Does the Source of Carrier Image Affect Steganographic Detectability?

A Comparative Study of Real vs. ML-Generated
Image Steganography

Project Proposal — Draft for Team Review

Nico, Nikolas, Abdul, Daria, Jimena, David

Department of Advanced Computing Sciences (DACS)
Maastricht University · February 2026

1. Introduction and Motivation

Image steganography—the practice of concealing secret data within digital images—has been studied extensively over the past two decades through techniques spanning Least Significant Bit (LSB) substitution in the spatial domain, Discrete Cosine Transform (DCT) embedding, DWT-based methods, and content-adaptive approaches [? ? ?]. The security of these methods depends heavily on the statistical properties of the carrier image: embedding introduces subtle distributional shifts that steganalysis classifiers exploit for detection [?].

Simultaneously, a revolution has occurred in image generation. Models such as Stable Diffusion [?], StyleGAN3 [?], DALL-E 3, and Midjourney now produce images that are perceptually indistinguishable from real photographs. Each generative paradigm—latent diffusion, GAN-based adversarial training, transformer-based token prediction—imposes distinct statistical fingerprints on its output. These fingerprints are already exploited by image deepfake and forgery detectors [? ?], but their interaction with steganographic embedding is **unexplored**.

This gap is consequential. If ML-generated images prove harder to steganalyze, adversaries could exploit synthetic carriers to evade detection—a concern for law enforcement and digital forensics. Conversely, generative artifacts might make embedded content easier to detect. Some preliminary work [?] has begun examining AI-generated images as steganographic carriers, but no systematic, controlled study comparing real vs. ML-generated images under identical embedding conditions has been conducted. This proposal aims to fill that gap.

2. Research Questions

Primary Research Question (RQ):

> *Does the source of the carrier image (real human-photographed vs. ML-generated) affect the detectability of image steganography when using identical embedding methods and payload sizes?*

Secondary Research Questions:

RQ1 (Payload Sensitivity)
How does payload size influence detectability across real and ML-generated images? Do the two carrier types diverge at different embedding rates?

RQ2 (Embedding Method)
Do spatial-domain (LSB) and frequency-domain (DCT) embedding behave differently depending on the carrier image's origin?

RQ3 (Encryption Effect)
Does encrypting the payload before embedding affect detectability, and does this interaction vary across carrier types?

RQ4 (Cross-Domain Generalization)
How well do steganalysis detectors trained on one domain (real or ML-generated images) generalize to the other?

3. Background and Related Work

3.1 Image Steganography Techniques

Image steganography methods fall into spatial-domain and frequency-domain categories. **LSB substitution** replaces the least significant bits of pixel channel values with message bits—high capacity but vulnerable to statistical attacks [?]. **DCT-based embedding** modifies discrete cosine transform coefficients of 8×8 image blocks, mirroring JPEG-domain steganography (F5/JSteg) [?], using Quantization Index Modulation (QIM) [?] for coefficient selection. Content-adaptive methods (WOW, HILL, S-UNIWARD) additionally minimize statistical distortion by targeting textured image regions.

3.2 Image Steganalysis

Training-free statistical methods exploit distributional properties introduced by embedding. The **chi-square attack** [?] tests whether pixel value pairs $(2k, 2k{+}1)$ have equalised frequencies—the statistical signature of LSB substitution. **RS Analysis** [?] analyses pixel group regularity to estimate the embedding rate $\hat{p}$ analytically, requiring no training data and generalizing across image domains by construction.

Classical ML approaches extract hand-crafted features and classify with statistical methods. Fridrich and Kodovský [?] introduced the **Spatial Rich Model (SRM)**, which extracts high-pass residual co-occurrence features (∼35,000 dimensions) and classifies with an ensemble of Fisher Linear Discriminants—a strong CPU-only baseline that handles both LSB and DCT embedding. Deep learning approaches (CNN-based and residual network architectures) have since advanced steganalysis further [?], but fall outside the scope of this cryptography-focused study.

3.3 ML-Generated Images

Modern image generation has achieved remarkable quality across several paradigms. **Latent diffusion models:** Stable Diffusion [?] applies diffusion in a compressed latent space, producing highly photorealistic images from text prompts. **GAN-based models:** StyleGAN3 [?] generates high-resolution images with well-characterized spectral artifacts—GAN fingerprints already exploited by deepfake detectors [? ?]. Each generative paradigm imposes distinct statistical signatures that may interact with steganographic embedding differently.

3.4 Steganography in AI-Generated Media

De et al. [?] investigated steganographic secret sharing via AI-generated photorealistic images, finding that minimum-entropy coupling can achieve statistically undetectable embedding. This is the closest work to our proposal, but does not systematically compare real vs. ML-generated images under controlled, identical embedding conditions—the gap our study directly addresses. Deepfake detection research confirms that ML-generated images have exploitable statistical differences from real photographs [? ?], supporting our hypothesis that the domain boundary will affect steganalysis performance.

