RESIDUAL TRANSFERABILITY IN NEURAL IMAGE WATERMARKING

Anonymous authors
Paper under double-blind review

ABSTRACT

Neural image watermarks can be forged by extracting watermark-bearing residuals from released images and transferring them to unrelated content. While prior work has demonstrated this vulnerability, what makes these residuals transferable remains poorly understood. We formalize this vulnerability with **residual transferability (RT)**, a metric that quantifies how well watermark evidence remains decodable after transfer across unrelated images. Through comparative analyses and controlled interventions, we find that common training-side variations do not account for the large RT differences across watermarking systems; instead, architectural design plays a central role. By contrasting high- and low-RT systems and validating their architectural differences through controlled interventions, we identify two mechanisms that strengthen the dependence of watermark evidence on the cover image, thereby suppressing the residual transferability. These findings provide concrete design guidance for developing more forgery-resistant watermarking architectures. Complementarily, for existing watermarking systems where architectural redesign is impractical, we introduce **CoverLock**, a plug-and-play strategy for existing watermarking systems that strengthens such image dependence without architectural redesign. Across representative watermarking systems exhibiting high residual transferability, CoverLock achieves a more favorable security–robustness trade-off than both traditional handcrafted defenses and learned classifier-based defenses.

1 INTRODUCTION

The growing prevalence of AI-generated images has made invisible watermarking (Zhu et al., 2018; Zhang et al., 2019; Fernandez et al., 2023; Lu et al., 2025; Souček et al., 2025) increasingly important for content provenance and attribution. By embedding imperceptible message-carrying perturbations, watermarking systems enable ownership verification and source attribution through subsequent decoding. Major technology companies, including Google (SynthID) (Deepmind, 2023), Meta (Meta, 2026), OpenAI (OpenAI), and Amazon (Amazon, 2024), have already integrated watermarking solutions into their generative AI ecosystems to improve transparency, trace harmful or misleading AI-generated content, and support content accountability.

However, recent studies have shown that current image watermarking systems remain vulnerable to security threats. Among these threats, residual-based forgery attack (Müller et al., 2024; Dong et al., 2025; Souček et al., 2026) has drawn increasing attention, as attackers can transfer a watermark from legitimate images to unauthorized content and induce false attribution. Recent studies further show that such attacks may require only one or a few watermarked images (Yang et al., 2024a; Souček et al., 2026), substantially reducing the barrier to forgery and challenging the reliability of watermark-based attribution. Such attacks recover watermark-bearing residuals through different estimation procedures, averaging a collection of watermarked images with natural-image references (Yang et al., 2024a) or using an auxiliary reference model (Souček et al., 2026).

Despite their demonstrated effectiveness and low attack cost, these studies primarily focus on constructing increasingly effective forgery attacks, while the intrinsic property that enables such cross-image transfer remains poorly understood. Elucidating the mechanisms underpinning the success of such attacks is essential for both developing principled defenses and informing the design of more secure watermarking systems.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Table 1: Comparison of representative watermarking methods in terms of training configurations, architectural designs, and residual transferability. Lower RT indicates a lower risk of residual-based watermark forgery. The computation of RT@100 is detailed in Sec. 3.1.

Method	Loss	Disc.	Noise	Arch.	Dataset	Resolution	Message Len.	RT
CIN (Ma et al., 2022)	MSE+MSE	✗	✓	INN	COCO	128	30	1.00
VINE (Lu et al., 2025)	BCE+MSE+LPIPS	✓	✓	CNN	OpenImages	512	100	0.99
MBRS (Jia et al., 2021)	MSE+MSE	✓	✓	CNN	ImageNet	128	256	0.98
HiDDeN (Zhu et al., 2018)	MSE+MSE	✓	✓	CNN	COCO	128	30	**0.28**
RivaGAN (Zhang et al., 2019)	BCE+Feature	✓	✓	CNN	Hollywood2	256	32	**0.08**

To this end, we first formalize residual transferability (RT) as an inherent property of watermarking systems: watermark residuals may remain decodable after being transferred across unrelated cover images. We then quantify this property using the RT metric, which measures a system's susceptibility to residual-based forgery independently of any specific residual estimation procedure. As summarized in Table 1, we select five representative watermarking schemes spanning different training configurations and architectural designs, and observe strikingly different levels of RT. Even methods trained with similar objectives and robustness-oriented augmentations can exhibit substantially different transferability. These observations motivate the following question:

What makes watermark residuals transferable across unrelated images?

To answer this question, we first conduct a comparative behavioral analysis across all evaluated watermarking systems, examining how watermark information is encoded and represented by their encoders and decoders. This helps us narrow down the plausible explanations for RT. RT reveals a clear distinction in the relationship between watermark information and image content: systems with high RT encode and decode watermark signals with limited reliance on the host image, whereas systems with low RT retain a strong coupling between watermark evidence and image content. This observation points to a possible shortcut mechanism: the model learns to encode watermark information through a largely *cover-independent pathway*, which in turn makes the residual transferable across unrelated images.

This interpretation naturally narrows the search space for understanding why some systems exhibit low RT. Through controlled interventions on low-RT systems, we identify two architectural mechanisms that strengthen the dependence between watermark information and image content, thereby substantially reducing residual transferability. These findings reveal concrete architectural principles for suppressing cover-independent watermark pathways and improving resistance to residual-based forgery.

Architectural redesign, however, can be method-specific and costly for existing watermarking systems. As a complementary option, we introduce **CoverLock**, a lightweight image-binding wrapper that incorporates image-dependent visual information into the decoding process without redesigning the core watermark architecture. Across three high-RT systems and both existing residual-based forgery attacks, CoverLock achieves a more favorable security–robustness trade-off than prior defenses. Our main contributions are summarized as follows.

- We introduce residual transferability (RT), a system-level property that measures the cross-image decodability of watermark residuals and characterizes susceptibility to residual-based forgery independent of the specific residual estimation method.
- Through controlled studies of training and architectural factors, we identify model architecture as a major factor governing RT. We further isolate two architectural design choices that substantially reduce residual transferability.
- As a complementary option for existing systems, we introduce **CoverLock**, an image-binding wrapper that mitigates RT without redesigning the watermark architecture. Across three high-RT systems, CoverLock achieves a more favorable security–robustness trade-off than prior defenses, reducing average forgery success to below 0.25% under both existing residual-based forgery attacks while attaining higher robustness.

2 Background and Problem Formulation

2.1 Image Watermarking

Given an image $x \in \mathbb{R}^{H\times W\times 3}$ of height H and width W, and a L-bit binary message $m \in \{0,1\}^L$, a watermark encoder E embeds m into x, while a decoder D recovers the message from the resulting watermarked image:

$$x^w = E(x, m), \qquad m' = D\big(x^w\big). \tag{1}$$

Here, x^w denotes the watermarked image, m' is the recovered message. During training, the encoder and decoder are typically optimized jointly using

$$\mathcal{L} = \lambda_{\mathrm{wm}}\ell_{\mathrm{wm}}\big(D(\tau(x^w)), m\big) + \lambda_{\mathrm{img}}\ell_{\mathrm{img}}(x^w, x), \tag{2}$$

where ℓ_{wm} and ℓ_{img} denote the message-recovery and image-fidelity losses, respectively, λ_{wm} and λ_{img} are their corresponding weights, and $\tau \sim \mathcal{T}$ denotes a sampled transformation used for robustness training (Zhu et al., 2018; Satheesh et al., 2026).

2.2 Residual-based Watermark Forgery

Residual-based watermark forgery (Yang et al., 2024a; Souček et al., 2026) aims to make an unrelated image appear to carry the watermark of released watermarked images. Given s reference watermarked images $\{x^w_{r,i}\}_{i=1}^s$ carrying the same message m, the attacker recovers approximate clean counterparts $\{\hat{x}_{r,i}\}_{i=1}^s$ and aggregates the estimated watermark residuals:

$$\hat{\bar{r}}_s = \frac{1}{s}\sum_{i=1}^{s}\left(x^w_{r,i} - \hat{x}_{r,i}\right). \tag{3}$$

The aggregated residual is then transferred to an unrelated target image x_t to construct a forged image:

$$x_f = \mathrm{clip}_{[0,1]}\left(x_t + \hat{\bar{r}}_s\right), \tag{4}$$

where $\mathrm{clip}_{[0,1]}(\cdot)$ clips pixel values to the valid image intensity range. During verification, the decoder may recognize x_f as carrying the same watermark as the reference images, leading to false ownership or provenance attribution. Existing residual-based forgery attacks mainly differ in how they reconstruct the clean counterparts or directly estimate the watermark residuals.

2.3 Threat Model

Adversarial goal. The adversary seeks to make an unrelated target image be verified as carrying the same watermark message as one or more released watermarked reference images. The forged image should preserve the semantic content and visual quality of the target image, while being falsely attributed to the source represented by the target watermark.

