Four Lousy Ghosts: Novel and Established Counter-Steganalysis Techniques vs ML-Capable Adversaries — To No Avail

Dylan Hart `(exec)`
Claude (Anthropic, `claude-opus-4-7`)

2026-05-02

Abstract

We extend the spatial-domain rounding-error-fingerprint observation underlying the Reverse JPEG Compatibility Attack [19] to QF=85 covers via attacker-aware CNN steganalysis, by evaluating four cost-map-based defenses for content-adaptive JPEG steganography in the *phantasm* project. Three are novel — **HYDRA** (ensemble-saliency-source assignment), **CHAMELEON** (stochastic per-stego wet-fraction), and **DOPPELGÄNGER** (variance-shaped costs from natural-cover statistics) — and the fourth, **PALIMPSEST**, is a parity-conditioned soft-penalty approximation of the heuristic side-informed scheme of Denemark and Fridrich [1], necessitated by phantasm's single-layer-binary STC pipeline. It is *not* the modern model-preserving / covariance-aware side-informed framing of Butora and Bas [2], which derives the embedding from the covariance between rounding error and embedding changes and which we have not tested.

All four tested defenses are refuted against an in-distribution EfficientNet-B0 adversary trained from ImageNet initialization on multi-passphrase phantasm output ($N \approx 2000$ cover/stego pairs, 15 epochs), with measured AUC ranging from 0.76 to 1.00. The strongest hypothesis (DOPPELGÄNGER) was the *most* catastrophically refuted: it rendered stegos *more* detectable than the unmodified J-UNIWARD baseline — even an *un-adapted* baseline detector that never saw DOPPELGÄNGER stegos during training detects them at AUC 0.992–0.994.

The simplest explanation across these four refutations is an **operator-level fingerprint** in the same family as RJCA's — the statistical artifact of decoding, modifying integer DCT coefficients, and re-encoding a JPEG — invariant to the cost-map manipulations the four constructions we test introduce. The *constructions* we refute are three WHERE-randomization variants (HYDRA, CHAMELEON, DOPPELGÄNGER) and one parity-conditioned approximation of heuristic HOW-with-rounding (the PALIMPSEST scheme); we are careful to distinguish "these four constructions hit a ceiling" (the empirical claim) from "the entire WHERE/HOW design space hits a ceiling" (the broader hypothesis those data points support but do not prove). The covariance-preserving side-informed family (Butora & Bas 2023) is **explicitly scoped out of the refutation** as the single most empirically tractable open question this paper leaves; cross-corpus generalization (ALASKA#2 / BOSSbase covers under our attacker recipe) is the second.

We publish this work as a falsification challenge — a precedent against which any future defense in the cost-map family, any model-preserving side-informed approach, or any operator-changing approach may credibly measure itself. The full reproducibility surface — trained adversary checkpoints, per-experiment branches, eval JSONs, and the `phantasm` v1.0.1 binary — is public.

Contents

1 Introduction

Phantasm [3] is a content-adaptive JPEG steganography tool written in Rust, with three independent layers of security:

- **L1.** Statistical undetectability of the stego from the cover against modern CNN-based steganalyzers.
- **L2.** Recovery of the embedding *positions* from a passphrase-keyed cryptographic state, which an attacker not in possession of the passphrase must brute-force from a 2^{256} space.
- **L3.** Authenticated decryption of the recovered payload through Argon2id-derived AEAD keys (XChaCha20-Poly1305) gated by an HMAC-SHA256-16 over the message, with HKDF [4] independent-extract separating AEAD and MAC keys.

Phantasm v1 explicitly **scopes down its L1 claim**: the tool degrades gracefully against off-the-shelf modern CNN steganalyzers (16.2% detection rate on phantasm stegos against an out-of-distribution JIN-SRNet [5] above a 7.1% cover false-positive floor, with 49.5% of stegos scoring lower $P(\text{stego})$ than their own cover) but **fails** against a phantasm-aware CNN trained on phantasm output (AUC 0.968 at the headline d500 scale [6]). The security argument for phantasm v1 rests on L2 and L3 — passphrase-keyed position recovery and authenticated decryption — not on L1.

This posture was not the project's original intent. It is the result of a four-experiment research arc, conducted in early 2026 between phantasm v0.4 and v1.0, in which we sequentially constructed and refuted four cost-map-based defenses against attacker-aware CNN steganalysis. This paper is the consolidated record of that arc, and of the structural lesson that emerged from it.

The lesson is unwelcome but, given the evidence, sturdy: **the detectable artifact in our regime is the embedding operator itself, not the operator's position distribution or rounding direction.** A CNN trained on phantasm output learns to recognize the statistical trace of "decode JPEG → modify DCT coefficient → re-encode JPEG" — across passphrases, across cost-map variants, across rounding-direction shaping. We refer to this as the **operator fingerprint**, and we argue that any cost-map-based defense in the WHERE / WET / HOW families — defenses that vary *where* or *with what direction* an unchanged operator is applied — must be expected to share the same ceiling.

We publish this work as a **falsification challenge**, with the result we report here as a measurable bar: *any future defense against an attacker-aware CNN steganalyzer in the cost-map family, and any operator-changing defense that genuinely escapes the operator-fingerprint ceiling, has a public, reproducible benchmark to clear — lower AUC than the corresponding phantasm baseline against the in-distribution adversary checkpoints we publish.* We invite that future work explicitly. The artifacts required to falsify our claim are all public (Section 9).

Two scope limits we name upfront so that the rest of the paper can be read against them, not behind them. First, every result reported here is on a single cover source: a 198-cover Picsum corpus (plus a 500-cover scaling extension). Cover-source mismatch is the central methodological concern in modern JPEG steganalysis, and we do not run a cross-corpus evaluation against ALASKA#2 or BOSSbase. Section 4.1 justifies the choice and Section 10 (Q2) lists the cross-corpus re-evaluation as the largest evaluation gap this paper leaves. Second, the attacker is a *vanilla* ImageNet-pretrained EfficientNet-B0 fine-tuned at $N \approx 2000$ multi-passphrase pairs; this is the "minimum credible attacker-aware adversary" from our prior cost-function arc, not a 2024-era

state-of-the-art steganalyzer (no Yousfi-Butora surgical modifications, modest data scale). Section 4.2 catalogues the implications: our DOPPELGÄNGER catastrophic-failure result is robust to attacker strength (it is independently confirmed by an un-adapted baseline detector at AUC 0.992–0.994), but the HYDRA result (AUC 0.76–0.78) is the most attacker-recipe-sensitive of the four and may shift under a stronger adversary.

The remainder of the paper is organized as follows. Section 2 reviews the relevant background. Section 3 introduces the four defenses in detail. Section 4 specifies the methodology. Section 5 reports the quantitative results. Section 6 develops the operator-fingerprint hypothesis and explains why PALIMPSEST is its sharpest confirmation. Section 7 sketches the five-defense roadmap from which two were selected for execution and three deferred, expanded with operator-changing approaches beyond the spec'd roadmap. Section 8 states the precedent-to-break framing as a formal challenge to the field. Section 9 documents the reproducibility surface. Section 10 enumerates seven open questions that constitute a concrete next-experiment list. Section 11 concludes.

2 Background

2.1 Content-adaptive JPEG steganography

We work in the standard *minimum-distortion* model of content-adaptive embedding [7]. A JPEG cover image C has DCT coefficients c_{ij} at non-zero AC positions; a per-coefficient cost $\rho_{ij} \in \mathbb{R}_{\geq 0}$ ranks the *detectability* of a ± 1 modification at position (i, j). The embedder, given a payload of m bits and a cost map ρ, produces a stego image S that encodes the payload while minimizing $\sum_{ij} \rho_{ij} \cdot |s_{ij} - c_{ij}|$. The operator solving this optimization is **Syndrome-Trellis Coding (STC)** [7], using the published DDE-Lab parity-check matrices ($h \in [7, 12]$, $w \in [2, 20]$) at phantasm's default 0.995× bits/L1 conditional-probability double-layer setting.

The cost map itself is computed by **J-UNIWARD** [8] — a wavelet-domain undetectability proxy that assigns low cost to coefficients whose modification produces minimal change in three 2-D wavelet sub-bands, and high cost to coefficients in smooth regions. J-UNIWARD has been the dominant cost function in the JPEG steganalysis literature since 2014.

Phantasm v1 ships J-UNIWARD as the only cost function. v0.x variants that experimented with UERD [9] and S-UNIWARD have been removed.

