Towards Provably Secure and Highly Robust Generative Image Steganography Leveraging Latent Diffusion Model

Chengsheng Yuan[1], Zhaonan Ji[1], Qi Cui[1], Zhili Zhou[2], Xinting Li[3*], Zhihua Xia[4†]

[1]Engineering Research Center of Digital Forensics, Ministry of Education, the School of Computer Science, Nanjing University of Information Science and Technology, Nanjing 210044, China
[2]Institute of Artificial Intelligence, Guangzhou University, Guangzhou 510006, China
[3]School of Foreign Languages, National University of Defense Technology, Nanjing 210039, China
[4]College of Cyber Security, Jinan University, Guangzhou 510632, China
{yuancs, jizn, cuiqi}@nuist.edu.cn, zhou zhili@163.com, lixt@tju.edu.cn, xia_zhihua@163.com

Abstract

Generative image steganography has attracted significant attention for its exceptional resistance to steganalysis. However, current generative steganography methods still face limitations in terms of the lack of provable security guarantees under statistical analysis and vulnerability to real-world, unforeseen channel attacks. To address these issues, this paper proposes a novel generative image steganography framework that leverages the Latent Diffusion Model (LDM). Notably, we have uncover a consistent trend: regardless of whether an image has undergone attacks such as compression or noise addition, the sign pattern of values in its latent vector encoded by the LDM remains largely invariant. Capitalizing on this trend, we have devised an adaptive distribution-preserving mapping (ADPM) mechanism, capable of converting a secret message into a latent vector that follows standard normal distribution in an adjustable way. Since both the secret latent vector and the latent vector randomly generated during regular image generation follow the same distribution, satisfying the optimal input conditions for the diffusion model, the proposed method can achieve provable security. Experimental results demonstrate the outstanding performance of our approach in terms of robustness, security, and extraction accuracy.

Introduction

With the rapid advancement of Artificial Intelligence Generated Content (AIGC), concerns regarding the protection of personal data and privacy have intensified. Steganography (Johnson and Jajodia 1998), also known as information hiding, a widely studied secret communication technology, aims to hide secret information, such as audio, text and images, within a cover. The cover medium can be digital media like text (Zhou et al. 2021), audio (Gopalan 2003), video (Meng et al. 2023) or images (Zhou et al. 2022). The receiver can only extract the secret information from the cover based on a shared agreement with the sender. Traditional steganography involves selecting a cover image and embedding secret information either through manual modification or by utilizing deep learning-based adaptive methods. However, steganographic images obtained by these methods often retain noticeable signs of modification, making them susceptible to suspicion from third parties and consequently compromising their security.

Recent advances in generative models have facilitated the emergence of generative image steganography (GIS) (Liu et al. 2022). By leveraging generative models, GIS constructs steganographic images directly from secret messages, producing outputs that are far more robust against advanced steganalysis. Hu et al. (Hu et al. 2018) proposed a steganographic method using deep convolutional generative adversarial networks (DCGANs), yet it faced challenges in achieving high-accuracy extraction of secret information, and the visual quality of the resulting steganographic images fell short of optimal standards. Later, the S2IRT method (Zhou et al. 2022), built upon the Glow model, enhanced the quality of generated images and facilitated precise extraction of secret information through a meticulously designed invertible mapping function. However, this method was susceptible to attacks like JPEG compression and noise. Meanwhile, the embedding of secret information causes the noise vector of the generated image to deviate from its original distribution, thereby compromising both the quality of the stego-image and the security of covert communication. The recent adoption of diffusion models (Ho, Jain, and Abbeel 2020; Song and Ermon 2019) has greatly enhanced the visual quality of generated images. Peng et al. proposed StegaDDPM (Peng et al. 2023), a method that leverages the denoising diffusion probabilistic model to generate steganographic images, maintaining high extraction accuracy even with higher steganographic capacity. However, constraints in the mapping method hinder the accurate recovery of secret information from the steganographic images after transmission through lossy channels. Subsequently, Zhou et al. (Zhou et al. 2025) incorporated the denoising diffusion implicit model into image steganography by embedding secret information within the frequency domain of Gaussian noise, which significantly improves the recovery accuracy of hidden information. However, this method exhibits limited robustness against common image attacks.

Motivated by these challenges, we propose a novel dif-

*Corresponding author
†Corresponding author

fusion model-based steganographic method that simultaneously achieves provable security and robust performance. The main contributions of this work are summarized as follows:

(1) Through our experimental investigations into the intrinsic robustness of the latent vector encoded by LDM, we have consistently observed that the sign of the values within this vector remains almost unaltered, regardless of whether the image has undergone attacks such as compression or noise addition. Motivated by this observation, we propose a robust generative image steganography scheme, enhancing the attack resistance of stego images.

(2) To increase the security of steganography, we devise an Adaptive Distribution-Preserving Mapping (ADPM) mechanism for embedding secret information, which simultaneously improves the quality of stego images and guarantees security of covert communication.

(3) Experimental results show that, the proposed method exhibits remarkable resilience to steganalysis tools and common image processing attacks.