Adversary capabilities. Our primary threat model follows practical residual-based forgery attacks (Souček et al., 2026). The adversary has access to one or a few released watermarked images carrying the same target message and may use external datasets or pretrained models to estimate transferable watermark evidence. The adversary does not know the deployed watermarking method or any protection mechanism, has no access to the parameters of the watermark encoder or decoder, and cannot query the encoder to generate additional watermarked images.

3 Understanding Residual Transferability

In this section, we systematically study residual transferability in neural image watermarking systems. First, we introduce an oracle transfer protocol that measures this property while removing the confounding effect of attack-specific residual estimation. Second, we conduct a comparative behavioral analysis across watermarking systems exhibiting different levels of residual transferability, from both the encoder and decoder perspectives. Guided by these observations, we then turn to systems with low residual transferability and perform targeted architectural interventions to identify the design choices underlying their low-transferability behavior.

3.1 MEASURING RESIDUAL TRANSFERABILITY

To measure watermark residual transferability without confounding from a particular residual-estimation attack, we use ground-truth (GT) residuals. For each message m_k from a set of M messages, we construct an oracle residual from s source images:

$$\bar{r}_s^{(k)} = \frac{1}{s} \sum_{i=1}^{s} \left(x_{r,i}^{w,(k)} - x_{r,i}^{(k)} \right), \tag{5}$$

where $x_{r,i}^{w,(k)}$ is obtained by embedding message m_k into the clean source image $x_{r,i}^{(k)}$.

We then transfer each residual $\bar{r}_s^{(k)}$ to a disjoint set of T target images $\{x_{t,j}\}_{j=1}^{T}$ and compute the average transfer accuracy over both messages and target images:

$$\bar{a}_s = \frac{1}{MT} \sum_{k=1}^{M} \sum_{j=1}^{T} \mathrm{BitAcc} \left(D\Big(\mathrm{clip}_{[0,1]} \left(x_{t,j} + \bar{r}_s^{(k)} \right) \Big), m_k \right). \tag{6}$$

Accordingly, we define residual transferability as

$$\mathrm{RT}@s = \max\left(2\bar{a}_s - 1,\, 0\right). \tag{7}$$

where $\mathrm{RT}@s$ provides a normalized measure of residual transferability in $[0, 1]$, with 0 corresponding to random guessing and larger values indicating stronger cross-image transfer. Using GT residuals removes the dependence on any particular residual-estimation procedure, allowing $\mathrm{RT}@s$ to more directly characterize transferability inherent to the watermarking system itself. Unless otherwise specified, we use $s = 100$ source images, $T = 100$ target images, and $M = 5$ messages in the following analysis. We additionally examine the effect of different source counts s in Appendix D.

3.2 BEHAVIORAL SIGNATURES OF HIGH RESIDUAL TRANSFERABILITY

Before examining the underlying design factors, we first characterize how high- and low-RT watermarking systems differ in their learned behavior from two complementary perspectives: watermark embedding and watermark decoding, as summarized in Figure 1.

Figure 1(a) visualizes residuals averaged over groups of five cover images. Averaging suppresses cover-dependent residual components and makes structures that persist across different covers easier to observe. High-RT methods exhibit substantially more consistent residual patterns across different image groups. This behavior is consistent with the presence of a stronger cover-agnostic component in the embedded watermark signal, whereas low-RT methods exhibit residuals that vary more strongly with the underlying cover.

We next examine the representations formed during decoding. Figure 1(b) compares the relative feature variation induced by the cover image and the embedded message at the decoder's final feature-extraction layer. Specifically, we measure cover-induced variation by fixing the message and varying the cover image, and message-induced variation by fixing the cover image and varying the message. The corresponding feature variances are then normalized to sum to 100%. High-RT decoders exhibit a substantially smaller relative contribution from cover variation, whereas low-RT systems retain a much larger cover-induced component. This contrast is consistent with a stronger dependence on the cover image during decoding in low-RT systems.

3.3 WHAT PREVENTS RESIDUAL TRANSFERABILITY?

The behavioral analysis above reveals a common signature of high RT, but does not explain which design choices prevent cover-agnostic watermark pathways from emerging. We therefore turn to two low-RT systems as positive cases and ask what keeps their watermark evidence dependent on the cover image.

By comparing the architectural differences across watermarking systems and ruling out common training-side factors as sufficient explanations for the RT gap (see Appendix E), we identify a distinctive design in HiDDeN that maintains spatial image–message interactions throughout the pipeline.

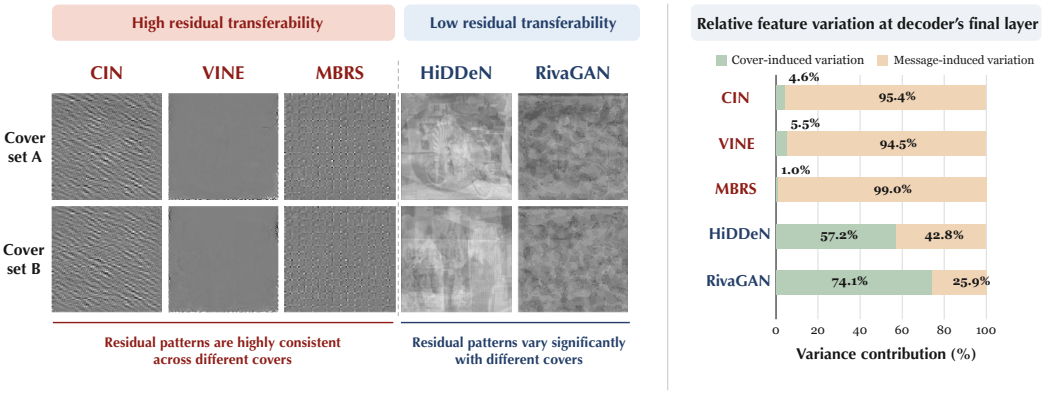


Figure 1: Behavioral signatures of high- and low-RT watermarking methods. **Left:** Average residual patterns computed from two different cover sets, each containing five images. **Right:** Relative feature variation induced by the cover image and the embedded message at the decoder's final layer.

During embedding of HiDDeN, the message is spatially broadcast and fused with local image features, while the decoder preserves spatial representations until a late global average pooling stage. We refer to this co-design as *Broadcast–GAP (BG)*. As shown in the left panel of Figure 2, removing BG from HiDDeN sharply increases RT, whereas introducing this design into high-RT watermarks substantially reduces RT.[1] These interventions identify BG as a design that promotes stronger dependence between watermark evidence and image content.

RivaGAN achieves low RT through a different mechanism. Its encoder contains content-adaptive attention that modulates the embedding pattern according to the input image. Replacing this adaptive mechanism with a fixed embedding pattern causes a large increase in RT, as shown in the left panel of Figure 2. This intervention identifies content adaptivity as another mechanism that binds the watermark signal to the cover image. The complete numerical results for the architectural interventions shown in the left panel are reported in Appendix Table 5.

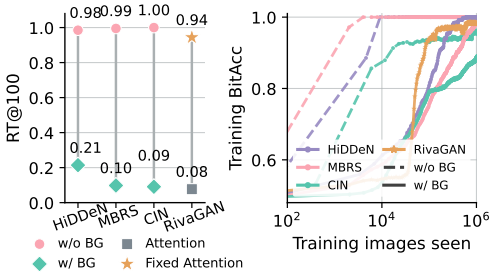


Figure 2: **Left:** RT under controlled architectural interventions with Broadcast–GAP (BG) and content-adaptive attention. **Right:** Training dynamics of high- and low-RT configurations.

The right panel of Figure 2 provides complementary evidence from training dynamics. High-RT configurations reach high training decoding accuracy after seeing relatively few images, consistent with learning an easier cover-independent shortcut for message recovery. In contrast, configurations containing the identified low-RT mechanisms require substantially more image observations before converging. This behavior is consistent with a harder learning regime in which message recovery depends more strongly on image-conditioned representations and therefore requires exposure to a broader range of image statistics.

4 CoverLock

Our analysis suggests that reducing residual transferability (RT) requires binding the embedded payload to the content of covers. Directly introducing such a dependency into an existing watermark scheme, however, typically requires model-specific redesign and retraining. We therefore propose **CoverLock**, a watermark-agnostic plug-in that wraps the message with a stable, image-conditioned binary code. CoverLock is trained independently of the underlying watermarking system and can be attached to an existing system without modifying its architecture or parameters.

[1]We also attempted to directly introduce BG into VINE (Lu et al., 2025), but the modified model did not converge reliably given VINE's already challenging optimization.