2.2 CNN steganalysis and the attacker-aware regime

The dominant tool of modern JPEG steganalysis is the deep convolutional neural network. Three architectures are particularly relevant here:

- **SRNet** [10] — a deep residual CNN that learns noise-residual features from spatial-domain JPEG decompressions.
- **JIN-SRNet** [5] — an SRNet variant trained on J-UNIWARD adversaries with multi-payload diversification, currently the strongest publicly released "off-the-shelf" J-UNIWARD detector.
- **EfficientNet-B0** [11] — a compound-scaled ImageNet backbone that, with appropriate input pre-processing, serves as a strong general-purpose JPEG steganalyzer when fine-tuned on attacker-defined stego data.

We distinguish two adversary classes:

1. **Off-the-shelf adversary.** A pre-trained detector (e.g. the DDE Lab JIN-SRNet release) with no exposure to phantasm output during training. This is the *cover-source mismatch* [12] regime.

2. **Attacker-aware adversary.** A detector with white-box access to phantasm's outputs at training time — typically as N cover/stego pairs spanning multiple passphrases, payload sizes, and quality factors. This corresponds to the realistic threat model in which the adversary has obtained phantasm v1 and can produce arbitrary training data.

Phantasm's L1 hardening claim — that v0.4-era defenses would withstand a phantasm-aware CNN — is the claim under test in this paper. Section 4.2 details our adversary recipe.

2.3 The phantasm v0.2–v0.3 cost-function research arc

The four experiments reported here are the second of two research arcs in phantasm's history. The first (v0.2–v0.3) tested whether varying the *cost function* itself — UERD vs. J-UNIWARD vs. cost mixing vs. random subsetting vs. payload truncation — could evade an attacker-aware EfficientNet-B0. It could not. Updates 1 through 8 of the prior research log [13] document the closure of that arc and the multi-config-passphrase attacker recipe used as the starting point for the current arc.

The second arc, reported here, fixes the cost function (J-UNIWARD) and the attacker recipe and varies *the cost map itself* — its distribution over coefficients (HYDRA, CHAMELEON), its statistical shaping by the cover (DOPPELGÄNGER), and its rounding direction (PALIMPSEST).

3 Four defenses

We summarize each defense's *hypothesis*, *construction*, and *expected mechanism by which the hypothesized hardening would manifest as reduced detection AUC against an attacker-aware CNN.* Detailed specs, implementation diffs, and per-experiment narratives are at `archive/research-arc/specs/` on the public phantasm repo, with per-experiment branches preserving full source for each construction.

3.1 HYDRA — ensemble-saliency-source assignment (novel)

Hypothesis. A phantasm-aware CNN trained against a fixed saliency-driven cost-concentration mechanism learns to recognize the specific low-saliency regions that detector ranks as hard-to-detect. If, instead of using a single saliency source, each stego is assigned (deterministically, from a hash of the cover stem and the passphrase) one of *three* steganalyzer architectures whose adversarial saliency maps define the cost map, the attacker must learn the union of three different "hard-to-detect" patterns rather than just one. The per-cover marginal cost map becomes a passphrase-conditioned random variable across architectures, breaking the single-architecture cost-fingerprint.

Construction. Three steganalyzer architectures — XuNet, an ImageNet-pretrained EfficientNet-B0, and a small custom shallow CNN — each compute a per-coefficient adversarial saliency map for the cover (gradient of the detector's stego-class output with respect to the DCT coefficient). For each (cover, passphrase) pair, a hash selects one architecture; its saliency map is converted to a PHCOST v3 sidecar and fed to phantasm via `--cost-function from-sidecar`, with `keep_fraction = 0.30` fixed.

Result. AUC 0.76–0.78 against the in-distribution attacker (checkpoint `hydra_effnet_N2000.pt`). **Refuted.** This is the *least* catastrophic refutation among the four — the ensemble saliency does modestly degrade an attacker-aware detector relative to a single fixed-saliency baseline, with all 9 measured conditions clustering between 0.7643 and 0.7785. But "modestly degrade from ∼0.99 to ∼0.78" is not a defense; it is statistical noise on a fundamentally broken signal.

3.2 CHAMELEON — stochastic per-stego wet-fraction (novel)

Hypothesis. A phantasm-aware CNN may learn the *fraction of non-zero AC coefficients that carry payload* as a feature: phantasm at default settings hits a near-constant ratio across covers. If the wet-fraction is randomized per stego — drawn from a passphrase-keyed distribution rather than fixed — the CNN's "expected payload density" feature is destroyed.

Construction. Per cover, the keep-fraction $k \sim \mathcal{U}(0.05, 0.50)$ is sampled and recorded in a sidecar JSON. The cost-concentration mechanism uses a fixed target steganalyzer (single-config `effnet-pretrained_s1`) to compute the saliency-driven cost map and treats only the lowest-cost k fraction of coefficients as embeddable, so that *which* and *how many* coefficients carry payload varies stego-to-stego.

Result. AUC 0.82–0.93 against the in-distribution attacker (checkpoint `effnet_chameleon_N2000.pt`); mean across 11 cross-test conditions is 0.84. **Refuted.** Worse, an attacker *trained* on CHAMELEON's randomized-keep stegos generalizes more strongly than an attack-mixture baseline trained on fixed-keep stegos: at keep = 0.10 the CHAMELEON-trained attacker reaches AUC 0.929 vs. 0.856 for the mixture baseline. Diversifying the wet-fraction during attacker training *strengthens* the attacker, not weakens the defense. The attacker's learned representation is keep-agnostic at the cover level.

3.3 DOPPELGÄNGER — variance-shaped costs (novel)

Hypothesis. The previous two defenses randomized *which* coefficients are modified by varying the cost map's source. DOPPELGÄNGER asks: what if the cost map *aligns with the cover's natural across-cover variance structure*, so that modifications fall in positions where similar covers naturally vary? If a CNN is trained to discriminate "this image was modified in DCT" from "this image was quantized once and never touched," but the modifications fall exactly in the positions where the cover's nearest natural neighbours also vary, the modifications should look — statistically — like ordinary across-cover variation rather than a foreign-operator fingerprint.

Construction. For each anchor cover, $K = 16$ *doppelgänger* covers are retrieved from a 500-cover Picsum reference pool by cosine similarity over a 73-dimensional feature vector (64-dim DCT energy + 8-dim luma histogram + 1 edge density). The anchor and its 16 doppelgängers are decoded to luma DCT coefficients, center-cropped to a common block grid, and stacked. Per-coefficient sample standard deviation σ_{ij} is computed across the $K + 1 = 17$ covers. The cost map is constructed as $\rho_{ij} = c_0 / \max(\sigma_{ij}, \sigma_{\text{floor}})$, with c_0 calibrated per cover so that the mean finite cost matches the mean finite J-UNIWARD base cost (preserving the perceptual budget). The DC coefficient is wet-paper-forbidden. The resulting sidecar is fed to phantasm via `--cost-function from-sidecar`.

Result. AUC 0.9998–1.000 against the in-distribution attacker (checkpoint `effnet_doppelganger_N2000.pt`). **Refuted catastrophically.** Two facts make this the worst result in the arc rather than just a failure:

- DOPPELGÄNGER stegos are detected at AUC 0.992–0.994 by an *unrelated, un-adapted* `s1` baseline detector that never saw a DOPPELGÄNGER stego during training. The defense made phantasm stegos *easier* to find with a generic detector than ordinary phantasm stegos are.
- Cross-test against a CHAMELEON-trained detector reaches AUC 0.985–0.987; cross-test against an attack-mixture detector reaches 0.94–0.96. There is no detector class against which the defense produces below-baseline detectability.

The variance-shaping concentrates modifications in high-variance DCT positions which, post-IDCT, fall in exactly the spatial regions where a CNN's learned residual filters most reliably discriminate "JPEG modified after quantization" from "JPEG quantized once and never touched." The $K = 16$ retrieval also introduces a *second* fingerprint on top of the operator fingerprint: the per-anchor doppelgänger-pool selection encodes a passphrase-deterministic cost-map signature unique to DOPPELGÄNGER, on top of the operator signature shared with all phantasm stegos.

DOPPELGÄNGER's catastrophic failure produced the post-mortem in Section 6.

3.4 PALIMPSEST — side-informed embedding (literature adaptation)

PALIMPSEST is the only defense in the four that is not a phantasm-original construction. We adapt the **heuristic** side-informed scheme of Denemark and Fridrich [1], specifically the rounding-direction-correct embedding variant. "Heuristic" is load-bearing: Butora and Bas [2] showed in 2023 that the *actual* discriminative signal in side-informed JPEG steganography is the **covariance between the rounding error and the embedding changes**, not the rounding-direction-correctness the 2015 scheme exploits — and they propose a model-preserving likelihood-ratio formulation that achieves state-of-the-art results against deep-learning detectors. PALIMPSEST tests the older heuristic, not the model-preserving family. Section 6.3 returns to this distinction.