Related Works

Embedded Image Steganography

Embedded image steganography entails altering pixel values of an image to hide secret information. The most traditional method is based on the Least Significant Bit (LSB) (Mielikainen 2006) to hide secret data. However, modifying the LSB of each pixel without discrimination can cause significant distortion to the image, potentially raising suspicions in secret communications. Later, adaptive image steganography (Filler, Judas, and Fridrich 2011; Holub and Fridrich 2012) reduced distortion via cost functions designed either manually or through deep learning. Following this direction, Yang et al. proposed a pioneering steganographic framework UTGAN (Yang et al. 2018), which harnesses Generative Adversarial Networks (GANs) to autonomously learn embedding costs. HiDDeN (Zhu et al. 2018) further innovated by utilizing an end-to-end training process, enabling an embedding strategy focused on local regions. AdaSteg (Pan et al. 2021) combined deep reinforcement learning with encryption noise techniques to implement an adaptive local image steganography method. Overall, cover-based image steganography inevitably leaves modification traces within the image, rendering it challenging to evade detection by sophisticated steganalysis tools.

Generative Image Steganography

To enhance security, Zhou et al. (Zhou et al. 2015) proposed a cover-selective image steganography method. During covert communication, the secret message is utilized to select the corresponding feature image from this indexed database to serve as the stego image. However, a significant limitation of this technique is the considerable time and expense required to establish a sufficiently indexed database. Subsequently, generative image steganography emerged, which conceals secret information during the image generation process. One strategy involves transforming secret information into features of the generated image. For example, You et al. (You et al. 2022) converted binary secret messages into facial expressions or other semantic features within the generated image, thereby facilitating the creation of compact sticker images. CtrGAN (Zhou et al. 2023) proposed a generative steganography framework, which encodes the given secret data into the semantic contours of objects in the generated image. While these methods offer promising solutions, they often necessitate retraining the generative model, which introduces additional computational overhead and potentially restricting embedding capacity.

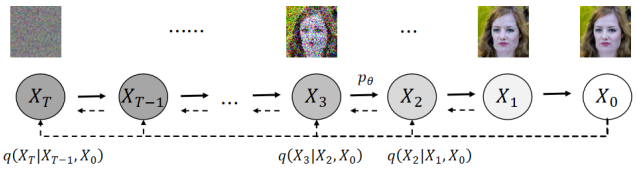

Figure 1: Schematic diagram of diffusion models without the non-Markov chain inference model.

Another approach is to map the secret information to Gaussian latent vectors used to generate the image. Zhou et al. (Zhou et al. 2022) designed a mapping mechanism leveraging the Glow model, which rearranges latent vector elements according to the guidance provided by the secret message. Peng et al. (Peng et al. 2023) explored the feasibility of using Denoising Diffusion Probabilistic Models (DDPM) for image steganography. Specifically, they designed a mapping rule in the final and penultimate steps of the diffusion process to convert binary secret information into decimal fractions, which were subsequently mapped onto latent vectors using the inverse cumulative distribution function of the Gaussian distribution. Both S2IRT (Zhou et al. 2022) and StegaDDPM (Peng et al. 2023) exhibit high accuracy in recovering secret information at elevated embedding capacities, yet they suffer from inadequate robustness. Moreover, S2IRT causes the secret-containing latent vectors to deviate from their original distribution, resulting in low-quality stego images that may raise suspicion among third parties.

Diffusion Models

Diffusion models (Ho, Jain, and Abbeel 2020; Song and Ermon 2019), as an advanced generative model, capture the distribution characteristics of target images from noisy data through training. Recently, due to their outstanding generative performance, diffusion models have been widely applied in various image processing fields, such as image synthesis (Song et al. 2020; Vahdat, Kreis, and Kautz 2021), image editing (Avrahami, Lischinski, and Fried 2022; Choi et al. 2021) and image reconstruction (Saharia et al. 2022; Wang, Yu, and Zhang 2022). A diffusion model consists of a forward and a backward process. The forward process is responsible for gradually adding Gaussian noise to an image to produce a nearly pure noisy image, while the backward process gradually removes the noise from the noisy image, ultimately generating a natural image. Denoising Diffusion Probabilistic Models (DDPM) (Ho, Jain, and Abbeel