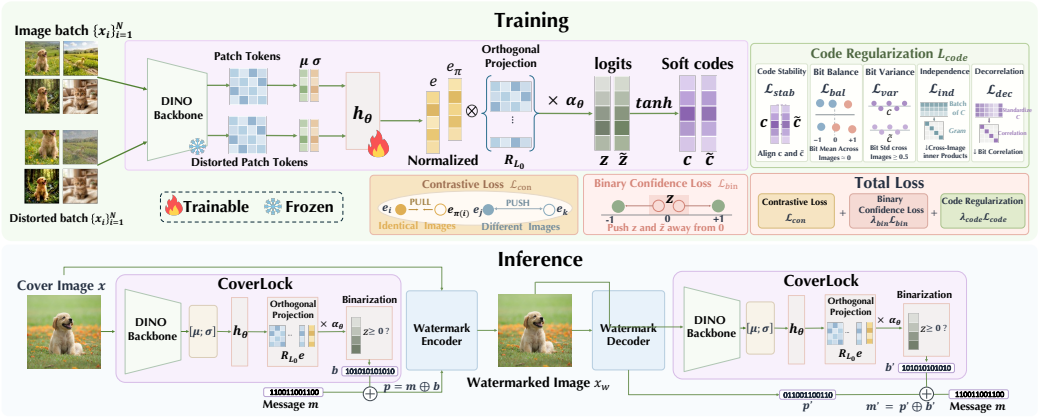


Figure 3: **Overview of CoverLock.** CoverLock derives a distortion-stable, image-conditioned binary code from frozen DINOv2 features and uses it to wrap and unwrap the watermark payload.

Plug-in image-conditioned wrapping. Let x denote a carrier image and $m \in \{0,1\}^L$ an L-bit message. CoverLock derives an image-conditioned binary code $b_\theta(x) \in \{0,1\}^L$ and uses it to wrap the message before watermark embedding:

$$p = m \oplus b_\theta(x), \qquad x^w = E(x, p), \tag{8}$$

where E denotes the original watermark encoder and $\oplus$ is bitwise XOR. At detection time, the original decoder first recovers the wrapped payload, while CoverLock recomputes the image-conditioned code from the received image:

$$\widehat{m} = D(\tau(x^w)) \oplus b_\theta(\tau(x^w)), \qquad \tau \sim \mathcal{T}. \tag{9}$$

Thus, the same user message is mapped to different embedded payloads depending on the carrier image, while the original watermark encoder and decoder remain unchanged.

Image-conditioned CoverLock hash. To generate a carrier-specific yet distortion-stable code, we extract dense patch features $\{q_p\}_{p=1}^P$ from the final layer of a frozen DINOv2 (Oquab et al., 2023) backbone. We summarize the patch features using their channel-wise mean and standard deviation, which capture complementary information about the global feature response and its spatial variation across image regions:

$$s(x) = [\mathrm{Mean}_p(q_p); \mathrm{Std}_p(q_p)]\,. \tag{10}$$

The resulting statistics are mapped by a lightweight trainable projector h_θ and ℓ_2-normalized to obtain the image embedding

$$e_\theta(x) = \mathrm{norm}_2\big(h_\theta(s(x))\big). \tag{11}$$

We then map the embedding to the hash space using a fixed orthogonal projection. Let $R_{L_0} \in \mathbb{R}^{D\times L_0}$ denote the base projection matrix, where $L_0 = 256$ and $R_{L_0}^\top R_{L_0} = I$. For a watermark system with payload length $L \le L_0$, we use a fixed subset of L projection directions, $R_L \in \mathbb{R}^{D\times L}$. The projected logits and corresponding soft codes are computed as

$$z_\theta(x) = \alpha_\theta R_{L_0}^\top e_\theta(x), \qquad c_\theta(x) = \tanh\big(z_\theta(x)\big), \tag{12}$$

where α_θ controls the scale of the projected logits. The binary code used for message wrapping is obtained by thresholding the logits:

$$b_\theta(x) = \mathbb{1}[z_\theta(x) \ge 0]. \tag{13}$$

The fixed orthogonal projection provides distinct decision directions for different bits without introducing trainable bit-specific predictors.

Watermark-agnostic contrastive training. CoverLock is trained solely from natural images in MSCOCO (Lin et al., 2014) and their transformed views, without involving the watermark encoder E, decoder D, or any forgery samples. For each training image x_i, we sample $\tau_i \sim \mathcal{T}$ and construct a transformed view $\widetilde{x}_i = \tau_i(x_i)$. The clean and transformed views of the same image form a positive pair, while views from different images act as negatives. Let $\{e_j\}_{j=1}^{2N}$ denote the normalized embeddings of the N clean images and their distorted counterparts, and let $\pi(j)$ denote the paired view of e_j. We optimize the symmetric contrastive objective

$$\mathcal{L}_{\text{con}} = -\frac{1}{2N}\sum_{j=1}^{2N}\log\frac{\exp(e_j^\top e_{\pi(j)}/\tau_c)}{\sum_{k\neq j}\exp(e_j^\top e_k/\tau_c)}, \tag{14}$$

where τ_c is the contrastive temperature. To further obtain stable and informative binary codes, we introduce a binary-confidence loss $\mathcal{L}_{\text{bin}}$ and a code regularizer $\mathcal{L}_{\text{code}}$. The overall objective is

$$\mathcal{L} = \mathcal{L}_{\text{con}} + \lambda_{\text{bin}}\mathcal{L}_{\text{bin}} + \lambda_{\text{code}}\mathcal{L}_{\text{code}}. \tag{15}$$

The transformation, individual code regularizers, and training details are provided in Appendix B.

5 Experiments

5.1 Setup

We evaluate CIN (Ma et al., 2022), VINE (Lu et al., 2025), and MBRS (Jia et al., 2021) as representative high-RT schemes identified as vulnerable in prior residual-based forgery studies (Souček et al., 2026). We consider two defense baselines. For classifier-based filtering, we train a ConvNeXt (Liu et al., 2022) binary classifier for each watermark family using genuine watermarked images and GT-residual forgeries. We select ConvNeXt because it achieves the strongest detection performance among the classification backbones we evaluated (see Appendix G). The second baseline follows MHDW (Lu et al., 2006), which, similar in spirit to our approach, binds watermark verification to image content through a robust media hash. In contrast, CoverLock constructs its image-dependent code through objectives jointly optimized for security and robustness.

Datasets. We evaluate all methods on three image datasets: DIV2K (Timofte et al., 2017), ImageNet (Russakovsky et al., 2015), and MS-COCO (Lin et al., 2014). These datasets have diverse image distributions and resolutions, allowing us to evaluate residual-based forgery across different visual domains. For each dataset, images used for residual estimation and those used as forgery targets are randomly sampled from two disjoint subsets.

Residual-based Forgery Attacks. We evaluate security against residual-based forgery attacks, where an attacker estimates the watermark residual from one or more released watermarked images and transfers it to an unrelated target image. We consider both single- and multi-reference settings: GT@1 uses one ground-truth residual, Yang et al. (2024a) uses 100 watermarked references, and WMForger (Souček et al., 2026) uses a single watermarked reference, following their original configurations unless otherwise specified.

Metrics. We report True Positive Rate (TPR) and Attack Success Rate (ASR) at the sample level, and Bit Accuracy (BitAcc) and Bit Error Rate (BER) at the bit level. TPR measures the fraction of legitimate watermarked images that are accepted, while ASR corresponds to the false-positive rate on forged images. BitAcc and BER measure the fractions of correctly and incorrectly decoded watermark bits, respectively. Verification thresholds are calibrated under the Bernoulli null model to FPR $< 10^{-6}$ for all watermark schemes. Higher TPR and BitAcc, and lower ASR and BER, are better.

5.2 Main Results of CoverLock

We first evaluate whether CoverLock can prevent residual-based watermark forgery across different watermarking systems, image domains, and attack methods.

Table 2 reports the ASR under GT@1, Yang et al. (2024a), and WmForger (Souček et al., 2026). CoverLock consistently achieves low ASR across all watermarking systems, datasets, and attacks,

Table 2: Forgery attack success rate (ASR, %) across evaluation datasets. Results are reported as mean ± std over three random seeds.

