C^2-STEG: GENERATIVE IMAGE STEGANOGRAPHY FOR INTEGRATED COVERT COMMUNICATION AND COPYRIGHT PROTECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Generative steganography and copyright watermarking, both subsumed under information hiding, have long been studied as separate paradigms. No single model supports covert communication and copyright authentication simultaneously. Direct integration of the two tasks confronts three fundamental challenges: (1) **Steganography capacity zero-sum game**: Simultaneously hiding secret images and watermarks makes secret residuals block watermark extraction, and watermark features leak steganographic traces, jointly compromising both tasks' security. (2) **Semantic representation deficiency**: Neither task is modeled at a semantic level, and joint two tasks in low-level numerical space causes mutual feature interference, degrading recovery fidelity. (3) **Conflicting security boundaries**: A single key for both streams makes secret data and watermarks indistinguishable under coercion: implausible secret denial (weakening steganography) or invalid ownership proof (invalidating authentication). To systematically address these issues, this paper makes the first attempt at a unified diffusion-based generative image steganography framework, **C^2-Steg**, which integrates covert communication and copyright protection in a single pipeline and is supported by the rigorous theoretical foundation. Specifically, the proposed C^2-Steg framework conducts a **shared Discrete Semantic Space** to elevate steganography from low-level numerical space to high-level semantic domain. Feature mutual interference and insufficient semantic representation induced by low-level steganography are thereby eliminated. Moreover, a **Disjoint Dual-key Mask Decoupling mechanism** is designed, by which the concealing space is spatially decoupled to overcome the zero-sum capacity trade-off caused by constrained steganography resources. Furthermore, the dual-key design resolves the conflicting security boundaries: **plausible deniability for steganography** under coercion (the watermark key alone recovers a legitimate watermark while secret images remain undetectable) and verifiability of copyright watermarks. Experiments demonstrate the C^2-Steg's outstanding performance in steganography fidelity, security, and copyright verifiability.

1 INTRODUCTION

Information hiding has served as a cornerstone of secure multimedia communication, with steganography Hu et al. (2024); Kombrink et al. (2024) and digital watermarking Hosny et al. (2024) as its two main branches. Owing to their widespread adoption, images are rendered ideal containers for both techniques. Image steganography aims to conceal secret data within cover images for undetectable communication. In contrast, copyright watermarking intends to embed robust and verifiable ownership signals for reliably proving provenance. Though both tasks share the information hiding goal, they have been studied separately: steganography emphasizes undetectability and high capacity, while watermarking requires robustness and verifiability.

With the maturation of image generation paradigms and the advancement of diffusion models, diffusion-based steganography and watermarking have been the focus of increasing research attention, with promising results being reported. Despite the shared objective of information hiding, a unified diffusion-based model enabling both covert communication and copyright authentication is absent

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.

from the literature. Directly merging the two paradigms, however, is non-trivial and confronts three fundamental challenges:

- **Steganography capacity zero-sum game:** Under conventional steganography pipeline, a unified latent space or pixel representation is shared by secret images and watermark signals. The residuals from secret image interfere with watermark extraction, while watermark concealing introduces statistical artifacts in steganography activity. This mutual interference yields the zero-sum trade-off: increasing one capacity degrades the other, so that the security of both is jointly weakened.
- **Semantic representation deficiency:** In most existing approaches, steganography and watermarking are confined to low-level numerical representations, such as pixel intensities or shallow feature maps. When jointly operated in such spaces, the feature representations become mutually entangled and destructive, severely compromising the reveal quality of both the concealed secret image and the watermark. This interference proves intractable unless higher-level semantic modeling is introduced.
- **Conflicting security boundaries:** Plausible deniability of concealed secret image is essential for steganography, while copyright verifiability is necessary for watermarking. When a single key governs both the secret data and the watermark, an attacker acquiring the key can extract both. With such a single-key design, either the secret is implausibly denied, by which steganography security is weakened, or the copyright verifiability is invalidated, thereby defeating the purpose of authentication.

To systematically address these issues, this paper makes the first attempt at a unified generative image steganography framework C^2**-Steg** that, for the first time, seamlessly **integrates covert communication and copyright protection** within one diffusion generation pipeline. Specifically, C^2-Steg conducts a **shared Discrete Semantic Space** to elevate steganography from low-level numerical space to high-level semantic domain and eliminate feature mutual interference and semantic representation deficiency in low-level space. To overcome the zero-sum capacity trade-off, a **Disjoint Dual-key Mask Decoupling mechanism** is introduced, where the steganography space is spatially partitioned and exclusive sub-regions are assigned to the secret and watermark respectively. Furthermore, this mechanism operates with two separate keys, a secret key and a watermark key, that grant access to their corresponding hidden information independently. Through this dual-key design, the conflicting security demands are reconciled: under coercive interrogation, only the watermark key need be disclosed, upon which a legitimate copyright watermark is recovered while the secret remains entirely undetectable. Consequently, **plausible deniability for steganography** is achieved, and the watermark simultaneously provides verifiability, fulfilling the authentication objective. Finally, the proposed method is supported by a rigorous theoretical foundation, including the distribution-preserving scheme and the collision probability upper bound. The main contributions are summarized as:

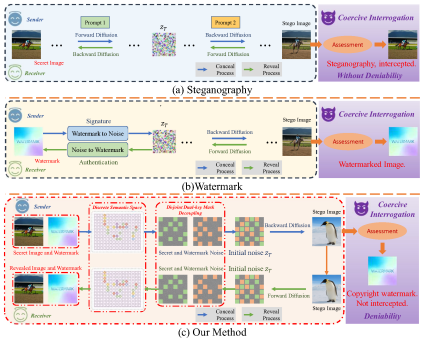

Figure 1: Comparison of overall frameworks across different pipelines. (a) Existing generative image steganography methods lack copyright protection and deniability. (b) Generative watermarking is similar to steganography in pipeline, yet direct integration is prevented. (c) In contrast, the proposed C^2-Steg extends the steganography pipeline toward high-level semantic steganography. Furthermore, deniability, absent from existing methods, is granted to steganography tasks.

- **Theoretical Dual-Functional Unified Framework:** We present C^2-Steg, the first theoretically grounded framework that unifies generative image steganography and copyright watermarking in a single diffusion model, thereby integrating two information hiding paradigms and conferring plausible deniability upon steganography .
- **Error-bounded Semantic Steganography:** We introduce a shared Discrete Semantic Space that advances steganography from low-level numerical space to high-level semantic domain with the closed-form revealing error upper-bound, effectively eliminating feature mutual interference and semantic representation deficiency.

- **Dual-key Decoupling:** To overcome the zero-sum trade-off and the tension between deniability and verifiability, we introduce a collision-bounded Disjoint Dual-key Mask Decoupling mechanism to partition the concealing space into separate secret and watermark regions, guaranteeing zero secret leakage upon watermark key disclosure.

2 Related Work

Existing image steganography methods comprise cover-based steganography Mielikainen (2006); Baluja (2017); Ke et al. (2024); Lu et al. (2021); Jing et al. (2021); Guan et al. (2022); Zhou et al. (2025); Li et al. (2024; 2025), which hides secret messages inside ordinary-looking covers to avoid detection, and generative steganography Yu et al. (2024); Yang et al. (2024); Xu et al. (2025), which has gained traction with the advancement of deep generative models. Generative steganography directly generates stego images from secret data, bypassing clean covers. Thus, this pipeline circumvents the inherent vulnerabilities of cover-based schemes.

Early generative methods largely relied on Generative Adversarial Networks (GANs) Zhu et al. (2017) and flow models Wei et al. (2022). Benefiting from state-of-the-art generation quality, diffusion model-based generative steganography Wei et al. (2023); Chen et al. (2025); Wang et al. (2026) has been established as the dominant approach in the generative steganography. CRoSS Yu et al. (2024) harnesses the translation ability of diffusion models to achieve controllable steganography without extra training. DiffStega Yang et al. (2024) provides universal, training-free coverless steganography by guiding diffusion inversion with a reference image. MDDM Xu et al. (2025) uses a grid coding strategy for the secret message, which enables precise extraction without requiring seeds or prompts.

Meanwhile, digital watermarking Zhu et al. (2024); Kou et al. (2025); Chung et al. (2026); Yang et al. (2026) has progressed from a posteriori concealing on existing images to generation-time concealing, where watermarks are integrated during generation. More related work are presented in **Appendix A**.

Motivated by the shared objective of concealing information, this paper proposes a theoretically grounded dual-functional framework, built upon diffusion models, that unifies image steganography and copyright watermarking in a shared discrete semantic space and further confers plausible deniability upon steganography tasks.

3 Proposed Methods

3.1 Framework

This paper propose a novel dual-functional generative image steganography framework, C^2**-Steg**, with its flowchart illustrated in Figure 2. For the first time, steganography and copyright watermarking are unified within a shared Discrete Semantic Space. During conceal process, the secret image x_s and watermark x_m are tokenized into discrete semantic tokens and transformed into standard Gaussian noise. Two non-overlapping masks (Secret and Watermark Masks), derived from a true key k_{sec} and a fake key k_{wm} through Disjoint Dual-key Mask Decoupling mechanism, determine where the token-derived noise is concealed, with remaining positions filled by random Gaussian noise. This noise initializes the diffusion process to generate the stego image x_{st}. At the receiver, the stego image is mapped back to Gaussian noise, and the corresponding masks extract the secret and watermark tokens, which are de-tokenized into the revealed secret image $\tilde{x}_s$ and watermark $\tilde{x}_m$. Under coercion, the fake key alone can be disclosed to extract the watermark, providing plausible deniability by presenting the stego image as a randomly generated, watermarked picture without concealed secret data. **The conceal and reveal algorithms are provided in Appendix B.**

3.2 Modeling within the Discrete Semantic Space

A pretrained VQGAN Sun et al. (2024) is composed of an encoder $E : \mathcal{X} \to \mathcal{S}$, a codebook $\mathcal{E} = \{\mathbf{e}_k\}_{k=0}^{K-1} \subset \mathbb{R}^d$, and a decoder $D : \mathcal{S} \to \mathcal{X}$, where $\mathcal{X} = \mathbb{R}^{3\times256\times256}$ denotes the image space, and $\mathcal{S} = \{0, 1, \ldots, K-1\}^L$ is the discrete semantic space with $K = 16384$ and $L = 1024$.