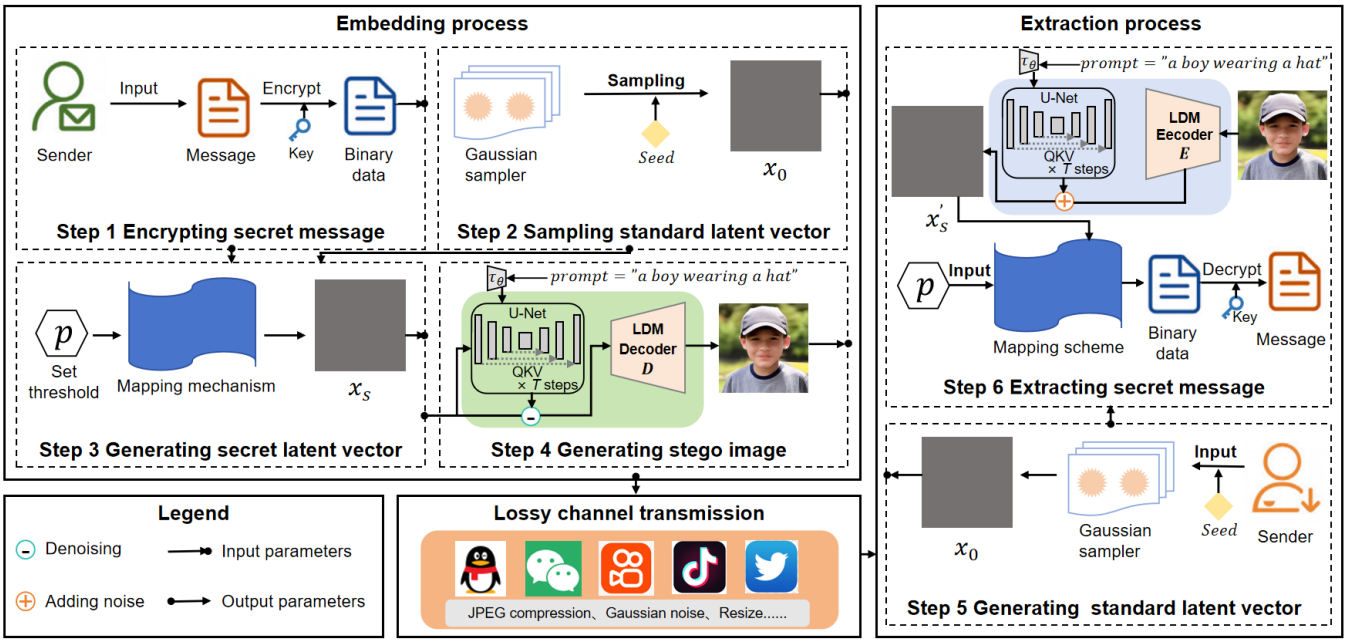

Figure 2: The framework of our proposed method.

2020; Choi et al. 2021) have more inference steps compared to other generative models, and generating an image typically takes more time. To speed up the sampling process, researchers have proposed various samplers. For example, the Denoising Diffusion Implicit Model (DDIM) is different from DDPM which requires iterative sampling step by step. As shown in Fig. 1, it supports sampling across any number of steps. Moreover, DPM-Solver++ (Lu et al. 2025) and DPM-solver (Lu et al. 2022) greatly reduce the time required to generate a high-quality image by solving ordinary differential equations.

Methodology

In this paper, we propose a provably secure and robust generative image steganography based on Diffusion Model. As shown in Fig. 2, the sender samples a latent vector x_0 that obeys standard positive distribution by using a random $seed$ and sets a threshold p. These elements are then input into the mapping mechanism ADPM alongside the binary secret data. Guided by the secret data, ADPM transforms x_0 into a secret latent vector x_s while preserving the original distribution. Subsequently, semantic information $prompt$ is set to guide image generation, and the reverse denoising process of the diffusion model is employed to convert x_s into a high-quality steganography image in a controlled manner. Finally, the stego image, the threshold p, the $prompt$ and the $seed$ are transmitted to the receiver through the lossy channel. After the receiver obtains them, the $seed$ is input into the Gaussian sampler to generate the standard latent vector x_0 firstly. Next, the stego image is converted into the corresponding latent vector by using the forward denoising process of the same diffusion model. Finally, it is input into the mapping sheme along with the standard latent vector x_0, the $prompt$ and the threshold p to recover the secret message.

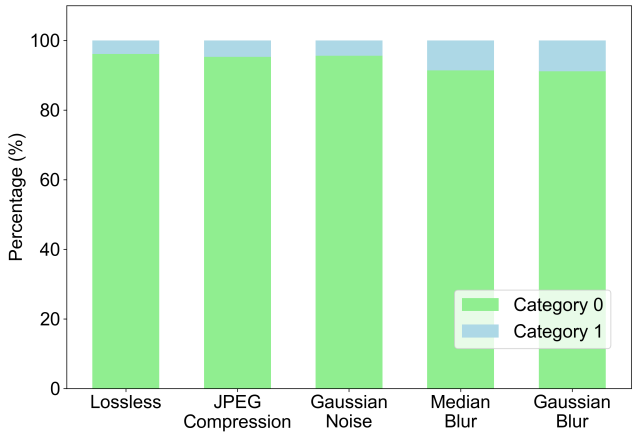

Figure 3: The sign statistics plot of the latent vector's value corresponding to the attacked reconstructed images (0 indicates that the latent vector's value of the attacked image have the same sign as the original image, while 1 indicates the opposite sign).

Exploration of The Consistent Trend

Diffusion models (DMs) (Dhariwal and Nichol 2021) consists of two fundamental processes: a forward noising process and a reverse denoising process. Traditional models typically operate directly in pixel space; however, optimizing high-performance DMs often requires hundreds of GPU days, and inference remains costly due to the need for sequential evaluations. To facilitate DM training on limited

computational resources while preserving their quality and flexibility, Rombach et al. (Rombach et al. 2022) implemented these models in the latent space of powerful pretrained autoencoders and introduced latent diffusion models (LDMs).