Hypothesis. Denemark and Fridrich show that when the embedder has access to the *pre-quantization* DCT coefficient (the "side information" recoverable when the JPEG was produced from a known higher-quality source), modifications can be routed to align with the natural quantization rounding direction. A modification of +1 at a coefficient whose pre-quantized value rounds *up* is statistically indistinguishable from the natural rounding noise the JPEG already contains; conversely, a +1 where the pre-quantized value rounds *down* is anomalous. Heuristic side-informed embedding thus aligns modifications with the natural rounding direction at a 1-bit asymmetric cost cap, and was, until the 2022–2023 model-preserving line of work appeared, widely regarded as the strongest known classical content-adaptive embedding against CNN steganalysis.

Construction. We implement side-informed cost asymmetry as a parity-conditioned soft-penalty modifier on top of J-UNIWARD, achieving **99.92% direction-correct modifications** in the executed configuration. The literature side-informed family typically reports near-100% direction-correctness — direction-correctness is largely a coding-theory question rather than an optimization one — so the more relevant question is what the residual 0.08% of modifications are. In our pipeline they are precisely the parity-conflicts that single-layer-binary STC forces: positions where the syndrome the STC encoder must satisfy disagrees with the parity of the cover coefficient, leaving no choice but to flip in the wrong direction. They are constraint-satisfactions, not adversarial choices. (A note on implementation constraint appears in Section 6.5 — phantasm's STC pipeline collapses asymmetric (c^+, c^-) to $\min(c^+, c^-)$ as a consequence of the single-layer-binary STC variant, requiring this soft-penalty workaround to recover the asymmetry the literature side-informed scheme

depends on. We are therefore testing a parity-conditioned *approximation* of Denemark–Fridrich 2015, not the full asymmetric ternary embedding the literature defines.)

Result. AUC 0.99 against the in-distribution attacker (checkpoint `effnet_attacker_s1.pt` for PALIMPSEST). **Refuted, in the heuristic-direction-correct sense.** A double-layer STC implementation of the literature scheme, and the model-preserving approach of Butora and Bas, both remain untested in our regime.

PALIMPSEST is the most important of the four results. It tells us that the operator-fingerprint hypothesis is not a peculiarity of randomization-based defenses; it survives even when the embedding is done in the rounding-direction-correct manner that is, in the classical literature, considered the strongest known content-adaptive attack. We discuss this in Section 6.3.

4 Methodology

4.1 Cover corpus

All experiments use a fixed **198-cover Picsum corpus** at QF=85, short-side ≈720 px, deterministically generated from Picsum seeds. Headline numbers in the phantasm README [14] additionally use a 500-cover scaling extension (the d500 corpus) for the attacker-aware EfficientNet-B0 (`juniward_d500_best.pt`, AUC 0.968).

Per-experiment AUC numbers in this paper are computed on the 198-cover corpus at default phantasm settings (≈0.2 bpnzac payload, J-UNIWARD costs, single-layer-binary STC at 0.995× bits/L1).

On the choice of Picsum. The canonical modern JPEG-steganalysis benchmark is ALASKA#2 [12] (∼80,000 covers at multiple QFs); BOSSbase 1.01 [15] is the historical benchmark. We use Picsum, a smaller seed-deterministic source, for two reasons: (a) reproducibility — every cover in our corpus is recoverable from a Picsum seed, and the manifests are public on the phantasm repository; (b) cost — training a multi-passphrase attacker recipe across multiple defenses on ALASKA#2 scale would require an order of magnitude more GPU time than was available for this work. The cost is a real one: cover-source mismatch is known to be central to modern steganalysis, and our Picsum-only results should be read with the explicit understanding that cross-corpus generalization (ALASKA#2 or BOSSbase covers, our attacker recipe) is **not measured**. We list the cross-corpus re-evaluation as Q2 in Section 10.

4.2 Attacker recipe

The standard adversary across all four experiments is an **EfficientNet-B0** initialized from ImageNet [16] and fine-tuned on $N \approx 2000$ **phantasm cover/stego pairs** (5 passphrases per cover across 364 covers from the d500 corpus, ∼1820 train + 272 test pairs after split-aware deduplication), spanning the experiment-specific defense configuration. Training runs for 15 epochs. Pre-processing is JPEG decode → spatial-domain RGB → ImageNet normalization, with the classification head replaced by a 2-class linear layer. Exact training hyperparameters (optimizer, learning rate, batch size, augmentation policy) are defined by the `train_adapted.py` driver and are reproducible from the per-experiment training scripts on the corresponding branches.

This recipe was empirically established in the v0.2–v0.3 cost-function arc as the "minimum credible attacker-aware adversary": at lower N or fewer passphrases, even a phantasm-aware EfficientNet-B0 fails to saturate. At $N = 2000$ with multi-passphrase exposure, it saturates near

1.00 AUC against fixed-cost-function phantasm and serves as a sensitive instrument for testing cost-map defenses.

Caveats on attacker strength. Two limitations of this recipe relative to the modern steganalysis state-of-the-art are worth naming explicitly:

- *No surgical architectural modifications.* Yousfi, Butora, Khvedchenya, and Fridrich [17] showed at WIFS 2020 that ImageNet-pretrained CNNs (including EfficientNet) need steganalysis-specific input-resolution and early-pooling/stride changes to outperform residual-front-end architectures like SRNet on classical benchmarks; the follow-up work [18] ("Improving EfficientNet for JPEG Steganalysis") gives the specific surgical modifications that make EfficientNet competitive. Our attacker uses **vanilla** EfficientNet-B0 with no such modifications. This is plausibly weaker than a best-effort 2024-era attacker.
- *Modest training-data scale.* $N = 2000$ multi-passphrase pairs is small relative to the 10K–100K-pair regimes typical in modern steganalysis literature. Q3 of Section 10 lists this as an open scaling question.

The headline DOPPELGÄNGER catastrophic result (AUC 0.987–0.994 against an *un-adapted* baseline detector that never saw DOPPELGÄNGER stegos during training; Section 5.2) does not depend on attacker strength: an unrelated baseline detects DOPPELGÄNGER stegos at near-perfect AUC regardless of adversary capability. That refutation stands. The HYDRA result (AUC 0.76–0.78 against the in-distribution attacker), by contrast, is the one most sensitive to attacker strength: a surgically-modified EfficientNet at higher N, or a dedicated SRNet/JIN-SRNet-class adversary, may move the in-distribution AUC up. We flag this explicitly so that the HYDRA result is read as "AUC 0.76–0.78 against the recipe defined here," not as a dispositive ceiling.

4.3 Evaluation protocol

For each defense, we hold out 20% of the corpus as a test set and report the test-set AUC of the in-distribution attacker. AUC is the area under the receiver-operating-characteristic curve over test-set cover/stego pairs. The reported range (e.g. "0.76–0.78" for HYDRA) is the spread across passphrases and across two training-data subsamples; the worst-case (highest AUC) is the adversary's success and is the value we credit.

All checkpoints and full eval JSONs are public (Section 9).

5 Results

The load-bearing result of this paper is the *cross-detector* generalization data in Section 5.2: a detector trained *only* on vanilla J-UNIWARD phantasm output — never having seen DOPPELGÄNGER or PALIMPSEST stegos during training — detects both defenses at AUC 0.987–0.994. This is what the four refutations are evidence *for*: a property shared across the cost-map manipulations we test rather than a property of any one defense's specific cost-map. The in-distribution per-defense AUC numbers in Section 5.3 corroborate this finding but are individually weaker evidence (in particular, the HYDRA result is the most attacker-recipe-sensitive of the four; see Section 4.2). We lead with the cross-detector evidence accordingly.

5.1 Reference baselines

For comparison:

- **Off-the-shelf JIN-SRNet on phantasm (J-UNIWARD, default).** 16.2% stego-detection rate vs. a 7.1% cover false-positive floor on the 198-cover Picsum corpus, with 49.5% of stegos scoring lower $P(\text{stego})$ than their own cover. Phantasm passes an off-the-shelf JIN-SRNet; this is the L1 graceful-degradation claim the v1 README makes.
- **Phantasm-aware EfficientNet-B0 on phantasm (J-UNIWARD, default).** AUC 0.968 (`juniward_d500_best.pt`) on the 500-cover d500 extension. This is the headline number in the phantasm README and the L1 ceiling against which the four defenses were tested.

