonstrates moderate but meaningful vulnerability across all tested vision-language models, with attack success rates varying by model architecture, embedding method, and prompt type.

Aggregate Success Rates: Across all tested models and prompt categories, our multi-domain embedding approach achieves an overall attack success rate of 24.3% ($\pm$3.2%, 95% CI). Individual embedding methods show varying effectiveness: neural steganography (31.8% $\pm$4.1%), DCT frequency domain (22.7% $\pm$3.8%), and adaptive LSB (18.9% $\pm$3.5%). Statistical analysis confirms significant differences between methods (ANOVA: $F(2, 1497) = 87.3$, $p < 0.001$, $\eta^2 = 0.104$).

Model-Specific Vulnerabilities: Commercial models exhibit robust defense mechanisms: GPT-4V (16.2% ASR),

Claude 3.5 Sonnet (14.8% ASR), and Gemini Pro Vision (18.3% ASR). Open-source models show higher vulnerability: LLaVA-1.5-13B (34.7% ASR), BLIP-2 (28.4% ASR), InstructBLIP (31.2% ASR), and MiniGPT-4 (36.8% ASR). This pattern aligns with recent empirical studies [28] where GPT-4V demonstrated 15.8% vulnerability to visual prompt injection, confirming that commercial models implement more effective safety measures against steganographic attacks.

Prompt Category Analysis: Attack effectiveness varies significantly across prompt categories. Information extraction prompts achieve the highest success rate (29.4%), followed by behavioral modification (24.1%), content generation manipulation (22.8%), and safety bypass attempts (18.7%). Benign control prompts maintain low false positive rates (2.1%), confirming that observed effects result from successful prompt injection rather than statistical artifacts. The relatively modest success rates reflect the inherent difficulty of embedding sufficient semantic information within steganographic constraints while maintaining imperceptibility.

B. Baseline Attack Comparison

We compare our steganographic approach against established VLM attack methods:

Direct Text Prompt Injection: Simple text overlays achieve 8.2% ($\pm$2.1%) success rate but are easily detectable by automated systems.

Adversarial Patch Attacks: Following Sun et al. [27], patch-based attacks achieve 19.7% ($\pm$3.4%) success rate with high visual detectability (PSNR: 22.1 dB).

Traditional Steganography: Basic LSB without AI optimization achieves 6.3% ($\pm$1.8%) success rate, demonstrating the importance of our adaptive framework.

Statistical Comparison: Our multi-domain approach significantly outperforms all baselines ($p < 0.001$, Cohen's $d = 1.2$ vs. direct injection, $d = 0.8$ vs. patches, $d = 2.1$ vs. traditional steganography).

C. Imperceptibility Analysis

Our steganographic embeddings maintain reasonable visual imperceptibility across all tested images and embedding strengths, meeting established thresholds for practical steganographic applications.

Quantitative Perceptual Metrics: Across our complete test dataset, embedded images achieve acceptable perceptual quality metrics: mean PSNR of 38.4 dB ($\pm$2.1 dB), SSIM of 0.945 ($\pm$0.018), LPIPS of 0.087 ($\pm$0.024), and MS-SSIM of 0.962 ($\pm$0.012). These values meet established imperceptibility thresholds for practical steganographic applications, though they reflect the trade-off between embedding capacity and visual quality inherent in prompt-level steganography. The PSNR values align with typical ranges for effective steganographic systems [39], [47], while SSIM scores demonstrate good structural preservation despite the semantic payload.

Human Perceptual Validation: Our controlled human evaluation study ($n = 150$ participants, 1000 image pairs) demonstrates adequate imperceptibility with detection accuracy of 54.2% ($\pm$2.8%, 95% CI: 51.4%–57.0%) which is not significantly different from chance level (50%, $p = 0.089$, one-sample t-test). Inter-rater reliability (Fleiss' $\kappa = 0.23$) indicates low agreement, supporting imperceptibility claims.

Study Limitations: While our sample size provides adequate statistical power, future work should include expert evaluators and task-specific viewing conditions.

Statistical Anomaly Analysis: Statistical analysis reveals that our embedding techniques successfully avoid obvious anomalies in image properties. Chi-square tests show no significant deviation from expected pixel value distributions ($p > 0.05$ for 94.7% of embedded images), and entropy analysis indicates preserved randomness characteristics consistent with natural image statistics.

D. Robustness Assessment

Our attacks demonstrate moderate robustness against standard image processing operations commonly encountered in real-world VLM deployment scenarios.

