ROBUST GENERATIVE IMAGE WATERMARKING VIA TEXTURE-AWARE INITIAL NOISE SELECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Robust watermark recovery remains challenging in generative image watermarking under lossy channel distortions. In inversion-based diffusion approaches, we find that initial noise, typically sampled randomly and treated as interchangeable, can substantially affect post-distortion latent recovery, even when the model, prompt, and generation configuration are fixed. More importantly, this variation can be anticipated before transmission: the gray-level co-occurrence matrix (GLCM) homogeneity of the generated image exhibits a consistent positive correlation with latent recovery error under both JPEG compression and additive Gaussian noise. Based on this observation, we propose a texture-aware initial noise selection mechanism that samples candidate initial noises and selects a candidate whose resulting watermarked image exhibits low GLCM homogeneity. The mechanism requires neither modification of the embedding mapping nor retraining of the generative model, and introduces only bounded, controllable sender-side generation overhead. Extensive experiments across multiple existing latent embedding schemes, payload capacities, prompts, and channel distortions demonstrate consistent improvements in payload extraction accuracy, with an average gain of 2.78 percentage points. Our work identifies initial noise selection as an effective design dimension for improving the robustness of inversion-based diffusion watermarking.

1 INTRODUCTION

Diffusion models now synthesize images whose quality rivals real photographs (Rombach et al., 2022; Saharia et al., 2022), and such content circulates widely across social platforms. This capability raises pressing concerns about content provenance, copyright protection, and the accountable use of generative models, motivating techniques that trace a generated image back to its source (Fernandez et al., 2023; Wen et al., 2023). Generative image watermarking addresses this need by embedding an invisible watermark into the image during the generation process itself (Yang et al., 2024; Gesny et al., 2026), so that provenance information travels with the image wherever it circulates (Hu et al., 2026). Once released, however, generated images routinely undergo lossy operations such as JPEG compression and noise corruption (Yang et al., 2024; Hu et al., 2026), so the embedded watermark must remain decodable after such degradations. Robustness against lossy channel attacks is therefore the central performance criterion of generative image watermarking.

For diffusion models, inversion-based approaches have emerged as an effective paradigm for generative image watermarking (Wen et al., 2023; Fang et al., 2025). These methods encode watermark information into the initial Gaussian latent before generation and recover it from the inverted latent at detection time (Wang et al., 2025a). Existing approaches improve robustness primarily through payload encoding and latent mapping strategies, including structured frequency-domain and redundant Gaussian constructions (Ci et al., 2024; Yang et al., 2024). More recent methods further explore spherical and distribution-preserving mappings (Hu et al., 2026; Yuan et al., 2026). Despite their different designs, this line of work has largely focused on *how* information is robustly embedded into the latent, while considerably less attention has been paid to *which* initial noise should be used for watermark recovery.

Meanwhile, a broader line of diffusion research has shown that the initial noise itself can substantially affect generation behavior. Existing studies have selected or optimized initial noise based on

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.

inversion stability (Qi et al., 2024), while other studies have optimized initial noise for text–image alignment (Guo et al., 2024), semantic generation quality (Zhou et al., 2025), and temporal consistency (Wu et al., 2024). These studies, however, focus on generation behavior and output quality rather than the recovery of embedded information after channel distortion. This raises a distinct question for inversion-based generative image watermarking: does the choice of initial noise also affect post-distortion latent recoverability?

Our analysis shows that the choice of initial noise substantially affects post-distortion latent recovery and that this variation can be anticipated from the generated image before transmission. In particular, GLCM homogeneity provides a consistent pre-distortion signal of latent recoverability under JPEG compression and additive Gaussian noise. Motivated by this observation, we propose texture-aware initial noise selection, which screens candidate noises according to the texture of their generated images and can be integrated into existing inversion-based watermarking pipelines without retraining.

Our contributions can be summarized as follows:

- We reveal that the choice of initial noise is a previously overlooked source of robustness variation in inversion-based generative image watermarking. More importantly, we show that this variation can be anticipated before distortion using the GLCM homogeneity of the generated image.
- We propose a lightweight texture-aware initial noise selection mechanism that screens candidate initial noises according to the GLCM homogeneity of their generated watermarked images, improving payload recovery without retraining or modifying the extraction procedure at the receiver.
- We validate the proposed selection mechanism across diverse embedding schemes and generation configurations, showing consistent robustness improvements while maintaining comparable generation quality.

2 Related Work

2.1 Image Watermarking

Image watermarking enables content authentication and provenance tracking by embedding imperceptible information into images. Conventional methods mainly modified images in the spatial or frequency domain (Cox et al., 1997; Podilchuk & Zeng, 1998), while later learning-based methods employed neural networks to learn more robust embedding and extraction procedures. HiDDeN (Zhu et al., 2018) improved robustness against common image processing operations through differentiable distortion layers, while StegaStamp (Tancik et al., 2020) further extended this idea to physical-channel perturbations such as printing and photographing. With the development of diffusion models, watermarks began to be integrated directly into generative models and their generation processes. Stable Signature (Fernandez et al., 2023) fine-tuned the decoder of a latent diffusion model so that generated images carried a fixed signature, while SleeperMark (Wang et al., 2025b) decoupled watermark information from semantic representations to improve watermark retention after model fine-tuning. Another line of work embedded watermark representations directly into the initial latent and improved post-distortion recovery by refining the payload-to-latent mapping. Tree-Ring (Wen et al., 2023) embedded structured watermark patterns in the frequency domain of the initial noise, Gaussian Shading (Yang et al., 2024) combined redundant encoding with Gaussian latent mapping, and Spherical Watermark (Hu et al., 2026) adopted spherical mapping to improve robustness to latent perturbations. Similar latent information mappings have also been explored in generative image steganography. RGSD (Hu et al., 2024) constructed payload representations through an orthogonal mapping, while ADPM (Yuan et al., 2026) exploited the stability of latent signs under channel distortions to design a distribution-preserving embedding mapping. Although these methods adopt different encoding structures, their improvements mainly focus on the payload-to-latent mapping, while the initial noise carrying these mappings is still typically sampled randomly from a standard Gaussian distribution. The effect of different noise samples on post-distortion recovery has therefore received considerably less attention.