In latent diffusion models (LDMs), both the noising and denoising procedures are carried out exclusively within a latent space that has a lower dimensionality compared to the original RGB image space. Given a RGB image $X \in \mathbb{R}^{H\times W\times 3}$, the encoder E of LDMs encodes X into a latent image z, and the decoder D of LDMs can reconstruct X from z. The formula is presented as follows:

$$z = E\left(X\right), X^{'} = D\left(z\right) \tag{1}$$

where the representation z has dimensions $\mathbb{R}^{\frac{H}{f}\times\frac{W}{f}\times C_f}$, where f represents the spatial downsampling factor and C_f specifies the number of channels within the compressed latent space. We conducted an exploration of the latent space using a dataset of 100 RGB images, each of dimensions $H = W = 512$ and $C = 3$. After processing through the encoder E, the spatial dimensions are reduced, resulting in a latent vector of dimensions $H^{'} = W^{'} = 64, C^{'} = 4$. By thoroughly examining the 16384 elements within the latent vectors z and $z^{'}$ corresponding to the original image X and the image after attacks $X^{'}$, we uncover a consistent trend: as demonstrated in Fig. 3, when the image undergoes JPEG compression, Gaussian noise addition, median blur and Gaussian blur, only a negligible fraction of the components within the latent vector exhibit a change in sign. The vast majority of these components maintain their original signs, even after enduring these perturbations.

Steganography Network

Our method necessitates the use of three pretrained submodels from the LDM: a Decoder D, an Encoder E and a Conditional Encoder τ_θ, while utilizing DPM-Solver++ for deterministic sampling. The covert communication process involves two participants: the sender and receiver. To generate stego images, the sender initiates by initializing the latent vector x_0 using a random $seed$. Subsequently, the sender utilizes the key to transform the secret message m into binary data m^e. It is crucial that x_0 adheres to the distribution $N\left(0, I\right)$, where I represents the identity matrix.

$$x_0 = GS\left(seed, N\left(0, I\right)\right), \tag{2}$$

where $GS\left(\cdot\right)$ denotes Gaussian sampler. To guarantee the security and robustness of image steganography, we devise an adaptive distribution-preserving mapping mechanism $H\left(\cdot\right)$ based on the consistent trend to hide the secret data into x_0. Moreover, given that the components of x_0 close to 0 are extremely susceptible to image attacks and the buildup of numerical errors during the diffusion process, we utilize a threshold p to control the regions for information embedding, thereby enhancing the accuracy of secret data recovery.

$$x_s^i = \begin{cases} H\left(x_0^i, m_j^e\right), & \text{if } \left|x_0^i\right| \geq p \\ x_0^i, & \text{if } \left|x_0^i\right| < p \end{cases} \tag{3}$$

where x_0^i and m_j^e represent the components of x_0 and m^e, respectively. Meanwhile, x_s^i denotes the component of x_s used to generate stego image. The value range of p is $[0, 1]$.

Algorithm 1: Generating stego I_s

Input:
Decoder of the pretrained LDM D, deterministic sampler f, sampling steps T, encryption function $F\left(\cdot\right)$, key, and $prompt$. Secret message m, $seed$, the threshold p, Gaussian sampler $GS\left(\cdot\right)$, and calculating length function $Len\left(\cdot\right)$.

Output:
I_s

$x_0 = GS\left(seed, N\left(0, I\right)\right), m^e = F\left(m, key\right)$
$L = Len\left(m^e\right), k = 0$
if $k < L$ **then**
  for $i \in \left\{1, 2, \cdots, H^{'} \times W^{'} \times C^{'}\right\}$ **do**
    if $\left|x_0^i\right| \geq p$ **then**
      $x_s^i = H\left(x_0^i, m_k^e\right)$
      $k = k + 1$
    else
      $x_s^i = x_0^i$
    end if
  end for
end if
$c = \tau_\theta\left(prompt\right)$
$x_T = f\left(x_s, c, 0, T\right)$
$I_s = D\left(x_T\right)$
return I_s

Subsequently, the sender configures the semantic $prompt$ for the generated stego image and employs the conditional encoder τ_θ to convert it into the input parameter c for the diffusion model.

$$c = \tau_\theta\left(prompt\right), \tag{4}$$

Then, x_s and c are fed into the diffusion model after setting the diffusion step T.

$$x_T = f\left(x_s, c, 0, T\right), \tag{5}$$

where f represents the deterministic sampler of diffusion model. The initial input x_s progresses through T denoising iterations to yield x_T, which is subsequently decoded by D into the final stego image I_s.

$$I_s = D\left(x_T\right), \tag{6}$$

The detailed procedure for generating the stego image I_s is outlined in Algorithm 1.

Message Hiding and Extraction

The process of hide stage Our method achieves the covert transmission of secret information through four

steps: preprocessing the secret message, sampling a standard Gaussian vector x_0, designing the adaptive distribution-preserving mapping mechanism $H(\cdot)$ to hide the secret data into x_0, and generating stego image I_s. During the preprocessing stage, the sender uses a key to encrypt the secret message m into binary data m^e. Then, using the $seed$ to sample latent vector x_0 following standard normal distribution. To ensure the security of the steganographic system, we embed m^e into x_0 in a distribution-preserving manner. Specifically, the sender sets a threshold p (where $p \in [0,1]$) to control the portion of x_0 eligible for information hiding. The m^e is embedded only in the regions of x_0 outside $(-p, p)$ to enhance robustness. The mathematical formulation of $H(\cdot)$ is shown in Eq. (7).

$$x_s^i = \left|x_0^i\right| \bullet sgn\left[\sin\left(\frac{2m_i^e+1}{2}\pi\right) - \frac{1}{2}e^{-(m_i^e)^2}\right], \quad (7)$$

where $|\cdot|$ indicates absolute value operations, $\cdot$ is used t o compute the product of two adjacent numbers, and $sgn(\cdot)$ denotes sign extraction operation that is defined in Eq. (8).