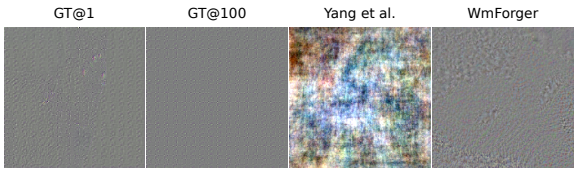
		GT@1 ↓			Yang et al. (2024a) ↓			WmForger (Souček et al., 2026) ↓		
Watermark	Dataset	ConvNeXt	MHDW	CoverLock	ConvNeXt	MHDW	CoverLock	ConvNeXt	MHDW	CoverLock
CIN	COCO	24.12 ±3.7	0.01 ±0.0	**0.00** ±0.0	36.16 ±2.6	2.17 ±0.6	**0.00** ±0.0	86.09 ±2.0	0.00 ±0.0	**0.00** ±0.0
	ImageNet	29.36 ±3.9	**0.00** ±0.0	0.03 ±0.0	36.37 ±1.7	3.19 ±0.9	**0.00** ±0.0	86.93 ±1.7	**0.00** ±0.0	**0.00** ±0.0
	DIV2K	18.72 ±4.6	**0.00** ±0.0	**0.00** ±0.0	29.94 ±1.9	4.72 ±0.6	**0.11** ±0.2	78.56 ±2.1	**0.00** ±0.0	**0.00** ±0.0
MBRS	COCO	2.41 ±1.4	11.37 ±4.4	**0.11** ±0.0	2.37 ±0.1	46.14 ±2.4	**0.00** ±0.0	96.28 ±0.1	4.76 ±1.7	**0.06** ±0.0
	ImageNet	1.01 ±0.6	10.81 ±5.0	**0.82** ±0.2	2.96 ±0.6	49.57 ±0.8	**0.00** ±0.0	91.78 ±1.3	3.68 ±1.6	**0.17** ±0.1
	DIV2K	**0.83** ±0.7	17.33 ±9.7	**0.83** ±0.1	2.44 ±0.2	55.11 ±0.9	**0.00** ±0.0	87.17 ±2.3	5.44 ±4.2	**0.28** ±0.2
VINE	COCO	14.44 ±4.7	1.16 ±0.4	**0.00** ±0.0	4.39 ±0.4	28.02 ±1.6	**0.00** ±0.0	22.34 ±4.1	0.06 ±0.1	**0.00** ±0.0
	ImageNet	14.52 ±4.2	0.49 ±0.3	**0.11** ±0.1	9.32 ±0.9	29.22 ±1.0	**0.02** ±0.0	25.66 ±2.7	0.06 ±0.1	**0.01** ±0.0
	DIV2K	4.61 ±1.2	1.56 ±1.0	**0.28** ±0.1	4.00 ±0.4	23.06 ±4.4	**0.00** ±0.0	13.61 ±2.8	0.06 ±0.1	**0.00** ±0.0
Average		12.226	4.747	**0.243**	14.216	26.800	**0.015**	65.379	1.560	**0.059**

with even its highest average ASR remaining below 0.25%. In contrast, the classifier-based baseline generalizes poorly across attacks, as the residual artifacts learned from GT-based forgeries can change substantially under different estimation procedures. This dependence on known attack patterns may lead to an iterative attacker–defender cycle. MHDW (Lu et al., 2006) also exhibits attack-dependent protection and becomes less effective when the attacker obtains cleaner residuals through stronger averaging or more accurate estimation. By making watermark verification explicitly dependent on image content, CoverLock avoids relying on attack-specific forgery signatures and provides more consistent protection across attack methods.

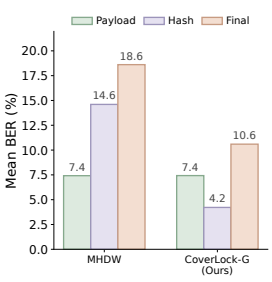
5.3 Analysis of CoverLock's Security and Robustness

We further analyze the behavior of different forgery defenses and the security–robustness trade-off of CoverLock.

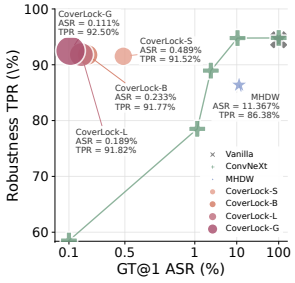
Figure 4(a) visualizes the residuals produced by different forgery procedures. Even for the same watermarking system, the residual patterns vary substantially with the estimation procedure and the number of source images. In particular, GT@1 and GT@100 exhibit different estimation quality, while Yang et al. (2024a) and WmForger (Souček et al., 2026) produce visually distinct residual structures. This diversity helps explain the limited cross-attack generalization of classifier-based defenses in Table 2, which rely on artifacts observed from known forgery patterns. Figure 4(b) decomposes the mean BER under distortions into payload and image-dependent hash components, while Table 3 reports the corresponding bit accuracy under each individual distortion. MHDW and CoverLock-G have the same payload BER of 7.4%, whereas CoverLock-G reduces the hash BER from 14.6% to 4.2%, resulting in a substantially lower final BER. This suggests that CoverLock's robustness advantage mainly comes from more reliable recovery of the image-dependent component rather than improved payload decoding. Figure 4(c) compares the resulting security–robustness trade-off for MBRS on COCO. For the classifier baseline, reducing ASR requires a substantial loss in TPR, while operating points with high TPR remain considerably more vulnerable to forgery. MHDW similarly provides a less



(a) Residual Patterns across Attacks



(b) BER Decomposition



(c) Security–Robustness

Figure 4: Attacks and CoverLock analysis.

Table 3: Bit accuracy (BitAcc, %) on COCO under Gaussian noise ($\sigma = 0.05$), brightness ($\times 1.2$), Gaussian blur ($\sigma = 1.0$), JPEG compression ($Q = 70$), contrast ($\times 1.2$), and center crop (80% area). Average is over the six distortions and bold indicates the best binding defense, excluding Vanilla.

Watermark	Method	Noise	Brightness	Blur	JPEG	Contrast	Crop	Average
CIN	Vanilla	99.98	99.92	100.00	96.28	100.00	100.00	99.36
	MHDW	76.49	91.08	79.83	82.19	91.74	83.20	84.09
	CoverLock	**93.35**	**97.13**	**95.58**	**91.52**	**97.45**	**92.55**	**94.60**
MBRS	Vanilla	69.16	98.52	94.13	100.00	99.06	94.41	92.55
	MHDW	59.92	88.57	83.81	90.30	89.43	76.32	81.39
	CoverLock	**66.43**	**96.16**	**91.62**	**96.65**	**96.75**	**88.82**	**89.40**
VINE	Vanilla	99.72	99.85	99.94	99.91	99.96	50.27	91.61
	MHDW	82.27	93.11	88.04	91.62	93.50	50.20	83.12
	CoverLock	**94.08**	**95.90**	**95.39**	**94.89**	**95.98**	**50.25**	**87.75**

favorable trade-off. In contrast, CoverLock variants remain concentrated in the low-ASR, high-TPR region. Increasing the DINOv2 backbone size from Small to Giant further reduces ASR while maintaining TPR at approximately 92%, with CoverLock-G lying on the favorable Pareto frontier.

6 Related work

Image Watermarking Existing image watermarking methods can be broadly categorized into *post-processing* and *in-generation* approaches. Post-processing methods embed watermarks into already generated images. Early approaches based on handcrafted embedding (e.g., LSB (Chopra et al., 2012), DWT-DCT (Al-Haj, 2007; Navas et al., 2008)) demonstrate feasibility but are vulnerable to common image transformations. Recent neural approaches improve robustness via learned encoder–decoder architectures with adversarial objectives (Zhu et al., 2018; Jia et al., 2021; Tancik et al., 2020; Zhang et al., 2019; Ma et al., 2022; Fang et al., 2022; Lu et al., 2025; Sander et al., 2025; Souček et al., 2025). These methods are compatible with existing generative systems without architectural modifications. In contrast, in-generation watermarking incorporates watermark signals directly during image synthesis rather than embedding them into already generated images (Yang et al., 2024b; Wen et al., 2023; Fernandez et al., 2023; Arabi et al., 2025; Wang et al., 2025; Chen et al., 2025; Lee & Cho, 2025; Liu et al., 2025). In this work, we focus on post-processing methods due to their model-agnostic and deployment-friendly properties.

Forgery Attack and Defense. Watermark forgery attacks estimate transferable watermark evidence from one or a few watermarked images and apply it to unrelated covers (Kutter et al., 2000; Yang et al., 2024a; Souček et al., 2026). Dong et al. (2025) instead assumes access to a large collection of watermarked images with the same message, which is outside our few-shot threat model. Prior defenses either bind watermarks to image content (Lu et al., 2006) or improve forgery resistance for in-generation watermarking by intervening in the diffusion process, e.g., conditioning latent watermarks on image semantics, modifying the denoising trajectory, or constraining intermediate latents (Zhang et al., 2026; Liu et al., 2026; Lai et al., 2026). In contrast, we focus on strengthening cover dependence for post-processing neural watermarking systems.

7 Conclusion

In this work, we study watermark forgery through *residual transferability* (RT), revealing that many deep watermarking systems can develop cover-agnostic watermark pathways whose signals transfer across unrelated images. By analyzing low-RT systems, we further identify coupled spatial interaction and content-adaptive embedding as two architectural mechanisms associated with stronger cover dependence. We also introduce CoverLock, a complementary solution for existing systems without redesigning or retraining the underlying watermark model. Across different watermarking systems, datasets, and attacks, CoverLock achieves a more favorable balance between forgery resistance and robustness than prior media-hash-based defenses, further highlighting cover dependence as an important design principle for secure deep image watermarking.