Several prior studies on diffusion models have shown that initial noises are not completely interchangeable. Qi et al. (Qi et al., 2024) found that different initial noises exhibited different inversion

stability, and accordingly performed noise selection and optimization to improve generation quality. InitNO (Guo et al., 2024) optimized the initial noise to improve text–image alignment. Golden Noise (Zhou et al., 2025) learned initial noises that were more favorable for semantic generation. FreeInit (Wu et al., 2024) adjusted the low-frequency components of the initial latent to improve temporal consistency in video diffusion. Together, these studies showed that the choice of initial noise could affect the generation or inversion behavior of diffusion models, but their objectives mainly focused on generation quality, semantic alignment, and generation stability rather than the recovery of embedded payloads after lossy channels. Therefore, when the generative model, text condition, and embedding mapping are all fixed, whether different initial noises systematically lead to different post-distortion latent recoverability remains underexplored.

3 Empirical Analysis

We first examine how latent recoverability varies with the sampled initial noise. For each prompt, we independently sample 10 initial noises and generate the corresponding images while keeping the diffusion model, text condition, and generation configuration unchanged. We then apply common image distortions to the generated images and recover their latents through DDIM inversion. Let x_0 denote the sampled initial noise and $\hat{x}_u$ is the latent recovered after distortion u. We measure recovery using the relative error

$$E_u = \frac{\|\hat{x}_u - x_0\|_2}{\|x_0\|_2}, \tag{1}$$

where lower values indicate more faithful recovery.

Across most tested distortions, the mean within-prompt difference between the maximum and minimum recovery errors among the 10 initial noises ranges from 0.13 to 0.28. This variation indicates that initial noises are not interchangeable with respect to post-distortion latent recovery.

Prior work provides a clue for further analyzing this variation. Staniszewski et al. (Staniszewski et al., 2026) found that latents obtained by DDIM inversion exhibit pronounced structural patterns in smooth image regions, suggesting an association between local image smoothness and inversion outcomes. Motivated by this observation, we consider image texture as a potential pre-distortion signal of latent recoverability.

To this end, we examine several texture statistics of the generated image and their association with E_u. Among the examined statistics, GLCM homogeneity shows the most consistent association under JPEG compression and additive Gaussian noise. We therefore use it to characterize image texture in the analysis below. A systematic comparison of texture statistics and GLCM spatial scales is presented in Section 5.2.

We follow the standard computation of GLCM homogeneity (Haralick et al., 1973; Oprisan & Oprisan, 2025). First, each generated image is converted to a grayscale image I with 256 gray levels. A displacement $(\Delta x, \Delta y)$ then pairs each pixel p with the pixel $p + (\Delta x, \Delta y)$, and the occurrences of each gray-level pair are counted:

$$P_{(\Delta x,\Delta y)}(i,j) = \#\left\{\, p \;:\; I(p) = i \text{ and } I\big(p + (\Delta x, \Delta y)\big) = j \,\right\}, \tag{2}$$

where $\#$ denotes the cardinality of a set, the first index i is the intensity of the reference pixel, and the second index j is that of the target pixel. Dividing each count by the total number of valid pairs $N_{(\Delta x,\Delta y)}$ yields the normalized GLCM,

$$G_{(\Delta x,\Delta y)}(i,j) = \frac{P_{(\Delta x,\Delta y)}(i,j)}{N_{(\Delta x,\Delta y)}}, \tag{3}$$

where $G_{(\Delta x,\Delta y)}(i,j)$ represents the fraction of valid pixel pairs with gray levels (i,j) at displacement $(\Delta x, \Delta y)$. The normalized GLCM is then scored by the homogeneity statistic of Haralick et al. (1973),

$$H(G) = \sum_{i=0}^{255}\sum_{j=0}^{255} \frac{G(i,j)}{1 + (i-j)^2}, \tag{4}$$

where the weight $1/(1 + (i-j)^2)$ decreases with the gray-level difference, so that lower $H(G)$ indicates richer local texture. Finally, the homogeneity is averaged over four displacements at pixel

distance d (horizontal, diagonal, vertical, and anti-diagonal), giving the texture score

$$\mathrm{gh}_d(I) = \tfrac{1}{4}\left[H(G_{(d,0)}) + H(G_{(d,d)}) + H(G_{(0,d)}) + H(G_{(-d,d)})\right]. \tag{5}$$

We set $d = 4$ and write $\mathrm{gh}_4(I)$ as $\mathrm{gh4}(I)$ in what follows. For each prompt, we rank the 10 candidates by $\mathrm{gh4}(I)$ from low to high and average E_u at each rank over the 800 prompts. The resulting relationship varies across distortion types. Under Gaussian blur, resizing, center crop, and brightness adjustment, recovery error shows no consistent trend with the $\mathrm{gh4}$ ranking. In contrast, under JPEG compression and additive Gaussian noise, recovery error consistently increases with $\mathrm{gh4}$ rank. Figure 1 shows this relationship at three levels of JPEG compression and three levels of additive Gaussian noise.

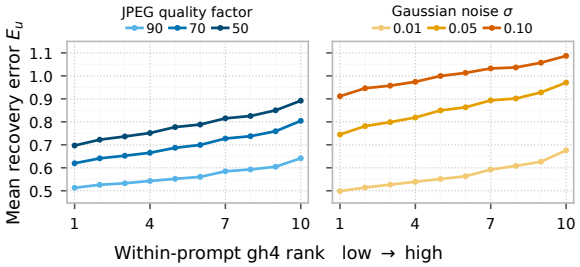


Figure 1: Within-prompt texture–recovery relationship. Candidates are ranked by $\mathrm{gh4}(I)$. Higher ranks consistently yield larger recovery errors under JPEG compression and additive Gaussian noise.

The study yields two main observations. First, different initial noises exhibit different post-distortion latent recoverability even when the model, prompt, and generation configuration are fixed. Second, under JPEG compression and additive Gaussian noise, this variation is consistently reflected by the pre-distortion $\mathrm{gh4}$ ordering: candidates with lower $\mathrm{gh4}$ yield smaller recovery errors across all six settings. Thus, $\mathrm{gh4}$ provides a pre-distortion signal for identifying initial noises with better latent recoverability under these distortions. This observation motivates our texture-aware initial noise selection method.

4 Method

Based on the analysis in Sec. 3, we introduce initial noise selection based on texture. As shown in Fig. 2, the sender generates candidate watermarked images from different initial noises and screens them according to their texture scores before transmission, while the original payload embedding and receiver extraction procedures remain unchanged.