Definition 1. (**Semantic Encoding**) For any secret image $x_s \in \mathcal{X}$, its semantic encoding $\mathbf{s} = E(x_s)$ is obtained through the following two steps:

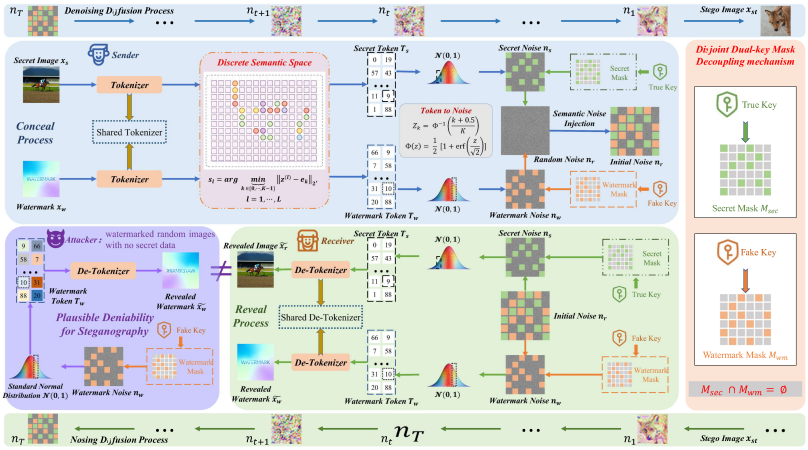

Figure 2: Pipeline of the proposed model. During concealing, the secret image x_s and watermark x_m are tokenized into discrete semantic tokens and converted to standard Gaussian noise. Two non-overlapping masks, derived from the true and fake keys, specify the concealing positions. This noise, containing the secret and watermark, initializes the diffusion process to generate the stego image x_{st}. During revealing, the secret image $\tilde{x}_s$ and watermark $\tilde{x}_m$ are recovered by inverting the conceal process. For deniability, under coercion, revealing with only the fake key extracts the watermark, supporting the claim that the stego image is merely a randomly watermarked image.

1. **Continuous Feature Extraction:** The convolutional encoder maps x_s to a feature map $\mathbf{Z} \in \mathbb{R}^{h\times w\times d}$, where $h = w = 32$ and d is the dimensionality of the code vectors. The vector $\mathbf{z}^{(\ell)} \in \mathbb{R}^d$ at each spatial position ℓ is termed the semantic embedding at that location, which captures the local visual semantics within the corresponding receptive field.

2. **Semantic Quantization:** At each position ℓ, the embedding $\mathbf{z}^{(\ell)}$ is mapped to the nearest codebook vector through nearest-neighbor lookup:

$$s_\ell = \arg\min_{k\in\{0,\dots,K-1\}} \|\mathbf{z}^{(\ell)} - \mathbf{e}_k\|_2, \quad \ell = 1,\dots,L. \tag{1}$$

The resulting index s_ℓ is the semantic token at position ℓ, and the entire sequence $\mathbf{s} = (s_1,\dots,s_L)$ constitutes the semantic token sequence of the image.

This formulation elevates the discretization process of VQGAN to the semantic level. Each codebook vector $\mathbf{e}_k$ represents a specific class of local visual patterns, and the index k serves as the semantic label for that pattern. The encoder E thus realizes a mapping from continuous images to discrete semantic representations.

3.3 Transform Semantic Tokens to Gaussian Noise

To map a discrete token s to a continuous value that matches the distribution of the standard normal distribution $\mathcal{N}(0,1)$, an equal-probability quantile mapping is employed. Let the codebook size be K, and let ϕ be the standard normal density. For each token $s \in \{0,\dots,K-1\}$, define the equal-probability interval:

$$I_s = \left(\Phi^{-1}\left(\frac{s}{K}\right),\ \Phi^{-1}\left(\frac{s+1}{K}\right)\right). \tag{2}$$

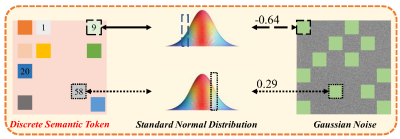

Figure 3: Mapping from Semantic Tokens to Gaussian Noise.

These intervals partition the real line into K segments, each carrying probability $1/K$ under the standard normal distribution.

3.3.1 Randomized Quantile Mapping

Definition 2 (Randomized Quantile Mapping)**.** For a token $s \in \{0,\dots,K-1\}$, independently sample:

$$U \sim \text{Uniform}\left(\frac{s}{K}, \frac{s+1}{K}\right), \qquad z = \Phi^{-1}(U). \tag{3}$$

Then z lies in the interval I_s. Conditioned on s, z follows the standard normal distribution truncated to I_s, denoted by μ_s, whose density is $f_s(z) = K\,\phi(z)\,\mathbf{1}_{z\in I_s}$. If the token prior is $\pi = (\pi_0, \ldots, \pi_{K-1})$, the marginal distribution of a single concealed coordinate is the mixture $P_z = \sum_{s=0}^{K-1} \pi_s \mu_s$. The standard normal distribution can be written as the uniform mixture $Q_z = \mathcal{N}(0,1) = \sum_{s=0}^{K-1} \frac{1}{K}\mu_s$.

Proposition 1 (Exact Preservation under Uniform Prior)**.** *If the token prior is uniform, i.e., $\pi_s = 1/K$ for all s, then ubstituting $\pi_s = 1/K$ into $P_z = \sum_{s=0}^{K-1} \pi_s \mu_s$ gives $P_z = \sum_{s=0}^{K-1} \frac{1}{K}\mu_s = Q_z$. Hence$P_z = Q_z = \mathcal{N}(0,1)$. That is, each concealed coordinate is exactly standard normal.*

Definition 3 (Token Prior Deviation)**.** For a token prior $\pi = (\pi_0, \ldots, \pi_{K-1})$, define:

$$\delta_{\text{token}} = \sum_{s=0}^{K-1} \left| \pi_s - \frac{1}{K} \right|. \tag{4}$$

This quantity measures the deviation of the token prior from uniformity.

3.3.2 Distribution Preservation

Lemma 1 (TV Bound for Mixture Measures)**.** *Let $\{\mu_s\}_{s=0}^{K-1}$ be arbitrary probability measures, and let $\pi = (\pi_0, \ldots, \pi_{K-1})$ and $u = (1/K, \ldots, 1/K)$ be probability distributions on $\{0, \ldots, K-1\}$. Then:*

$$d_{\text{TV}}\left(\sum_{s=0}^{K-1} \pi_s \mu_s,\ \sum_{s=0}^{K-1} u_s \mu_s \right) \le \sum_{s=0}^{K-1} \left| \pi_s - \frac{1}{K} \right|. \tag{5}$$

Theorem 1 (Distribution Preservation under Randomized Quantile Mapping)**.** *Assume that the secret token sequence and the watermark token sequence are independent and identically distributed according to the token prior π. For each token, the R repetition copies are sampled independently according to Definition 2. Let the secret mask and the watermark mask be disjoint, with:*

$$\mathbf{M}_{\text{sec}} \cap \mathbf{M}_{\text{wm}} = \emptyset, \qquad |\mathbf{M}_{\text{sec}}| = |\mathbf{M}_{\text{wm}}| = LR. \tag{6}$$

Let $A = \mathbf{M}_{\text{sec}} \cup \mathbf{M}_{\text{wm}}$ and $N_h = |A| = 2LR$. Fill the remaining coordinates with i.i.d. $\eta_i \sim \mathcal{N}(0,1)$. Construct $\mathbf{x}_T$ by placing the concealed values at the positions specified by the masks. Let P be the distribution of $\mathbf{x}_T$, and let $Q = \mathcal{N}(\mathbf{0}, \mathbf{I}_N)$. Then:

$$\boxed{d_{\text{TV}}(P, Q) \le N_h\, \delta_{\text{token}}.} \tag{7}$$

In particular, if the token prior is uniform, i.e., $\pi_s = 1/K$, then $\delta_{\text{token}} = 0$, and hence:

$$P = Q, \qquad d_{\text{TV}}(P, Q) = 0. \tag{8}$$

If diffusion inversion introduces an error $d_{\text{TV}}(\hat{P}, P) \le \varepsilon_{\text{inv}}$, then:

$$d_{\text{TV}}(\hat{P}, Q) \le N_h\, \delta_{\text{token}} + \varepsilon_{\text{inv}}. \tag{9}$$

3.3.3 Repetition Coding and Diversity Reception

For each token s, its R repetition copies must be sampled independently. That is, for $r = 1, \ldots, R$, sample:

$$U_{s,r} \sim \text{Uniform}\left(\frac{s}{K}, \frac{s+1}{K} \right), \qquad z_{s,r} = \Phi^{-1}(U_{s,r}). \tag{10}$$

Thus all concealed coordinates are independent and identically distributed according to P_z. This independence is essential for the product-measure argument in Theorem 1. If the repetition copies reuse the same randomness, independence is broken and the total-variation bound no longer holds.

At the receiver, after inversion, the R observations for a token are decoded as follows. For each observation y_r, $u_r = \Phi(y_r)$ is computed. The token index is then estimated as $s_r = \lfloor K\, u_r \rfloor$ and majority voting is applied over the R estimates to obtain $\hat{s}$. This diversity reception scheme exploits the independent randomness of the R copies to reduce the probability that noise pushes an observation across an interval boundary.