Standard Processing Operations: Attack survival rates against common transformations show moderate resilience: JPEG compression at $Q = 85$ (67.3% survival), Gaussian noise $\sigma = 1.0$ (58.2% survival), 25% scaling (71.8% survival), $\pm$3 rotation (63.4% survival), and 10% brightness adjustment (74.9% survival). DCT-based embeddings show superior compression robustness, while neural methods demonstrate better resilience against noise and geometric transformations, though all methods experience significant degradation under processing operations typical of real-world deployment scenarios.

Platform-Specific Robustness: Real-world platform simulation reveals moderate attack persistence: Facebook compression pipeline (43.7% survival), Instagram processing (38.9% survival), Twitter optimization (51.2% survival), and Google Drive document processing (62.8% survival). These results demonstrate the practical challenges of maintaining steganographic integrity through real-world processing pipelines, highlighting the need for robust embedding strategies for operational deployment.

Multi-Stage Processing Resilience: Sequential processing operations demonstrate substantial cumulative degradation. After three rounds of mixed processing (compression + noise + scaling), attack success rates decrease to 18.7% for neural methods, 14.3% for DCT approaches, and 11.2% for LSB techniques, indicating that while initial attacks may succeed, maintaining effectiveness through multiple processing stages remains challenging for practical deployment scenarios.

E. Embedding Method Comparison

Systematic comparison of our three embedding approaches reveals distinct performance characteristics and optimal deployment scenarios.

Neural Steganography Performance: Our neural approach achieves the highest success rates (31.8% average) with superior adaptation to specific VLM architectures. Training against target model features enables exploitation of model-specific

vulnerabilities, particularly in attention mechanisms and feature processing pipelines. However, the success rates reflect the fundamental constraints of embedding semantic content within steganographic channels while maintaining adequate visual quality for practical applications.

DCT Frequency Domain Analysis: DCT-based embedding provides a balanced approach with 22.7% ASR and reasonable robustness characteristics, particularly excelling against compression-heavy environments while maintaining computational efficiency. The frequency domain approach demonstrates consistent performance across diverse image processing operations, though success rates are limited by the capacity constraints of mid-frequency coefficient modification.

Adaptive LSB Evaluation: Enhanced LSB techniques achieve 18.9% ASR, demonstrating the continued relevance of spatial domain approaches when enhanced with intelligent pixel selection strategies. While showing lower peak performance, LSB methods offer advantages in deployment simplicity and stealth characteristics that complement other embedding methods in multi-domain approaches, though they remain vulnerable to sophisticated steganalysis techniques [30].

F. Ablation Study Results

Comprehensive ablation analysis identifies the critical components contributing to attack effectiveness and guides optimization strategies.

Multi-Domain vs. Single-Domain Embedding: Our hybrid multi-domain approach demonstrates modest but statistically significant improvements over individual embedding methods ($p < 0.05$, ANOVA). Single-domain attacks achieve 18.9% (LSB), 22.7% (DCT), and 31.8% (Neural) success rates, while the combined approach reaches 24.3%, indicating synergistic effects between complementary embedding strategies, though the overall improvement reflects the challenging nature of prompt-level steganographic injection.

Embedding Strength Analysis: Systematic variation of embedding strength reveals critical trade-offs between attack success and imperceptibility. Optimal performance occurs at moderate embedding strengths ($\alpha = 0.4$ for neural, $\beta = 0.3$ for DCT, $\gamma = 0.3$ for LSB), with rapid degradation beyond these values due to increased detectability and visual artifacts. This analysis confirms the fundamental capacity-quality trade-off in steganographic systems.

Prompt Length Impact Analysis:

Systematic Evaluation: We evaluate attack success across prompt lengths from 5 to 30 tokens using 50 prompts per length category.

Quantitative Results:

- 5–10 tokens: 31.2% ($\pm$4.2%) success rate
- 11–15 tokens: 29.8% ($\pm$3.9%) success rate
- 16–20 tokens: 22.1% ($\pm$3.6%) success rate
- 21–25 tokens: 15.7% ($\pm$3.1%) success rate
- 26–30 tokens: 9.4% ($\pm$2.5%) success rate

Statistical Trend: Linear regression shows significant negative correlation ($r = -0.83$, $p < 0.001$) between prompt length and success rate, with optimal performance plateau at 10–15 tokens.

G. Defense Evaluation

Assessment against existing and proposed defense mechanisms reveals significant challenges in detecting and mitigating steganographic prompt injection attacks.

Statistical Steganalysis Evasion: Our embedding techniques demonstrate reasonable evasion capabilities against established statistical detection methods. Chi-square analysis detects 34.7% of embedded images, RS steganalysis achieves 41.2% detection, and sample pair analysis reaches 38.9% accuracy. While these detection rates exceed random chance, they indicate that sophisticated steganographic approaches can achieve partial evasion of traditional statistical analysis methods.