4.1 Embedding with Texture-Aware Noise Selection

Before entering the texture-aware initial noise selection process, the sender first encrypts the payload and specifies a semantic prompt for the watermarked image to be generated. Given a payload m, the sender uses a shared secret key k to obtain the encrypted payload

$$m_e = \mathrm{Enc}(m, k), \tag{6}$$

where $\mathrm{Enc}(\cdot)$ denotes the payload encryption operation used by the underlying scheme. Meanwhile, the sender converts the semantic prompt into a diffusion condition using the condition encoder τ

$$c = \tau(\text{prompt}), \tag{7}$$

which guides the subsequent generation of candidate watermarked images. The sender then proceeds with texture-aware initial noise selection, which consists of five stages: **Initial Noise Sampling**, **Payload Embedding**, **Candidate Generation**, **Texture Scoring**, and **Candidate Selection**.

Initial Noise Sampling. At the a-th attempt, the sender uses a new random seed s_a to generate an initial noise through a Gaussian sampler

$$x_0^{(a)} = \mathrm{GS}(s_a), \qquad x_0^{(a)} \sim \mathcal{N}(0, I), \tag{8}$$

where $\mathrm{GS}(\cdot)$ denotes a Gaussian sampler controlled by the random seed. Given s_a, the sampler generates noise that follows a standard Gaussian distribution, which is then used as the initial noise for image generation in the diffusion model.

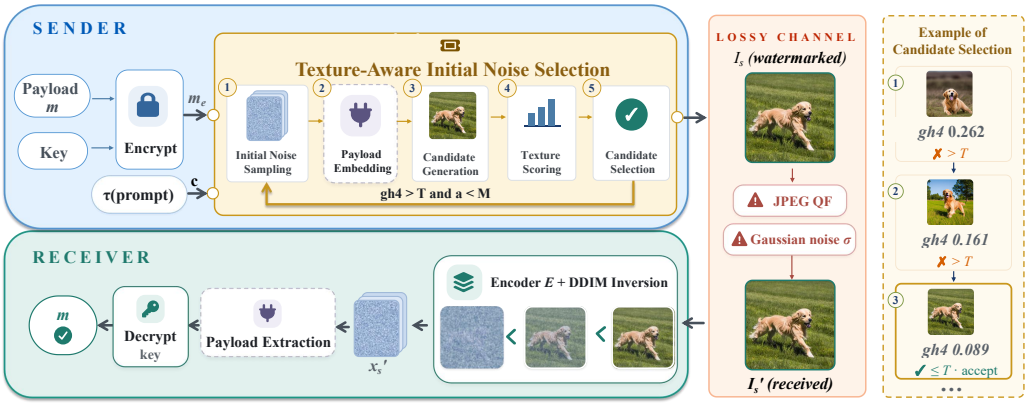


Figure 2: Overview of the watermarking pipeline with our texture-aware initial noise selection mechanism.

Payload Embedding. The underlying watermarking scheme then embeds the payload m_e into the initial noise using its original embedding mapping

$$x_s^{(a)} = \mathrm{Embed}\left(x_0^{(a)}, m_e\right). \tag{9}$$

where $\mathrm{Embed}(\cdot)$ denotes the payload embedding operation of the latent embedding scheme.

Candidate Generation. Guided by c, the embedded latent $x_s^{(a)}$ is denoised for L iterations using deterministic DDIM sampling (Song et al., 2021), yielding the final latent representation. The resulting latent is then decoded into image space by the pretrained Variational Autoencoder (VAE) decoder (Kingma & Welling, 2014) used in the Latent Diffusion Model (LDM) (Rombach et al., 2022), producing a candidate watermarked image

$$I_s^{(a)} = D\left(f(x_s^{(a)}, c)\right), \tag{10}$$

where f denotes deterministic DDIM sampling with L denoising steps and D denotes the pretrained VAE decoder in the LDM.

Texture Scoring. For each generated candidate watermarked image $I_s^{(a)}$, the sender computes its texture score before transmission

$$\mathrm{gh}_d^{(a)} = \mathrm{gh}_d\left(I_s^{(a)}\right), \tag{11}$$

where $\mathrm{gh_d}(\cdot)$ is defined in Sec. 3.

Candidate Selection. We set an acceptance threshold T to determine whether the current candidate has sufficient potential to resist distortion. When $\mathrm{gh}_d^{(a)} \le T$, the candidate exhibits texture characteristics that are more favorable for latent recovery. The search therefore terminates immediately and the current candidate is accepted as the final watermarked image. When $\mathrm{gh}_d^{(a)} > T$, the candidate does not meet the predefined acceptance criterion. The sender then samples a new initial noise and repeats payload embedding, image generation, and texture evaluation to search for another candidate with a lower texture score.

To prevent the search from continuing indefinitely and to limit the additional computation introduced by repeated image generation, we further set a maximum number of attempts M. If a candidate satisfying $\mathrm{gh}_d^{(a)} \le T$ is found before the search reaches M attempts, the search terminates at the first candidate that meets the threshold. If no candidate meets the threshold after M attempts, the candidate with the lowest $\mathrm{gh_a}$ among the generated candidates is selected. The final watermarked image is defined as

$$I_s = \begin{cases} I_s^{\left(\min\{a \mid \mathrm{gh}_d^{(a)} \le T\}\right)}, & \text{if such an } a \le M \text{ exists,} \\ I_s^{\left(\arg\min_{1 \le i \le M} \mathrm{gh}_d^{(i)}\right)}, & \text{otherwise.} \end{cases} \tag{12}$$

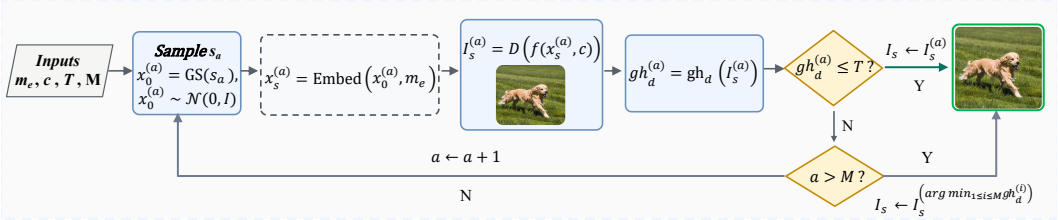


Figure 3: The proposed texture-aware initial noise selection mechanism.

This rule prioritizes the first candidate that satisfies the acceptance threshold and falls back to the generated candidate with the lowest texture score when the search budget is exhausted.

Figure 3 illustrates the complete image generation procedure with our texture-aware initial noise selection mechanism.