AI USE STATEMENT

LLM tools were used to assist with manuscript writing and editing, as well as with minor modifications to Python utility scripts, such as plotting and visualization scripts. All generated content and code modifications were reviewed and verified by the authors, who take full responsibility for the final content of this work.

ETHICS STATEMENT

This work analyzes watermark forgery attacks introduced in prior work, with the goal of understanding their underlying mechanisms and developing effective defenses. Our study focuses on security analysis and on deriving design insights for more forgery-resistant image watermarking systems.

REPRODUCIBILITY STATEMENT

We provide detailed experimental settings, model configurations, training procedures, evaluation protocols, and additional ablation results in the appendices. The datasets and pretrained watermarking models used in our evaluation are publicly available. We will release the implementation of our CoverLock to facilitate reproduction of the reported results.

REFERENCES

Ali Al-Haj. Combined dwt-dct digital image watermarking. *Journal of computer science*, 3(9): 740–746, 2007.

Amazon. Watermark detection for amazon titan image generator now available in amazon bedrock. https://aws.amazon.com/cn/about-aws/whats-new/2024/04/watermark-detection-amazon-titan-image-generator-bedrock/, 2024.

Kasra Arabi, R. Teal Witter, Chinmay Hegde, and Niv Cohen. SEAL: Semantic aware image watermarking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16196–16205, 2025.

Yuzhuo Chen, Zehua Ma, Han Fang, Weiming Zhang, and Nenghai Yu. TAG-WM: Tamper-aware generative image watermarking via diffusion inversion sensitivity. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16723–16732, 2025.

Deepshikha Chopra, Preeti Gupta, Gaur Sanjay, and Anil Gupta. Lsb based digital image watermarking for gray scale image. *IOSR Journal of Computer Engineering*, 6(1):36–41, 2012.

Google Deepmind. Synthid: Identifying ai-generated content with synthid. https://deepmind.google/technologies/synthid/, 2023.

Ziping Dong, Chao Shuai, Zhongjie Ba, Peng Cheng, Zhan Qin, Qinglong Wang, and Kui Ren. Wmcopier: Forging invisible image watermarks on arbitrary images. *arXiv preprint arXiv:2503.22330*, 2025.

Han Fang, Zhaoyang Jia, Zehua Ma, Ee-Chien Chang, and Weiming Zhang. PIMoG: An Effective Screen-shooting Noise-Layer Simulation for Deep-Learning-Based Watermarking Network. In *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 2267–2275, Lisboa Portugal, October 2022. ACM.

Pierre Fernandez, Guillaume Couairon, Hervé Jégou, Matthijs Douze, and Teddy Furon. The stable signature: Rooting watermarks in latent diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22466–22477, 2023.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. URL https://arxiv.org/abs/1512.03385.

Zhaoyang Jia, Han Fang, and Weiming Zhang. Mbrs: Enhancing robustness of dnn-based watermarking by mini-batch of real and simulated jpeg compression. In *Proceedings of the 29th ACM international conference on multimedia*, pp. 41–49, 2021.

Martin Kutter, Sviatoslav V Voloshynovskiy, and Alexander Herrigel. Watermark copy attack. In *Security and Watermarking of Multimedia Contents II*, volume 3971, pp. 371–380. SPIE, 2000.

Jiewei Lai, Lan Zhang, Chen Tang, Pengcheng Sun, Zhaopeng Zhang, Yunhao Wang, and Hui Jin. Csguard: Toward forgery-resistant watermarking in diffusion models via compressed sensing constraint. *arXiv preprint arXiv:2605.01479*, 2026.

Sung Ju Lee and Nam Ik Cho. Semantic watermarking reinvented: Enhancing robustness and generation quality with fourier integrity. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 18759–18769, 2025.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pp. 740–755. Springer, 2014.

Qingyu Liu, Yitao Zhang, Zhongjie Ba, Chao Shuai, Peng Cheng, Tianhang Zheng, and Zhibo Wang. Attack-resistant watermarking for aigc image forensics via diffusion-based semantic deflection. In *International Conference on Learning Representations*, 2026.

Zhenguang Liu, Chao Shuai, Shaojing Fan, Ziping Dong, Jinwu Hu, Zhongjie Ba, and Kui Ren. Harnessing frequency spectrum insights for image copyright protection against diffusion models, 2025. URL https://arxiv.org/abs/2503.11071.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s, 2022. URL https://arxiv.org/abs/2201.03545.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019.

Chun-Shien Lu, Shih-Wei Sun, Chao-Yong Hsu, and Pao-Chi Chang. Media hash-dependent image watermarking resilient against both geometric attacks and estimation attacks based on false positive-oriented detection. *IEEE Transactions on Multimedia*, 8(4):668–685, August 2006. doi: 10.1109/TMM.2006.876300.

Shilin Lu, Zihan Zhou, Jiayou Lu, Yuanzhi Zhu, and Adams Kong. Robust watermarking using generative priors against image editing: From benchmarking to advances. In *International Conference on Learning Representations*, volume 2025, pp. 83902–83936, 2025.

Rui Ma, Mengxi Guo, Yi Hou, Fan Yang, Yuan Li, Huizhu Jia, and Xiaodong Xie. Towards Blind Watermarking: Combining Invertible and Non-invertible Mechanisms. In *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 1532–1542, Lisboa Portugal, October 2022. ACM. doi: 10.1145/3503161.3547950.

Meta. Content seal: State-of-the-art invisible watermarking. https://facebookresearch.github.io/content-seal/, 2026. Accessed: 2026-07-30.

Andreas Müller, Denis Lukovnikov, Jonas Thietke, Asja Fischer, and Erwin Quiring. Black-box forgery attacks on semantic watermarks for diffusion models. *arXiv preprint arXiv:2412.03283*, 2024.

K. A. Navas, Mathews Cheriyan Ajay, M. Lekshmi, Tampy S. Archana, and M. Sasikumar. DWT-DCT-SVD based watermarking. In *2008 3rd International Conference on Communication Systems Software and Middleware and Workshops (COMSWARE '08)*, pp. 271–274. IEEE, January 2008.

OpenAI. Advancing content provenance for a safer, more transparent ai ecosystem. URL https://openai.com/index/advancing-content-provenance/.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Russell, Armand Joulin, Piotr Bojanowski, and Mathilde Caron. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2023.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.

Tom Sander, Pierre Fernandez, Alain Oliviero Durmus, Teddy Furon, and Matthijs Douze. Watermark anything with localized messages. In *International Conference on Learning Representations*, 2025.

Anirudh Satheesh, Michael-Andrei Panaitescu-Liess, Andrew Xu, Georgios Milis, Heng Huang, Zikui Cai, and Furong Huang. Compositional adversarial training for robust visual watermarking, 2026. URL https://arxiv.org/abs/2605.16720.

Tomáš Souček, Sylvestre-Alvise Rebuffi, Pierre Fernandez, Nikola Jovanović, Hady Elsahar, Valeriu Lacatusu, Tuan Tran, and Alexandre Mourachko. Transferable black-box one-shot forging of watermarks via image preference models. *Advances in Neural Information Processing Systems*, 38:50251–50274, 2026.

Tomáš Souček, Pierre Fernandez, Hady Elsahar, Sylvestre-Alvise Rebuffi, Valeriu Lacatusu, Tuan Tran, Tom Sander, and Alexandre Mourachko. Pixel seal: Adversarial-only training for invisible image and video watermarking. *arXiv preprint arXiv:2512.16874*, 2025.

Matthew Tancik, Ben Mildenhall, and Ren Ng. Stegastamp: Invisible hyperlinks in physical photographs. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2117–2126, 2020.

Radu Timofte, Eirikur Agustsson, Luc Van Gool, Ming-Hsuan Yang, Lei Zhang, , et al. NTIRE 2017 Challenge on Single Image Super-Resolution: Dataset and Study. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 114–122, July 2017.

Zilan Wang, Junfeng Guo, Jiacheng Zhu, Yiming Li, Heng Huang, Muhao Chen, and Zhengzhong Tu. Sleepermark: Towards robust watermark against fine-tuning text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8213–8224, 2025.

Yuxin Wen, John Kirchenbauer, Jonas Geiping, and Tom Goldstein. Tree-ring watermarks: Fingerprints for diffusion images that are invisible and robust. *arXiv preprint arXiv:2305.20030*, 2023.

Pei Yang, Hai Ci, Yiren Song, and Mike Zheng Shou. Can simple averaging defeat modern watermarks? *Advances in Neural Information Processing Systems*, 37:56644–56673, 2024a.

Zijin Yang, Kai Zeng, Kejiang Chen, Han Fang, Weiming Zhang, and Nenghai Yu. Gaussian Shading: Provable Performance-Lossless Image Watermarking for Diffusion Models, May 2024b. Comment: 17 pages, 11 figures, accepted by CVPR 2024.