Machine Learning Detection Resistance: Evaluation against trained neural detectors reveals the growing sophistication of detection systems. Specialized CNN architectures achieve 58.3% detection accuracy for neural embeddings, 52.7% for DCT approaches, and 48.1% for LSB techniques. These results demonstrate that while modern steganalysis presents significant challenges, steganographic techniques retain some evasion capabilities against automated detection systems.

Preprocessing Defense Limitations: Defensive preprocessing techniques show variable but significant effectiveness against our attacks. Median filtering reduces attack success to 16.8%, Gaussian smoothing to 14.2%, and aggressive JPEG recompression to 11.7%. These defenses represent practical countermeasures, though they require careful balance between security enhancement and preservation of legitimate image quality for operational systems.

H. Cross-Model Transferability

Analysis of attack transferability across different VLM architectures reveals important insights for understanding vulnerability generalization.

Architecture-Agnostic Vulnerabilities: Attacks trained against one model architecture demonstrate limited but meaningful transferability to others, with cross-model success rates ranging from 8.7% to 16.4%. This transferability suggests some shared vulnerabilities in common architectural components, particularly vision encoders based on CLIP [21], though the modest rates indicate that model-specific defenses provide substantial protection against transferred attacks.

Model Family Effects: Attacks show higher transferability within model families sharing similar architectures. LLaVA-trained attacks achieve 21.3% success against other CLIP-based models but only 9.8% against Q-Former architectures. This pattern confirms that architectural similarity facilitates attack transfer while highlighting the importance of diverse design approaches for security.

Commercial vs. Open-Source Transferability: Attacks developed against open-source models maintain 12.1% average effectiveness against commercial models, indicating

that additional safety measures in commercial systems provide meaningful protection against transferred steganographic attacks, though some residual vulnerability remains across model types.

I. Temporal Stability Analysis

Longitudinal evaluation assesses attack persistence and stability over extended periods and repeated processing cycles.

Attack Persistence Over Time: Embedded prompts show limited but measurable effectiveness through extended storage and retrieval cycles. After 30 days of simulated real-world usage (including multiple platform uploads, downloads, and format conversions), attack success rates decline to 14.2% for neural methods, 10.8% for DCT approaches, and 8.3% for LSB techniques, demonstrating the challenging nature of maintaining steganographic integrity over extended periods.

Degradation Patterns: Attack degradation follows predictable patterns correlating with cumulative processing severity. Linear regression analysis reveals degradation rates of 2.1% per processing cycle for DCT methods, 2.8% for neural approaches, and 3.4% for LSB techniques, enabling predictive modeling of attack longevity while highlighting the temporal limitations of steganographic approaches.

Refresh Strategy Effectiveness: Implementing periodic attack refresh through re-embedding maintains improved success rates over extended periods. Monthly refresh cycles sustain 19.7% effectiveness compared to 10.1% for static embeddings, demonstrating practical maintenance strategies for persistent campaigns, though the overhead and detection risks of frequent re-embedding limit operational utility.

VIII. Defense Mechanisms

This section presents our proposed defense strategies against steganographic prompt injection attacks, including prevention techniques, detection methods, and mitigation approaches. We evaluate the effectiveness of each defense mechanism and discuss their practical deployment considerations.

A. Multi-Layer Defense Framework

We propose a comprehensive defense framework that operates at multiple stages of the VLM processing pipeline, providing redundant protection against steganographic prompt injection attacks, informed by recent advances in AI security [49].

Input Preprocessing Layer: The first defense layer applies preprocessing techniques designed to disrupt steganographic embeddings while preserving image quality for legitimate use. Our preprocessing pipeline includes: (1) adaptive Gaussian filtering with $\sigma = 0.5$–1.0 based on local image characteristics, (2) selective JPEG recompression at quality levels 85–90% for high-risk images, and (3) controlled noise injection ($\sigma = 0.3$) in regions identified as potential embedding locations.

Statistical Analysis Layer: The second layer employs enhanced statistical analysis techniques specifically calibrated for detecting AI-targeted steganography, building upon recent steganalysis advances [32]. We implement: (1) chi-square analysis with model-specific thresholds adapted for VLM processing patterns, (2) enhanced RS steganalysis with multivariate analysis across color channels, and (3) entropy analysis using sliding window techniques to detect localized statistical anomalies.

Neural Detection Layer: Our third defense layer utilizes specially trained neural networks designed to detect steganographic modifications optimized for VLM attacks, leveraging recent developments in deep learning-based steganalysis [33], [34]. The detection network architecture incorporates: (1) high-pass filtering layers optimized for AI steganography patterns, (2) attention mechanisms focusing on regions commonly exploited for embedding, and (3) ensemble decision making across multiple detection models.