3.3.4 Maximum Likelihood Detection and Error Probability

Given an observation y, the maximum likelihood (ML) detector is:

$$\hat{s} = \arg \max_{s \in \{0,\ldots,K-1\}} p(y \mid s), \tag{11}$$

where:

$$p(y \mid s) = \int_{I_s} \phi(z)\, \mathcal{N}(y - z; 0, \sigma^2)\, \mathrm{d}z = \frac{1}{\sqrt{2\pi}\sigma} \int_{I_s} \phi(z) \exp\left(-\frac{(y-z)^2}{2\sigma^2} \right) \mathrm{d}z. \tag{12}$$

Under a uniform token prior, the ML rule simplifies to the interval decision $\hat{s} = \lfloor K\,\Phi(y) \rfloor$. That is, y is assigned to the token whose equal-probability interval contains $\Phi(y)$.

The single-copy error probability for token s is:

$$P_e(s) = \int_{I_s} \phi(z) \left[Q\left(\frac{b_s^+ - z}{\sigma} \right) + Q\left(\frac{z - b_s^-}{\sigma} \right) \right] \mathrm{d}z, \tag{13}$$

where $b_s^- = \Phi^{-1}\left(\frac{s}{K}\right)$, $b_s^+ = \Phi^{-1}\left(\frac{s+1}{K}\right)$, and $Q(\cdot)$ is the standard normal tail probability. For R independent copies with majority voting, the final error probability is:

$$P_e^{(R)} = \sum_{k=\lceil R/2 \rceil}^{R} \binom{R}{k} P_e^k (1 - P_e)^{R-k}, \tag{14}$$

where P_e is the average single-copy error probability. If P_e is small, $P_e^{(R)} \approx \binom{R}{\lceil R/2 \rceil} P_e^{\lceil R/2 \rceil}$.

3.4 Disjoint Dual-key Mask Decoupling

Candidate pool partition. Let the full coordinate set be $\Omega = \{0, 1, \ldots, N-1\}$. A deterministic public partition function is fixed as $\mathrm{Partition} : \Omega \to \{\mathrm{sec}, \mathrm{wm}\}$, and the two candidate pools are defined as:

$$\Omega_{\mathrm{sec}} = \{i \in \Omega : \mathrm{Partition}(i) = \mathrm{sec}\}, \qquad \Omega_{\mathrm{wm}} = \{i \in \Omega : \mathrm{Partition}(i) = \mathrm{wm}\}. \tag{15}$$

The partition satisfies $\Omega_{\mathrm{sec}} \cap \Omega_{\mathrm{wm}} = \emptyset$, $\Omega_{\mathrm{sec}} \cup \Omega_{\mathrm{wm}} = \Omega$.

A simple and concrete instantiation is the parity partition $\Omega_{\mathrm{sec}} = \{i : i \bmod 2 = 0\}$, $\Omega_{\mathrm{wm}} = \{i : i \bmod 2 = 1\}$. For $N = 3 \times 256 \times 256 = 196608$, both pools have size $N/2 = 98304$.

Mask generation. The secret and watermark masks are generated as:

$$\mathbf{M}_{\mathrm{sec}} = \mathrm{MaskGen}(k_s, LR, \Omega_{\mathrm{sec}}), \qquad \mathbf{M}_{\mathrm{wm}} = \mathrm{MaskGen}(k_w, LR, \Omega_{\mathrm{wm}}). \tag{16}$$

When $L = 1024$ denotes the number of tokens and $R = 5$ the repetition factor, so that each mask requires $LR = 5120$ coordinates. Since $|\Omega_{\mathrm{sec}}| = |\Omega_{\mathrm{wm}}| = 98304$, the capacity is sufficient.

The deterministic algorithm $\mathrm{MaskGen}(k, n, \Omega_{\mathrm{pool}})$ proceeds as: the full SHA-256 digest of k is computed and used as a 256-bit seed for a cryptographically secure pseudorandom number generator. Rejection sampling is then applied to sequentially generate n distinct coordinates belonging to Ω_{pool}, yielding an ordered set $\mathbf{M} \subset \Omega_{\mathrm{pool}}$. Because the algorithm is deterministic, both communicating parties independently produce identical masks. Since Ω_{sec} and Ω_{wm} are disjoint, the resulting masks satisfy $\mathbf{M}_{\mathrm{sec}} \cap \mathbf{M}_{\mathrm{wm}} = \emptyset$ by construction.

3.5 Semantic Noise Injection and Stego Generation

Background noise is sampled as $\boldsymbol{\eta} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_N)$, and the coordinates specified by the masks are overwritten as follows:

$$\mathbf{x}_T[i] = \begin{cases} z_{\mathrm{sec}}^{(j)}, & i = \mathbf{M}_{\mathrm{sec}}[j], \\ z_{\mathrm{wm}}^{(j)}, & i = \mathbf{M}_{\mathrm{wm}}[j], \\ \eta_i, & \text{otherwise.} \end{cases} \tag{17}$$

After reshaping $\mathbf{x}_T$ to $(1, 3, 256, 256)$, the deterministic BELM Wang et al. (2024) $\mathbf{G}$ is executed to generate the stego image x_{st} as follows:

$$x_{st} = \mathbf{G}(\mathbf{x}_T). \tag{18}$$

Finally, the pixel values are scaled from $[-1, 1]$ to $[0, 1]$.

3.6 Secret Extraction and Inherent Deniability

3.6.1 Receiver-side Processing

Upon receiving stego image x_{st}, the receiver first performs inversion to obtain $\hat{\mathbf{x}}_T = \mathbf{G}^{-1}(x_{st})$. According to the key k it holds (k_{sec} or k_{wm}), the corresponding mask $\mathbf{M}$ is regenerated, and the received values are extracted as $\mathbf{y} = \hat{\mathbf{x}}_T[\mathbf{M}]$. These values are partitioned into L groups of size R. Within each group, the mean $\bar{y}_i$ is computed, and an ML decision recovers the token $\hat{s}_i$. Finally, the token sequence is decoded by VQGAN to obtain the revealed secret image $\tilde{x}_s$ and watermark $\tilde{x}_m$.

Method	Nvidia RTX 4090		Nvidia RTX 3090		Stego images		Secret Images		Copyright Watermarks	
	Time(s)	Memory(GB)	Time(s)	Memory(GB)	PSNR↓	SSIM↓	PSNR↑	SSIM↑	PSNR↑	SSIM↑
CRoSS	14.41	7.43	39.38	8.58	19.03	0.66	21.25	0.71	-	-
DiffStega	20.90	8.35	41.00	9.02	**18.61**	**0.59**	23.29	0.77	-	-
DiffStega‡	20.90	8.35	41.00	9.02	19.73	0.65	**23.92**	**0.79**	-	-
Ours	10.57	10.39	33.94	11.43	**9.27 (9.34↓)**	**0.23 (0.36↓)**	**24.37 (0.45↑)**	**0.81 (0.02↑)**	**24.28**	**0.82**

Table 1: Numerical comparisons with generative steganography methods on the UniStega dataset.

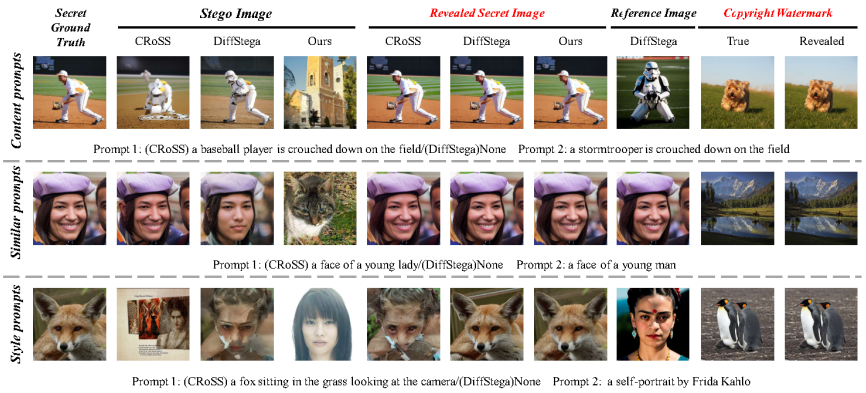

Figure 4: Visualization comparison results of the proposed model versus other methods. The text prompts are utilized in CRoSS and DiffStega and the reference image is used in DiffStega.

3.6.2 Deniability

Because both the secret and the watermark are embedded in the same stego image, C^2-Steg possesses inherent plausible deniability, which is absent from existing generative steganography. Under coercive interrogation, only the watermark key k_{wm} is disclosed, from which only the watermark x_m can be revealed. Given that no secret data are contained in the watermark and that, through the Disjoint Dual-key Mask Decoupling mechanism, the watermark concealing region $\mathbf{M}_{\text{wm}}$ is spatially decoupled from the secret concealing region $\mathbf{M}_{\text{sec}}$, no secret information is recoverable with this key. The presence of concealed secret image can thus be denied, and the security of the secret image is maintained.

4 Experiments

4.1 Experimental Setting

The proposed framework utilizes the VQGAN Sun et al. (2024) with a codebook size of 16384 and a downsample ratio of 8 as the Tokenizer and De-Tokenizer, and leverages BELM Wang et al. (2024) to generate stego images. Implemented in PyTorch, the framework is trained and evaluated on UniStega Yang et al. (2024), DIV2K Agustsson & Timofte (2017) and Stega260 Yu et al. (2024) datasets with the resolution of 256×256. Further details are in the **Appendix D**.

4.2 Quality Analysis

Quantitative results.

Numerical results comparing the proposed framework with other generative steganography models are summarized in Table 1 and Table 2, with evaluations conducted across three datasets. Quantitative results on the UniStega dataset are shown in Table 1, while results on DIV2K and Stega260 are listed in Table 2. On UniStega, the proposed model achieves a 0.45dB PSNR gain for recovered secret data with a slight SSIM improvement, demonstrating clear progress. Meanwhile, when evaluating secret–stego pairs, a 9.34 dB PSNR drop and a 36% SSIM decrease are obtained relative to baselines, greatly decreasing the visual similarity between stego and secret images. The proposed model also reconstructs high-quality copyright watermarks, offering copyright protection and endowing the system with deniability. Similar trends are evident on DIV2K and Stega260 datasets. These results indicate that the C^2-Steg outperforms competitors in steganography quality and recovery fidelity, while resolving the absence of deniability in existing methods.