Kevin Alex Zhang, Lei Xu, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. Robust invisible video watermarking with attention. *arXiv preprint arXiv:1909.01285*, 2019.

Xin Zhang, Zijin Yang, Kejiang Chen, Linfeng Ma, Weiming Zhang, and Nenghai Yu. Sembind: Binding diffusion watermarks to semantics against black-box forgery attacks. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026.

Jiren Zhu, Russell Kaplan, Justin Johnson, and Li Fei-Fei. Hidden: Hiding data with deep networks. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 657–672, 2018.

A LIMITATION

Although CoverLock achieves low forgery success and a favorable balance between forgery resistance and robustness under the distortions considered in our evaluation, this balance may change under more severe or unseen image transformations. Its behavior under broader distribution shifts and adaptive attacks specifically targeting the image-binding mechanism also remains to be studied. CoverLock provides a lightweight way to augment existing watermarking systems without redesigning their core architectures, but introduces an additional image-binding module and associated training and deployment overhead. Finally, while our analysis covers several representative neural watermarking architectures and identifies two mechanisms associated with low RT, other architectures may achieve cover dependence through different design principles.

B COVERLOCK IMPLEMENTATION AND TRAINING DETAILS

B.1 ARCHITECTURE AND ORTHOGONAL HASHING

Frozen feature extractor. CoverLock uses a frozen DINOv2 ViT (Oquab et al., 2023) as the image feature extractor. Unless otherwise specified, we use ViT-g/14; the CoverLock-S/B/L/g variants replace it with the corresponding DINOv2 Small, Base, Large, and Giant backbones while leaving the remaining architecture unchanged. All images are resized to 224×224, and patch tokens are extracted from the final transformer layer. This yields a 16×16 grid of $P = 256$ patch features,

$$Q(x) = \{q_p\}_{p=1}^{P}, \qquad q_p \in \mathbb{R}^d, \tag{16}$$

where d depends on the selected backbone ($d = 1536$ for ViT-g/14).

We summarize the patch features using their channel-wise mean and standard deviation across spatial tokens

$$\mu(x) = \frac{1}{P}\sum_{p=1}^{P} q_p, \qquad \sigma(x) = \mathrm{Std}_p[q_p], \tag{17}$$

and concatenate them to form

$$s(x) = [\mu(x); \sigma(x)] \in \mathbb{R}^{2d}. \tag{18}$$

Projection head. The pooled representation is mapped to a D-dimensional image embedding using a lightweight trainable projection head,

$$h_\theta(s(x)) = W_2\,\mathrm{GELU}\left(W_1\,\mathrm{LN}(s(x)) + b_1\right) + b_2, \tag{19}$$

followed by ℓ_2 normalization,

$$e_\theta(x) = \frac{h_\theta(s(x))}{\|h_\theta(s(x))\|_2}. \tag{20}$$

We use a hidden dimension of 512 and an output embedding dimension $D = 512$.

Fixed orthogonal projection. CoverLock maintains a native code length of $L_0 = 256$. We first sample a Gaussian matrix

$$G \in \mathbb{R}^{D \times L_0} \tag{21}$$

using a fixed seed and orthogonalize its columns once using QR decomposition. This produces a fixed projection matrix

$$R_{L_0} \in \mathbb{R}^{D \times L_0}, \qquad R_{L_0}^\top R_{L_0} = I, \tag{22}$$

which is stored as a non-trainable buffer.

For a watermarking system with payload length $L \le L_0$, a seed-determined fixed permutation selects L columns from R_{L_0},

$$R_L = [r_{\pi(1)}, \ldots, r_{\pi(L)}] \in \mathbb{R}^{D \times L}. \tag{23}$$

The resulting hash logits are

$$z_\theta^{(L_0)}(x) = \alpha_\theta R_L^\top e_\theta(x), \tag{24}$$

where $\alpha_\theta > 0$ is a learned logit scale that controls the magnitude of the projected responses. Since the image embedding and projection directions are normalized, this scale allows the model to adapt the confidence of the continuous hash logits without changing their signs, facilitating the subsequent soft-code regularization and binary-margin objective. The corresponding binary image-dependent code is obtained by thresholding the logits,

$$b_\theta^{(L)}(x) = \mathbb{1}\left[z_\theta^{(L)}(x) \geq 0\right]. \tag{25}$$

B.2 Code Regularization and Binary Confidence

The contrastive objective introduced in the main paper structures the continuous image embedding $e_\theta(x)$ by bringing different views of the same carrier together while separating different carriers. In addition, we regularize the projected codes to encourage distortion stability, balanced and diverse bit usage, cross-carrier independence, and low inter-bit correlation. A separate binary-confidence objective further encourages the projected logits to remain away from the binarization threshold.

For a minibatch of N clean images $\{x_i\}_{i=1}^N$, we independently sample $\tau_i \sim \mathcal{T}$ and construct distorted views

$$\widetilde{x}_i = \tau_i(x_i). \tag{26}$$

During CoverLock training, the code-level objectives are applied to the native L_0-dimensional projection. We therefore define

$$z_i = z_\theta^{(L_0)}(x_i), \qquad \widetilde{z}_i = z_\theta^{(L_0)}(\widetilde{x}_i), \tag{27}$$

with the corresponding soft codes

$$c_i = \tanh(z_i), \qquad \widetilde{c}_i = \tanh(\widetilde{z}_i). \tag{28}$$

Code stability. Different views of the same carrier should produce consistent codes. We therefore minimize

$$\mathcal{L}_{\text{stab}} = \frac{1}{NL_0} \sum_{i=1}^{N} \|c_i - \widetilde{c}_i\|_2^2 . \tag{29}$$

Bit balance. To prevent individual bits from being biased toward one binary value over the training distribution, we encourage each bit to have zero mean across carriers:

$$\mathcal{L}_{\text{bal}} = \frac{1}{L_0} \left\| \frac{1}{N} \sum_{i=1}^{N} c_i \right\|_2^2 . \tag{30}$$

Bit variance. Bit balance alone does not prevent a bit from remaining close to zero for all carriers. We therefore encourage each bit to maintain sufficient variation across images:

$$\mathcal{L}_{\text{var}} = \frac{1}{2L_0} \sum_{b=1}^{L_0} \left[\max\left(0, \nu - \text{Std}_i[c_{ib}]\right) + \max\left(0, \nu - \text{Std}_i[\widetilde{c}_{ib}]\right) \right], \tag{31}$$

where $\nu = 0.5$.

Cross-carrier independence. Codes from different carriers should remain dissimilar. Let

$$C = [c_1^\top; \ldots; c_N^\top] \in \mathbb{R}^{N \times L_0}. \tag{32}$$

We penalize pairwise similarities between different carriers:

$$\mathcal{L}_{\text{ind}} = \text{mean}\left[\text{offdiag}\left(\frac{CC^\top}{L_0} \right)^2 \right]. \tag{33}$$

Bit decorrelation. Orthogonal projection directions do not necessarily guarantee uncorrelated outputs over the image distribution. We therefore standardize each bit over the minibatch,

$$\overline{c}_{ib} = \frac{c_{ib} - \mu_b}{\sigma_b + \epsilon}, \tag{34}$$

where μ_b and σ_b are the batch mean and standard deviation of bit b, and $\epsilon = 10^{-4}$ is used for numerical stability. Let $\overline{C} \in \mathbb{R}^{N \times L_0}$ denote the resulting standardized code matrix. We minimize its off-diagonal bit correlations:

$$\mathcal{L}_{\text{dec}} = \text{mean}\left[\text{offdiag}\left(\frac{\overline{C}^\top \overline{C}}{N}\right)^2\right]. \tag{35}$$

The complete code regularization objective is

$$\begin{aligned}\mathcal{L}_{\text{code}} = \lambda_{\text{stab}}\mathcal{L}_{\text{stab}} + \lambda_{\text{bal}}\mathcal{L}_{\text{bal}} + \lambda_{\text{var}}\mathcal{L}_{\text{var}} \\ + \lambda_{\text{ind}}\mathcal{L}_{\text{ind}} + \lambda_{\text{dec}}\mathcal{L}_{\text{dec}}.\end{aligned} \tag{36}$$

Binary confidence. Since the final image-dependent code is obtained by thresholding the projected logits at zero, logits close to the decision boundary are particularly susceptible to bit flips under image distortions. We therefore encourage both clean and distorted logits to remain away from zero:

$$\mathcal{L}_{\text{bin}} = \frac{1}{2NL_0}\sum_{i=1}^{N}\sum_{b=1}^{L_0}\left[\text{softplus}\left(\rho - |z_{ib}|\right) + \text{softplus}\left(\rho - |\widetilde{z}_{ib}|\right)\right], \tag{37}$$

where $\rho = 1$ denotes the binary margin.