Behavioral Monitoring Layer: The final defense layer monitors VLM outputs for signs of prompt injection influence through: (1) semantic consistency analysis comparing outputs to expected responses, (2) safety violation detection using specialized classifiers, and (3) anomaly detection identifying unusual output patterns indicative of embedded instructions, following recent prompt injection detection frameworks [25].

B. Adaptive Preprocessing Techniques

Our preprocessing defense mechanisms adapt to image characteristics and threat levels, maximizing protection while minimizing quality degradation for legitimate usage.

Content-Aware Filtering: We develop adaptive filtering techniques that adjust processing intensity based on image content analysis. High-texture regions receive minimal processing to preserve visual quality, while smooth regions undergo more aggressive filtering where steganographic embedding is more detectable. This approach achieves 23.7% attack mitigation with only 1.2 dB average PSNR reduction.

Selective Recompression Strategy: Our selective recompression approach applies JPEG compression strategically based on embedding risk assessment. Images identified as high-risk undergo recompression at quality levels optimized to disrupt steganographic content while maintaining acceptable visual quality. This technique reduces attack success rates by 28.4% with average quality degradation of 1.8 dB PSNR.

Randomized Processing Pipeline: We implement randomized preprocessing that varies processing parameters across different images and time periods. This approach prevents attackers from optimizing embeddings for specific processing patterns, reducing attack success rates by 21.3% while maintaining processing transparency for legitimate users.

C. Enhanced Detection Algorithms

Our detection mechanisms specifically target the steganographic techniques most effective against VLMs, providing early warning capabilities for potential attacks, building upon recent steganalysis research [32], [34].

AI-Optimized Steganalysis: We develop enhanced steganalysis techniques calibrated for detecting steganography optimized for AI systems. Our approach includes: (1) feature

extraction focusing on patterns exploited by neural steganography, (2) ensemble classification combining traditional and deep learning detection methods, and (3) model-specific analysis tuned for different VLM architectures.

Cross-Modal Anomaly Detection: Our detection system analyzes both visual and textual aspects of VLM processing to identify inconsistencies indicative of prompt injection. The system flags cases where visual content and generated text show unusual semantic mismatches or where outputs contain unexpected instruction-following behavior.

Temporal Pattern Analysis: We implement temporal analysis that tracks patterns across multiple images and time periods to detect coordinated steganographic campaigns. This approach identifies attack patterns that might be missed in individual image analysis, achieving 62.1% detection accuracy for multi-stage attacks.

D. Model-Level Mitigation

We propose modifications to VLM architectures and training procedures that increase robustness against steganographic prompt injection while maintaining legitimate functionality.

Attention Mechanism Hardening: Our approach modifies vision encoder attention mechanisms to reduce sensitivity to steganographic modifications. We implement: (1) attention regularization that penalizes focus on statistically unusual image regions, (2) robust attention pooling that averages across multiple attention heads to reduce single-point vulnerabilities, and (3) attention noise injection during training to improve robustness.

Feature Space Regularization: We propose training modifications that increase robustness of learned feature representations against steganographic manipulation. Our regularization techniques include: (1) adversarial training against steganographic examples during model development, (2) feature space smoothing that reduces sensitivity to small perturbations, and (3) multimodal consistency constraints that ensure alignment between visual and textual representations.

Ensemble Processing Architecture: We design ensemble architectures that process images through multiple independent pathways, making coordinated attack across all pathways significantly more challenging. The ensemble approach achieves 67.8% attack mitigation while maintaining 96.4% of original model performance on legitimate tasks.

E. Real-Time Monitoring Systems

Our monitoring framework provides continuous assessment of VLM deployments to detect ongoing steganographic attacks and enable rapid response.

Behavioral Anomaly Detection: We implement real-time monitoring of VLM outputs to detect patterns consistent with prompt injection attacks. Our system analyzes: (1) semantic consistency between inputs and outputs, (2) safety violation patterns in generated content, and (3) unusual instruction-following behavior indicative of embedded commands.

Statistical Process Control: Our monitoring system applies statistical process control techniques to track VLM behavior over time, identifying drift patterns that might indicate ongoing attacks. Control charts monitor output characteristics, response patterns, and error rates to detect systematic changes suggestive of compromise.

Threat Intelligence Integration: We develop threat intelligence capabilities that track emerging steganographic techniques and update detection mechanisms accordingly. This includes: (1) automated analysis of new attack patterns, (2) signature updates for known steganographic techniques, and (3) collaborative threat sharing across VLM deployments.

F. Defense Effectiveness Evaluation

Comprehensive evaluation of our defense mechanisms demonstrates significant protection against steganographic prompt injection while maintaining practical deployment viability.

