No Prompt, No Leaks: A Robust Generative Steganography Framework via Prompt-Free Diffusion

Jingwen Cai, Fen Xiao, Shuhua Deng, Xieping Gao, *Member, IEEE,*

***Abstract*—Generative image steganography synthesizes stego images directly from secret information to achieve inherent security advantages. Latent Diffusion Models (LDMs) have recently emerged as a fundamental image steganography framework that modulates secret latent representations with text prompts. Limited by the inflexibility of text prompts, these methods still struggle to generate high-quality stego images and accurately recover secret images. In this work, we propose a prompt-free diffusion image steganography framework that integrates style semantic priors to control more robust and reliable stego image generation. Specifically, a Cascaded Affine Coupling Module (CACM) establishes a bijective, deterministic mapping between a secret image and its latent representation. Then, style semantics are integrated into the diffusion process to control latent representation and ensure visual imperceptibility in the generated stego images. To mitigate trajectory deviations stemming from the unconditioned reverse process, a predictor-corrector mechanism is introduced to iteratively refine the generation trajectory via feedback from the current and predicted next states. Extensive experimental results show that the proposed method achieves competitive performance compared to state-of-the-art methods in terms of security, secret image reconstruction accuracy and controllability.**

***Index Terms*—Generative image steganography, diffusion model, robust steganography, prompt-free.**

I. Introduction

Image steganography conceals secret information within a cover image to enable secure and imperceptible transmission. It has been widely deployed in various security-critical scenarios, including copyright protection [1], digital watermarking [2], and secure data communication [3]. Existing methods inherently involve a three-way trade-off among payload capacity, security against steganalysis, and robustness against channel distortions and malicious attacks. Traditional image steganography typically embeds secret data directly into a cover image by minimizing a predefined distortion function. However, this embedding process inevitably introduces artifacts into the stego images that can be easily detected by advanced steganalysis tools, thereby severely limiting the security of the steganographic system [4].

Jingwen Cai, Fen Xiao and Shuhua Deng are with the MOE Key Laboratory of Intelligent Computing and Information Processing, Xiangtan University, Xiangtan 411105, China (E-mail: Jenniewer_cici@163.com; xiaof@xtu.edu.cn; shuhuadeng@xtu.edu.cn). (Corresponding authors: Fen Xiao and Shuhua Deng)

Xieping Gao is with the College of Information Science and Engineering, Hunan Normal University, Changsha 410006, China (E-mail: xpgao@hunnu.edu.cn).

Manuscript received May 15, 2026.

(a) (b) (c) (d) (e)

Fig. 1. Impact of prompt type on stego image visual quality given the same secret image and semantic content. (a) Secret image. (b) Coarse-grained text prompt (a tree). (c) Fine-grained text prompt (A massive ancient oak tree with deeply textured bark, complex branching structure, dense overlapping foliage, standing in a vast meadow at sunset). (d) Image prompt. (e) Reference image.

Generative steganography leverages significant advances in generative models to directly synthesize the stego image from secret information without relying on the cover image [5]. This property theoretically enhances the undetectability of the stego images. Early studies employed generative adversarial networks (GANs) [6], and flow-based models [7], which mapped secret data to noise vectors [8]–[11]. These methods can maintain high fidelity in recovered images. They suffer from insufficient robustness to cope with noise in real-world scenarios. To address this limitation, recent studies have introduced diffusion models into generative steganography [12]. Latent Diffusion Models (LDMs) [13] have attracted significant attention due to their inherent noise robustness, powerful generative capabilities, and computational efficiency. In this paradigm, the secret image is embedded into the latent representation, in which both the visual content of the stego image and the secret recovery are guided by user-provided text prompts [14]–[17].