4.2 Payload Extraction

The proposed initial noise selection mechanism operates only at the sender and requires no modification at the receiver. Given a received watermarked image I'_s after transmission through a lossy channel, the receiver first maps the image to the latent space using the pretrained VAE encoder in the LDM and then recovers the latent through deterministic DDIM inversion corresponding to the generation process

$$x'_s = f^{-1}(E(I'_s)), \tag{13}$$

where E denotes the pretrained VAE encoder in the LDM and f^{-1} denotes the deterministic DDIM inversion process corresponding to the generation procedure at the sender.

After obtaining the recovered latent x'_s, the receiver applies the extraction rule of the original latent embedding scheme to recover the encrypted payload

$$\hat{m}_e = \mathrm{Extract}(x'_s), \tag{14}$$

where $\mathrm{Extract}(\cdot)$ denotes the payload extraction operation of the original scheme. The recovered payload is then obtained as $\hat{m} = \mathrm{Dec}(\hat{m}_e, k)$ using the same shared secret key k as the sender, where $\mathrm{Dec}(\cdot)$ denotes the decryption operation corresponding to the payload encryption procedure at the sender.

Since initial noise selection is completed before image transmission, the receiver does not need to conduct the candidate search or selection process. It follows the original image encoding, DDIM inversion, and payload extraction procedures without modification.

5 Experimental Results

5.1 Experimental Setup

Model and Generation Configuration. All experiments are based on the pretrained Stable Diffusion v1.5 (Rombach et al., 2022) with the DDIM sampler. We set CFG $= 7.5$ for generation and CFG $= 1$ with an empty prompt for inversion. The latent variable has shape $1 \times 4 \times 64 \times 64$, the generated images have a resolution of 512×512, and all experiments are conducted on an NVIDIA RTX 4090 GPU.

Prompt Sets. We use two prompt sets with different levels of complexity. The first consists of relatively short MS COCO-style prompts (Lin et al., 2014), typically containing 6–10 words with concise semantics. The second is drawn from Gustavosta/Stable-Diffusion-Prompts (Gustavosta, 2022), typically containing 20–50 words with richer descriptions of style, lighting, and composition. We randomly sample 5000 prompts from each set.

Embedding Schemes and Comparison Protocol. Our method operates on initial noise selection and is independent of the specific payload embedding mapping. To evaluate its applicability

across diverse latent embedding constructions, we consider four schemes, including the generative watermarking methods Gaussian Shading (GS) (Yang et al., 2024) and Spherical Watermark (SW) (Hu et al., 2026), both evaluated at 256 and 4096 bits, and the generative steganography methods ADPM (Yuan et al., 2026) at 5217 and 11384 bits and RGSD (Hu et al., 2024) at 16384 bits. For each scheme, we compare the original version (base) with its version augmented by texture-aware initial noise selection (+Sel), while keeping the embedding, generation, and extraction procedures unchanged. The selection parameters are set to $M = 10$, $T = 0.10$, and GLCM distance $d = 4$.

Channel Conditions and Evaluation Metrics. We use lossless PNG storage as the reference condition and evaluate robustness under JPEG compression and additive Gaussian noise, with JPEG quality factors $\mathrm{QF} \in \{90, 70, 50\}$ and Gaussian noise levels $\sigma \in \{0.01, 0.05, 0.10\}$. Payload recovery is measured by extraction accuracy (Acc), defined as the proportion of correctly recovered payload bits. Generation quality is evaluated using CLIP Score (Radford et al., 2021) and MUSIQ (Ke et al., 2021) for text–image alignment and perceptual quality, respectively.

5.2 Validation of the Texture Indicator

The analysis in Section 3 establishes the relationship between GLCM homogeneity and latent recovery after distortion. We further examine whether this relationship depends on the GLCM spatial scale and compare homogeneity with alternative statistics, including gradient magnitude, high frequency energy, LBP entropy (Ojala et al., 2002), and other GLCM descriptors such as energy, correlation, and contrast (Haralick et al., 1973). For each statistic, we compute within-prompt Pearson correlations with latent recovery error under six channel distortions and report the mean absolute correlation across distortions.

Table 1: Within-prompt Pearson correlation of candidate pre-attack image statistics with recovery error. Bold marks the strongest $|r|$ in each column.

Statistic	**JPEG90**	**JPEG70**	**JPEG50**	**Noise.01**	**Noise.05**	**Noise.10**	**avg $\lvert r\rvert$**
GLCM homogeneity across distances d							
gh_1	0.502	0.564	0.581	0.715	0.639	0.549	0.592
gh_2	0.551	0.611	0.621	0.746	0.661	0.566	0.626
gh_4	**0.621**	**0.681**	**0.682**	**0.769**	**0.694**	0.600	**0.674**
gh_8	0.612	0.665	0.661	0.720	0.668	0.585	0.652
gh_{16}	0.568	0.608	0.599	0.636	0.605	0.538	0.592
Alternative statistics							
Gradient magnitude	-0.189	-0.376	-0.474	-0.244	-0.633	$\mathbf{-0.722}$	0.440
GLCM energy (d=4)	0.419	0.339	0.294	0.441	0.280	0.222	0.333
LBP entropy	-0.392	-0.310	-0.262	-0.434	-0.249	-0.181	0.305
GLCM correlation (d=4)	0.072	0.195	0.262	0.089	0.361	0.446	0.238
High-frequency energy	-0.010	-0.092	-0.136	-0.053	-0.263	-0.321	0.146
GLCM contrast (d=4)	-0.016	-0.166	-0.248	-0.036	-0.362	-0.482	0.218

As shown in Table 1, GLCM homogeneity remains consistently correlated with recovery error across spatial scales, with $d = 4$ showing the strongest overall association. Alternative statistics exhibit weaker or less consistent correlations. We therefore use $\mathrm{gh}4$ as the screening statistic in the subsequent experiments.

5.3 Effectiveness of the Noise Selection Mechanism

We evaluate the effect of the proposed texture-aware initial noise selection on payload recovery across GS, SW, ADPM, and RGSD. We compare extraction accuracy with and without selection across different payload capacities, prompt sets, JPEG compression, and Gaussian noise attacks.

As shown in Table 2, across all lossy-channel settings, texture-aware initial noise selection increases the average extraction accuracy from 88.76% to 91.54%, corresponding to an average gain of 2.78 percentage points. Among the 84 attacked settings, selection improves extraction accuracy in 78 cases and matches the corresponding base scheme in the remaining six, with no degradation observed. The benefit can be particularly substantial in challenging settings. For example, for SW with

Table 2: Extraction accuracy for GS (Yang et al., 2024), SW (Hu et al., 2026), ADPM (Yuan et al., 2026) and RGSD (Hu et al., 2024) with and without texture-aware initial noise selection (+Sel). Bold marks the higher value within each base/+Sel pair.