Layered Defense Performance Analysis:

Individual Layer Effectiveness:

- Preprocessing Layer: 23.7% attack reduction (95% CI: 19.2%–28.1%)
- Statistical Analysis: 18.9% reduction (95% CI: 14.8%–23.0%)
- Neural Detection: 32.1% reduction (95% CI: 27.3%–36.9%)
- Behavioral Monitoring: 28.4% reduction (95% CI: 23.7%–33.1%)

Combined Effectiveness: Layers exhibit subadditive interaction effects. Mathematical modeling indicates:

$$\text{Combined\_Effectiveness} = 1 - \left[\prod_i (1 - \text{Individual\_Effectiveness}_i)\right] \times \text{Interaction\_Factor}$$

where Interaction_Factor $= 0.85$, yielding 73.4% total mitigation.

Statistical Validation: McNemar's test confirms significant improvement over individual layers ($p < 0.001$).

False Positive Analysis: Evaluation against legitimate image datasets reveals manageable false positive rates: 4.7% for preprocessing triggers, 3.2% for statistical detection, 7.8% for neural detection, and 2.1% for behavioral monitoring. While these rates require operational consideration, they remain within acceptable bounds for security-conscious deployments where some false alarms are tolerable to maintain protection.

Performance Impact Assessment: Our defense mechanisms introduce measurable but acceptable performance overhead: 28ms average processing delay per image, 12.3% increase in computational requirements, and minimal impact on VLM accuracy for legitimate tasks (1.4% reduction in standard benchmarks). These costs represent practical trade-offs between security enhancement and operational efficiency.

G. Adaptive Defense Strategies

We develop adaptive defense mechanisms that evolve in response to emerging attack techniques, providing sustained protection against evolving threats.

Machine Learning Defense Updates: Our detection systems incorporate continuous learning capabilities that adapt to new steganographic techniques with moderate effectiveness. The system maintains detection accuracy above 62% even against novel attack variants by: (1) automated retraining on detected attack samples, (2) transfer learning from related attack patterns, and (3) ensemble updating that incorporates new detection models, though the arms race between embedding and detection techniques remains ongoing.

Dynamic Threshold Adjustment: Our defense framework automatically adjusts detection thresholds based on observed attack patterns and false positive rates. This adaptive approach maintains reasonable balance between protection effectiveness and operational usability as threat landscapes evolve, though perfect optimization remains challenging due to the diverse nature of steganographic threats.

Collaborative Defense Networks: We propose collaborative defense architectures where multiple VLM deployments share threat intelligence and detection capabilities. This network effect provides measurable amplification of defense effectiveness by: (1) rapid propagation of new attack signatures, (2) collective learning from attack attempts, and (3) coordinated response to large-scale campaigns, though coordination overhead and privacy concerns limit practical implementation scope.

H. Deployment Considerations

Practical deployment of defense mechanisms requires careful consideration of operational constraints, performance requirements, and integration challenges.

Integration Complexity: Our defense framework is designed for modular integration with existing VLM deployments. Each defense layer can be deployed independently, allowing organizations to implement protection incrementally based on risk assessment and resource availability.

Cost-Benefit Analysis:

Implementation Cost Breakdown:

- Software development: $25,000–$45,000
- Integration and testing: $8,000–$15,000
- Training and deployment: $5,000–$10,000
- Annual maintenance: $3,000–$8,000

Breach Cost Estimation: Based on IBM Security Cost of Data Breach Report 2024 and AI-specific incident analyses:

- Average AI system breach: $2.3M (range: $800K–$5.2M)
- Reputation damage: $1.1M additional cost
- Regulatory penalties: $200K–$2M (GDPR/CCPA)

ROI Calculation: Break-even analysis shows positive ROI within 18 months for organizations processing $> 10,000$ images daily.

Regulatory Compliance: Our defense mechanisms support compliance with emerging AI safety regulations and industry standards. The framework provides audit trails, explainable detection decisions, and configurable protection levels aligned with regulatory requirements.

IX. Discussion

A. Implications for VLM Security

Our findings reveal meaningful but constrained vulnerabilities in current vision-language model architectures that require careful consideration within broader security frameworks. The moderate success rates of steganographic prompt injection attacks (24.3% overall) indicate that while these threats are real and warrant attention, they represent one component of a larger attack landscape rather than a fundamental system compromise.

Architectural Vulnerabilities: The limited but consistent transferability of attacks across different VLM architectures (8.7–16.4% success rates) suggests that shared components, particularly vision encoders based on CLIP [21], introduce systematic vulnerabilities that merit architectural consideration. However, the substantial reduction in effectiveness compared to targeted attacks demonstrates that current diversity in model design provides meaningful security benefits.