5.2 Cross-detector generalization (the strongest evidence)

The most compelling evidence for the operator-fingerprint hypothesis is *cross-detector* AUC: how well a detector trained on one defense generalizes to detect a *different* defense's stegos. If each defense produced a distinct cost-map fingerprint, cross-detector AUC should drop sharply. It does not:

Test set	**Detector trained on...**	**AUC**
DOPPELGÄNGER stegos	un-adapted `s1` baseline (never saw DOPPELGÄNGER)	0.992–0.994
DOPPELGÄNGER stegos	CHAMELEON-trained	0.985–0.987
DOPPELGÄNGER stegos	attack-mixture	0.94–0.96
PALIMPSEST stegos	un-adapted `s1` baseline	0.987
PALIMPSEST stegos	DOPPELGÄNGER-trained	0.975
PALIMPSEST stegos	attack-mixture	0.924

The cleanest data point: a detector trained only on a *single-config J-UNIWARD baseline* — never having seen DOPPELGÄNGER or PALIMPSEST during training — detects both at AUC 0.987–0.994. Whatever each defense's stegos contain that the CNN reads as "stego," it is a property *shared* with vanilla J-UNIWARD phantasm stegos. The cost map is not the discriminative feature. This result is also robust to attacker-recipe choices in a way the in-distribution per-defense numbers below are not: the un-adapted baseline detector was trained on *vanilla* phantasm stegos at the same recipe long before the four defenses existed; its cross-detection is independent of how well our in-distribution attacker recipe captures DOPPELGÄNGER or PALIMPSEST specifically.

5.3 Per-defense in-distribution AUC

For completeness, the in-distribution per-defense AUC against an attacker retrained from scratch on each defense's stegos:

All numbers vs. an attacker-aware EfficientNet-B0 (ImageNet-pretrained, $N = 2000$ multi-passphrase pairs, 15 epochs). The **headline finding** from this table: none of the four defenses substantially moves the attacker-aware AUC away from the J-UNIWARD-default ceiling, and DOPPELGÄNGER raises it. As the load-bearing-evidence note above flags, the HYDRA 0.76–0.78 figure in particular should be read as "against the recipe defined in Section 4.2" — a stronger attacker (surgically modified EfficientNet, JIN-SRNet, larger N) may push HYDRA closer to the other three. The DOPPELGÄNGER, CHAMELEON, and PALIMPSEST in-distribution numbers are independently corroborated by the cross-detector evidence above.

Defense	Hypothesis class	Origin	Measured AUC	Verdict
HYDRA	WHERE-randomization	novel	0.76–0.78	refuted (partial)
CHAMELEON	WET-fraction randomization	novel	0.82–0.93	refuted
DOPPELGÄNGER	WHERE / variance-shaped	novel	0.99–1.00	refuted (catastrophic)
PALIMPSEST	heuristic HOW-with-rounding (Denemark–Fridrich 2015 approx.)	literature	0.99	refuted

5.4 Cross-defense convergence

Three of the four defenses (HYDRA, CHAMELEON, DOPPELGÄNGER) randomize *where* an unchanged STC modify-and-re-encode operator places its modifications, by varying the cost map fed to STC across three different cost-map sources (multi-detector adversarial saliency, saliency with stochastic keep-fraction, cross-cover variance estimate). The fourth (PALIMPSEST) randomizes *with what rounding direction* the same STC operator places them, using a side-informed J-UNIWARD-derived cost shape. The four defenses span at least three distinct cost-map sources and two distinct hypothesized hardening mechanisms, were implemented and debugged by independent toolchains, and were evaluated against an attacker retrained from scratch for each defense.

Despite that independence, all four converge on AUC ≥ 0.76, with three converging at AUC ≥ 0.99. The hypothesis that explains this convergence — that a *single shared property of all four defenses* is detectable, irrespective of cost-map design — is the operator-fingerprint hypothesis we develop in Section 6.

6 The operator-fingerprint hypothesis

6.1 Statement

We state the hypothesis as it emerged from the DOPPELGÄNGER post-mortem and was confirmed by PALIMPSEST. The underlying *mechanism* — modification-and-re-encoding of an already-quantized JPEG leaving a CNN-detectable statistical trace — is closely related to the Reverse JPEG Compatibility Attack of Butora and Fridrich [19], which makes the analogous observation for QF=99–100 covers and proposes a model-based universal detector. RJCA does not directly apply at our QF=85 operating point, but the *kind* of artifact it isolates (spatial-domain rounding-error structure caused by integer modify-then-re-quantize) is exactly the kind of artifact our attacker-aware CNN appears to be learning. The empirical claim we make is therefore not a new mechanism but a new *scope*: that an analogous artifact remains learnable at QF=85 by an attacker-aware CNN, and that the four specific cost-map manipulations we test do not remove it.

Operator-fingerprint hypothesis. *For a JPEG content-adaptive embedding scheme of the form* decode JPEG → modify integer DCT coefficient → re-encode JPEG*, evaluated on cover/stego pairs at typical web-distribution quality factors (QF ≈ 85), the dominant detectable signature available to an attacker-aware CNN steganalyzer trained on the scheme's output is* ***not the location at which modifications are placed****, nor the heuristic rounding-direction-correctness explored in Denemark and Fridrich's 2015 side-informed scheme. The dominant signature is the* ***statistical artifact of the modification-and-re-encoding operation itself****, applied to the JPEG produced by libjpeg's standard quantization pipeline. This artifact is invariant under the cost-map redistributions*

(HYDRA, CHAMELEON, DOPPELGÄNGER) and the heuristic rounding-direction shaping (our PALIMPSEST approximation) that are the standard tools of cost-map defense. We do not *claim it is invariant under model-preserving covariance-aware side-informed embedding (Butora–Bas 2023), which we have not tested.*

6.2 Why all three WHERE-randomization defenses converge

For HYDRA, CHAMELEON, and DOPPELGÄNGER, the only thing varied across stegos (relative to J-UNIWARD-default phantasm) is *which coefficients* receive a ± 1 modification. The modification itself — its sign distribution, its integer step size, its spatial-domain consequence after IDCT — is unchanged.

A CNN that has learned to discriminate "this 8×8 block was quantized once by libjpeg" from "this 8×8 block was quantized, modified by ± 1, and re-encoded" will, in test time, consult that block-level discriminator regardless of whether the modification was placed by a J-UNIWARD lowest-cost rule, a HYDRA ensemble-saliency mixture, a CHAMELEON wet-fraction randomized mixture, or a DOPPELGÄNGER variance-shaped mixture. The cost-map defenses change the spatial distribution of modified blocks; they do not change the per-block detectability.

This explains why all three converge at AUC ≥ 0.76, with DOPPELGÄNGER worse than baseline: its variance-shaping concentrates modifications in the very spatial regions where the per-block discriminator is most sensitive.

6.3 Why PALIMPSEST also converges

PALIMPSEST is the harder case. Side-informed J-UNIWARD with rounding-direction-correct embedding does, in principle, align modifications with the *natural* quantization rounding direction. The literature claim is that this aligns the post-modification block with the distribution a fresh re-quantization would have produced.

The post-mortem on PALIMPSEST identifies why this claim does not help in our regime. Even with 99.92% direction-correct modifications, **every modification still shifts the integer DCT coefficient by** ± 1 (equivalently $\pm q$ in the de-quantized representation) *on top of* the cover's existing quantization rounding error. Crucially, libjpeg's quantization on the already-quantized cover is the identity — there is no fresh quantization noise to absorb the modification into — so the modification *is* the fingerprint. The "with-rounding" framing equalizes the modification against a *hypothetical alternative quantizer that quantizes the pre-modification value in the modified direction*, but it does not equalize against the actual libjpeg-quantized cover the CNN compares to. The CNN is trained on cover/stego pairs in which the cover is the libjpeg-quantized JPEG and the stego is the post-modification JPEG; the discriminator learns the modification-and-re-encoding artifact directly. Side-informed rounding does not produce a modification that is statistically identical to the cover's own quantization noise, because the cover does not have the modification. It produces a modification that *would have been* statistically natural in a counterfactual JPEG.

Counterfactual indistinguishability is not enough. The CNN is trained on the actual JPEG, not the counterfactual.

This is also the cleanest cross-validation of the operator-fingerprint hypothesis: PALIMPSEST stegos are detected at AUC 0.987 by an *un-adapted single-config baseline* detector trained only on vanilla J-UNIWARD phantasm output (Section 5.2). The detector has never seen a side-informed modification at all; the only feature available to it is the operator fingerprint, which heuristic side-informed rounding does not remove.

Independent confirmation in the literature, and what remains untested. The critique we develop in this subsection — that heuristic rounding-direction-correctness equalizes against a counterfactual rather than the actual quantized cover — is, as of 2023, also the critique made by Butora and Bas [2] of the broader heuristic side-informed literature. They argue that the discriminative signal in side-informed JPEG steganography is the **covariance between the rounding error and the embedding changes**, not the rounding direction per se, and they derive a model-preserving likelihood-ratio formulation that achieves state-of-the-art performance against deep-learning detectors on BOSSbase and BOWS2. Our PALIMPSEST refutation does *not* refute their formulation: PALIMPSEST tests a parity-conditioned approximation of the 2015 heuristic, and their model-preserving approach is mechanistically different (preservation of a decompressed-image covariance model, rather than alignment of per-coefficient rounding direction). A faithful evaluation of the model-preserving family against our attacker recipe is open work; we list it explicitly in Section 10 (Q5).