Although text prompts offer enhanced controllability over the stego image, the semantic gap makes it difficult to ensure visual quality. Describing complex layouts, poses, shapes, and forms remains highly challenging and thus often leads to degradation in the generated stego image [18]. As shown in Fig. 1, three stego images generated by CRoSS [12] for the same secret image and share semantic content (“tree”), yet their visual quality differs dramatically depending on the type of text prompt. As shown in Fig. 1(b), coarse-grained prompt ”a tree” fails to sufficiently guide the generation process, which introduces obvious artifacts into the resulting stego image. While finer-grained prompts (e.g.a sentence), shows in Fig. 1(c), can improve stego image quality to some extent, better aligning the generated images with user intent still requires numerous trial-and-error prompt edits. This process is not only often impractical but also inevitably increases

the generation cost [18], [19]. To reduce manual prompt engineering, we also follow the prior work [20] by using a reference image for structural guidance, as shown in Fig. 1(e). However, the resulting stego image shown in Fig. 1(d) leaks content both from the reference image and the secret image, which increases the risk of being detected by a steganalyzer.

Beyond the above security risks, the transmission of text prompts introduces additional reliability issues. Since the diffusion model requires the prompt as a conditioning signal for secret recovery, the prompt must be transmitted alongside the stego image. However, as a semantic key transmitted over practical channels, it is vulnerable to noise and tampering attacks [21]. Given the high sensitivity of diffusion models, even minor perturbations to the prompt can severely disrupt the recovery process [15]. This disruption is further aggravated by iterative accumulation, which gradually drives the recovery trajectory away from its intended path [22], [23] and ultimately leads to failed secret reconstruction. Moreover, relying on a reference image as the recovery condition introduces lossy transmission [24], which further exacerbates the system's unreliability in real-world scenarios.

To address the above security and reliability challenges, we propose a prompt-free diffusion steganographic framework for robust image steganography. We first discuss replacing text prompts with semantic priors to guide realistic stego image generation. These priors, extracted from reference datasets via a Vision Transformer (ViT) [25], are used to avoid content-specific overfitting in the stego process, thereby improving both generation security and visual quality. Although this design eliminates the transmission risk associated with text prompts, the absence of explicit semantic guidance inevitably increases the uncertainty of the reverse trajectory. An unconditional way is introduced to stabilize the recovery process in the prompt-free setting. This framework offers several advantages, including user-controllable stego generation and enhanced robustness against real-world transmission distortions. The main contributions of this work are as follows.

1) We introduce a prompt-free stego image generation scheme with semantic priors. A cascaded affine coupling module (CACM) is designed to establish a deterministic mapping between the secret image and its latent representation. Subsequently, semantic priors obtained from reference datasets via ViT are incorporated into this representation to guide authentic stego image generation. This approach substantially enhances both flexibility and security.
2) The proposed framework incorporates a predictor-corrector mechanism that iteratively refines the recovery trajectory without external conditioning. This mechanism uses residual feedback from the current and predicted next steps to correct the reverse trajectory, which stabilizes latent representation extraction even for degraded stego image.
3) Extensive experiments demonstrate that the proposed framework outperforms state-of-the-art (SOTA) image steganography methods. Quantitative and qualitative results further show that the framework enables flexible control over stego image generation while preserving strong security against steganalysis. It also achieves superior recovery performance even in real transmission scenarios.

The rest of this paper is organized as follows. Section II introduces some classical and diffusion-based generative image steganography. Section III introduces the proposed method in detail. Section IV presents extensive experimental results and analysis. Section V summarizes the paper.

II. Related Works

A. Modification-Based Image Steganography

Traditional image steganography refers to the practice of concealing a secret image within a cover image in an unnoticed manner. Baluja [26] pioneered the use of deep neural networks for image steganography and later extended this framework [27] to conceal two secret images within a single cover image. Later studies incorporated GANs [6] into image steganography [28], where adversarial training is employed to embed a secret image into a cover image. The realistic appearance of the stego image ensures its security. SteganoGAN [29] introduces an adversarial network alongside the encoder and decoder networks for information embedding and extraction. This design enhances resistance to steganalysis.