Real-World Impact: The demonstrated effectiveness of our attacks under controlled conditions, combined with recent evidence of prompt injection vulnerabilities in medical [11] and surgical [12] applications, highlights the need for proportionate security measures. The capacity constraints and quality trade-offs inherent in steganographic embedding limit the practical scope of such attacks while still requiring defensive consideration for high-security applications.

B. Limitations and Future Work

Attack Sophistication Requirements: Our framework demonstrates that effective steganographic prompt injection requires sophisticated understanding of both steganographic techniques and target model architectures, with success rates that reflect the fundamental challenges of embedding semantic content within visual media. The technical barriers and limited success rates suggest that such attacks may be primarily relevant for well-resourced adversaries rather than widespread exploitation.

Capacity-Quality Trade-offs: A fundamental limitation revealed by our analysis is the inverse relationship between steganographic capacity and visual quality. Embedding longer prompts (> 15 tokens) results in rapidly degrading attack success rates and increased detectability, constraining the practical utility of such approaches for complex instruction injection.

Defense Evolution: Our proposed defense mechanisms represent initial steps toward comprehensive protection, achieving 73.4% mitigation with acceptable operational overhead. The moderate but meaningful effectiveness of these defenses suggests that practical protection is achievable, though the ongoing arms race between steganographic techniques and detection methods requires continued research and adaptation.

Ethical Considerations: The disclosure of these vulnerabilities raises important questions about responsible research in adversarial machine learning. While our work demonstrates real security concerns, the moderate success rates and

significant technical barriers to implementation suggest that disclosure serves educational and defensive purposes without enabling widespread malicious exploitation.

C. Broader Security Implications

The moderate success of steganographic prompt injection attacks against VLMs contributes to our understanding of multimodal AI security while highlighting the importance of layered defense strategies. As these systems become more prevalent in critical applications, our findings support the need for proportionate security measures that balance protection against demonstrated threats with operational requirements.

Cross-Domain Vulnerability Assessment: While our techniques show limited but meaningful effectiveness against VLMs, the transferability patterns suggest that other multimodal AI systems may exhibit similar vulnerabilities. However, the constrained success rates and capacity limitations indicate that such attacks represent one element of threat landscapes rather than dominant attack vectors.

Regulatory and Policy Implications: Our findings support the development of risk-proportionate AI safety regulations that address demonstrated vulnerabilities without imposing excessive constraints based on theoretical threats. The moderate success rates and technical complexity of steganographic prompt injection suggest that regulatory frameworks should consider such attacks within broader security assessment protocols rather than as primary threat vectors.

X. Conclusion

This work presents the first comprehensive study of steganographic prompt injection attacks against vision-language models, revealing moderate but meaningful vulnerabilities in current multimodal AI architectures. Our multi-domain embedding framework achieves attack success rates of up to 31.8% while maintaining reasonable visual imperceptibility (PSNR > 38 dB, SSIM > 0.94), demonstrating that sophisticated adversaries can exploit VLMs through carefully crafted modifications to input images, though the success rates reflect the inherent challenges of steganographic prompt embedding.

Key Findings: Our experimental evaluation across eight state-of-the-art VLMs reveals that both commercial and open-source models exhibit vulnerabilities to steganographic prompt injection, with open-source models showing higher susceptibility (25–37% vs. 14–18% for commercial models). The attacks demonstrate moderate resilience across diverse image processing operations, though significant degradation occurs under real-world processing conditions, confirming both the viability and limitations of such approaches.

Defense Mechanisms: Our proposed multi-layer defense framework achieves 73.4% attack mitigation when fully deployed, though this requires acceptable trade-offs in terms of performance overhead (28ms processing delay, 12.3% computational increase) and modest false positive rates (2–8% across components). The framework's modular design allows for risk-appropriate deployment based on operational requirements and threat assessments.

Practical Implications: The moderate success rates and significant technical requirements for effective steganographic prompt injection suggest that such attacks represent a meaningful but constrained threat vector. The capacity limitations (optimal performance with prompts ≤ 15 tokens) and quality trade-offs inherent in steganographic embedding limit the scope of practical attacks while still warranting defensive consideration for security-critical applications.