$$sgn(x) = \begin{cases} 1, & \text{if } x > 0 \\ -1, & \text{if } x < 0 \end{cases} \quad (8)$$

Provable security analysis This framework adheres to the information-theoretic definition of steganographic security established by Cachin (Cachin 2004). To achieve absolute security, the Kullback-Leibler (KL) divergence between the probability distributions of the cover image and the stego image should satisfy:

$$D_{KL}(P(I_c) \parallel P(I_s)) = 0, \quad (9)$$

where $P(I_c)$ and $P(I_s)$ represent the probability distributions of the cover image and the stego image respectively. In generative image steganography, the cover image and stego image synthesized by generative models are derived from the random latent variable z_0 and the latent variable z_s mapped from the secret information, respectively. Therefore, in order to ensure perfect security in generative image steganography, the probability distributions of z_0 and z_s should satisfy that their relative entropy equals 0, as expressed by the following equation:

$$D_{KL}(P(z_0) \parallel P(z_s)) = 0, \quad (10)$$

In this paper, since the random latent vector x_0 used to generate the cover image is sampled from standard Gaussian noise, all its elements $x_0^i \sim N(0,1)$. Moreover, $m_i^e \in \{0,1\}$ and $P(m_i^e = 0) = P(m_i^e = 1) = 0.5$, so the $sgn(\cdot)$ function takes values -1 and 1 with equal probability 0.5 in Eq. (7). Then, x_i^s can be expressed as:

$$x_s^i = \begin{cases} \left|x_0^i\right|, & \text{if } m_i^e = 0 \\ -\left|x_0^i\right|, & \text{if } m_i^e = 1 \end{cases} \quad (11)$$

Therefore, when $\left|x_0^i\right| \geq p$, the mean and variance of x_i^s can be expressed as:

$$E\left(x_s^i\right) = E\left(\pm\left|x_0^i\right|\right) = E\left(\pm x_0^i\right) = \pm E\left(x_0^i\right) = 0, \quad (12)$$

$$Var\left(x_s^i\right) = (\pm 1)^2 \times Var\left(x_0^i\right) = 1, \quad (13)$$

where $E(\cdot)$ and $Var(\cdot)$ represent the mean function and variance function, respectively. When $\left|x_0^i\right| < p$, $x_s^i = x_0^i$, so $x_s^i \sim N(0,1)$. Since the distributions of x_0 and x_s are the same, so the distributions of generated cover images and stego images are the same, too, i.e., $D_{KL}(P(I_c) \parallel P(I_s)) = 0$, then provably secure generative image steganography is achieved.

The process of extraction stage During the secret data extraction phase, the receiver initially utilizes the $seed$ to sample the standared Gaussian latent vector x_0 and converts stego image to latent vector $x_T^{'}$ using the encoder E of LDM. Then, the diffusion step T, the semantic $prompt$ and $x_T^{'}$ are input into LDM together, and the secret latent vector $x_s^{'}$ is obtained after forward noising. The receiver extracts the secret data from $x_s^{'}$ based on p and the length l of the shared secret message. Specifically, through the analysis of the mapping sheme $H(\cdot)$, the receiver determines that if a component of $x_s^{'}$ is greater than 0, the embedded secret message bit is 0; conversely, if it is less than 0, the embedded secret message bit is 1. Guided by p, the components of $x_s^{'}$ are sequentially evaluated until the secret information of length l is fully extracted. Finally, the receiver employs the key to decrypt the original data from the binary secret message.

Experimental Results

Implementation Details

In the experiments, we selected the publicly accessible pretrained Latent Diffusion Model, Stable Diffusion v1.5, along with the deterministic sampler DPM-Solver++, to undertake the task of generating steganographic images. The latent vector parameters for generating the stego image were configured as $\frac{H}{f} = \frac{C}{f} = 64,\ C_f = 4,\ f = 8$. Therefore, the size of generated images is 512×512. All experiments were conducted on an NVIDIA RTX 3090 GPU. The proposed approach operates directly on the pretrained LDM architecture, requiring no additional model fine-tuning or retraining. Our experimental validation employs the identical dataset configuration as RGSD-Stego (Hu et al. 2024) to assess the method's efficacy. They are LAION-10K (Schuhmann et al. 2022), MS-COCO (Lin et al. 2014) and Flicker8K (Hodosh, Young, and Hockenmaier 2013).

Our experimental framework utilizes a well-rounded suite of evaluation metrics, including hiding capacity, extraction accuracy, robustness and security.

Hiding Capacity We assess the hiding capacity by determining the aggregate quantity of binary bits that can be embedded in a single image, measured in bits.

Extraction Accuracy The extraction accuracy is determined by executing a bitwise XOR operation on the original binary secret and the retrieved secret message. Its representation is shown as follows:

$$Acc = \frac{1}{L}\sum_{i=1}^{L}\left(1 - m_i^e \oplus m_i^{e'}\right), \quad (14)$$

where L denotes the length of secret message, and m_i^e and $m_i^{'}$ represent the components of original secret message m^e and recovered secret message $m^{'}$, respectively.