3.5 Cross-Domain Generalization

Cross-domain generalization is a well-documented challenge: steganalysis models trained on one domain suffer significant performance degradation when tested on another [?]. The specific domain shift from real photographs to ML-generated images has not been studied in the steganalysis context—a gap our study directly addresses through conditions C, D, and E.

4. Experimental Design

4.1 Overview

We employ a 2 × 2 × 3 × 2 **factorial design** with four factors: carrier source (real vs. ML-generated), embedding method (LSB vs. DCT), payload rate (three levels), and detector (RS Analysis vs. SRM). For each combination, we produce matched cover–stego pairs and evaluate detectability across five training–testing conditions.

4.2 Dataset Construction

Real images. We sample **500 images** from established photographic datasets: **RAISE** [?] (250 images, high-quality RAW DSLR), **COCO** [?] (150 images, natural photographs), and **Flickr30k** (100 images, diverse everyday scenes). All images normalized to 512×512 px, RGB, 8-bit, lossless PNG. A BRISQUE quality gate (≤50) ensures perceptual consistency.

ML-generated images. We generate a matched set of **500 images** using two representative models:

- **Stable Diffusion v2.1** [?]—latent diffusion, photorealistic text-to-image, run locally via `diffusers` (MPS).
- **StyleGAN3** [?]—alias-free GAN with well-characterized spectral artifacts.

Each model contributes 250 images. ML-generated images use the same semantic prompts as the real image content (e.g., "a dog in a park") and are normalized to identical format specifications. The 500+500 dataset improves statistical power for cross-domain experiments.

4.3 Embedding Methods and Payload Rates

LSB substitution (spatial domain): Replace the k least significant bits of each 8-bit pixel channel value with pseudorandom message bits. Pixel selection uses a PRNG keyed by a shared secret, applied across all RGB channels.

DCT-based embedding (frequency domain): Segment each channel into non-overlapping 8×8 pixel blocks, compute the 2D DCT, and embed bits by quantizing selected mid-frequency coefficients using QIM [?]. Optionally, the payload is pre-encrypted with AES-256-CBC (addressing RQ3).

Table 1: Payload rate levels for both embedding methods.

Level	LSB config	DCT config
Low	$k=1$, 25% pixels	10% coefficients
Medium	$k=1$, 50% pixels	25% coefficients
High	$k=2$, 50% pixels	50% coefficients

4.4 Steganalysis Detectors

We use two detectors chosen to stay within the scope of classical signal processing and statistics, avoiding deep learning:

1. **RS Analysis** [?]: A fully training-free statistical test. Analyses pixel group regularity to estimate the embedding rate $\hat{p}$ analytically. Requires no training data and generalizes across image domains by construction—any difference in detection rate between real and ML-generated images reflects only the carriers' statistical properties, not classifier bias.
2. **SRM + FLD Ensemble** [?]: Spatial Rich Model. Extracts high-pass residual co-occurrence features (~35,000 dimensions) and classifies with an ensemble of Fisher Linear Discriminants—a classical statistical classifier, not a neural network. CPU-only. Included because it handles DCT-domain embedding better than training-free methods and its hand-crafted features are expected to generalize across the real/ML domain boundary.

The chi-square attack [?] is applied as a supplementary analytical check on LSB results. SRM is trained with 3-fold cross-validation; training-free detectors produce a continuous score directly.

4.5 Training–Testing Conditions

Table 2: Cross-domain experimental conditions.

	Train	Test	Purpose
A	Real	Real	Baseline
B	ML-gen	ML-gen	Within-domain ML
C	Real	ML-gen	Real→ML transfer
D	ML-gen	Real	ML→Real transfer
E	Mixed	Both	Domain-agnostic

4.6 Evaluation Metrics

Detection: ROC-AUC (primary), accuracy at optimal threshold (Youden's J), Equal Error Rate (EER), and

FPR at 5% FNR. **Image quality:** PSNR, SSIM, and FSIM. **Payload integrity:** Bit Error Rate (BER); target = 0 for lossless PNG. **Statistics:** Two-way ANOVA (carrier × method) with payload as covariate; Wilcoxon signed-rank tests for pairwise comparisons; Bonferroni correction; Cohen's *d* effect sizes.

5. Hypotheses and Expected Results

H1 (Distributional Difference)
ML-generated images will exhibit different steganalysis detectability due to learned statistical regularities (GAN spectral peaks, diffusion noise patterns, smoother textures) that interact with embedding distortion differently than natural photographs.

H2 (Payload Divergence)
Detectability divergence between carrier types will increase with payload size; at low embedding rates, both may be similarly difficult to steganalyze.

H3 (Method Sensitivity)
DCT embedding will show greater sensitivity to carrier origin than LSB, because DCT coefficients are more directly shaped by the generative process.

H4 (Cross-Domain Drop)
Detectors will suffer 10–25% AUC degradation in cross-domain conditions (C, D) versus within-domain (A, B), paralleling findings in image deepfake detection [**?**].