Recently, Invertible Neural Networks (INNs) have attracted attention in image steganography for establishing effective invertible mappings between secret and stego images [30]. HiNet [31] models the image concealing and revealing as the forward and backward processes of an invertible network, which drastically increases both the hiding security and recovering accuracy. DeepMIH [32] extends this paradigm to multi-image hiding by using the preceding stego image as guidance for embedding subsequent secret images. Subsequent works adopted different strategies to improve INN-based steganography. For instance, iSCMIS [33] incorporates spatial and channel attention for secure region-guided embedding, while MIGIIHNet [34] employs mutual information estimation between the lost information and the stego image to achieve accurate recovery. However, these methods all rely on modifying the cover image to hide the secret image, which inevitably leaves detectable traces and renders the stego images vulnerable to well-designed steganalysis methods [35], [36].

B. Diffusion-Based Generative Image Steganography

Diffusion models [37], [38] have recently emerged as a powerful generative framework for image generation [39]–[41]. Among them, LDMs have attracted increasing attention in generative steganography [42]–[45] due to their superior image generation quality. Most existing methods rely on Denoising Diffusion Implicit Model (DDIM) inversion [37] to construct the mapping from the secret image to the latent space, while the semantic content of the stego image is specified by text prompts.

CRoSS [12] and subsequent works [14] employ a text prompt to establish the mapping from the secret image to the latent space, while another text prompt is used to guide stego