Robustness When digital images are transmitted via lossy channels, distortion is an inevitable issue that severely degrades the extraction accuracy of hidden message. To evaluate the robustness of the proposed method under such conditions, we test the bit extraction accuracy under five types of common attacks: resize, JPEG compression, median blur, Gaussian noise, and Gaussian blur.

Security In image steganography, security is determined by whether a third party can distinguish between cover images and stego images. Unlike traditional modification-based image steganography, generative steganography operates without requiring a cover image. In our experimental setup, we define the cover image as being synthesized directly from a randomly sampled standard Gaussian vector x_0. To classify the cover and stego images, we employ three steganographic analyzers SRNet (Boroumand, Chen, and Fridrich 2018), XuNet (Xu, Wu, and Shi 2016), and SiaStegNet (You, Zhang, and Zhao 2020). Moreover, we employ the detection error rate P_e to evaluate the resilience of the proposed method against steganography analysis, which is defined as follows:

$$P_e = min_{P_{FA}} \frac{1}{2} \left(P_{FA} + P_{MD} \right), \tag{15}$$

where P_{FA} represents the false alarm rate of the steganalysis system, while P_{MD} represents the miss detection rate.

To demonstrate the advantages of our proposed method, we compared it with four existing SOTA methods, namely S2IRT (Zhou et al. 2022), GSD (Zhou et al. 2025), LDStega (Peng et al. 2024) and RGSD-Stego (Hu et al. 2024). To ensure fair comparison, we integrate all four benchmark mapping methods into our framework, generating stego images under identical experimental settings.

Dataset	p	0.4	0.5	0.6	0.8	1.0
	Capacity (bits)	11384	10218	9088	6999	5217
Flicker8K		97.68	98.55	99.26	99.65	99.84
LAION-10K	Acc(%)	95.14	96.16	97.83	98.80	98.98
MS-COCO		96.77	97.74	98.82	99.38	99.69

Table 1: The steganography capacity and extraction accuracy when saving the stego image in PNG format under different thresholds.

Adjustability of Steganography Capacity

To validate the effectiveness of the proposed method under different thresholds, Table 1 presents the steganographic capacity under five different threshold settings and the recovery accuracy of secret information when stego images are saved in PNG format. It is evident that accuracy rates surpass 97% across all three datasets. Moreover, since PNG format is widely used in daily applications, our method demonstrates superior practicality. Additionally, Table 2 demonstrates the robustness of the proposed method on the Flicker8K dataset under different thresholds. The results demonstrate that the stego images can still extract the secret information with over 90% accuracy in most cases after various attacks.

Attack	Factor	p 0.4	0.5	0.6	0.8	1.0
Identity	/	97.73	98.57	99.28	99.67	99.86
Resize	0.5	96.07	97.22	98.43	99.03	99.55
	0.75	97.11	98.14	99.01	99.46	99.75
	1.25	97.47	98.44	99.19	99.59	99.81
	1.5	97.54	98.49	99.23	99.61	99.82
JPEG Compression	90	96.82	98.15	98.88	99.32	99.67
	70	94.96	97.13	97.85	98.34	99.13
	50	93.19	96.07	96.74	97.41	98.51
Gaussian Blur	3	97.06	98.11	98.99	99.43	99.74
	5	96.34	97.52	98.59	99.13	99.61
	7	94.91	96.33	97.74	98.56	99.28
Median Blur	3	96.60	97.48	98.59	99.17	99.62
	5	93.66	94.82	96.65	97.94	98.84
	7	89.70	91.35	93.82	96.18	97.57
Gaussian Noise	0.01	97.31	98.45	99.15	99.47	99.76
	0.05	93.78	96.73	97.37	97.58	98.69
	0.1	90.29	94.04	94.95	95.66	97.38

Table 2: The comparison of steganographic capacity and robustness for distinct thresholds.

Method	Capacity (bits)	Image size
S2IRT (Zhou et al. 2022)	4096	64×64
GSD (Zhou et al. 2025)	4096	64×64
LDStega (Peng et al. 2024)	4096	256×256
RGSD-Stego (Hu et al. 2024)	16384	512×512
Ours	11384	512×512

Table 3: Performance comparison of the proposed method with SOTA work in terms of hiding capacity and generated stego image size.

Comparison with Existing Methods

Hiding Capacity & Stego Image Size Initially, we conduct a comprehensive comparison between the proposed framework and existing SOTA generative steganographic methods, evaluating both hiding capacity and the size of stego images. As shown in Table 3, our approach substantially outperforms S2IRT, GSD and LDStega in terms of embedding capacity, though it shows slightly inferior performance compared to RGSD-Stego.

Method	Identity	Resize				JPEG Compression			Gaussian Blur			Median Blur			Gaussian Noise		
	/	0.5	0.75	1.25	1.5	90	70	50	3	5	7	3	5	7	0.01	0.05	0.1
S2IRT	87.17	77.66	79.64	80.59	80.83	79.06	76.53	75.02	78.37	74.92	72.52	79.60	78.15	76.20	80.17	76.15	73.78
GSD	88.44	85.86	87.53	88.11	88.23	87.31	85.13	83.01	87.48	86.34	84.13	86.51	81.77	76.18	87.92	84.18	79.66
LDStega	98.52	89.21	90.86	90.89	90.92	98.25	96.43	94.41	95.83	92.64	90.96	85.97	83.62	80.16	97.65	93.27	90.38
RGSD-Stego	98.91	96.82	98.29	98.72	98.81	97.79	94.99	92.21	97.36	92.32	84.95	98.25	**97.29**	**95.11**	98.46	93.85	88.29
Ours	**99.28**	**98.43**	**99.01**	**99.19**	**99.23**	**98.88**	**97.85**	**96.74**	**98.99**	**98.59**	**97.74**	**98.59**	96.65	93.82	**99.15**	**97.37**	**94.95**

Table 4: Comparison of the extraction accuracy of different methods against different types of image attacks on the Flicker8K dataset.