The closest existing literature on what we are calling the *operator fingerprint* is the **Reverse JPEG Compatibility Attack** of Butora and Fridrich [19], which detects "any type of steganography" at JPEG quality factors 99–100 from the spatial-domain rounding-error signature induced by modifying an already-quantized DCT coefficient and re-encoding. The QF=99–100 regime is where RJCA is universally powerful; at our QF=85 operating point, the rounding-error signature is much weaker, and RJCA does not apply directly. Levecque, Butora, and Bas [20] have since extended the line by identifying *incompatible blocks* — DCT blocks with no spatial-domain pre-image — as a specific consequence of the same operator. Our framing is therefore best understood as the extension of the RJCA-class observation to the QF=85 regime via attacker-aware CNNs: not a discovery of a novel class of fingerprint, but a demonstration that an analogous fingerprint remains learnable under conditions where RJCA itself does not apply, and that no cost-map manipulation in our four-experiment scope reduces it.

6.4 What is left

If the operator-fingerprint hypothesis is correct, the space of *viable* defenses against attacker-aware CNN steganalysis on JPEG covers is constrained — though not as narrowly as a strong reading of our four refutations might suggest:

- Defenses that vary the *position distribution* of an unchanged modify-and-re-encode operator (HYDRA, CHAMELEON, DOPPELGÄNGER-class) are not viable in our regime.
- Defenses that vary the *heuristic rounding direction* of an unchanged modify-and-re-encode operator (Denemark–Fridrich-style, PALIMPSEST-class) are not viable in our regime *as we approximated them.*
- *Open*: defenses that preserve a covariance model of the decompressed JPEG (Butora–Bas-class) are mechanistically different from what PALIMPSEST tested and remain untested in our regime. Whether they escape the operator-fingerprint ceiling identified here is an empirical question we do not answer.
- What remains structurally most likely to break the ceiling is a defense that **changes the embedding operator itself** — not its position distribution, not its rounding direction, but the operation it performs. We sketch candidates for such operator-changing defenses in Section 7.

We emphasize: the hypothesis is *empirically supported by four data points*, *consistent with* the spatial-domain rounding-error signature characterized by RJCA at QF=99–100 [19], and *consistent*

with the structure of CNN steganalysis. It is not proven, and the model-preserving side-informed family remains a mechanistically open avenue. We invite falsification.

6.5 Two architectural findings (load-bearing for any cost-map work in this codebase)

Two implementation-level findings emerged during this arc that constrain the interpretation of the results:

1. **Natural row-major vs. zigzag DCT order.** Phantasm's `JpegComponent::coefficients` and `quant_table` are stored in *natural* row-major order, despite inline comments earlier claiming zigzag. PHCOST sidecar slots match natural order. Documented in `phantasm-image/src/side_info.rs` on the `experiment/palimpsest` branch.

2. **Single-layer-binary STC pipeline collapses asymmetric costs.** Phantasm's pipeline collapses asymmetric (c^+, c^-) pairs to $\min(c^+, c^-)$ and modifies via coeff $\oplus$ 1, so the modification *direction* is forced by the parity of the cover coefficient. The asymmetric (c^+, c^-) on which the literature side-informed STC depends is *not direction-selective* in our pipeline. PALIMPSEST's 99.92% direction-correctness was achieved by a parity-conditioned soft-penalty workaround. A true asymmetric implementation requires `phantasm_stc::DoubleLayerEncoder` wired through the orchestrator, which is post-v1 work.

These findings do not change the AUC conclusion — PALIMPSEST at 99.92% direction-correctness is still refuted at AUC 0.99 — but they constrain how strongly we can claim "this is exactly the literature side-informed scheme." We claim, conservatively, that PALIMPSEST captures the *operative mechanism* of side-informed embedding (rounding-direction-correct modifications) under phantasm's STC pipeline, and refute it. A double-layer STC implementation of the literature scheme would make the same claim more precisely.

7 The five-defense roadmap and what remains

Following the DOPPELGÄNGER post-mortem and the PALIMPSEST confirmation, we triaged five candidate defense families under the constraint imposed by the operator-fingerprint hypothesis. The status of each at v1.0 ship time is summarized below; full specs are at `archive/research-arc/specs/`.

7.1 DOPPELGÄNGER and PALIMPSEST (executed)

Both refuted, as reported. They produced, respectively, the post-mortem (Section 6.1) and the operator-fingerprint ceiling confirmation (Section 6.3).

7.2 OUROBOROS — learned-generator GAN with accumulating discriminator pool (deferred)

OUROBOROS proposes a generator-discriminator pair in which the generator E_θ produces stego images directly (in pixel or DCT space) and the discriminator D_t is the current best attacker-aware CNN, with the discriminator pool accumulating across training rounds to prevent generator-mode collapse.

Status: deferred. The diff-embed-researcher reference implementation hit a *parity-loss plateau* at flip ≈ 0.5 — the generator could not recover the parity bits the embedder needs — which is a

known prerequisite wall in this defense family. Resume conditions: GHOSTWRITER refutes (i.e. patch synthesis also fails the operator-fingerprint test, which would update the prior on operator-changing defenses), or new theoretical insight on the diff-STC parity wall.

7.3 GHOSTWRITER — patch synthesis instead of coefficient modification (deferred)

GHOSTWRITER proposes a true *operator change*: instead of modifying DCT coefficients of the cover, the embedder **synthesizes** small image patches (e.g. 16×16 spatial pixel regions) conditioned on a learned natural-image prior, encoding payload bits in the *choice* of synthesized patch rather than in any direct modification of cover coefficients. The operator is no longer "decode JPEG $\rightarrow$ modify integer DCT $\rightarrow$ re-encode JPEG" — it is **"synthesize a natural-looking patch consistent with local context, paste it in, encode the JPEG once."** No post-quantization integer modification ever takes place.

This is, on the operator-fingerprint hypothesis, structurally the most promising of the deferred defenses: by replacing the modify-and-re-encode operator with a synthesize-and-encode operator, the *specific* fingerprint identified in Section 6 is removed by construction. Whether GHOSTWRITER works at all reduces to whether the *new* operator (a learned patch synthesizer) introduces a different, equally-detectable fingerprint of its own.

Mechanism sketch. A small generative model — a denoising autoencoder, a tiny latent-diffusion model, or a context-conditional flow — is trained on the same JPEG-corpus distribution as the cover source, so that its sampled patches match the corpus's quantization-noise statistics. Payload bits are encoded through:

- **Latent-space dithering.** The generator's latent vector is partitioned into an "appearance" subspace (driven by local context) and a "payload" subspace (driven by the bit stream via a passphrase-keyed permutation). The payload bits select among 2^k valid candidate latents whose decoded outputs are perceptually equivalent under a chosen distance metric.
- **Rejection sampling.** Many candidate patches are sampled; the bit-encoded patch is the first whose hash matches the target syndrome. This is computationally heavier but trivially composable with STC-like syndrome coding.

Either path requires that the synthesis preserve **lossless recovery** of the encoded bits across a JPEG re-encoding round trip — the patch must survive QF=85 quantization without bit loss. This is the harder of the two technical gates.

Phase 1 gate criteria. Concretely, before any attacker-aware evaluation is run, GHOSTWRITER must demonstrate:

1. **Synthesis quality.** A Fréchet Inception Distance below some target threshold (e.g. FID ≤ 20 on 16×16 patches sampled from natural Picsum JPEGs at QF=85).
2. **Bit recovery.** Reliable round-trip recovery of payload bits from a synthesized + re-encoded patch — at minimum, $\geq 99\%$ recovery on a representative QF=85 corpus, with the recovery error rate fully absorbable by the existing Reed-Solomon ECC layer (`phantasm-ecc`).
3. **Capacity.** Aggregate per-cover capacity of at least 1024 bytes — enough to carry a phantasm v1 envelope payload in a single share.

Any one of these gates failing kills the experiment without ever invoking the attacker-aware adversary.