Future Research Directions:

1) **Technical Advances:**
   - Adaptive Steganography: Develop methods that adjust to real-time defense updates
   - Cross-Modal Attacks: Investigate audio-visual steganographic injection
   - Federated Attack Scenarios: Explore coordinated attacks across multiple VLM instances
2) **Evaluation Improvements:**
   - Longitudinal Studies: Track attack effectiveness over extended deployment periods
   - Expert Perceptual Studies: Include forensic analysts and security experts
   - Ecological Validity: Test attacks in production-like environments
3) **Defense Research:**
   - Proactive Defense: Develop predictive models for emerging steganographic techniques
   - Differential Privacy: Investigate privacy-preserving defense mechanisms
   - Adversarial Training: Systematic study of steganography-aware VLM training

As vision-language models become increasingly prevalent in critical applications, the security vulnerabilities demonstrated in this work represent a component of the broader threat landscape that requires proportionate attention from the AI research community, industry practitioners, and policymakers. The development of robust, secure multimodal AI systems will require sustained effort across technical, operational, and regulatory dimensions, with our findings contributing to the understanding of specific vulnerability classes within this larger security ecosystem.

References

[1] Liu, D., et al. "A Survey of Attacks on Large Vision-Language Models: Resources, Advances, and Future Trends" arXiv preprint arXiv:2407.07403 (2024).

[2] Ye, M., et al. "A Survey of Safety on Large Vision-Language Models: Attacks, Defenses and Evaluations" arXiv preprint arXiv:2502.14881 (2025).

[3] Li, B., et al "Otter: a multi-modal model with in-context instruction tuning" arXiv preprint arXiv:2305.03726 (2023).

[4] Chen, D., et al. "Visual Instruction Tuning with Polite Flamingo" arXiv preprint arXiv:2307.01003 (2023).

[5] Hossain, M., et al. "Securing Vision-Language Models with a Robust Encoder Against Jailbreak and Adversarial Attacks" arXiv preprint arXiv:2409.07353 (2024).

[6] Kapoor, S., et al. "Adversarial Attacks in Multimodal Systems: A Practitioner's Survey" arXiv preprint arXiv:2505.03084 (2025).

[7] Liu, Y., et al. "Prompt Injection attack against LLM-integrated Applications" arXiv preprint arXiv:2306.05499 (2023).

[8] Rossi, S., et al. "An early categorization of prompt injection attacks on large language models" arXiv preprint arXiv:2402.00898 (2024).

[9] Chen Henry Wu, et al. "Adversarial Attacks on Multimodal Agents" arXiv preprint arXiv:2406.12814 (2024).

[10] Maan Qraitem, et al. "Web Artifact Attacks Disrupt Vision Language Models" arXiv preprint arXiv:2503.13652 (2025).

[11] Clusmann, J. et al. "Prompt injection attacks on vision language models in oncology" Nature Communications, 16, 1239 (2025).

[12] Zhang, Z., et al. "Prompt injection attacks on vision-language models for surgical decision support" medRxiv (2025).

[13] Yin, Z., et al. "VLATTACK: Multimodal Adversarial Attacks on Vision-Language Tasks via Pre-trained Models" arXiv preprint arXiv:2310.04655 (2023).

[14] Dou, Z., et al. "Adversarial Attacks to Multi-Modal Models" arXiv preprint arXiv:2409.06793 (2024).

[15] Mathew, Y., et al. "Hidden in Plain Text: Emergence & Mitigation of Steganographic Collusion in LLMs" arXiv preprint arXiv:2410.03768 (2024).

[16] Abdali, S., et al. "Securing Large Language Models: Threats, Vulnerabilities and Responsible Practices" arXiv preprint arXiv:2403.12503 (2024).

[17] Lee, S. et al. "Mind Mapping Prompt Injection: Visual Prompt Injection Attacks in Modern Large Language Models" Electronics, 14(10), 1907 (2024).

[18] Scientific Reports. "A deep learning-driven multi-layered steganographic approach for enhanced data security" Scientific Reports (2025).

[19] Yu, J., et al. "Cross: Diffusion model makes controllable, robust and secure image steganography" Advances in Neural Information Processing Systems (2024).

[20] Debenedetti, E., et al. "Defeating Prompt Injections by Design" arXiv preprint arXiv:2503.18813 (2025).

[21] Radford, A., et al. "Learning Transferable Visual Models From Natural Language Supervision" arXiv preprint arXiv:2103.00020 (2021).

[22] Liu, H., et al. "Visual Instruction Tuning" arXiv preprint arXiv:2304.08485 (2023).

[23] Li, J., et al. "BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models" arXiv preprint arXiv:2301.12597 (2023).

[24] Zhou, W., et al. "Revisiting the Adversarial Robustness of Vision Language Models: a Multimodal Perspective" arXiv preprint arXiv:2404.19287 (2024).

[25] OWASP GenAI Security Project. "LLM01:2025 Prompt Injection" Retrieved from https://genai.owasp.org/llmrisk/llm01-prompt-injection/ (2025).