Extraction Accuracy & Robustness We evaluated the robustness of our proposed method against four existing SOTA methods, namely S2IRT (Zhou et al. 2022), GSD (Zhou et al. 2025), LDStega (Peng et al. 2024) and RGSD-Stego (Hu et al. 2024) on the Flicker8K dataset. As shown in Table 4, the extraction accuracy of secret information declines as the attack intensity on images increases. However, our method demonstrates more robust performance compared to the other four methods, showing significantly less degradation under most operations. For example, when the JPEG compression intensity increased from 90 to 50, the accuracy of S2IRT, RGSD-Stego, and LDStega all declined by approximately 4 percentage points, while GSD suffered a more severe drop of nearly 10 percentage points. In contrast, our method exhibited superior robustness, with only a marginal decrease of about 2 percentage points. Furthermore, under most image attack scenarios, our method achieves higher secret information extraction accuracy compared to the other four methods, demonstrating superior robustness. Notably, when subjected to JPEG compression, which is the most common attack in real life, our method maintains approximately 97% accuracy, demonstrating superior practical utility.

Method	SRNet	XuNet	SiaStegNet
S2IRT (Zhou et al. 2022)	0.507	0.502	0.495
GSD (Zhou et al. 2025)	0.503	0.492	0.490
LDStega (Peng et al. 2024)	0.498	0.486	0.501
RGSD-Stego (Hu et al. 2024)	0.492	0.491	0.504
Ours	0.499	0.501	0.498

Table 5: The comparison of security of different steganographic methods.

Security The security of steganography includes both visual security and resistance to steganalysis detection. Figure 4 presents stego images under four different scenarios, demonstrating their high visual quality that makes them nearly indistinguishable from natural images to the naked eyes. Moreover, as indicated in Table 5, we utilize three steganographic analyzers to classify the cover and stego images on the Flicker8K dataset. Evidently, the detection error rate P_e hovers around 0.5, suggesting that steganalyzers are unable to reliably differentiate between cover and stego images. This outcome further validates the distribution-preserving property of our method.

Figure 4: Stego images generated using different text prompts for four distinct scenarios: person, marine life, flowers, and dogs (shown in four respective columns).

Conclusion

In this paper, we propose a provably secure and highly robust generative steganography framework that utilizes Latent Diffusion Models (LDM). By experimental exploration of the latent vector encoded by the encoder of LDM, we uncover a consistent pattern: irrespective of whether an image undergoes attacks such as compression or noise addition, the sign of the values within its latent vector remains almost unchanged. Leveraging this intrinsic robustness, we design an adaptive distribution-preserving mapping mechanism that enables transform secret messages into latent vectors adhering to the standard Gaussian distribution. Through this approach, both the latent vectors used to generate stego images and natural images maintain identical distributions, thereby enhancing security. Experimental results show that, our method effectively evades third-party detection while maintaining high extraction accuracy under various common attacks, demonstrating superior robustness.

Acknowledgments

This work is supported by the National Natural Science Foundation of China under grant (U22B2062; U23B2023;

62102189; 62372125), the National Social Sciences Foundation of China under grant 2022-SKJJ-C-082, the Guangdong Natural Science Funds for Distinguished Young Scholar under Grant 2023B1515020041.

References

Avrahami, O.; Lischinski, D.; and Fried, O. 2022. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 18208–18218.

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

Cachin, C. 2004. An information-theoretic model for steganography. *Information and computation*, 192(1): 41–56.

Choi, J.; Kim, S.; Jeong, Y.; Gwon, Y.; and Yoon, S. 2021. Ilvr: Conditioning method for denoising diffusion probabilistic models. *arXiv preprint arXiv:2108.02938*.

Dhariwal, P.; and Nichol, A. 2021. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34: 8780–8794.

Filler, T.; Judas, J.; and Fridrich, J. 2011. Minimizing additive distortion in steganography using syndrome-trellis codes. *IEEE Transactions on Information Forensics and Security*, 6(3): 920–935.

Gopalan, K. 2003. Audio steganography using bit modification. In *2003 International Conference on Multimedia and Expo. ICME'03. Proceedings (Cat. No. 03TH8698)*, volume 1, I–629. IEEE.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Hodosh, M.; Young, P.; and Hockenmaier, J. 2013. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47: 853–899.

Holub, V.; and Fridrich, J. 2012. Designing steganographic distortion using directional filters. In *2012 IEEE International workshop on information forensics and security (WIFS)*, 234–239. IEEE.

Hu, D.; Wang, L.; Jiang, W.; Zheng, S.; and Li, B. 2018. A novel image steganography method via deep convolutional generative adversarial networks. *IEEE access*, 6: 38303–38314.