		GS				SW				ADPM				RGSD	
Capacity (bits)		256		4096		256		4096		5217		11384		16384	
Prompts	**Attack**	base	+Sel	base	+Sel	base	+Sel	base	+Sel	base	+Sel	base	+Sel	base	+Sel
COCO	PNG	100.00	100.00	95.20	**95.61**	100.00	100.00	98.48	**98.95**	98.39	**98.75**	93.96	**94.75**	97.30	**97.56**
	JPEG QF=90	100.00	100.00	93.55	**94.67**	100.00	100.00	95.67	**97.61**	96.66	**98.11**	91.40	**93.39**	94.48	**96.19**
	JPEG QF=70	100.00	100.00	90.62	**92.73**	99.86	**100.00**	88.52	**94.33**	93.26	**96.39**	87.30	**90.52**	89.30	**92.86**
	JPEG QF=50	99.95	**100.00**	88.37	**90.81**	99.39	**99.98**	82.66	**90.84**	90.53	**94.56**	84.15	**88.15**	85.38	**89.79**
	Gauss σ=0.01	100.00	100.00	93.22	**94.70**	99.99	**100.00**	94.94	**97.86**	96.13	**98.15**	91.02	**93.71**	93.94	**96.39**
	Gauss σ=0.05	99.79	**99.98**	85.95	**89.06**	97.67	**99.88**	76.87	**87.58**	87.59	**92.45**	81.40	**86.11**	81.77	**87.23**
	Gauss σ=0.10	98.95	**99.73**	79.69	**82.69**	90.48	**96.95**	66.42	**74.67**	80.77	**85.71**	74.62	**78.76**	73.68	**78.51**
Gustavosta	PNG	100.00	100.00	94.64	**94.92**	99.99	**100.00**	97.38	**97.47**	97.89	**98.31**	92.72	**93.32**	96.40	**96.72**
	JPEG QF=90	100.00	100.00	92.18	**93.23**	99.55	**99.81**	92.94	**94.37**	95.05	**96.53**	88.41	**90.52**	91.75	**93.69**
	JPEG QF=70	99.93	**99.97**	87.55	**89.87**	98.65	**99.32**	84.29	**87.70**	89.98	**92.52**	82.41	**85.87**	84.36	**87.77**
	JPEG QF=50	99.81	**99.88**	84.02	**86.95**	97.20	**98.32**	77.61	**82.09**	86.22	**89.39**	78.40	**82.46**	79.51	**83.44**
	Gauss σ=0.01	100.00	100.00	91.88	**93.48**	99.37	**99.89**	92.35	**95.33**	94.99	**97.06**	88.39	**91.41**	91.61	**94.47**
	Gauss σ=0.05	99.54	**99.82**	80.70	**84.30**	94.45	**97.53**	72.88	**78.23**	82.71	**87.06**	74.98	**80.12**	75.61	**80.29**
	Gauss σ=0.10	98.11	**98.87**	73.67	**77.05**	84.28	**89.72**	62.81	**66.33**	75.34	**79.43**	68.21	**72.81**	68.34	**71.90**

a 4096-bit payload under Gaussian noise with $\sigma = 0.05$ on the COCO prompt set, the extraction accuracy increases from 76.87% to 87.58%, a gain of 10.71 percentage points. Overall, texture-aware initial noise selection consistently improves payload recovery across different embedding schemes under lossy channels.

5.4 Generation Quality Evaluation

To evaluate whether texture-aware initial noise selection affects generation quality, we use CLIP Score to measure text–image alignment and MUSIQ to assess perceptual quality. As shown in Table 3, the CLIP Score and MUSIQ remain at comparable levels before and after selection. Figure 4 further shows representative images generated with initial noise selection under diverse textual prompts.

Table 3: Generation quality with and without selection (↑: higher is better).

	CLIP ↑		MUSIQ ↑	
	base	+Sel	base	+Sel
COCO	0.32	0.32	68.96	68.98
Gustavosta	0.35	0.35	70.37	70.66



Figure 4: Representative images generated with texture-aware initial noise selection under diverse textual prompts. Top: MS COCO prompts. Bottom: Gustavosta/Stable-Diffusion-Prompts.

5.5 Generalization Across Backbones and Samplers

In practical deployment, diffusion models may adopt different backbones and samplers. We therefore further examine the behavior of initial noise selection under two such changes: replacing the

Table 4: Extraction accuracy across generation backbones and samplers, with and without selection. Bold values indicate the higher accuracy within each pair.

(a) Across backbones (DDIM sampler)

	GS				SW			
	JPEG 50		Noise 0.10		JPEG 50		Noise 0.10	
Backbone	base	+Sel	base	+Sel	base	+Sel	base	+Sel
SD1.5	97.71	**98.98**	91.42	**94.47**	94.27	**97.86**	80.29	**87.89**
SD2.1	97.63	**98.96**	91.39	**94.42**	94.32	**97.76**	80.36	**87.93**

(b) Across samplers (SD1.5 backbone)

	GS				SW			
	JPEG 50		Noise 0.10		JPEG 50		Noise 0.10	
Sampler	base	+Sel	base	+Sel	base	+Sel	base	+Sel
DDIM	97.71	**98.98**	91.42	**94.47**	94.27	**97.86**	80.29	**87.89**
DPM-Solver++	97.67	**98.94**	91.38	**94.46**	94.28	**97.83**	80.29	**87.90**

Table 5: Screening outcome and sender-side cost under different stopping rules: (a) varying M with T = 0.10; (b) varying T with M = 10. Bold identifies the reference operating point.

(a) Budget sensitivity ($T = 0.10$)

Budget M	Selected gh4 ↓	Attempts
1	0.1788	1.000
5	0.1337	3.998
10	0.1219	7.231

(b) Threshold sensitivity ($M = 10$)

Threshold T	Selected gh4 ↓	Attempts
0.20	0.1465	2.308
0.15	0.1318	4.177
0.10	0.1219	7.231

backbone from Stable Diffusion v1.5 to Stable Diffusion 2.1, and replacing the sampler from DDIM to DPM-Solver++ (Lu et al., 2022). As shown in Table 4, the extraction-accuracy gains on both GS and SW remaine stable after changing either the diffusion backbone or the sampler, indicating that the selection strategy is insensitive to the tested generation configurations.