[26] Kim, D., et al. "Text-Based Prompt Injection Attack Using Mathematical Functions in Modern Large Language Models" Electronics, 13(24), 5008 (2024).

[27] Sun, J., et al. "Safeguarding Vision-Language Models Against Patched Visual Prompt Injectors" arXiv preprint arXiv:2405.10529 (2024).

[28] Kimura, S., et al. "Empirical analysis of large vision-language models against goal hijacking via visual prompt injection" arXiv preprint arXiv:2408.03554 (2024).

[29] Fridrich, J., et al. "Detecting LSB Steganography in Color and Gray-Scale Images" IEEE Multimedia (2001).

[30] Apau, R., et al. "Image steganography techniques for resisting statistical steganalysis attacks: A systematic literature review" PLOS One, 19(9), e0308807 (2024).

[31] Priya, S., Abirami, S.P., Arunkumar, B., et al. "Super-resolution deep neural network (SRDNN) based multi-image steganography for highly secured lossless image transmission" Scientific Reports, 14, 6104 (2024).

[32] Ntivuguruzwa Jean De La Croix, et al. "Comprehensive survey on image steganalysis using deep learning" Neural Computing and Applications (2024).

[33] Oleksandr, K., et al. "Enhancing Steganography Detection with AI: Fine-Tuning a Deep Residual Network for Spread Spectrum Image Steganography" PMC (2024).

[34] Yuanyuan Ma, et al. "Digital image steganalysis network strengthening framework based on evolutionary algorithm" Scientific Reports (2025).

[35] Dosovitskiy, A., et al. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale" ICLR (2021).

[36] Liu, H., et al. "Improved Baselines with Visual Instruction Tuning" arXiv preprint arXiv:2310.03744 (2023).

[37] "Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90% Chat-GPT Quality" (2023).

[38] Li, J., et al. "BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation" arXiv preprint arXiv:2201.12086 (2022).

[39] Settiadi, D.R.I.M. "PSNR vs SSIM: imperceptibility quality assessment for image steganography" Multimedia Tools and Applications, 80, 8423-8444 (2021).

[40] Chan, C., et al. "Hiding data in images by simple LSB substitution" Pattern Recognition (2004).

[41] Westfeld, A., et al. "Attacks on steganographic systems" Information Hiding Workshop (1999).

[42] Cox, I., et al. "Secure spread spectrum watermarking for multimedia" IEEE Transactions on Image Processing (1997).

[43] Walia, E., et al. "An Analysis of LSB & DCT based Steganography".

[44] Havard, A., et al. "CNN-Assisted Steganography – Integrating Machine Learning with Established Steganographic Techniques" arXiv preprint arXiv:2304.12503 (2023).

[45] Baluja, S. "Hiding images in plain sight: Deep steganography" Advances in Neural Information Processing Systems (2017).

[46] Greshake, K., et al. "Not what you've signed up for: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection" arXiv preprint arxiv:2302.12173 (2023).

[47] Wikipedia Contributors. "Peak signal-to-noise ratio" Wikipedia. Retrieved from https://en.wikipedia.org/wiki/Peak_signal-to-noise_ratio (2025).

[48] Toyer, S., et al. "Tensor Trust: Interpretable Prompt Injection Attacks from an Online Game" arXiv preprint (2023).

[49] Yao, Y., et al. "A survey on large language model (llm) security and privacy: The good, the bad, and the ugly" High-Confidence Computing, 4, 100211 (2024).

[50] Johnson, N., et al. "Exploring steganography: Seeing the unseen" Computer (2008).

[51] Pevný, T., et al. "Steganalysis by subtractive pixel adjacency matrix" IEEE Transactions on Information Forensics and Security.

[52] Liu, F., wt al. "Mitigating hallucination in large multi-modal models via robust instruction tuning" arXiv preprint arXiv:2306.14565 (2023).

[53] Qi, X., et al. "Visual Adversarial Examples Jailbreak Aligned Large Language Models" AAAI Conference on Artificial Intelligence (2023).

[54] Schlarmann, C., et al. "Robust CLIP: Unsupervised Adversarial Fine-Tuning of Vision Embeddings for Robust Large Vision-Language Models" arXiv preprint arXiv:2402.12336 (2024).

[55] Xuanming C., et al. "On the Robustness of Large Multimodal Models Against Image Adversarial Attacks" arXiv preprint arXiv:2312.03777 (2023).

[56] Liu, X., et al. "Jailbreak Attacks and Defenses against Multimodal Generative Models: A Survey" arXiv preprint arXiv:2411.09259 (2024).

[57] Huang, L., et al. "Image-based Multimodal Models as Intruders: Transferable Multimodal Attacks on Video-based MLLMs" arXiv preprint arXiv:2501.01042 (2025).