Hu, X.; Li, S.; Ying, Q.; Peng, W.; Zhang, X.; and Qian, Z. 2024. Establishing robust generative image steganography via popular stable diffusion. *IEEE Transactions on Information Forensics and Security*.

Johnson, N. F.; and Jajodia, S. 1998. Exploring steganography: Seeing the unseen. *Computer*, 31(2): 26–34.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, 740–755. Springer.

Liu, X.; Ma, Z.; Ma, J.; Zhang, J.; Schaefer, G.; and Fang, H. 2022. Image disentanglement autoencoder for steganography without embedding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2303–2312.

Lu, C.; Zhou, Y.; Bao, F.; Chen, J.; Li, C.; and Zhu, J. 2022. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *Advances in Neural Information Processing Systems*, 35: 5775–5787.

Lu, C.; Zhou, Y.; Bao, F.; Chen, J.; Li, C.; and Zhu, J. 2025. Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models. *Machine Intelligence Research*, 1–22.

Meng, L.; Jiang, X.; Sun, T.; Zhao, Z.; and Xu, Q. 2023. A robust coverless video steganography based on the similarity of inter-frames. *IEEE Transactions on Multimedia*, 26: 5996–6011.

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

Pan, W.; Yin, Y.; Wang, X.; Jing, Y.; and Song, M. 2021. Seek-and-hide: adversarial steganography via deep reinforcement learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11): 7871–7884.

Peng, Y.; Hu, D.; Wang, Y.; Chen, K.; Pei, G.; and Zhang, W. 2023. Stegaddpm: Generative image steganography based on denoising diffusion probabilistic model. In *Proceedings of the 31st ACM International Conference on Multimedia*, 7143–7151.

Peng, Y.; Wang, Y.; Hu, D.; Chen, K.; Rong, X.; and Zhang, W. 2024. LDStega: Practical and Robust Generative Image Steganography based on Latent Diffusion Models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 3001–3009.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Saharia, C.; Ho, J.; Chan, W.; Salimans, T.; Fleet, D. J.; and Norouzi, M. 2022. Image super-resolution via iterative refinement. *IEEE transactions on pattern analysis and machine intelligence*, 45(4): 4713–4726.

Schuhmann, C.; Beaumont, R.; Vencu, R.; Gordon, C.; Wightman, R.; Cherti, M.; Coombes, T.; Katta, A.; Mullis, C.; Wortsman, M.; et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35: 25278–25294.

Song, Y.; and Ermon, S. 2019. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32.

Song, Y.; Sohl-Dickstein, J.; Kingma, D. P.; Kumar, A.; Ermon, S.; and Poole, B. 2020. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*.

Vahdat, A.; Kreis, K.; and Kautz, J. 2021. Score-based generative modeling in latent space. *Advances in neural information processing systems*, 34: 11287–11302.

Wang, Y.; Yu, J.; and Zhang, J. 2022. Zero-shot image restoration using denoising diffusion null-space model. *arXiv preprint arXiv:2212.00490*.

Xu, G.; Wu, H.-Z.; and Shi, Y.-Q. 2016. Structural design of convolutional neural networks for steganalysis. *IEEE Signal Processing Letters*, 23(5): 708–712.

Yang, J.; Liu, K.; Kang, X.; Wong, E. K.; and Shi, Y.-Q. 2018. Spatial image steganography based on generative adversarial network. *arXiv preprint arXiv:1804.07939*.

You, W.; Zhang, H.; and Zhao, X. 2020. A Siamese CNN for image steganalysis. *IEEE Transactions on Information Forensics and Security*, 16: 291–306.

You, Z.; Ying, Q.; Li, S.; Qian, Z.; and Zhang, X. 2022. Image generation network for covert transmission in online social network. In *Proceedings of the 30th ACM International Conference on Multimedia*, 2834–2842.

Zhou, Q.; Wei, P.; Qian, Z.; Zhang, X.; and Li, S. 2025. Improved Generative Steganography Based on Diffusion Model. *IEEE Transactions on Circuits and Systems for Video Technology*.

Zhou, X.; Peng, W.; Yang, B.; Wen, J.; Xue, Y.; and Zhong, P. 2021. Linguistic steganography based on adaptive probability distribution. *IEEE Transactions on Dependable and Secure Computing*, 19(5): 2982–2997.

Zhou, Z.; Dong, X.; Meng, R.; Wang, M.; Yan, H.; Yu, K.; and Choo, K.-K. R. 2023. Generative steganography via auto-generation of semantic object contours. *IEEE Transactions on Information Forensics and Security*, 18: 2751–2765.

Zhou, Z.; Su, Y.; Li, J.; Yu, K.; Wu, Q. J.; Fu, Z.; and Shi, Y. 2022. Secret-to-image reversible transformation for generative steganography. *IEEE Transactions on Dependable and Secure Computing*, 20(5): 4118–4134.

Zhou, Z.; Sun, H.; Harit, R.; Chen, X.; and Sun, X. 2015. Coverless image steganography without embedding. In *Cloud Computing and Security: First International Conference, ICCCS 2015, Nanjing, China, August 13-15, 2015. Revised Selected Papers 1*, 123–132. Springer.

Zhu, J.; Kaplan, R.; Johnson, J.; and Fei-Fei, L. 2018. Hidden: Hiding data with deep networks. In *Proceedings of the European conference on computer vision (ECCV)*, 657–672.