Failure modes worth naming explicitly — each represents an "adjacent ceiling" GHOSTWRITER could hit:

- **Generator self-fingerprint.** A growing body of work in generative-model forensics suggests that modern image generators (GANs, autoencoders, small diffusions) leave architecture-level statistical traces detectable by a CNN trained specifically for that purpose. If GHOSTWRITER's synthesizer's outputs are detectable as *synthetic* rather than natural, the operator fingerprint is *replaced* by a generator fingerprint rather than dissolved. Verifying generator-output naturalness against a generative-forensics classifier is a sensible additional Phase 1 gate.
- **Local-context coherence.** Synthesized patches that ignore cover lighting / texture / contour produce obvious seam artifacts. The synthesizer must condition on enough surrounding context to maintain coherence; this conditioning is itself a fingerprint candidate (e.g., the conditioning model's preferences for certain textures may be learnable).
- **Capacity vs. naturalness trade-off.** Encoding many bits per patch reduces the size of the candidate-latent space, which reduces synthesis flexibility, which produces less-natural patches. Whether the trade-off curve crosses a usable capacity threshold *before* it crosses a detectable-quality threshold is an open empirical question.
- **JPEG re-encoding bit loss.** A patch synthesized in spatial pixels and then JPEG-quantized at QF=85 loses precision in ways that may flip encoded bits. Either the encoder operates directly in the post-quantization integer-DCT domain (which re-introduces the modify-after-quantize fingerprint we are trying to avoid), or the encoder operates in pixel space with explicit QF=85 round-trip compensation (which constrains what the synthesizer can express).

GHOSTWRITER is structurally most-likely-to-succeed *if* the operator-fingerprint hypothesis is correct, and structurally most-likely-to-reveal-an-adjacent-ceiling *if* the hypothesis is missing a class of fingerprint we have not yet characterized. Either outcome is informative.

7.4 PRISM — multi-cover diversification (deprioritized)

PRISM proposes hiding a single payload across multiple covers in such a way that no single cover is detectable in isolation. The original spec framed this as a breakthrough; we re-spec'd it as an honest baseline (since multi-cover schemes are already in the classical literature) and deprioritized.

7.5 Operator-changing approaches beyond the spec'd roadmap

The five-defense roadmap was scoped for what we could plausibly prototype within the v1 timeframe; it is not a survey of the operator-changing design space. Several directions outside the roadmap deserve serious attention from future work, each with a distinct relationship to the operator-fingerprint hypothesis.

Diffusion-based steganography. Sample stegos from a JPEG-corpus-trained diffusion model conditioned on the cover and the payload — the diffusion process becomes the embedding operator. The fingerprint hypothesis predicts this should work *if* the diffusion model's outputs are statistically indistinguishable from naturally-encoded JPEGs at the chosen QF. Existing literature on diffusion

fingerprinting suggests this *if* may be load-bearing: diffusion outputs frequently carry detectable spectral artifacts at high frequencies. A diffusion-based stego scheme that survives JPEG-domain attacker-aware CNNs is, on present evidence, an open research question, not a known solution.

Source-recompression embedding. When the embedder has access to a *higher-QF* version of the cover, modify in the high-QF domain and re-quantize once at the target QF=85. The fresh quantization absorbs the modification fingerprint into the quantization noise that *would have been there anyway* — the operator becomes "modify high-QF + quantize once," not "modify quantized + re-quantize." This is essentially the threat model PALIMPSEST's *Path A* assumed (uncompressed source available) but combined with a re-quantization step. Whether the side-info benefit recovers under this stronger threat model is empirically testable and not yet tested at our adversary recipe.

Cross-domain transformation. Modify in a non-DCT domain (e.g. learned latent space, wavelet, JPEG-XL transform) and project back to the JPEG output via re-encoding. The projection itself becomes a candidate fingerprint, but it may dissolve the *integer-DCT-modification* fingerprint specifically. The operator becomes "encode JPEG once from a different representation" — closer to source-recompression than to modify-and-re-encode. The risk is that the new domain's modifications survive projection in detectable form.

Generator-discriminator pairs broader than OUROBOROS. The ASDL-GAN [21] and ADV-EMB [22] line of work in DCT-domain generative steganography, and UT-GAN [23] in cost-learning, are existing research directions whose modern re-implementation against attacker-aware CNNs would be informative on its own merits. The GAN-stego line did not stop in 2020: a continuing body of work has proposed side-information-augmented GAN cost learners, diversified-discriminator training schemes, and combinations of UT-GAN with hand-crafted residual-feature adjustments. We have not evaluated this concurrent literature against our specific attacker recipe, and we do not catalogue specific 2023–2025 papers here without first reading and verifying their reported results — but a faithful survey of the continuing GAN-stego line, paired with an attacker-aware re-evaluation, is well-aligned with both OUROBOROS' deferred spirit and the concurrent-work expectations of any reviewer in this area. OUROBOROS' deferral was driven by a specific parity-loss-plateau wall in our diff-embed prototype; a clean re-implementation of ASDL-GAN or UT-GAN against a phantasm-aware adversary recipe would test the broader question of whether *any* GAN-style operator escapes the ceiling, independent of phantasm's specific pipeline constraints.

Distributed multi-cover embedding with information-theoretic bounds. PRISM was deprioritized as a single-image defense because it inherits the cost-map ceiling. But a *multi-cover* formulation — distribute payload across N covers via secret sharing or forward error correction such that detection of one cover provides no information about whether *any* cover carries data — sidesteps the per-cover detectability question entirely under a different threat model. The interesting question is whether multi-cover schemes can be combined with operator-changing single-cover defenses to produce a layered defense whose ceiling is strictly below either component's individual ceiling. This is an information-theoretic question with an empirical evaluation component.

We list these directions to make explicit that our four refutations close *cost-map randomization* and *side-informed rounding*, not the entire defense design space. The space of operator-changing approaches is plausibly large, and the hypothesis we publish does not predict *every* such approach must succeed — it predicts that approaches that *retain* the modify-and-re-encode operator must fail.

7.6 What an operator-changing defense must demonstrate

For any defense in the operator-changing class to *credibly* break the precedent reported here, we argue it must measurably demonstrate, on the public phantasm corpus and against the public adversary checkpoints:

1. **AUC strictly below the corresponding baseline.** Lower than J-UNIWARD-default phantasm against `juniward_d500_best.pt`, and lower than PALIMPSEST against `effnet_attacker_s1.pt`.
2. **Robustness to retraining.** The new adversary, retrained on the new defense's output at the same $N = 2000$ multi-passphrase recipe, also fails to saturate.
3. **Honest scope.** The defense's claim explicitly addresses the operator-fingerprint hypothesis — i.e. the defense's authors identify *what about the operator they changed*, not just *what about the cost map they changed*.

These three criteria are the bar; we publish them not as gatekeeping but as the minimum reasonable specification of "this is no longer the same regime."

8 Precedent to break

We restate, as a formal challenge to the field:

> **The Four Lousy Ghosts ceiling (in-regime).** As of 2026-05-02, no cost-map-based defense against the specific attacker-aware CNN steganalysis recipe defined in Section 4 has, on the public Picsum phantasm corpus at QF=85, measurably escaped AUC 0.76. Three randomization-based defenses (WHERE, WET, WHERE/variance-shaped) and one parity-conditioned approximation of heuristic side-informed rounding (HOW) have been tested and refuted. The hypothesized common cause is the operator-level fingerprint of the modify-and-re-encode operation. We invite falsification.

Different *kinds* of breakthrough that would credibly break the precedent:

1. **A cost-map-family breakthrough.** Any future cost-map defense — a tweak to PALIMPSEST, a new variant of HYDRA or CHAMELEON, a defense not in the five-spec roadmap — that achieves AUC strictly below the corresponding baseline against the public adversary checkpoints, with a retrained-attacker robustness check, is a *positive* result and breaks the ceiling we identify.
2. **A model-preserving side-informed evaluation.** A faithful evaluation of the Butora–Bas covariance-preserving side-informed family [2] against our attacker recipe, or a double-layer-STC implementation of the Denemark–Fridrich 2015 scheme that does not require a parity-conditioned approximation, would close the gap left by PALIMPSEST. If either reaches AUC below the J-UNIWARD-default baseline, our HOW-with-rounding verdict softens further.
3. **An operator-changing breakthrough.** Any operator-changing defense (GHOSTWRITER, OUROBOROS, diffusion-based steganography on JPEG covers, model-preserving generative embedders) that achieves the same standard and explicitly addresses the operator-fingerprint hypothesis breaks the ceiling.
4. **A theoretical breakthrough.** Any new analysis that disproves the operator-fingerprint hypothesis — by showing a defense in our regime that the hypothesis wrongly predicted should fail, but that succeeds — is the strongest result and we invite it.

A *fifth* form of revision we would welcome but which does not constitute a precedent-break is a **stronger-attacker re-evaluation**. If a surgically-modified EfficientNet [18] or a JIN-SRNet [5] adversary at higher N reveals that one of our four AUC numbers was an artifact of EfficientNet-B0's specific inductive biases rather than a property of the operator fingerprint, the relevant defense returns to live status and the recipe-vs-ceiling distinction in Section 4.2 sharpens. If instead the stronger attacker raises in-distribution AUC further, the precedent gets *more* strongly confirmed, not broken. Either outcome is informative; neither is a breakthrough. We list it explicitly so that the natural follow-up "but is HYDRA's 0.76 robust to a better attacker?" has a clear place in the research roadmap that isn't in the breakthrough list.