Dataset	Stego Image		Revealed Image		Copyright Watermarks	
	PSNR↓	SSIM↓	PSNR↑	SSIM↑	PSNR↑	SSIM↑
DIV2K	9.16	0.18	23.94	0.78	24.09	0.79
Stega260	9.31	0.22	24.39	0.81	24.51	0.82

Table 2: Numerical Results on Further Datasets.

Ratio f	Stego Image		Revealed Image		Copyright Watermarks	
	PSNR↓	SSIM↓	PSNR↑	SSIM↑	PSNR↑	SSIM↑
16	9.26	0.24	20.84	0.68	20.79	0.68
8	9.27	0.23	24.37	0.81	24.28	0.82

Table 3: Effect of the downsample ratio f.

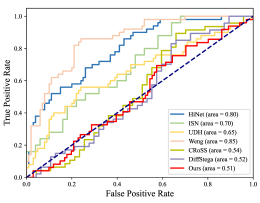

Figure 5: Security assessed by StegExpose.

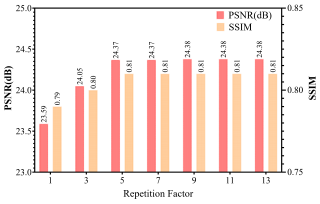

Figure 6: Ablation on the Repetition Factor R.

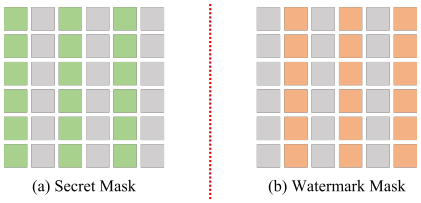

Figure 7: Illustration of Candidate Positions for the Secret and Watermark Masks.

Qualitative Results. Qualitative comparisons of stego images, revealed secret images, and copyright watermarks are presented in Figure 4. In the evaluation, natural images are randomly selected as copyright watermarks for the proposed method. As shown in the figure, a substantial reduction in the similarity between stego images and secret images is achieved compared with prompt-guided image editing methods, while reasonable improvements are attained in the quality of revealed secret images. High-quality copyright watermarks are also revealed, a capability absent from all competing approaches. This functionality equips the steganography system with deniability. Under adversarial coercion, the fake key alone can be surrendered to extract the copyright watermark; the stego image can then be claimed to be a randomly generated, watermarked image, a statement fully substantiated by the high fidelity of the revealed watermark. These results demonstrate that the proposed method simultaneously enhances the visual quality of stego images and the security of secret images, while addressing the absence of deniability.

4.3 Security Analysis

Naturalness and Imperceptibility. The naturalness and visual security of images are assessed using NIQE Mittal et al. (2012) , which operates without reference images or human feedback. As reported in Table 4, a 0.11 decrease in NIQE is achieved by the proposed model relative to other generative steganographic methods. This improvement demonstrates that the proposed model offers superior naturalness and imperceptibility.

	Original image	ISN	HiNet	Cross	DiffStega	**Ours**
NIQE↓	5.14	11.28	**5.35**	5.60	5.58	**5.47**

Table 4: NIQE scores for various models.

Steganography analysis. Anti-steganalysis performance is regarded as a crucial metric in image steganography security, defined by the probability that stego images are distinguished by steganalytic detectors. The evaluation is conducted using the publicly available tool StegExpose Boehm (2014) and three deep steganalysis networks: SRNetBoroumand et al. (2018), XuNet Xu et al. (2016), and YeNet Ye et al. (2017). A lower detection accuracy and a reduced AUC value correspond to higher security. The detailed evaluation outcomes are presented in Table 5 and Figure 5, respectively. Through these results, the proposed model is shown to outperform existing SOTA methods in terms of anti-steganalysis capability.

	Cover-based Method				Generative Method		
	Weng	UDH	ISN	HiNet	CRoSS	DiffStega	Ours
SRNet	81.47	76.34	54.86	53.52	51.73	51.68	**50.62 (1.06↓)**
XuNet	85.31	79.82	55.68	55.54	52.10	52.03	**50.96 (1.07↓)**
YeNet	86.24	80.26	55.42	55.37	53.12	52.86	**51.07 (1.79↓)**

Table 5: The detection accuracy (%) of various methods as evaluated by SRNet, XuNet, and YeNet.

4.4 Ablation Study

Effect of Repetition Factor R. To mitigate token mismatch errors introduced during extraction, the token-to-noise mapping for each token is repeated R times in the proposed framework. Figure 6 reports the performance across different Repetition Factor R. It is observed that the system is most susceptible to errors when $R = 1$. This vulnerability arises from the narrow span of each quantized subinterval, which is easily distorted and leads to erroneous recovery of discrete semantic tokens in the extraction phase. The effect of errors diminishes as R increases, and for $R \geq 5$, the performance gain becomes marginal, so $R = 5$ is set as the default. These results validate the effectiveness of the token-to-noise mapping module integrated within the proposed framework.

Effect of the downsample ratio f. The proposed method uses VQGAN as the tokenizer to extract discrete semantic tokens from secret images and copyright watermarks, where the downsampling ratio f determines the resulting token count. Table 3 reports the ablation results over commonly used downsampling factors. For $f = 16$, the proposed method achieves steganography performance comparable to CRoSS, while yielding a greater reduction in secret–stego similarity. As f decreases, the performance improves; at $f = 8$, the proposed method outperforms CRoSS and DiffStega and retains a strong capability to lower secret–stego similarity. These findings demonstrate that the method effectively balances steganography quality and security.

Effectiveness of Dual-Key Deniability. Two non-overlapping masks are generated using a true key and a fake key to separately conceal the secret image and the copyright watermark. Deniability is provided by the fake key together with the concealed watermark. As shown in Figure 7, the secret mask occupies odd-indexed nodes, while the watermark mask occupies even-indexed nodes. Since disjoint candidate positions are predefined for the two masks, they are inherently non-overlapping. Hence, the secret concealing positions remain undisclosed when the fake key is disclosed, demonstrating the effective spatial isolation of the masks. These results validate the dual-key deniability and secret information cannot be revealed solely with the fake key.

Effectiveness of Token-to-Noise Mapping. To meet the initial noise requirements of diffusion models, discrete semantic tokens are transformed into standard Gaussian noise through a transformation before concealing to satisfy the initial noise requirements of diffusion models. Figure 8 illustrates the statistical distributions of the initial noise before and after concealing the secret images and copyright watermarks. It is observed that concealing the secret images and watermarks exerts a negligible impact on the noise distribution. Under simultaneous concealing, only minor discrepancies are noticed in a few outliers, and the overall distribution remains unaltered. These results validate the effectiveness of the token-to-noise mapping in preserving the original noise standard normal distribution $\mathcal{N}(\mathbf{0}, \mathbf{I}_N)$. More experiments are presented in the **Appendix E**.

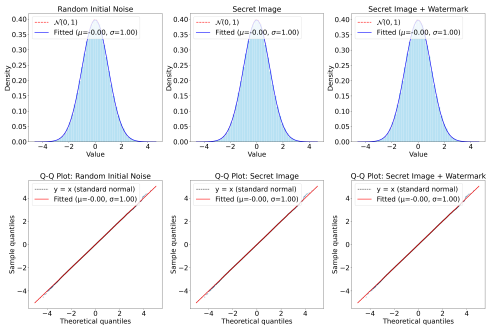

Figure 8: Statistical distributions of initial noise before and after concealing secret images and copyright watermarks

5 Conclusion

This paper proposes C^2-Steg, a unified diffusion-based generative image steganography framework with a rigorous theoretical foundation, in which steganography and watermarking are unified for the first time. Within a shared discrete semantic space, the secret image and watermark are tokenized into semantic tokens, elevating steganography from a low-level numerical space to a high-level semantic domain and thereby addressing feature interference and insufficient semantic representation. The introduced Disjoint Dual-key Mask Decoupling mechanism maps these tokens into standard Gaussian noise concealed within the initial noise of the diffusion model, achieving seamless integration with diffusion models and resolving the steganography resource zero-sum game. Moreover, plausible deniability is provided by the dual-key mechanism: under coercive interrogation, the stego image is identified solely as a randomly generated image with a watermark, fully denying the existence of the secret image. Extensive experiments demonstrate superior stego image quality and overall security.

6 AI USE STATEMENT

Large language models (LLMs) were employed solely for manuscript polishing and limited code debugging. Specifically, LLMs assisted with language refinement to improve clarity and readability, and with debugging of non-core utility scripts. All research ideation, experimental design, implementation of the core methodology, data analysis, and interpretation of results were performed entirely by the authors. All LLM-generated outputs were reviewed, verified, and edited by the authors, who assume full responsibility for the content of this work.

7 REPRODUCIBILITY STATEMENT

To ensure reproducibility, all datasets, model configurations, training scripts, evaluation code, and result files will be publicly released upon acceptance. Comprehensive documentation covering data preprocessing, hyperparameter settings, and experimental protocols necessary for reproducing the main results reported in this paper will also be provided. During the review period, detailed descriptions of the experimental setup and methodology have been provided in the main text and appendix to facilitate assessment. All reported experimental results are authentic and reproducible.

REFERENCES

Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 126–135, 2017.

Shumeet Baluja. Hiding images in plain sight: Deep steganography. *Advances in Neural Information Processing Systems*, 30, 2017.

Grigory Bartosh, Dmitry Vetrov, and Christian A. Naesseth. Neural flow diffusion models: Learnable forward process for improved diffusion modelling. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Benedikt Boehm. Stegexpose - a tool for detecting lsb steganography, 2014.

Mehdi Boroumand, Mo Chen, and Jessica Fridrich. Deep residual network for steganalysis of digital images. *IEEE Transactions on Information Forensics and Security*, 14(5):1181–1193, 2018.