We use

$$\begin{aligned}&\lambda_{\text{stab}} = 2, \quad &&\lambda_{\text{bal}} = 5, \quad &&\lambda_{\text{var}} = 5, \\ &\lambda_{\text{ind}} = 1, \quad &&\lambda_{\text{dec}} = 0.05, \quad &&\lambda_{\text{bin}} = 0.0125, \\ &\lambda_{\text{code}} = 0.25.\end{aligned} \tag{38}$$

Together with the contrastive objective defined in the main paper, the complete CoverLock training objective is

$$\mathcal{L} = \mathcal{L}_{\text{con}} + \lambda_{\text{bin}}\mathcal{L}_{\text{bin}} + \lambda_{\text{code}}\mathcal{L}_{\text{code}}. \tag{39}$$

B.3 Training Configuration

CoverLock is trained solely from natural images and their distorted versions. For each image x_i, the second view $\widetilde{x}_i$ is generated by sampling a transformation from the training distortion distribution $\mathcal{T}$, including JPEG compression (quality 45–90), Gaussian blur ($\sigma = 0.3$–2.0), Gaussian noise ($\sigma = 0.005$–0.02), Crop that retains 70%–95% of the image area, and brightness or contrast adjustment with factors in $[0.7, 1.3]$. We optimize only the lightweight projection head and the learnable logit scale using AdamW (Loshchilov & Hutter, 2019) with a learning rate of 3×10^{-5}. The DINOv2 backbone and orthogonal projection matrix remain fixed. We use a batch size of 256, contrastive temperature $\tau_c = 0.1$, and train the models reported in the main paper for 200k optimization steps. All CoverLock variants are trained on COCO train2017, which contains 118,287 images. We use 94,629 images (80%) for training and reserve the remaining 23,658 images for validation.

B.4 Training Dynamics of CoverLock

Figure 5 shows the optimization trajectories of the overall CoverLock objective and its individual components. The total, contrastive, and code-regularization losses decrease rapidly early in training and then improve more gradually, while the stability and balance objectives also decrease steadily. The independence and decorrelation losses exhibit a backbone-dependent late-stage rebound after their initial reduction; this rebound becomes less pronounced as the backbone size increases and is largely absent for CoverLock-G, suggesting that higher-capacity features may better accommodate these statistical code constraints during joint optimization. The variance and binary-confidence terms follow distinct non-monotonic trajectories as the projected codes evolve during training. Despite these different dynamics, all four variants remain well-behaved throughout optimization.

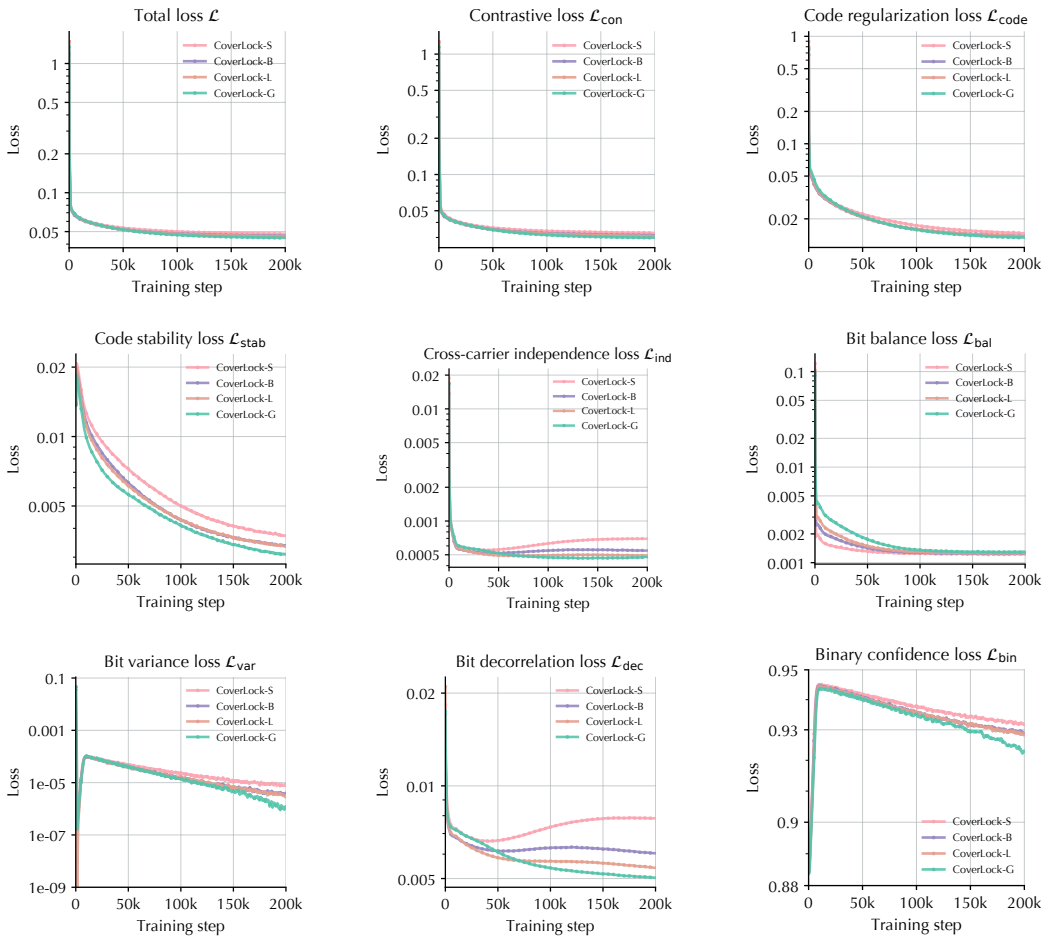


Figure 5: **Training dynamics of CoverLock.** Training curves of the overall objective and its individual loss components across four CoverLock variants. While the components exhibit different optimization trajectories, they eventually reach stable regimes under joint optimization.

C Apparent Signs of Image Dependence Can Be Misleading

As shown in Section 3.3, explicitly combining image and message features does not necessarily prevent a cover-agnostic watermark pathway from emerging. Similarly, an image-dependent-looking residual does not necessarily indicate true image dependence. A single residual can retain visible structures of the host image simply due to image-specific reconstruction errors introduced by the encoder, rather than watermark evidence used for decoding.

This motivates the residual averaging used in the main paper. Averaging residuals from images carrying the same message suppresses image-specific reconstruction errors while retaining watermark components shared across covers. As shown in Figure 6, averaging only five residuals substantially removes the apparent image content and more clearly reveals the shared structure of high-RT methods, providing a more reliable visualization of cover-agnostic watermark signals.

D Sensitivity to the Number of Source Images

As shown in Table 4, the overall RT patterns remain consistent across different source-set sizes. High-RT methods remain highly transferable, while HiDDeN and RivaGAN consistently exhibit substantially lower RT, indicating that the observed differences are not specific to a particular source-

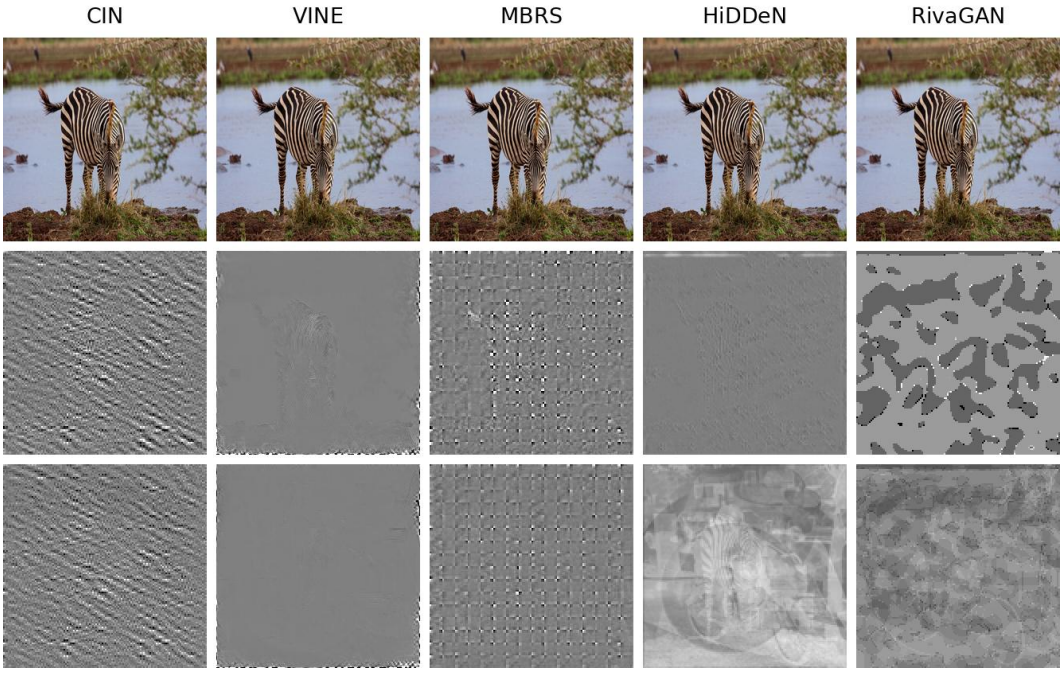


Figure 6: **Comparison of single-image and averaged watermark residuals.** From top to bottom, we show the watermarked image, the residual from a single image, and the residual averaged over five images carrying the same message for different watermarking methods.