H5 (Asymmetric Transfer)
Real→ML transfer (C) will perform worse than ML→Real (D), because ML images' distinctive statistical profile may mask or alter the embedding artifacts the detector learned from natural images.

H6 (Encryption Effect)
Payload encryption will not significantly affect detectability, since AES output is statistically similar to a pseudorandom bitstream. Any measurable difference would indicate payload structure contributes to detection.

6. Timeline and Feasibility

The project spans **7 weeks** using team members' M4 Pro MacBooks for local model inference and CPU-based detection.

Table 3: Project timeline (7 weeks).

Week	Activity
1	Dataset collection (RAISE/COCO/Flickr30k); ML image generation (SD v2.1 + StyleGAN3)
2	LSB + DCT implementation; embedding at all payload rates ± AES-256; quality checks
3–4	RS Analysis + SRM detection (CPU); conditions A–E
5	Cross-domain experiments; statistical analysis
6	Paper writing, visualization, revision
7	Buffer / final revision / submission

Feasibility notes. Stable Diffusion runs locally via `diffusers` on MPS (approx. 3–5 h for 250 images); StyleGAN3 via NVIDIA's PyTorch implementation (2–3 h for 250 images). RS Analysis and chi-square attack require no training—seconds per image with NumPy. SRM + FLD ensemble (scikit-learn) runs in under 45 minutes total on CPU for 1,000 images. **Total detection compute is under 6 hours**, compared to days with neural network approaches. All tools are open-source.

7. Ethical Considerations

No human subjects are involved. Real image datasets (RAISE, COCO, Flickr30k) are publicly available under permissive research licenses. ML-generated images are produced using open-source models (Stable Diffusion) and publicly available architectures (StyleGAN3). Our contribution is analytical—we study detection of known steganographic techniques rather than developing new evasion methods. Dual-use implications (potential exploitation of ML-generated carriers) will be discussed explicitly in the paper.

References

[1] F. A. P. Petitcolas, R. J. Anderson, and M. G. Kuhn, "Information hiding—a survey," *Proc. IEEE*, vol. 87, no. 7, pp. 1062–1078, 1999.

[2] A. Cheddad, J. Condell, K. Curran, and P. McKevitt, "Digital image steganography: Survey and analysis of current methods," *Signal Process.*, vol. 90, no. 3, pp. 727–752, 2010.

[3] M. Hussain, A. W. A. Wahab, Y. I. B. Idris, A. T. S. Ho, and K. H. Jung, "Image steganography in spatial domain: A survey," *Signal Process.: Image Commun.*, vol. 65, pp. 46–66, 2018.

[4] J. Fridrich and J. Kodovský, "Rich models for steganalysis of digital images," *IEEE Trans. Inf. Forensics Security*, vol. 7, no. 3, pp. 868–882, 2012.

[5] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proc. IEEE CVPR*, pp. 10684–10695, 2022.

[6] T. Karras, M. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Alias-free generative adversarial networks," in *Proc. NeurIPS*, vol. 34, pp. 852–863, 2021.

[7] A. De, W. Kinzel, and I. Kanter, "Steganographic secret sharing via AI-generated photorealistic images," *EURASIP J. Wireless Commun. Netw.*, art. 108, 2022.

[8] J. Fridrich, M. Goljan, and R. Du, "Detecting LSB steganography in color and grayscale images," *IEEE Multimedia*, vol. 8, no. 4, pp. 22–28, 2001.

[9] A. Westfeld and A. Pfitzmann, "Attacks on steganographic systems," in *Proc. 3rd Int. Workshop Information Hiding*, LNCS 1768, pp. 61–76, 1999.

[10] N. Provos and P. Honeyman, "Hide and seek: An introduction to steganography," *IEEE Security Privacy*, vol. 1, no. 3, pp. 32–44, 2003.

[11] B. Chen and G. W. Wornell, "Quantization index modulation: A class of provably good methods for digital watermarking and information embedding," *IEEE Trans. Inf. Theory*, vol. 47, no. 4, pp. 1423–1443, 2001.

[12] S. Y. Wang, O. Wang, R. Zhang, A. Owens, and A. A. Efros, "CNN-generated images are surprisingly easy to spot. . . for now," in *Proc. IEEE CVPR*, pp. 8695–8704, 2020.

[13] R. Corvi, D. Cozzolino, G. Zingarini, G. Poggi, K. Nagano, and L. Verdoliva, “On the detection of synthetic images generated by diffusion models,” in *Proc. IEEE ICASSP*, pp. 1–5, 2023.

[14] D.-T. Dang-Nguyen, C. Pasquini, V. Conotter, and G. Boato, “RAISE: A raw images dataset for digital image forensics,” in *Proc. ACM MMSys*, pp. 219–224, 2015.

[15] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, “Microsoft COCO: Common objects in context,” in *Proc. ECCV*, LNCS 8693, pp. 740–755, 2014.

[16] Y. Luo et al., “Deep learning for steganalysis of diverse data types: A review of methods, taxonomy, challenges and future directions,” *Neurocomputing*, Elsevier, 2024.