Karl Chahine and Hyeji Kim. Neural cover selection for image steganography. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Haoyu Chen, Yunqiao Yang, Nan Zhong, and Kede Ma. Hiding images in diffusion models by editing learned score functions. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 18663–18673, 2025.

Jun Woo Chung, Yingjie Lao, and Weijie Zhao. Robust watermarking on gradient boosting decision trees. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 20624–20633, 2026.

Zhenyu Guan, Junpeng Jing, Xin Deng, Mai Xu, Lai Jiang, Zhou Zhang, and Yipeng Li. Deepmih: Deep invertible network for multiple image hiding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1):372–390, 2022.

Guande He, Kaiwen Zheng, Jianfei Chen, Fan Bao, and Jun Zhu. Consistency diffusion bridge models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Khalid M Hosny, Amal Magdi, Osama ElKomy, and Hanaa M Hamza. Digital image watermarking using deep learning: A survey. *Computer science review*, 53:100662, 2024.

Kun Hu, Mingpei Wang, Xiaohui Ma, Jia Chen, Xiaochao Wang, and Xingjun Wang. Learning-based image steganography and watermarking: A survey. *Expert Systems with Applications*, 249:123715, 2024.

Junpeng Jing, Xin Deng, Mai Xu, Jianyi Wang, and Zhenyu Guan. Hinet: Deep image hiding by invertible network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4733–4742, 2021.

Xiaolong Ke, Huanqi Wu, and Wenzhong Guo. Stegformer: Rebuilding the glory of autoencoder-based steganography. In *AAAI Conference on Artificial Intelligence*, 2024.

Meike Helena Kombrink, Zeno Jean Marius Hubert Geradts, and Marcel Worring. Image steganography approaches and their detection strategies: A survey. *ACM Computing Surveys*, 57(2):1–40, 2024.

Feifei Kou, Yuhan Yao, Siyuan Yao, Jiahao Wang, Lei Shi, Yawen Li, and Xuejing Kang. Iwrn: A robust blind watermarking method for artwork image copyright protection against noise attack. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 370–378, 2025.

Fengyong Li, Yang Sheng, Xinpeng Zhang, and Chuan Qin. iscmis:spatial-channel attention based deep invertible network for multi-image steganography. *IEEE Transactions on Multimedia*, 26: 3137–3152, 2024.

Fengyong Li, Hao Liu, Xinpeng Zhang, and Chuan Qin. Stegflow: Flow-based high-frequency distribution mapping network for multi-image steganography. *IEEE Transactions on Multimedia*, 2025.

Shao-Ping Lu, Rong Wang, Tao Zhong, and Paul L Rosin. Large-capacity image steganography based on invertible neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10816–10825, 2021.

Jarno Mielikainen. Lsb matching revisited. *IEEE signal processing letters*, 13(5):285–287, 2006.

Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal processing letters*, 20(3):209–212, 2012.

Sumeet Ramesh Motwani, Mikhail Baranchuk, Martin Strohmeier, Vijay Bolina, Philip Torr, Lewis Hammond, and Christian Schroeder de Witt. Secret collusion among AI agents: Multi-agent deception via steganography. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Gabriel Nobis, Maximilian Springenberg, Marco Aversa, Michael Detzel, Rembert Daems, Roderick Murray-Smith, Shinichi Nakajima, Sebastian Lapuschkin, Stefano Ermon, Tolga Birdal, Manfred Opper, Christoph Knochenhauer, Luis Oala, and Wojciech Samek. Generative fractional diffusion models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Ilker Oguz, Niyazi Ulas Dinc, Mustafa Yildirim, Junjie Ke, Innfarn Yoo, QIFEI WANG, Feng Yang, Christophe Moser, and Demetri Psaltis. Optical diffusion models for image generation. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

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.

Peize Sun, Yi Jiang, Shoufa Chen, Shilong Zhang, Bingyue Peng, Ping Luo, and Zehuan Yuan. Autoregressive model beats diffusion: Llama for scalable image generation, 2024.

Bram Wallace, Akash Gokul, and Nikhil Naik. Edict: Exact diffusion inversion via coupled transformations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22532–22541, 2023.

Fangyikang Wang, Hubery Yin, Yue-Jiang Dong, Huminhao Zhu, Chao Zhang, Hanbin Zhao, Hui Qian, and Chen Li. BELM: Bidirectional explicit linear multi-step sampler for exact inversion in diffusion models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Jiannian Wang, Yao Lu, and Guangming Lu. Large-capacity and receiver authenticable generative image steganography. In *Forty-third International Conference on Machine Learning*, 2026.

Ping Wei, Ge Luo, Qi Song, Xinpeng Zhang, Zhenxing Qian, and Sheng Li. Generative steganographic flow. In *2022 IEEE International Conference on Multimedia and Expo*, pp. 1–6. IEEE, 2022.

Ping Wei, Qing Zhou, Zichi Wang, Zhenxing Qian, Xinpeng Zhang, and Sheng Li. Generative steganography diffusion, 2023.

Guanshuo Xu, Han-Zhou Wu, and Yun-Qing Shi. Structural design of convolutional neural networks for steganalysis. *IEEE Signal Processing Letters*, 23(5):708–712, 2016.

Zihao Xu, Zihan Li, Chuan Zhang, et al. Mddm: Practical message-driven generative image steganography based on diffusion models. In *Forty-second International Conference on Machine Learning*, 2025.

Yiwei Yang, Zheyuan Liu, Jun Jia, Zhongpai Gao, Yunhao Li, Wei Sun, Xiaohong Liu, and Guangtao Zhai. Diffstega: Towards universal training-free coverless image steganography with diffusion models. In *33rd International Joint Conference on Artificial Intelligence (IJCAI)*, 2024.

Yunfei Yang, Xiaojun Chen, Yuexin Xuan, Zhendong Zhao, Xin Zhao, and He Li. Deeptracer: Tracing stolen model via deep coupled watermarks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 27711–27718, 2026.

Jian Ye, Jiangqun Ni, and Yang Yi. Deep learning hierarchical representations for image steganalysis. *IEEE Transactions on Information Forensics and Security*, 12(11):2545–2557, 2017.

Jiwen Yu, Xuanyu Zhang, Youmin Xu, and Jian Zhang. Cross: Diffusion model makes controllable, robust and secure image steganography. *Advances in Neural Information Processing Systems*, 36, 2024.

Junchao Zhou, Yao Lu, Jie Wen, and Guangming Lu. Efficient and separate authentication image steganography network. In *Forty-second International Conference on Machine Learning*, 2025.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international Conference on Computer Vision*, pp. 2223–2232, 2017.

Xingyu Zhu, Guanhui Ye, Xiapu Luo, and Xuetao Wei. Rethinking mesh watermark: Towards highly robust and adaptable deep 3d mesh watermarking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 7784–7792, 2024.

A Detailed Related Work

A.1 Image Steganography and Watermark

Existing image steganography methods comprises cover-based steganography Mielikainen (2006); Baluja (2017); Ke et al. (2024); Lu et al. (2021); Guan et al. (2022); Zhou et al. (2025); Li et al. (2024; 2025), which hides secret messages inside ordinary-looking covers to avoid detection, and generative steganography Yu et al. (2024); Yang et al. (2024); Xu et al. (2025), which has gained traction with the advancement of deep generative models. Generative steganography directly generates stego images from secret data, bypassing clean covers. Thus, this pipeline circumvents the inherent vulnerabilities of cover-based steganography schemes and nullifies the typical pairwise comparison exploited by steganalysis, severely degrading detection reliability of the steganalysis methods.

Early methods largely relied on Generative Adversarial Networks (GANs) Zhu et al. (2017) and flow models Wei et al. (2022). Benefiting from state-of-the-art generation quality, diffusion model-based generative steganography Chen et al. (2025); Wang et al. (2026) has been established as the dominant approach in the generative steganography. GSD Wei et al. (2023) conceals secret data into the frequency domain of the initial Gaussian noise and CRoSS Yu et al. (2024) harnesses the translation ability of diffusion models to achieve controllable steganography without extra training. DiffStega Yang et al. (2024) provides universal, training-free coverless steganography by guiding diffusion inversion with a reference image. MDDM Xu et al. (2025) uses a grid coding strategy for the secret message, which enables precise extraction without requiring seeds or prompts.

Meanwhile, digital watermarking Zhu et al. (2024); Kou et al. (2025); Chung et al. (2026); Yang et al. (2026) has progressed from a posteriori concealing on existing images to generation-time concealing, where watermarks are integrated during generation. Recently, diffusion models are also adopted for enhanced robustness and invisibility.

However, reliance on prompt-based image editing pipelines causes existing generative steganography to introduce abundant secret-related artifacts into the stego image while entirely omitting copyright protection. To overcome these deficiencies, this paper introduces a dual-functional scheme that unifies secret image concealing and copyright watermarking within a unified discrete semantic space, and further equips the system with strong plausible deniability.

A.2 Diffusion Models

Diffusion models, which are founded on a progressive denoising Markov process, have been established as the leading generative modeling paradigm. A latent diffusion model was later introduced by Rombach et al. Rombach et al. (2022), where the generation procedure is conducted in a low-dimensional latent space, thereby significantly reducing the associated computational burden. The capabilities of diffusion models have been further extended by recent innovations Chahine & Kim (2024); Oguz et al. (2024); He et al. (2024); Bartosh et al. (2024); Nobis et al. (2024); Motwani et al. (2024). To address inversion inaccuracies, EDICT was proposed by Wallace et al. Wallace et al. (2023), enabling mathematically exact diffusion inversion through coupled transformations. Subsequently, the BELM framework was introduced by Wang et al. Wang et al. (2024), in which heuristic exact inversion methods are treated as special cases, and optimal samplers are designed by minimizing truncation error.

This paper presents a novel generative steganography framework that advances the field from prompt-conditioned image editing to high-level semantic steganography. The unification of covert communication and copyright watermarking is, for the first time, realized within a single architecture. In addition to providing copyright protection for stego images, the proposed approach equips the steganography system with deniability, a capability absent in existing state-of-the-art methods.