5.6 Screening Behavior and Generation Cost

We evaluate the effects of the screening budget M and acceptance threshold T on screening effectiveness and sender side generation cost. As shown in Table 5, increasing M or lowering T reduces the gh4 of the selected images, but requires more generation attempts. The main experiments use $M = 10$ and $T = 0.10$ to provide stronger screening while limiting the search to at most 10 generations. Under this setting, the average number of generations is 7.231. In practical use, the sender side generation cost can be controlled by adjusting M and T, allowing screening effectiveness to be balanced against the available computation budget.

6 Conclusion

In this work, we show that robustness in inversion-based generative image watermarking depends not only on how the payload is mapped into the latent, but also on which initial noise is used. Our analysis reveals that generated-image texture provides a pre-distortion cue for this variation, with GLCM homogeneity consistently associated with latent recovery error under JPEG compression and additive Gaussian noise. This observation motivates a texture-aware initial noise selection mechanism that screens candidate noises while leaving the underlying embedding and extraction procedures unchanged. Across existing latent embedding schemes, the resulting selector consistently improves payload recovery while maintaining comparable generation quality and a controllable generation cost. More broadly, these results establish initial noise selection as a complementary robustness dimension for inversion-based generative image watermarking.

AI USE STATEMENT

In this work, we used generative AI tools for language editing and polishing of the manuscript, and for assistance in implementing and debugging the experimental and visualization code, and for assisting with literature search and identification of relevant work. We have not used generative AI tools for generating research ideas or producing the experimental results. We have reviewed all AI-assisted work. All AI-assisted code was executed by the authors, and its outputs were checked against the raw experimental logs. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work studies robust watermarking for generative images, with the goal of supporting content provenance and copyright protection for generated content. Our experiments use publicly available generative models and public prompt sets, and involve no human subjects, personal data, or newly released datasets. While robust watermarking could in principle be used to track generated content at scale, we believe the primary impact of this work is to improve the accountability of generative models, and we are not aware of ethical concerns beyond those already associated with generative image watermarking in general.

REPRODUCIBILITY STATEMENT

All details required to reproduce our results are described in the paper. The selection mechanism is fully specified by Section 4, with the parameter settings given in Section 5. The complete per-configuration results underlying all reported averages are provided in Appendix A. The mechanism requires no training and introduces no learned components. Source code is available at `https://anonymous.4open.science/r/texture-aware--1A7D/`.

REFERENCES

Hai Ci, Pei Yang, Yiren Song, and Mike Zheng Shou. RingID: Rethinking Tree-Ring watermarking for enhanced multi-key identification. In *Proceedings of the European Conference on Computer Vision*, pp. 338–354, 2024.

Ingemar J. Cox, Joe Kilian, F. Thomson Leighton, and Talal Shamoon. Secure spread spectrum watermarking for multimedia. *IEEE Transactions on Image Processing*, 6(12):1673–1687, 1997.

Han Fang, Kejiang Chen, Zijin Yang, Bosen Cui, Weiming Zhang, and Ee-Chien Chang. CoSDA: Enhancing the robustness of inversion-based generative image watermarking framework. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 2888–2896, 2025.

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.

Enoal Gesny, Eva Giboulot, Teddy Furon, and Vivien Chappelier. Guidance watermarking for diffusion models. In *International Conference on Learning Representations*, 2026.

Xiefan Guo, Jinlin Liu, Miaomiao Cui, Jiankai Li, Hongyu Yang, and Di Huang. InitNO: Boosting text-to-image diffusion models via initial noise optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9380–9389, 2024.

Gustavosta. Stable-Diffusion-Prompts, 2022. `https://huggingface.co/datasets/Gustavosta/Stable-Diffusion-Prompts`, accessed August 2026.

Robert M. Haralick, K. Shanmugam, and Itshak Dinstein. Textural features for image classification. *IEEE Transactions on Systems, Man, and Cybernetics*, 3(6):610–621, 1973.

X. Hu, S. Li, Q. Ying, W. Peng, X. Zhang, and Z. Qian. Establishing robust generative image steganography via popular Stable Diffusion. *IEEE Transactions on Information Forensics and Security*, 19:8094–8108, 2024.

Xiaoxiao Hu, Jiaqi Jin, Sheng Li, Wanli Peng, Xinpeng Zhang, and Zhenxing Qian. Spherical Watermark: Encryption-free, lossless watermarking for diffusion models. In *International Conference on Learning Representations*, 2026.

Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. MUSIQ: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5148–5157, 2021.

Diederik P. Kingma and Max Welling. Auto-encoding variational Bayes. In *International Conference on Learning Representations*, 2014.

T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: Common objects in context. In *Proceedings of the European Conference on Computer Vision*, pp. 740–755, 2014.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. DPM-Solver++: Fast solver for guided sampling of diffusion probabilistic models. *arXiv preprint arXiv:2211.01095*, 2022.

Timo Ojala, Matti Pietikäinen, and Topi Mäenpää. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(7):971–987, 2002.

Sorinel A. Oprisan and Ana Oprisan. Scaling laws for Haralick texture features of linear gradients. *PeerJ Computer Science*, 11:e2856, 2025.

Christine I. Podilchuk and Wenjun Zeng. Image-adaptive watermarking using visual models. *IEEE Journal on Selected Areas in Communications*, 16(4):525–539, 1998.

Zipeng Qi, Lichen Bai, Haoyi Xiong, and Zeke Xie. Not all noises are created equally: Diffusion noise selection and optimization. *arXiv preprint arXiv:2407.14041*, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 8748–8763, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L. Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In *Advances in Neural Information Processing Systems*, volume 35, pp. 36479–36494, 2022.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021.

Łukasz Staniszewski, Łukasz Kuciński, and Kamil Deja. There and back again: On the relation between noise and image inversions in diffusion models. In *International Conference on Learning Representations*, 2026.

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.

Hanyi Wang, Han Fang, Shi-Lin Wang, and Ee-Chien Chang. ROAR: Reducing inversion error in generative image watermarking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19742–19751, 2025a.

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*, pp. 8213–8224, 2025b.

Yuxin Wen, John Kirchenbauer, Jonas Geiping, and Tom Goldstein. Tree-Ring watermarks: Invisible fingerprints for diffusion images. In *Advances in Neural Information Processing Systems*, volume 36, pp. 58047–58063, 2023.