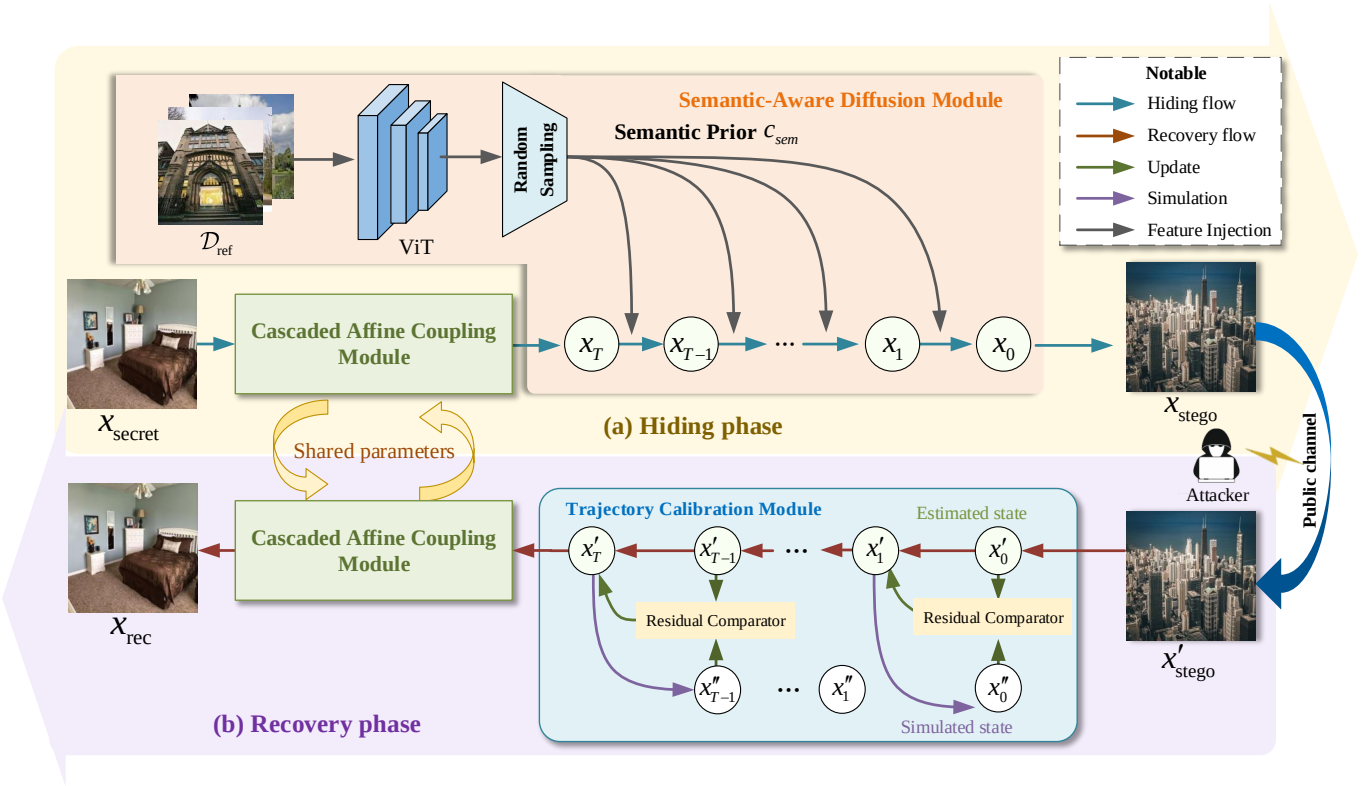


Fig. 2. The proposed prompt-free generative steganography framework mainly consists of three modules: CACM, SADM, and TCM. These modules are responsible for secret image embedding/decoding, stego image generation, and latent recovery, respectively.

image generation. Although this mechanism ensures that the secret image can be recovered only with the correct prompts, text prompts have limited expressive ability for constructing the secret-to-latent mapping. Only part of the secret image can be concealed, which still compromises system security. To improve the hiding capability in the latent space without relying on text prompts, subsequent studies introduce VQGAN compression [15] and scrambling transformations [16]. However, these methods still rely heavily on text prompts during stego image generation. Due to the limited ability of text prompts to describe complex visual details, the generated images may suffer from degraded visual quality [18], [19]. DiffStega [17] introduces pre-determined passwords to generate reference images. These reference images provide auxiliary guidance for stego image generation and improve the visual details. However, the resulting stego images may contain content from both the reference and secret images. The significant content overlap between the secret and stego images, which compromises the confidentiality and the undetectability of the steganographic system.

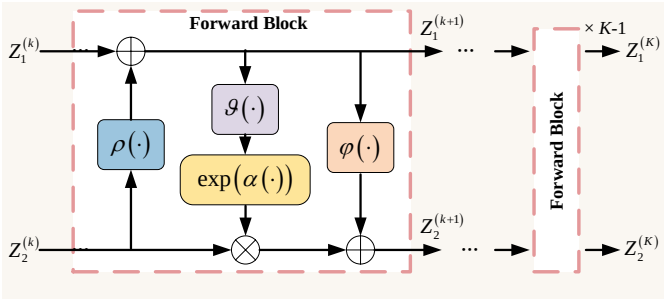
In addition, stego images are often affected by lossy transmission in real-world scenarios, such as JPEG compression and Gaussian noise. These distortions may degrade the recovery accuracy. Existing work [46] has focused mainly on robustness against known channel attacks. However, secret recovery in these methods still relies heavily on text prompts. Text prompts must be transmitted together with the stego image to guide secret recovery. During transmission, they are vulnerable to noise interference and tampering attacks [21]. Even minor perturbations to the transmitted text prompts can introduce errors into the recovery trajectory. Diffusion models exhibit inherent numerical instability and high sensitivity to perturbations [24]. Through iterative accumulation in the recovery process, these errors may cause the recovered trajectory to deviate significantly from the original trajectory [22], [23]. Consequently, the reliance on text prompts poses a major challenge to robust recovery.