B Algorithm

Algorithm 1 Conceal Process

Input: Secret image x_s, copyright watermark x_m, secret key k_s, watermark key k_w, repetition factor R, VQGAN encoder E and decoder D, diffusion generator G (BELM).
Output: Stego image x_{st}.

Semantic Tokenization:
$\mathbf{s}_s \leftarrow E(x_s) \in \{0, \ldots, K-1\}^L, \quad \mathbf{s}_m \leftarrow E(x_m) \in \{0, \ldots, K-1\}^L$, with $K = 16384,\ L = 1024$.
Randomized Quantile Mapping:
Define equal-probability intervals $I_s = \left(\Phi^{-1}(s/K),\ \Phi^{-1}((s+1)/K)\right)$ for $s = 0, \ldots, K-1$, where Φ is the standard normal CDF.
For each token $s_{s,i}$ ($i = 1..L$) and each repetition $r = 1..R$, independently sample $U_{i,r} \sim \mathrm{Uniform}\left(\frac{s_{s,i}}{K}, \frac{s_{s,i}+1}{K}\right)$ and set $z_{i,r} = \Phi^{-1}(U_{i,r})$. Concatenate to form $\mathbf{z}_{\mathrm{sec}}$ of length $N_s = L \times R$.
Similarly, form $\mathbf{z}_{\mathrm{wm}}$ from watermark tokens, of length $N_m = L \times R$.
Disjoint Dual-Key Mask Generation with Pre-partitioned Pools:
Let $\Omega = \{0, 1, \ldots, N-1\}$ with $N = 3 \times 256 \times 256 = 196608$.
Pre-partition Ω into two disjoint candidate pools, e.g., $\Omega_{\mathrm{sec}} = \{i \in \Omega : i \bmod 2 = 0\}$ and $\Omega_{\mathrm{wm}} = \{i \in \Omega : i \bmod 2 = 1\}$, with $\Omega_{\mathrm{sec}} \cap \Omega_{\mathrm{wm}} = \emptyset$.
$\mathbf{M}_{\mathrm{sec}} \leftarrow \mathrm{MaskGen}(k_s, N_s, \Omega_{\mathrm{sec}})$ with $|\mathbf{M}_{\mathrm{sec}}| = N_s$.
$\mathbf{M}_{\mathrm{wm}} \leftarrow \mathrm{MaskGen}(k_w, N_m, \Omega_{\mathrm{wm}})$ with $|\mathbf{M}_{\mathrm{wm}}| = N_m$.
(By construction, $\mathbf{M}_{\mathrm{sec}} \cap \mathbf{M}_{\mathrm{wm}} = \emptyset$.)
Noise Initialization:
Sample background noise $\eta \sim \mathcal{N}(0, \mathbf{I}_N)$.
Construct initial noise $\mathbf{x}_T \in \mathbb{R}^N$ as:

$$\mathbf{x}_T[i] = \begin{cases} \mathbf{z}_{\mathrm{sec}}[j], & i = \mathbf{M}_{\mathrm{sec}}[j], \\ \mathbf{z}_{\mathrm{wm}}[j], & i = \mathbf{M}_{\mathrm{wm}}[j], \\ \eta_i, & \text{otherwise.} \end{cases}$$

Stego Image Generation:
Reshape $\mathbf{x}_T$ to $(1, 3, 256, 256)$.
$x_{st} \leftarrow G(\mathbf{x}_T)$.
Scale pixel values to $[0, 1]$.
return x_{st}.

C Mathematical Proofs

C.1 Properties of the Transformation from Semantic Tokens to Gaussian Noise

Let the codebook size be K, let Φ be the standard normal cumulative distribution function (CDF), and let ϕ be its density. For each token $s \in \{0, \ldots, K-1\}$, define the equal-probability interval

$$I_s = \left(\Phi^{-1}\left(\frac{s}{K}\right),\ \Phi^{-1}\left(\frac{s+1}{K}\right)\right). \tag{19}$$

These intervals partition the real line into K segments, each carrying probability $1/K$ under the standard normal distribution.

Definition 4 (Randomized Quantile Mapping)**.** For a token $s \in \{0, \ldots, K-1\}$, independently sample

$$U \sim \mathrm{Uniform}\left(\frac{s}{K}, \frac{s+1}{K}\right), \qquad z = \Phi^{-1}(U). \tag{20}$$

Then z lies in I_s. Conditioned on s, z follows the standard normal distribution truncated to I_s, denoted by μ_s, whose density is

$$f_s(z) = K\,\phi(z)\,\mathbf{1}_{z \in I_s}. \tag{21}$$

Property 1 (Conditional Distribution)**.** For every token $s \in \{0, \ldots, K-1\}$, the conditional distribution of the concealed value given s is the standard normal distribution truncated to I_s, *i.e.*,

$$z \mid s \sim \mu_s,$$

with density $f_s(z) = K\phi(z)\mathbf{1}_{z \in I_s}$.

Algorithm 2 Reveal Process

Input: Stego image x_{st}, a key k (either secret key k_s or watermark key k_w), repetition factor R, VQGAN decoder D, diffusion inverse G^{-1} (BELM inversion).
Output: Revealed secret image $\tilde{x}_s$ (if $k = k_s$) or revealed watermark $\tilde{x}_m$ (if $k = k_w$).

Inversion to Noise:
$\hat{\mathbf{x}}_T \leftarrow G^{-1}(x_{st})$, reshape to vector length N.
Mask Regeneration from Pre-partitioned Candidate Pools:
$N_{\text{req}} \leftarrow L \times R$.
if $k = k_s$ **then**
  $\mathbf{M} \leftarrow \text{MaskGen}(k_s, N_{\text{req}}, \Omega_{\text{sec}})$.
else
  $\mathbf{M} \leftarrow \text{MaskGen}(k_w, N_{\text{req}}, \Omega_{\text{wm}})$.
end if
# Extract Received Values: #
$\mathbf{y} \leftarrow \hat{\mathbf{x}}_T[\mathbf{M}]$ (length N_{req}).
Per-Copy Interval Decision and Majority Voting:
Partition $\mathbf{y}$ into L groups of size R: $\mathbf{y}_i = (y_{i,1}, \ldots, y_{i,R})$.
for $i \leftarrow 1$ to L **do**
  for $r \leftarrow 1$ to R **do**
    Compute $u_{i,r} \leftarrow \Phi(y_{i,r})$.
    Estimate token index $s_{i,r} \leftarrow \lfloor K\, u_{i,r} \rfloor$.
  end for
  Set $\hat{s}_i$ by majority voting over $\{s_{i,1}, \ldots, s_{i,R}\}$.
end for
Assemble token sequence $\hat{\mathbf{s}} = (\hat{s}_1, \ldots, \hat{s}_L)$.
Image Reconstruction:
$\tilde{x} \leftarrow D(\hat{\mathbf{s}})$.
if $k = k_s$ **then**
  $\tilde{x}$ as $\tilde{x}_s$.
else
  $\tilde{x}$ as $\tilde{x}_m$.
end if
return $\tilde{x}$.

Algorithm 3 MaskGen: Deterministic Mask Generation from a Candidate Pool

Input: Key k, required number of coordinates n, candidate pool $\Omega_{\text{pool}} \subseteq \Omega$ with $|\Omega_{\text{pool}}| \geq n$.
Output: Ordered set $\mathbf{M} = (m_1, \ldots, m_n) \subseteq \Omega_{\text{pool}}$ with $m_j \neq m_{j'}$ for $j \neq j'$.

Compute $h = \text{SHA256}(k)$ and use the full 256-bit digest as the seed of a CSPRNG.
Initialize an empty ordered set $\mathbf{M} = \emptyset$.
while $|\mathbf{M}| < n$ **do**
  Sample a candidate coordinate i uniformly from Ω_{pool} using the CSPRNG.
  if $i \notin \mathbf{M}$ **then**
    Append i to $\mathbf{M}$.
  end if
end while

return $\mathbf{M}$.

Proof. Since $U \sim \text{Uniform}(s/K, (s+1)/K)$ and Φ is strictly increasing, $z = \Phi^{-1}(U)$ lies in I_s. For any Borel set $A \subseteq I_s$,

$$\Pr[z \in A \mid s] = \Pr[U \in \Phi(A) \mid s] = K\, \lambda(\Phi(A)),$$

where λ is the Lebesgue measure. Changing variables $z = \Phi^{-1}(u)$, $\mathrm{d}u = \phi(z)\mathrm{d}z$, gives

$$\Pr[z \in A \mid s] = K \int_{\Phi(A)} \mathrm{d}u = K \int_A \phi(z)\, \mathrm{d}z.$$

Hence the conditional density is $f_s(z) = K\phi(z)\mathbf{1}_{z \in I_s}$. □

Property 2 (Marginal Distribution and Uniform Prior)**.** Let the token prior be $\pi = (\pi_0, \ldots, \pi_{K-1})$. The marginal distribution of a single concealed coordinate is the mixture

$$P_z = \sum_{s=0}^{K-1} \pi_s \mu_s. \tag{22}$$

If the token prior is uniform, *i.e.*, $\pi_s = 1/K$ for all s, then

$$P_z = \sum_{s=0}^{K-1} \frac{1}{K} \mu_s = \mathcal{N}(0,1). \tag{23}$$

That is, each concealed coordinate is exactly standard normal.