Tianxing Wu, Chenyang Si, Yuming Jiang, Ziqi Huang, and Ziwei Liu. FreeInit: Bridging initialization gap in video diffusion models. In *Proceedings of the European Conference on Computer Vision*, pp. 378–394, 2024.

Zijin Yang, Kai Zeng, Kejiang Chen, Han Fang, Weiming Zhang, and Nenghai Yu. Gaussian Shading: Provable performance-lossless image watermarking for diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12162–12171, 2024.

Chengsheng Yuan, Zhaonan Ji, Qi Cui, Zhili Zhou, Xinting Li, and Zhihua Xia. Towards provably secure and highly robust generative image steganography leveraging latent diffusion model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 1498–1506, 2026.

Zikai Zhou, Shitong Shao, Lichen Bai, Shufei Zhang, Zhiqiang Xu, Bo Han, and Zeke Xie. Golden noise for diffusion models: A learning framework. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 17688–17697, 2025.

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*, pp. 657–672, 2018.

A Complete Extraction-Accuracy Results

Tables 6 and 7 report the complete extraction-accuracy results across all evaluated embedding schemes and payload capacities, extending the configurations shown in Table 2.

Table 6: Extraction accuracy for GS (Yang et al., 2024) and SW (Hu et al., 2026) across payload capacities with and without texture-aware initial-noise selection. Each base and +Sel pair compares an embedding scheme with the same scheme augmented by selection. Bold values indicate the higher accuracy within each pair.

		GS						SW									
Capacity (bits)		256		1024		4096		256		512		1024		2048		4096	
Prompts	**Attack**	base	+Sel	base	+Sel	base	+Sel	base	+Sel	base	+Sel	base	+Sel	base	+Sel	base	+Sel
COCO	PNG	100.00	100.00	99.96	99.96	95.20	**95.61**	100.00	100.00	99.99	99.99	99.99	99.99	99.86	**99.89**	98.48	**98.95**
	JPEG QF=90	100.00	100.00	99.85	**99.93**	93.55	**94.67**	100.00	100.00	99.89	**99.98**	99.95	**99.96**	99.04	**99.60**	95.67	**97.61**
	JPEG QF=70	100.00	100.00	99.38	**99.75**	90.62	**92.73**	99.86	**100.00**	98.30	**99.90**	99.24	**99.63**	95.59	**98.11**	88.52	**94.33**
	JPEG QF=50	99.95	**100.00**	98.58	**99.42**	88.37	**90.81**	99.39	**99.98**	95.80	**99.44**	97.07	**98.89**	92.21	**96.05**	82.66	**90.84**
	Gauss σ=0.01	100.00	100.00	99.77	**99.92**	93.22	**94.70**	99.99	**100.00**	99.79	**100.00**	99.88	**99.96**	98.34	**99.62**	94.94	**97.86**
	Gauss σ=0.05	99.79	**99.98**	97.43	**98.89**	85.95	**89.06**	97.67	**99.88**	92.09	**98.40**	93.13	**97.26**	88.18	**93.21**	76.87	**87.58**
	Gauss σ=0.10	98.95	**99.73**	93.62	**96.06**	79.69	**82.69**	90.48	**96.95**	84.09	**92.98**	83.02	**88.16**	77.59	**81.92**	66.42	**74.67**
Gustavosta	PNG	100.00	100.00	99.88	**99.94**	94.64	**94.92**	99.99	**100.00**	99.94	**100.00**	99.86	**99.93**	98.82	**99.46**	97.38	**97.47**
	JPEG QF=90	100.00	100.00	99.50	**99.82**	92.18	**93.23**	99.55	**99.81**	98.84	**99.69**	99.04	**99.10**	97.50	**97.89**	92.94	**94.37**
	JPEG QF=70	99.93	**99.97**	98.30	**99.30**	87.55	**89.87**	98.65	**99.32**	95.34	**98.20**	95.64	**97.01**	91.71	**93.65**	84.29	**87.70**
	JPEG QF=50	99.81	**99.88**	96.84	**98.54**	84.02	**86.95**	97.20	**98.32**	92.16	**96.23**	91.33	**93.91**	85.66	**88.47**	77.61	**82.09**
	Gauss σ=0.01	100.00	100.00	99.50	**99.82**	91.88	**93.48**	99.37	**99.89**	98.22	**99.86**	99.01	**99.47**	97.48	**98.01**	92.35	**95.33**
	Gauss σ=0.05	99.54	**99.82**	95.16	**97.53**	80.70	**84.30**	94.45	**97.53**	88.72	**93.55**	86.47	**90.39**	79.75	**84.85**	72.88	**78.23**
	Gauss σ=0.10	98.11	**98.87**	89.21	**92.87**	73.67	**77.05**	84.28	**89.72**	77.80	**83.35**	73.01	**77.96**	67.45	**71.45**	62.81	**66.33**

Table 7: Extraction accuracy for ADPM (Yuan et al., 2026) and RGSD (Hu et al., 2024) across payload capacities with and without texture-aware initial-noise selection. Each base and +Sel pair compares an embedding scheme with the same scheme augmented by selection. Bold values indicate the higher accuracy within each pair.

		ADPM										RGSD	
Capacity (bits)		5217		6999		9088		10218		11384		16384	
Prompts	**Attack**	base	+Sel	base	+Sel	base	+Sel	base	+Sel	base	+Sel	base	+Sel
COCO	PNG	98.39	**98.75**	97.78	**97.88**	96.09	**96.72**	95.13	**95.99**	93.96	**94.75**	97.30	**97.56**
	JPEG QF=90	96.66	**98.11**	96.03	**97.02**	93.71	**95.56**	92.41	**94.69**	91.40	**93.39**	94.48	**96.19**
	JPEG QF=70	93.26	**96.39**	92.63	**94.85**	89.83	**92.95**	88.15	**91.90**	87.30	**90.52**	89.30	**92.86**
	JPEG QF=50	90.53	**94.56**	89.77	**92.75**	86.80	**90.56**	85.05	**89.41**	84.15	**88.15**	85.38	**89.79**
	Gauss σ=0.01	96.13	**98.15**	95.56	**97.03**	93.29	**95.77**	91.94	**94.86**	91.02	**93.71**	93.94	**96.39**
	Gauss σ=0.05	87.59	**92.45**	86.88	**90.78**	84.17	**88.29**	82.42	**87.38**	81.40	**86.11**	81.77	**87.23**
	Gauss σ=0.10	80.77	**85.71**	79.84	**83.53**	77.51	**80.52**	75.73	**79.69**	74.62	**78.76**	73.68	**78.51**
Gustavosta	PNG	97.89	**98.31**	96.85	**96.99**	95.27	**95.95**	94.40	**94.66**	92.72	**93.32**	96.40	**96.72**
	JPEG QF=90	95.05	**96.53**	93.55	**94.93**	91.15	**93.56**	90.38	**92.09**	88.41	**90.52**	91.75	**93.69**
	JPEG QF=70	89.98	**92.52**	87.79	**90.78**	84.89	**88.89**	84.42	**87.31**	82.41	**85.87**	84.36	**87.77**
	JPEG QF=50	86.22	**89.39**	83.65	**87.54**	80.72	**85.34**	80.28	**83.67**	78.40	**82.46**	79.51	**83.44**
	Gauss σ=0.01	94.99	**97.06**	93.64	**95.40**	91.06	**94.24**	90.35	**92.73**	88.39	**91.41**	91.61	**94.47**
	Gauss σ=0.05	82.71	**87.06**	80.11	**84.91**	77.59	**82.81**	77.29	**81.32**	74.98	**80.12**	75.61	**80.29**
	Gauss σ=0.10	75.34	**79.43**	72.82	**77.16**	70.56	**74.93**	70.24	**73.49**	68.21	**72.81**	68.34	**71.90**