We publish this work explicitly as a benchmark to be broken. The phantasm v1.0.1 binary, the in-distribution adversary checkpoints, the per-experiment branches, and the full eval JSONs are all public (Section 9). Anyone with a workstation and a few weeks of GPU time can reproduce our four refutations and, hopefully, produce a fifth result that is not a refutation.

9 Reproducibility

All artifacts required to reproduce the results, attempt to falsify the operator-fingerprint hypothesis, or build a new defense against the same adversary checkpoints, are public.

9.1 phantasm v1.0.1 binary and source

- **Source:** `github.com/exec/phantasm` (main branch, tag `v1.0.1`)
- **Binaries:** GitHub Release `v1.0.1` — Linux x86_64, macOS arm64
- **Tests:** 204 passing under `cargo test --workspace`

9.2 Per-experiment branches

Each defense's full implementation, training scripts, and per-experiment narrative (`scratch/<codename>/STATE.md`) is preserved as a branch on the public repo:

- `experiment/hydra` — HYDRA construction + scripts + eval
- `experiment/chameleon` — CHAMELEON construction + scripts + eval
- `experiment/doppelganger` — DOPPELGÄNGER construction + scripts + eval
- `experiment/palimpsest` — PALIMPSEST construction (parity-conditioned soft-penalty implementation) + scripts + eval

9.3 Specifications and post-mortem

`archive/research-arc/` on the main branch contains:

- `specs/2026-04-25-doppelganger.md`
- `specs/2026-04-25-doppelganger-postmortem.md` (load-bearing)
- `specs/2026-04-25-palimpsest.md`
- `specs/2026-04-25-ouroboros.md`

- `specs/2026-04-25-ghostwriter.md`
- `specs/2026-04-25-prism.md`
- `specs/PALIMPSEST-verdict.md`
- `README.md` (consolidated index)

9.4 Adversary checkpoint pool

The trained CNN steganalyzers used to produce the AUC numbers are released as a single GitHub Release on the phantasm repo:

- **Release:** `research-checkpoints-v1`
- **Total:** 244 MB / 20 PyTorch checkpoints
- **Contents:** Per-experiment in-distribution attackers, single-config baselines (`*_s1.pt`), phantasm-aware fine-tunes, and the d500-scale headline-corpus attacker (`juniward_d500_best.pt`).
- **Per-checkpoint:** training-config JSON, SHA256SUMS bundled.
- **Architectures included:** EfficientNet-B0, SRNet, XuNet, ViT-Tiny, small custom shallow CNN.
- **Architectures *not* redistributed (see CITATIONS.md for sources):** JIN-SRNet (DDE Lab Binghamton release), Aletheia EfficientNet-B0 (Daniel Lerch's Aletheia project [24]).

9.5 Earlier research log

The v0.2/v0.3 cost-function arc (Updates 1–8) that established the multi-config-passphrase attacker recipe used in this paper is preserved at `archive/ML_STEGANALYSIS.md` on the main branch.

10 Open questions and concrete next experiments

The work reported here closes specific defense families against a specific attacker recipe. It opens at least as many questions as it closes, and we list the most empirically tractable ones explicitly to give future work concrete starting points.

Q1. Is the operator-fingerprint hypothesis necessary, or only sufficient? Section 6 argues that the modify-and-re-encode artifact is a sufficient cause for our four refutations. It does not prove that the artifact is the *only* cause. A defense that retains the operator AND succeeds — by mechanism we have not anticipated — would refute the hypothesis. A defense that changes the operator AND fails would *also* refute it (since the hypothesis predicts operator-changing approaches at least *can* succeed). Both directions are evidence-relevant.

Q2. Does the ceiling depend on cover-source distribution? All four refutations were measured on the 198-cover Picsum corpus (plus the d500 extension). Picsum is one specific cover source; ALASKA#2 [12], BOSSbase [15], and ImageNet-scale corpora have different statistics. Whether the operator-fingerprint generalises across cover sources is a direct, low-cost re-evaluation: train one of our four in-distribution attackers on ALASKA#2 stegos produced with the same defense, and compare AUC. Cover-source mismatch is known to be central to modern steganalysis [12], and the absence of a cross-corpus evaluation is the single biggest evaluation gap in this paper.

Q3. What is the scaling behaviour of the attacker recipe? Our $N = 2000$ multi-passphrase attacker is the "minimum credible attacker-aware adversary" empirically established in the v0.2-v0.3 arc. AUC saturation behaviour at $N = 10{,}000$, $N = 100{,}000$, or with multi-architecture ensembles is not measured. A defense that defeats the $N = 2000$ recipe but loses at $N = 20{,}000$ has been demonstrated on a corpus the actual adversary will not use. Conversely, a defense that fails at $N = 2000$ but where AUC plateaus rather than saturates suggests the recipe is near the *attacker's* ceiling, not the defense's.

Q4. Does the fingerprint generalise across attacker architectures? Our headline attacker is EfficientNet-B0; SRNet variants (`srnet_N1000.pt`) and XuNet variants (`xunet_N2000.pt`) are included in `research-checkpoints-v1` but were not the headline test for the four defenses. ViT-Tiny, Swin transformers, and modern hybrid CNN-transformer detectors are not tested at all under the attacker-aware recipe. A defense whose ceiling is EfficientNet-specific would be evidence that the operator fingerprint is more about EfficientNet's inductive biases than about the operator itself.

Q5. Can the operator fingerprint be characterised formally, and does the model-preserving side-informed family escape it? Section 6 articulates the fingerprint qualitatively: "the statistical artifact of decoding, modifying integer DCT, and re-encoding." A precise characterisation — e.g., the spectrum of the residual between cover and stego in a chosen wavelet basis, the per-block divergence under a learned natural-image prior, or a closed-form distribution of post-modification quantization-error deltas — would make the hypothesis falsifiable in a stronger sense. The most empirically tractable specific question in this neighbourhood is whether the **model-preserving side-informed scheme of Butora and Bas [2]** — which derives the embedding from the covariance between rounding error and embedding changes — escapes the operator-fingerprint ceiling under our attacker recipe at QF=85. PALIMPSEST does *not* test their formulation; a faithful re-implementation against `juniward_d500_best.pt` and at least one SRNet-class adversary would close the largest open gap left by this paper, and is the single experiment we would most prioritize running next.

Q6. Does GHOSTWRITER's Phase 1 gate suffice? GHOSTWRITER's specified gate (FID + bit recovery + capacity) verifies the synthesizer is *individually* capable. It does not verify that the synthesizer's outputs survive a *generative-forensics-aware* adversary. A modern attacker training on phantasm output knows phantasm uses synthesised patches; an adversary specifically trained to detect synthesis-of-this-flavour is the relevant adversary. An additional Phase 1 gate ("synthesised-patch naturalness against a generative-forensics classifier") may be warranted before any expensive attacker-aware evaluation.

Q7. Does layered defense escape the single-defense ceiling? Each of our four defenses is evaluated standalone. A composition — e.g. PALIMPSEST applied to a GHOSTWRITER-

synthesised cover, or CHAMELEON's stochastic keep-fraction applied to a diffusion-generated stego — is not measured. Combinations may produce ceilings strictly below either component's standalone ceiling, *or* may produce additively-detectable composite fingerprints. We have no data on which.

These seven questions are the next-experiment list. Each one has a defined methodology, a public adversary checkpoint pool against which to evaluate, and a falsifiable outcome. None requires breakthrough theoretical insight to begin running.

11 Conclusions

Four content-adaptive cost-map defenses against attacker-aware CNN steganalysis on JPEG covers — three novel, one a parity-conditioned approximation of the heuristic side-informed scheme of Denemark and Fridrich — have been refuted under a single recipe of attacker-aware EfficientNet-B0 trained on $N \approx 2000$ multi-passphrase phantasm pairs at QF=85 on a Picsum cover corpus. Two families of defense — *WHERE-randomization* (HYDRA, CHAMELEON, DOPPELGÄNGER) and *heuristic HOW-with-rounding* (the PALIMPSEST approximation) — hit the same empirical ceiling in our regime. The operator-fingerprint hypothesis offers the simplest explanation: the attacker-aware CNN learns the statistical artifact of *decoding, modifying integer DCT, and re-encoding* an already-quantized JPEG, and that artifact is invariant under the cost-map and rounding-direction shaping our four defenses introduce. This artifact is in the same family as the spatial-domain rounding-error signature that Butora and Fridrich's Reverse JPEG Compatibility Attack [19] characterizes at QF=99–100; our work extends the empirical observation to QF=85 via attacker-aware CNNs.