Proof. The mixture representation follows from the law of total probability. For the uniform case, substituting $\pi_s = 1/K$ gives

$$P_z = \sum_{s=0}^{K-1} \frac{1}{K} \mu_s.$$

Since the intervals $\{I_s\}_{s=0}^{K-1}$ partition the real line and μ_s is the standard normal restricted to I_s with density $K\phi(z)\mathbf{1}_{z\in I_s}$, we have

$$\sum_{s=0}^{K-1} \frac{1}{K} \mu_s(A) = \sum_{s=0}^{K-1} \frac{1}{K} \int_A K\phi(z)\mathbf{1}_{z\in I_s}\,\mathrm{d}z = \int_A \phi(z) \sum_{s=0}^{K-1} \mathbf{1}_{z\in I_s}\,\mathrm{d}z = \int_A \phi(z)\,\mathrm{d}z.$$

Hence $P_z = \mathcal{N}(0,1)$. □

Property 3 (Mean and Variance)**.** Under a uniform token prior, the marginal distribution P_z has mean 0 and variance 1. Under a general token prior π, the mean is

$$\mathbb{E}[z] = \sum_{s=0}^{K-1} \pi_s\, \mathbb{E}[\mu_s],$$

and the variance is

$$\mathrm{Var}(z) = \sum_{s=0}^{K-1} \pi_s\, \mathrm{Var}(\mu_s) + \sum_{s=0}^{K-1} \pi_s \left(\mathbb{E}[\mu_s] - \mathbb{E}[z]\right)^2.$$

In particular, if π is uniform, $\mathbb{E}[z] = 0$ and $\mathrm{Var}(z) = 1$.

Proof. For a uniform prior, $P_z = \mathcal{N}(0,1)$ by Property 2, so the mean and variance are immediate. For a general prior, the mean and variance of a mixture are given by the law of total expectation and the law of total variance. □

Property 4 (Total Variation Bound)**.** Let $\pi = (\pi_0, \ldots, \pi_{K-1})$ be the token prior, and define

$$\delta_{\mathrm{token}} = \sum_{s=0}^{K-1} \left| \pi_s - \frac{1}{K} \right|.$$

Then the total variation distance between the marginal distribution P_z and the standard normal $Q_z = \mathcal{N}(0,1)$ satisfies

$$d_{\mathrm{TV}}(P_z, Q_z) \le \delta_{\mathrm{token}}.$$

If π is uniform, $\delta_{\mathrm{token}} = 0$ and $d_{\mathrm{TV}}(P_z, Q_z) = 0$.

Proof. This is a direct application of Lemma 1 in the main text, with μ_s being the truncated normal distributions and $u_s = 1/K$. □

Property 5 (Support and Interval Length)**.** For each token s, the support of the conditional distribution μ_s is the equal-probability interval I_s. Its length is

$$|I_s| = \Phi^{-1}\left(\frac{s+1}{K}\right) - \Phi^{-1}\left(\frac{s}{K}\right) = \frac{1}{K\,\phi(\xi_s)}$$

for some $\xi_s \in I_s$ by the mean value theorem. The minimum interval length over s satisfies

$$\min_s |I_s| \ge \frac{\sqrt{2\pi}}{K},$$

with equality approached near the origin as $K \to \infty$.

Proof. The support is immediate from Definition 4. The length formula follows from the mean value theorem applied to Φ^{-1} on $[s/K, (s+1)/K]$:

$$\Phi^{-1}\left(\frac{s+1}{K}\right) - \Phi^{-1}\left(\frac{s}{K}\right) = (\Phi^{-1})'(\eta_s) \cdot \frac{1}{K} = \frac{1}{K\,\phi(\Phi^{-1}(\eta_s))},$$

for some $\eta_s \in (s/K, (s+1)/K)$. Setting $\xi_s = \Phi^{-1}(\eta_s) \in I_s$ gives the exact expression. Since $\phi(z) \le \phi(0) = 1/\sqrt{2\pi}$, we obtain

$$|I_s| \ge \frac{1}{K\phi(0)} = \frac{\sqrt{2\pi}}{K}.$$

For large K, the intervals near the origin have length approximately $\sqrt{2\pi}/K$. □

Let K be the codebook size, $u_k = \frac{k+0.5}{K}$, and $z_k = \Phi^{-1}(u_k) = \sqrt{2}\,\mathrm{erf}^{-1}(2u_k - 1)$, where $\Phi(z) = \frac{1}{2}[1 + \mathrm{erf}(z/\sqrt{2})]$ is the standard normal cumulative distribution function (CDF) and $\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$ is its density.

C.2 Setup and Assumptions

Let the codebook size be K, let Φ be the standard normal cumulative distribution function, and let ϕ be its density. For each token $s \in \{0, \dots, K-1\}$, define the equal-probability interval

$$I_s = \left(\Phi^{-1}\left(\frac{s}{K}\right),\ \Phi^{-1}\left(\frac{s+1}{K}\right)\right).$$

Let μ_s be the standard normal distribution truncated to I_s, whose density is

$$f_s(z) = K\,\phi(z)\,\mathbf{1}_{z \in I_s}.$$

Let the token prior be $\pi = (\pi_0, \dots, \pi_{K-1})$, and let the uniform prior be $u_s = 1/K$. Define

$$\delta_{\text{token}} = \sum_{s=0}^{K-1} \left|\pi_s - \frac{1}{K}\right|.$$

Randomized quantile mapping. For a token s, independently sample

$$U \sim \mathrm{Uniform}\left(\frac{s}{K}, \frac{s+1}{K}\right), \qquad z = \Phi^{-1}(U).$$

Then $z \mid s \sim \mu_s$. If the token prior is π, the marginal distribution of a single concealed coordinate is the mixture

$$P_z = \sum_{s=0}^{K-1} \pi_s \mu_s.$$

The standard normal distribution can be written as the uniform mixture

$$Q_z = \mathcal{N}(0,1) = \sum_{s=0}^{K-1} \frac{1}{K} \mu_s.$$

Assumptions.

1. The secret token sequence and the watermark token sequence are independent and identically distributed, both following the prior π.
2. The R repetition copies of each token are sampled independently.
3. The secret mask and the watermark mask are disjoint, *i.e.*,
$$\mathbf{M}_{\text{sec}} \cap \mathbf{M}_{\text{wm}} = \emptyset, \qquad |\mathbf{M}_{\text{sec}}| = |\mathbf{M}_{\text{wm}}| = LR.$$
4. The masks are independent of the noise values.
5. The background coordinates are independently sampled from $\mathcal{N}(0,1)$.

Let the set of concealed coordinates be

$$A = \mathbf{M}_{\text{sec}} \cup \mathbf{M}_{\text{wm}}, \qquad N_h = |A| = 2LR.$$

The set of background coordinates is $C = \Omega \setminus A$, with $|C| = N - N_h$.

C.3 Proof of Lemma 1

Lemma 1 (TV Bound for Mixture Measures) Let $\{\mu_s\}_{s=0}^{K-1}$ be arbitrary probability measures, and let $\pi = (\pi_0, \ldots, \pi_{K-1})$ and $u = (1/K, \ldots, 1/K)$ be probability distributions on $\{0, \ldots, K-1\}$. Then

$$d_{\mathrm{TV}}\left(\sum_{s=0}^{K-1} \pi_s \mu_s,\ \sum_{s=0}^{K-1} u_s \mu_s\right) \le \sum_{s=0}^{K-1} \left|\pi_s - \frac{1}{K}\right|. \tag{24}$$

Proof. For any measurable set A,

$$\begin{aligned}\left|\sum_{s=0}^{K-1} \pi_s \mu_s(A) - \sum_{s=0}^{K-1} u_s \mu_s(A)\right| &= \left|\sum_{s=0}^{K-1} (\pi_s - u_s)\mu_s(A)\right| \\ &\le \sum_{s=0}^{K-1} |\pi_s - u_s|\, \mu_s(A) \\ &\le \sum_{s=0}^{K-1} |\pi_s - u_s|. \end{aligned} \tag{25}$$

Taking the supremum over A yields

$$d_{\mathrm{TV}}\left(\sum_{s} \pi_s \mu_s, \sum_{s} u_s \mu_s\right) \le \sum_{s=0}^{K-1} \left|\pi_s - \frac{1}{K}\right|.$$

This completes the proof. □ □

C.4 Proof of Theorem 1

Theorem 1 (Distribution Preservation under Randomized Quantile Mapping) Under the same assumptions as above, construct the initial noise $\mathbf{x}_T$ according to the randomized quantile mapping. Let P be the distribution of $\mathbf{x}_T$, and let $Q = \mathcal{N}(\mathbf{0}, \mathbf{I}_N)$. Then

$$\boxed{d_{\mathrm{TV}}(P, Q) \le N_h\, \delta_{\mathrm{token}}.} \tag{26}$$

In particular, if the token prior is uniform, *i.e.*, $\pi_s = 1/K$, then $\delta_{\mathrm{token}} = 0$, and hence

$$P = Q, \qquad d_{\mathrm{TV}}(P, Q) = 0.$$

If diffusion inversion introduces an error $d_{\mathrm{TV}}(\hat{P}, P) \le \varepsilon_{\mathrm{inv}}$, then

$$d_{\mathrm{TV}}(\hat{P}, Q) \le N_h\, \delta_{\mathrm{token}} + \varepsilon_{\mathrm{inv}}.$$

Proof. The proof consists of four steps.

Step 1: Distribution of a single concealed coordinate. Fix a token s and conditioned on s, there is:

$$U \sim \mathrm{Uniform}\left(\frac{s}{K}, \frac{s+1}{K}\right), \qquad z = \Phi^{-1}(U).$$

Since Φ is strictly increasing, z lies in the interval I_s. Its conditional density is

$$f_s(z) = K\, \phi(z)\, \mathbf{1}_{z \in I_s},$$

which is the standard normal distribution truncated to I_s, denoted by μ_s. When the token prior is π, the marginal distribution of a single concealed coordinate is the mixture

$$P_z = \sum_{s=0}^{K-1} \pi_s \mu_s.$$

On the other hand, the standard normal distribution can be written as the uniform mixture

$$Q_z = \mathcal{N}(0, 1) = \sum_{s=0}^{K-1} \frac{1}{K} \mu_s.$$

Step 2: Total variation distance for a single coordinate. By Lemma 1, it could directly obtain

$$d_{\mathrm{TV}}(P_z, Q_z) = d_{\mathrm{TV}}\left(\sum_{s} \pi_s \mu_s, \sum_{s} \frac{1}{K} \mu_s\right) \le \sum_{s=0}^{K-1} \left|\pi_s - \frac{1}{K}\right| = \delta_{\mathrm{token}}.$$

Method	Nvidia RTX 4090		Nvidia RTX 3090	
	Time(s)	Memory(GB)	Time(s)	Memory(GB)
CRoSS	14.41	7.43	39.38	8.58
DiffStega	20.90	8.35	41.00	9.02
DiffStega[‡]	20.90	8.35	41.00	9.02
Ours	10.57	10.39	33.94	11.43

Table 6: Computational Complexity Comparison with generative steganography methods on the UniStega dataset.