Table 4: Sensitivity of residual transferability to the number of source images. RT@s is computed using the averaged residual from s source images.

Model	**BitAcc** ↑	**PSNR (dB)** ↑	**RT@1** ↓	**RT@10** ↓	**RT@100** ↓	**RT@1000** ↓
HiDDeN	0.924±0.05	33.85±1.87	0.374±0.20	0.301±0.16	0.275±0.16	0.275±0.16
CIN	1.000±0.00	41.55±0.90	1.000±0.00	1.000±0.00	1.000±0.00	1.000±0.00
MBRS	1.000±0.00	41.61±1.69	0.940±0.08	0.955±0.07	0.982±0.04	0.982±0.04
RivaGAN	0.919±0.08	42.05±0.16	0.081±0.11	0.088±0.11	0.082±0.11	0.085±0.11
VINE	1.000±0.00	35.47±2.83	0.964±0.05	0.981±0.04	0.986±0.03	0.985±0.04

set size. Considering both evaluation cost and the need to adequately reflect forgery risk, we use 100 source images as the default setting in our main analysis.

E Detailed Architectural Analysis

E.1 Common training-side choices

We test whether common training-side choices can explain the RT gap between HiDDeN (Zhu et al., 2018) and MBRS (Jia et al., 2021). We vary the training dataset, message length, loss weights and robustness-oriented noise simulation while keeping the remaining settings fixed. As shown in Figure 7, these changes shift the exact RT values but do not alter the qualitative behavior: HiDDeN consistently remains in the low-RT regime, whereas MBRS remains highly transferable. This persistence across training settings suggests that the large RT gap is unlikely to be explained by these factors alone, motivating our investigation of architectural differences.

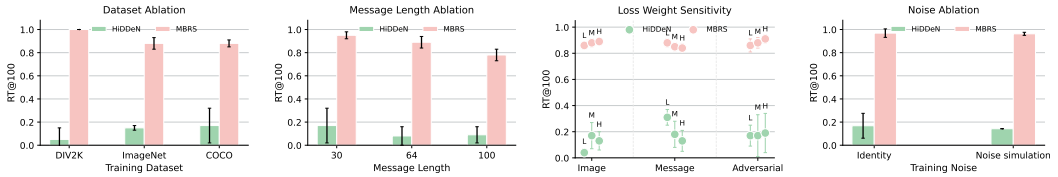


Figure 7: **Residual transferability (RT@100) under different training settings.** We vary the training dataset, message length, loss weight, and robustness-oriented noise simulation while keeping the remaining settings fixed. Results report the mean RT score over three independently trained models with different random seeds, with error bars indicating one standard deviation.

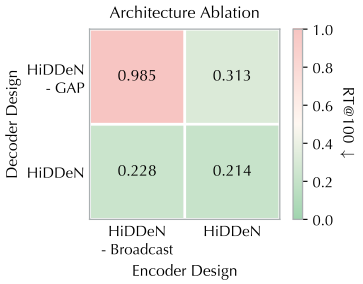


Figure 8: **Coupling between broadcast and GAP.** Removing both components sharply increases RT.

Table 5: Identifying architectural sources of low residual transferability. RT results are reported as mean ± standard deviation over three random seeds. Lower RT indicates stronger cover dependence.

Method	Variant	Watermark Quality		Residual Transferability		
		BitAcc ↑	PSNR ↑	RT@1 ↓	RT@10 ↓	RT@100 ↓
HiDDeN	Original	0.998	28.23	0.249±0.180	0.261±0.135	0.214±0.135
	− Broadcast–GAP	1.000	41.47	**0.909±0.084**	**0.977±0.057**	**0.985±0.049**
MBRS	Original	1.000	42.32	0.987±0.029	0.993±0.023	0.995±0.020
	+ Broadcast–GAP	0.964	27.07	**0.210±0.115**	**0.124±0.109**	**0.098±0.102**
CIN	Original	1.000	41.55	1.000±0.000	1.000±0.000	1.000±0.000
	+ Broadcast–GAP	0.934	21.71	**0.158±0.157**	**0.103±0.126**	**0.091±0.117**
VINE	Original	1.000	35.47	0.964±0.052	0.981±0.040	0.986±0.034
	+ Broadcast–GAP	0.600	6.97	*Training failure*		
RivaGAN	Original	0.977	42.12	0.071±0.100	0.073±0.103	0.077±0.109
	Fixed Attention	0.982	42.11	**0.922±0.107**	**0.946±0.094**	**0.944±0.096**

E.2 Broadcast and GAP as a Coupled Design

To clarify why we treat broadcast and global average pooling (GAP) as a unified Broadcast–GAP (BG) design rather than two independent architectural choices, we conduct a 2×2 ablation on HiDDeN. Specifically, we independently remove spatial message broadcast from the encoder and GAP from the decoder, while keeping all other components and training settings unchanged. Since the preceding analysis shows that robustness-oriented distortion training does not account for the qualitative RT gap, we omit simulated distortions in the following architectural intervention experiments to reduce optimization difficulty and enable cleaner controlled comparisons. All original and modified models in these experiments are retrained under the same distortion-free setting.

As shown in Figure 8, the original HiDDeN with both broadcast and GAP achieves an RT@100 of 0.214. Removing broadcast alone results in a nearly unchanged RT@100 of 0.228, while removing GAP alone increases it moderately to 0.313. In contrast, simultaneously removing both components causes RT@100 to rise sharply to 0.985. This pronounced interaction shows that the low-RT behavior cannot be attributed to either broadcast or GAP independently. Instead, the two components function as a coupled architectural design: broadcast spatially couples the message with local image features during embedding, while GAP requires the decoder to aggregate spatially distributed evidence for message recovery. We therefore refer to their combination as the *Broadcast–GAP (BG)* design throughout the paper.

F Verification Calibration under CoverLock

Watermark verification commonly sets its decision threshold using an independent Bernoulli null model. Since CoverLock introduces an additional image-dependent binary code into the verification representation, we examine whether it alters the null distribution and therefore requires threshold recalibration.

Figure 9 compares this theoretical distribution with empirical match counts from 50,000 held-out ImageNet images. Both vanilla MBRS and MBRS with CoverLock-G produce zero false accepts

and remain well aligned with the binomial model, with CoverLock-G showing particularly close agreement.

This behavior is consistent with CoverLock's statistical objectives in Section B.2. In the idealized case where the image-dependent hash b is uniform over $\{0,1\}^L$ and independent of the original decoded bits $\hat{p}$, the decoded message $m' = \hat{p} \oplus b$ is also uniform, yielding a $\mathrm{Binomial}(L, 0.5)$ match-count distribution. These objectives encourage the learned binary code to satisfy the statistical conditions underlying the Bernoulli null model, consistent with the close empirical agreement observed in Figure 9. Accordingly, the original verification threshold can be retained without additional recalibration after introducing CoverLock.

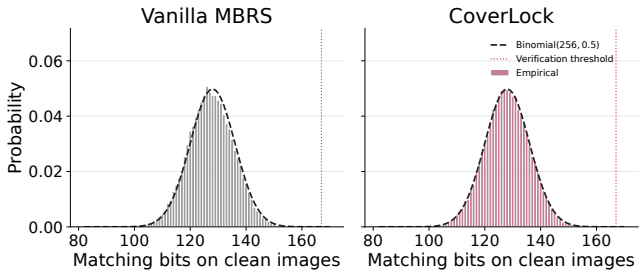


Figure 9: **Verification calibration on clean images.**

Table 6: Validation ROC-AUC (%) for classifier backbones.

Backbone	**ROC-AUC** ↑
DINOv2-G + MLP	72.43
ResNet-50	97.72
ConvNeXt-Tiny	**99.71**

G Classifier Backbone Selection

To select the backbone for the classifier-based defense used in the main experiments, we compare DINOv2-G (Oquab et al., 2023) with an MLP head, ResNet-50 (He et al., 2015), and ConvNeXt-Tiny Liu et al. (2022) on MBRS and MSCOCO val2017. Each classifier is evaluated on 500 genuine watermarked images and 500 GT@1 residual-transfer forgeries, under both clean and distorted settings. An image is accepted only when it passes both the standard watermark verification and the classifier.

As shown in Table 6, ConvNeXt-Tiny provides the strongest overall trade-off between genuine acceptance and forgery rejection among the evaluated backbones, particularly under distortions. We therefore adopt ConvNeXt-Tiny as the classifier backbone in the main experiments.