III. Method

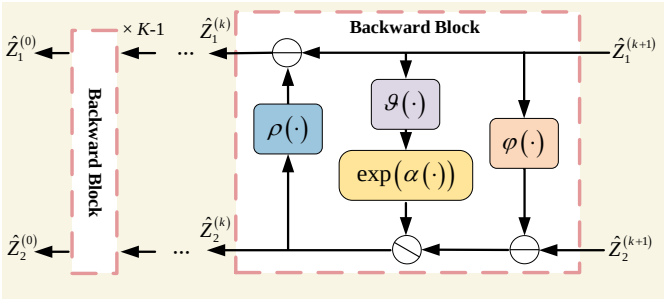
As shown in Fig. 2, the proposed prompt-free generative image steganography framework consists of two stages. In the hiding phase, the secret image x_{secret} is deterministically mapped by Cascaded Affine Coupling Module (CACM) to the latent representation x_T, which serves as the initial state of the diffusion process. In the Semantic-Aware Diffusion Module (SADM), a sampled semantic prior c_{sem} extracted from $\mathcal{D}_{\text{ref}}$ by ViT is injected into the diffusion process to guide stego image generation. In the recovery phase, the received stego image x'_{stego} is inverted through the Trajectory Calibration Module (TCM), which calibrates the recovery trajectory and obtains the latent representation x'_T. The recovered image x_{rec} is reconstructed from the x'_T through the inverse transformation of CACM.

A. Cascaded Affine Coupling Module

The CACM is designed to transform the secret image x_{secret} into a latent representation x_T that matches the input of



(a) Forward Block



(b) Backward Block

Fig. 3. Architecture of the proposed CACM. The forward block maps the secret image to the latent representation, and the backward block reconstructs the image through the inverse transformation.

the diffusion model. As shown in Fig. 3, CACM comprises forward and backward blocks with shared parameters, which are used for secret embedding and recovery, respectively. The initial feature $Z^{(0)} = x_{\text{secret}}$ is first split into two parts along the channel dimension, denoted as $Z_1^{(0)}$ and $Z_2^{(0)}$. In the k-th forward block, $Z_2^{(k-1)}$ first guides the update of $Z_1^{(k-1)}$. The updated feature $Z_1^{(k)}$ then generates the scale and shift terms to adaptively compress $Z_2^{(k-1)}$ through an affine coupling transformation. The forward block is formulated as follows:

$$Z_1^{(k)} = Z_1^{(k-1)} \oplus \tau(Z_2^{(k-1)}), \tag{1}$$

$$Z_2^{(k)} = Z_2^{(k-1)} \odot \exp(\gamma(\vartheta(Z_1^{(k)}))) \oplus \eta(Z_1^{(k)}), \tag{2}$$

where $\gamma(\cdot)$ denotes the sigmoid function. The operator $\odot$ denotes element-wise multiplication, and $\oplus$ denotes element-wise addition. $\tau(\cdot)$, $\vartheta(\cdot)$, and $\eta(\cdot)$ denote three subnetworks that share the same architecture but do not share parameters. Specifically, $\tau(\cdot)$ is used to predict the additive update of $Z_1^{(k)}$. $\vartheta(\cdot)$ and $\eta(\cdot)$ are used to parameterize the scaling and shifting terms of the affine transformation on $Z_2^{(k)}$, respectively. Given the input ϕ_0, the subnetwork is formulated as follows:

$$\begin{aligned} &\mathcal{H}(\phi_0) = \text{Concat}(\phi_0, \phi_1, \phi_2, \phi_3, \phi_4), \\ &\phi_i = \text{LReLU}(\text{Conv}_{3\times3}(\phi_{i-1})), \quad i = 1, \ldots, 4. \end{aligned} \tag{3}$$

where $\text{Concat}(\cdot)$ denotes concatenation along the channel dimension. After the K-th forward block, the final latent representation x_T is obtained by

$$x_T = \text{Down}\left(\text{Concat}(Z_1^{(K)}, Z_2^{(K)})\right), \quad x_T \in \mathbb{R}^{4\times64\times64} \tag{4}$$

where $\text{Down}(\cdot)$ denotes the downsampling operation. x_T is then used as the input to the subsequent diffusion model for stego image generation.