Several defense families remain explicitly *untested* and are not refuted by this work:

- **Model-preserving / covariance-aware side-informed steganography** (Butora and Bas 2023 [2]) — mechanistically different from the heuristic rounding-direction-correctness PALIMPSEST tests, and currently the literature's strongest attack on the same problem we are working on.
- **A faithful double-layer-STC implementation of the Denemark–Fridrich 2015 scheme** that does not require our parity-conditioned soft-penalty approximation.
- **Operator-changing defenses** (GHOSTWRITER, OUROBOROS, diffusion- or generative-prior-based steganography), which are out of scope in our four-experiment evaluation.

Phantasm v1 commits, accordingly, to an L1-scoped-down posture and stakes its security argument on the L2 (passphrase-keyed position recovery) and L3 (Argon2id + AEAD + HMAC + HKDF independent-extract) cryptographic envelope, with L1 framed as a graceful-degradation property against off-the-shelf adversaries rather than a hardness claim against phantasm-aware ones.

This paper is published as a falsification challenge, not as a final word. The artifacts are public; the in-regime empirical ceiling for the families we tested is measurable; and an explicit list of untested families and stronger-attacker re-evaluations is in Section 10. We invite the future work that breaks the precedent — or that demonstrates the precedent is recipe-specific by re-evaluating our defenses against a stronger attacker recipe and producing different numbers.

Author contributions

Dylan Hart designed and implemented the phantasm tool, conceived the four-defense research arc, executed all training and evaluation runs (including the GPU-bound multi-passphrase EfficientNet-B0 attacker training), and is the primary author of the codebase, the specs at `archive/research-arc/specs/`, the post-mortem, and the final editorial decisions on this paper.

Claude (Anthropic, `claude-opus-4-7`) collaborated on the design of HYDRA, CHAMELEON, and DOPPELGÄNGER (the three novel defenses) in the brainstorming and spec-writing phase, drafted earlier versions of the post-mortem and PALIMPSEST verdict from which Section 6 is distilled, performed the citation-verification pass that produced `CITATIONS.md`, and drafted this paper. Final scope, claims, and editorial control rest with the human author. The collaboration log is preserved in the public phantasm repository's commit history.

Note on LLM authorship. Several major peer-reviewed venues (Nature, Science, ICML, NeurIPS, ACM publications under recent policy updates) now explicitly disallow LLM systems as named authors. The author block above reflects this paper's status as a self-published research artifact on the author's personal site, where the disclosure norm we apply is full transparency about Claude's contribution rather than the venue-specific norm of acknowledging it without naming. For any peer-reviewed submission of this work, Claude's contributions described above would move to the Acknowledgements section per the relevant venue's policy.

Acknowledgements

Thanks to the authors of the open steganalysis-research toolchain — DDE Lab Binghamton, the Aletheia project, the J-UNIWARD and STC authors — without whose published artifacts (SRNet, JIN-SRNet, J-UNIWARD reference, STC parity-check tables, Aletheia EfficientNet-B0) this work would have required several person-years of additional groundwork.

Thanks also to Picsum (`picsum.photos`) for the deterministic-seeded JPEG corpus used as cover-source mismatch reference throughout.

The full and verified citation list with DOIs is available at `CITATIONS.md` on the phantasm repository. Citations in this paper were re-checked against canonical sources prior to publication.

References

[1] T. Denemark, J. Fridrich. "Side-informed steganography with additive distortion." *IEEE International Workshop on Information Forensics and Security (WIFS)*, 2015, pp. 1–6. doi:10.1109/WIFS.2015.7368589.

[2] J. Butora, P. Bas. "Side-Informed Steganography for JPEG Images by Modeling Decompressed Images." *IEEE Transactions on Information Forensics and Security*, 2023. doi:10.1109/TIFS.2023.3268884. arXiv:2211.05530.

[3] D. Hart. *phantasm: content-adaptive JPEG steganography in Rust.* Software, version 1.0.1, 2026. `https://github.com/exec/phantasm`.

[4] H. Krawczyk, P. Eronen. "HMAC-based Extract-and-Expand Key Derivation Function (HKDF)." *IETF RFC 5869*, May 2010. doi:10.17487/RFC5869.

[5] J. Butora, Y. Yousfi, J. Fridrich. "How to Pretrain for Steganalysis." *Proceedings of the 9th ACM Workshop on Information Hiding and Multimedia Security (IH&MMSec)*, 2021, pp. 143–148. doi:10.1145/3437880.3460395.

[6] phantasm v1.0.1 STATUS. `https://github.com/exec/phantasm/blob/main/STATUS.md`.

[7] T. Filler, J. Judas, J. Fridrich. "Minimizing Additive Distortion in Steganography Using Syndrome-Trellis Codes." *IEEE Transactions on Information Forensics and Security*, vol. 6, no. 3, pp. 920–935, 2011. doi:10.1109/TIFS.2011.2134094.

[8] J. Holub, J. Fridrich, T. Denemark. "Universal distortion function for steganography in an arbitrary domain." *EURASIP Journal on Information Security*, vol. 2014, no. 1, 2014. doi:10.1186/1687-417X-2014-1.

[9] L. Guo, J. Ni, W. Su, C. Tang, Y.-Q. Shi. "Using Statistical Image Model for JPEG Steganography: Uniform Embedding Revisited." *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 12, pp. 2669–2680, 2015. doi:10.1109/TIFS.2015.2473815.

[10] M. Boroumand, M. Chen, J. Fridrich. "Deep Residual Network for Steganalysis of Digital Images." *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 5, pp. 1181–1193, 2019. doi:10.1109/TIFS.2018.2871749.

[11] M. Tan, Q. V. Le. "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks." *Proceedings of the International Conference on Machine Learning (ICML)*, 2019, pp. 6105–6114. arXiv:1905.11946.

[12] R. Cogranne, Q. Giboulot, P. Bas. "ALASKA#2: Challenging Academic Research on Steganalysis with Realistic Images." *IEEE International Workshop on Information Forensics and Security (WIFS)*, 2020, pp. 1–5. doi:10.1109/WIFS49906.2020.9360896.

[13] phantasm v0.2–v0.3 cost-function research log (`archive/ML_STEGANALYSIS.md`, Updates 1–8). `https://github.com/exec/phantasm/blob/main/archive/ML_STEGANALYSIS.md`.

[14] phantasm v1.0.1 README. `https://github.com/exec/phantasm/blob/main/README.md`.

[15] P. Bas, T. Filler, T. Pevný. "'Break Our Steganographic System': The Ins and Outs of Organizing BOSS." *Information Hiding* (Lecture Notes in Computer Science, vol. 6958), 2011, pp. 59–70. doi:10.1007/978-3-642-24178-9_5.

[16] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei. "ImageNet: A large-scale hierarchical image database." *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009, pp. 248–255. doi:10.1109/CVPR.2009.5206848.

[17] Y. Yousfi, J. Butora, E. Khvedchenya, J. Fridrich. "ImageNet Pre-trained CNNs for JPEG Steganalysis." *IEEE International Workshop on Information Forensics and Security (WIFS)*, 2020, pp. 1–6. doi:10.1109/WIFS49906.2020.9360897.

[18] Y. Yousfi, J. Butora, J. Fridrich, C. Fuji Tsang. "Improving EfficientNet for JPEG Steganalysis." *Proceedings of the 9th ACM Workshop on Information Hiding and Multimedia Security (IH&MMSec)*, 2021, pp. 149–157. doi:10.1145/3437880.3460397.

[19] J. Butora, J. Fridrich. "Reverse JPEG Compatibility Attack." *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 1444–1454, 2020. doi:10.1109/TIFS.2019.2940904.

[20] É. Levecque, J. Butora, P. Bas. "Finding Incompatible Blocks for Reliable JPEG Steganalysis." *IEEE Transactions on Information Forensics and Security*, 2024. doi:10.1109/TIFS.2024.3470650.

[21] W. Tang, S. Tan, B. Li, J. Huang. "Automatic Steganographic Distortion Learning Using a Generative Adversarial Network." *IEEE Signal Processing Letters*, vol. 24, no. 10, pp. 1547–1551, 2017. doi:10.1109/LSP.2017.2745572.

[22] W. Tang, B. Li, S. Tan, M. Barni, J. Huang. "CNN-Based Adversarial Embedding for Image Steganography." *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 8, pp. 2074–2087, 2019. doi:10.1109/TIFS.2019.2891237.

[23] J. Yang, D. Ruan, J. Huang, X. Kang, Y.-Q. Shi. "An Embedding Cost Learning Framework Using GAN." *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 839–851, 2020. doi:10.1109/TIFS.2019.2922229.

[24] D. Lerch-Hostalot. "Aletheia: an open-source toolbox for steganalysis." *Journal of Open Source Software*, vol. 9, no. 93, art. 5982, 2024. doi:10.21105/joss.05982.