As shown in Tables 6 and 7, the trend observed in Table 2 holds across all evaluated payload capacities. Among the 168 attacked settings, texture-aware initial noise selection improves extraction accuracy in 162 cases and matches the corresponding base scheme in the remaining six, with no degradation observed. The gain is smallest at the lowest payload capacity, where the base accuracy is already close to saturation, and generally increases with both payload capacity and distortion strength. For example, under Gaussian noise with $\sigma = 0.10$ on the Gustavosta prompt set, the gain of

GS increases from 0.76 percentage points at 256 bits to 3.38 percentage points at 4096 bits. Overall, the complete results confirm that texture-aware initial noise selection consistently improves payload recovery across embedding schemes, payload capacities, prompt sets, and lossy channel distortions.

B Impact of Selection under Non-Target Distortions

We further evaluate the effect of texture-aware initial noise selection under the four distortion families for which the study found no consistent gh4 ordering. Figure 5 shows the within-prompt relationship between the gh4 ranking and latent recovery error, and Figure 6 compares the corresponding extraction accuracy with and without selection. This evaluation assesses whether applying the selection mechanism beyond its target distortions provides any additional benefit or introduces a penalty to payload recovery.

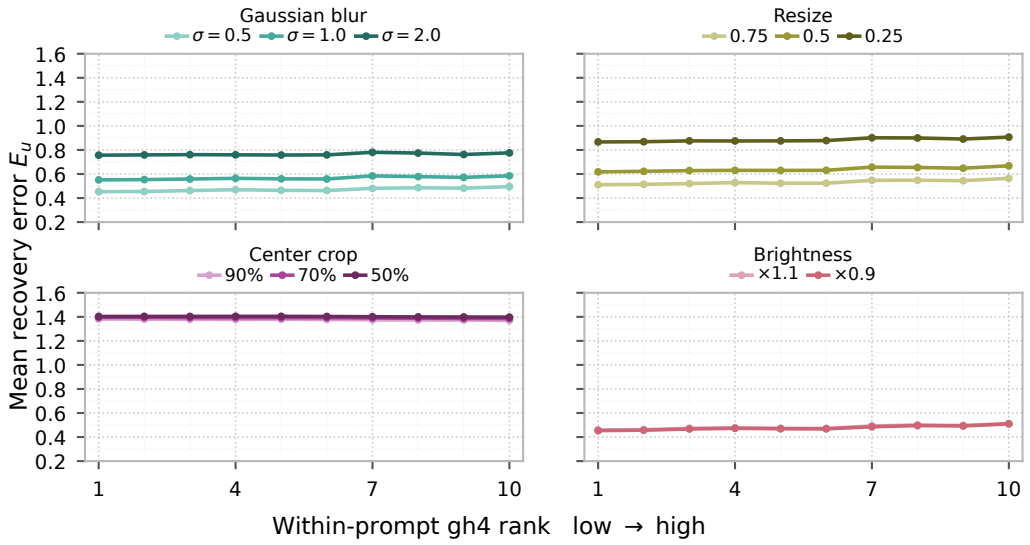


Figure 5: Within-prompt rank-wise recovery error under held-out distortions.

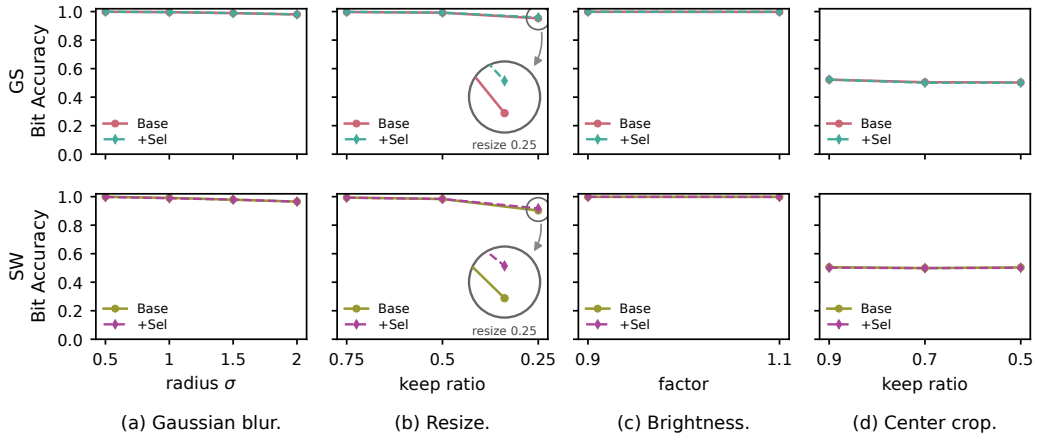


Figure 6: Extraction accuracy under held-out distortions with increasing strength from left to right.

As shown in Figure 5, the rank-wise recovery curves under Gaussian blur, resizing, center crop, and brightness adjustment are nearly flat, indicating that gh4 provides little separation in latent recoverability under these distortions. Consistent with this behavior, Figure 6 shows only marginal im-

provements from selection across these distortion types. Nevertheless, applying the selection mechanism does not reduce extraction accuracy relative to the corresponding base schemes. The benefit of texture-aware initial noise selection is therefore concentrated on the target distortions where gh4 is informative, without degrading payload recovery under the other evaluated distortions.