In the recovery process, the reconstructed latent representation $\hat{x}_T$ initializes the input feature $\hat{Z}^{(K)}$ of the K-th backward block. The subsequent propagation proceeds from the $(k+1)$-th block to the k-th block. The k-th backward block is formulated as follows:

$$\hat{Z}_2^{(k-1)} = (\hat{Z}_2^{(k)} \ominus \eta(\hat{Z}_1^{(k)})) \oslash \exp(\gamma(\vartheta(\hat{Z}_1^{(k)}))), \tag{5}$$

$$\hat{Z}_1^{(k-1)} = \hat{Z}_1^{(k)} \ominus \tau(\hat{Z}_2^{(k-1)}), \tag{6}$$

where $\oslash$ denotes element-wise division, and $\ominus$ represents element-wise subtraction. After the last backward block, the secret image is reconstructed as $\hat{Z}^{(0)} = x_{\text{rec}}$.

B. Semantic-Aware Diffusion Module

The SADM is designed to perform semantic-conditioned DDIM sampling from the latent representation x_T to the stego image x_{stego}. Semantic priors are first extracted from reference images and injected into the prediction network ϵ_θ, which provides semantic control over the generating trajectory toward the target semantics.

Given a reference dataset $\mathcal{D}_{\text{ref}}$, a semantic prior set $\mathcal{C}_{\text{sem}} = \{\text{ViT}(x) \mid x \in \mathcal{D}_{\text{ref}}\}$ is constructed by applying the ViT to the reference images. A semantic prior c_{sem} is then randomly sampled from $\mathcal{C}_{\text{sem}}$ and injected into the prediction network ϵ_θ. c_{sem} is fused with intermediate feature maps to provide semantic guidance and suppress potential artifacts.

The semantically guided sampling process can be formulated as the following ordinary differential equation (ODE):

$$\frac{dx_t}{dt} = \left(f(t) + \frac{g^2(t)}{2\sigma_t^2}\right) x_t - \frac{\alpha_t g^2(t)}{2\sigma_t^2} \epsilon_\theta(x_t, t, c_{\text{sem}}), \tag{7}$$

where $t \in [0, T]$, $f(t) := \frac{d\log\alpha_t}{dt}$, $g^2(t) := \frac{d\sigma_t^2}{dt} - 2\frac{d\log\alpha_t}{dt}\sigma_t^2$.

At each step, DPM-Solver++ [47] approximates the exact solution at t_i from the state $x_{t_{i-1}}$. This approximation is derived by applying a Taylor expansion around $\lambda_{t_{i-1}}$ as follows:

$$\begin{aligned} x_{t_i} = &\frac{\sigma_{t_i}}{\sigma_{t_{i-1}}} x_{t_{i-1}} \\ &+ \sigma_{t_i} \sum_{n=0}^{w-1} \epsilon_\theta^{(n)}(x_{t_{i-1}}, t_{i-1}, c_{\text{sem}}) \int_{\lambda_{t_{i-1}}}^{\lambda_{t_i}} e^{\lambda} \frac{(\lambda - \lambda_{t_{i-1}})^n}{n!} d\lambda \\ &+ \mathcal{O}(h_i^{w+1}). \end{aligned} \tag{8}$$

We use the first-order case ($w = 1$), and the corresponding update rule is written as follows:

$$x_{t_i} = \frac{\sigma_{t_i}}{\sigma_{t_{i-1}}} x_{t_{i-1}} - \alpha_{t_i}(e^{-h_i} - 1)\epsilon_\theta(x_{t_{i-1}}, t_{i-1}, c_{\text{sem}}), \tag{9}$$

where $h_i = \lambda_{t_i} - \lambda_{t_{i-1}}$ represents the step size in the log-SNR space.

The secret latent representation x_T is used as the initial state of the generation process, and the stego image x_{stego} is progressively generated from it. Algorithm 1 summarizes the stego image generation procedure of the proposed method.

Algorithm 1 Stego Image Generation with Semantic Guidance

Input: Secret image x_{secret}, reference dataset $\mathcal{D}_{\text{ref}}$, number of affine coupling blocks K, time steps $\{t_i\}_{i=1}^{N}$.
Output: Stego image x_{stego}.