If $\pi_s = 1/K$, then $\delta_{\text{token}} = 0$, and hence $P_z = Q_z$.

Step 3: Total variation distance for product measures. Since all concealed coordinates are independent and identically distributed according to P_z, all background coordinates are independent and identically distributed according to Q_z, and the concealed and background coordinates are independent, the overall distributions are

$$P = P_z^{\otimes N_h} \otimes Q_z^{\otimes (N-N_h)}, \qquad Q = Q_z^{\otimes N}.$$

The total variation distance for product measures satisfies subadditivity:

$$d_{\text{TV}}(P_1 \otimes P_2, Q_1 \otimes Q_2) \le d_{\text{TV}}(P_1, Q_1) + d_{\text{TV}}(P_2, Q_2).$$

Applying this repeatedly gives

$$d_{\text{TV}}\big(P_z^{\otimes N_h}, Q_z^{\otimes N_h}\big) \le N_h \, d_{\text{TV}}(P_z, Q_z) \le N_h \, \delta_{\text{token}}.$$

The background part is identical, so its total variation distance is 0. Therefore,

$$d_{\text{TV}}(P, Q) \le N_h \, \delta_{\text{token}}.$$

If the token prior is uniform, then $\delta_{\text{token}} = 0$, so $P = Q$ and $d_{\text{TV}}(P, Q) = 0$.

Step 4: Inversion error. The actual receiver obtains

$$\hat{\mathbf{x}}_T = G^{-1}(G(\mathbf{x}_T)).$$

Let its distribution be $\hat{P}$, and suppose that

$$d_{\text{TV}}(\hat{P}, P) \le \varepsilon_{\text{inv}}.$$

By the triangle inequality,

$$d_{\text{TV}}(\hat{P}, Q) \le d_{\text{TV}}(\hat{P}, P) + d_{\text{TV}}(P, Q) \le \varepsilon_{\text{inv}} + N_h \, \delta_{\text{token}}.$$

This completes the proof. □ □

D Experimental Setting

Benchmarks and settings. The VQGAN Sun et al. (2024) with a codebook size of 16384 and a downsampling ratio of 8 is adopted as both Tokenizer and De-Tokenizer, and a pre-trained DDPM is employed for stego image generation. The framework is implemented in PyTorch and is trained and evaluated on the UniStega Yang et al. (2024), DIV2K Agustsson & Timofte (2017) and Stega260 Yu et al. (2024) datasets. All images are set to a resolution of 256×256, which is maintained in the generated stego images. For a rigorous assessment of the proposed method, a comprehensive comparison is performed against existing image steganography approaches, including CRoSS Yu et al. (2024) and DiffStega Yang et al. (2024).

Evaluation Metrics. The quality of secret–recovery pairs is assessed using the Peak Signal-to-Noise Ratio (PSNR) and the Structural Similarity Index (SSIM), with higher values indicating superior performance. The naturalness of stego images is further evaluated by the Naturalness Image Quality Evaluator (NIQE) Mittal et al. (2012).

E More Results

Complexity and Efficiency Metrics. Computational efficiency was assessed on NVIDIA GeForce RTX 4090 and 3090 GPUs. As listed in Table 6, the proposed method markedly lower time costs (10.33s on the 4090, 7.06s on the 3090) and slightly improved memory consumption (2.04GB and

Method	Clean	Gaussian Blur	Gaussian Noise	JPEG compression	Poisson
ISN	21.73	5.16	3.97	4.33	4.05
HiNet	46.55	9.32	9.97	10.13	10.29
CRoSS	21.25	20.08	**18.97**	20.20	19.27
DiffStega	23.29	**20.85**	18.62	**21.16**	**20.15**
Ours	24.37	**21.43 (0.58↑)**	**21.18 (2.21↑)**	**21.97 (0.81↑)**	**21.01 (0.86↑)**

Table 7: Quantitative results of revealed images on the UniStega dataset with degraded stego images.

2.41GB, respectively) compared to DiffStega. In comparison with CRoSS, reductions in computation time of 3.84s and 5.44s are attained, respectively. These results indicate that the proposed method can be readily deployed on DDPM and diffusion model compatible hardware because its computational footprint is inherited from that architecture.

Robustness Analysis. The robustness of the proposed framework was assessed under multiple image degradations, including Gaussian blur, Gaussian noise, JPEG compression, and Poisson noise. For Gaussian blur, stego images were convolved with a Gaussian kernel of size 3, with the standard deviation set to $\sigma = 0.1 \times std$, where std denotes the standard deviation of the stego image. Gaussian noise was similarly parameterized by $\sigma = 0.1 \times std$. JPEG compression was applied with a quality factor of $Q = 75$. As reported in Table 7, the proposed model achieved superior PSNR values compared to existing state-of-the-art methods, confirming its strong resilience against the tested attacks.

Visualization Results. Figure 9 reports the performance of the proposed C^2-Steg against generative steganography methods on UniStega dataset with three prompts. Compared to CRoSS Yu et al. (2024)and DiffStega Yang et al. (2024), the proposed method yields markedly more natural and imperceptible stego images. It also permits large content deviations of the secret image, so that the similarity between the stego and the secret is suppressed. Style prompts cause CRoSS and DiffStega to produce high deviation metrics as well, but their reconstruction quality is critically harmed by the entanglement of style and content. High-fidelity watermark recovery, unavailable in prior works, grants the system deniability. Under coercion, revealing only the fake key yields a valid watermark, thereby substantiating the claim that the stego image is merely a randomly watermarked image containing no secrets.Visual results validate that the proposed model outperforms previous SOTA solutions in both efficacy and security.

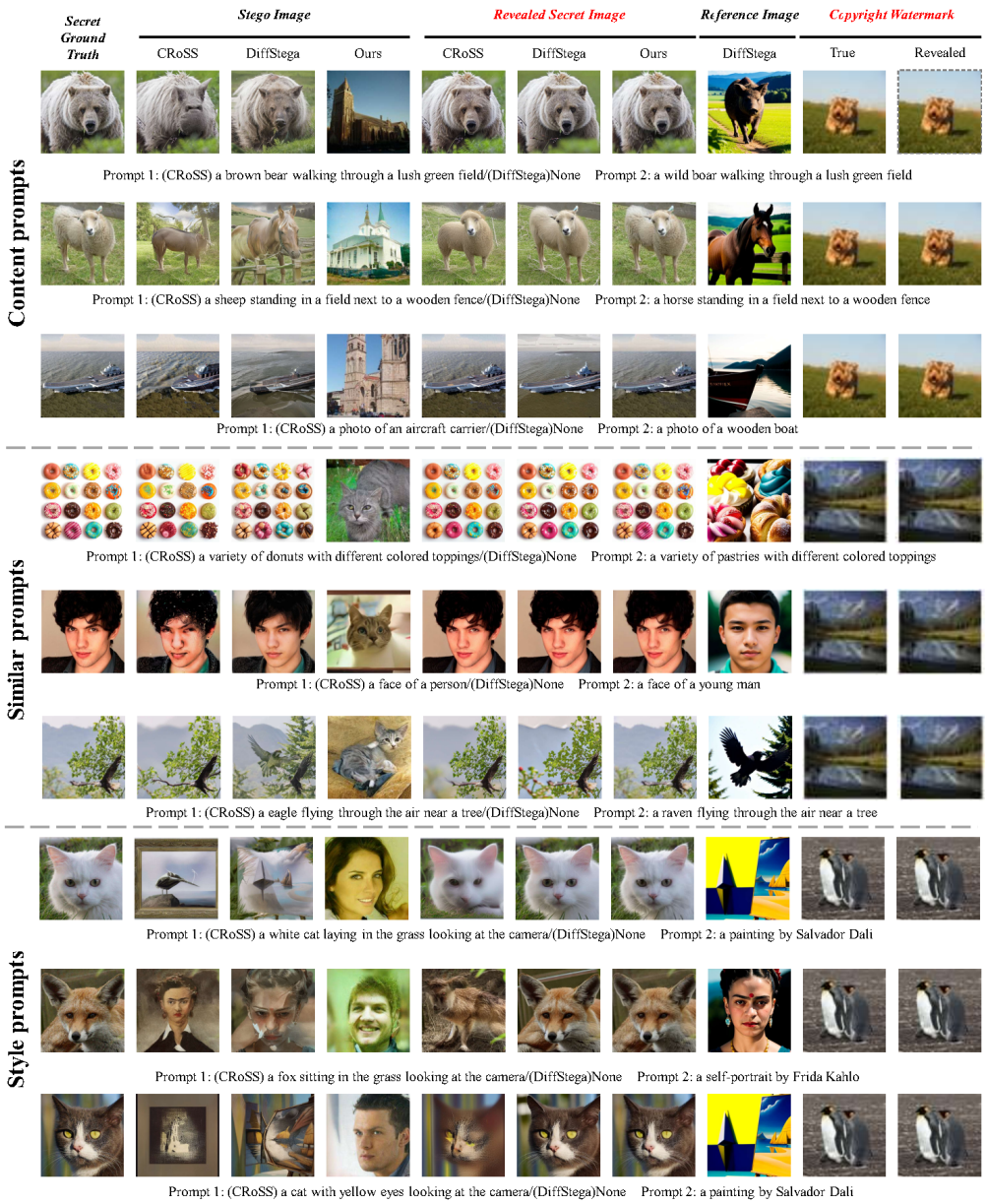

Figure 9: Visualization comparison results of the proposed model versus other methods. These prompts are utilized in CRoSS and DiffStega and the reference image is used as the image condition in DiffStega.