Stage 1: Invertible latent mapping and encoding
Initialize $Z^{(0)} \leftarrow x_{\text{secret}}$, and split $Z^{(0)}$ into $[Z_1^{(0)}, Z_2^{(0)}]$.
for $k = 1, \ldots, K$ **do**
  $Z_1^{(k)} \leftarrow Z_1^{(k-1)} \oplus \tau\left(Z_2^{(k-1)}\right)$.
  $Z_2^{(k)} \leftarrow Z_2^{(k-1)} \odot \exp\left(\gamma\left(\vartheta\left(Z_1^{(k)}\right)\right)\right) \oplus \eta\left(Z_1^{(k)}\right)$.
end for
Assemble secret latent $\hat{x}_T \leftarrow [Z_1^{(K)}, Z_2^{(K)}]$.

Stage 2: Semantic prior sampling
$\mathcal{C}_{\text{sem}} \leftarrow \text{ViT}\left(\mathcal{D}_{\text{ref}}\right)$.
$c_{\text{sem}} \leftarrow \text{RandomSample}\left(\mathcal{C}_{\text{sem}}\right)$.

Stage 3: Semantic-conditioned diffusion generation
$x_{t_N} \leftarrow \hat{x}_T$.
for $i = N, \ldots, 1$ **do**
  $\hat{\epsilon}_i \leftarrow \epsilon_\theta\left(x_{t_i}, t_i, c_{\text{sem}}\right)$.
  $h_i \leftarrow \lambda_{t_{i-1}} - \lambda_{t_i}$.
  $x_{t_{i-1}} \leftarrow \text{DPM-Solver++}\left(x_{t_i}, \hat{\epsilon}_i, h_i\right)$.
end for
$x_{\text{stego}} \leftarrow \text{Decode}\left(x_{t_0}\right)$.
return x_{stego}.

Algorithm 2 Secret Image Recovery without Condition

Input: Stego image x_{stego}, inversion steps N, time steps $\{t_i\}_{i=1}^{N}$, number of affine coupling blocks K.
Output: Recovered secret image x_{rec}.

Stage 1: Unconditional inversion with TCM
$x'_{t_0} \leftarrow x_{\text{stego}}$
for $i = 1, \ldots, N$ **do**
  $x'_{t_{i-1}} \leftarrow \frac{\sigma_{t_{i-1}}}{\sigma_{t_i}}\left(x'_{t_i} + \alpha_{t_i}(e^{-h_i} - 1)\epsilon_\theta(x'_{t_i}, t_i)\right)$
  repeat
    $x''_{t_i} \leftarrow \frac{\sigma_{t_i}}{\sigma_{t_{i-1}}} x'_{t_{i-1}} - \alpha_{t_i}(e^{-h_i} - 1)\epsilon_\theta(x'_{t_{i-1}}, t_{i-1})$
    $x'_{t_{i-1}} \leftarrow x'_{t_{i-1}} - \vartheta \nabla_{x'_{t_{i-1}}} \mathcal{R}_i$
    $\text{UPDATE}(x'_{t_{i-1}}, x''_{t_i}, x'_{t_i})$
  until converged
end for
$\hat{x}_T \leftarrow x'_{t_N}$

Stage 2: Inverse latent mapping and decoding
Split $\hat{x}_T$ into $[\hat{Z}_1^{(K)}, \hat{Z}_2^{(K)}]$
for $k = K, \ldots, 1$ **do**
  $\hat{Z}_2^{(k-1)} \leftarrow (\hat{Z}_2^{(k)} \ominus \eta(\hat{Z}_1^{(k)})) \oslash \exp(\gamma(\vartheta(\hat{Z}_1^{(k)})))$
  $\hat{Z}_1^{(k-1)} \leftarrow \hat{Z}_1^{(k)} \ominus \tau(\hat{Z}_2^{(k-1)})$
end for
$x_{\text{rec}} \leftarrow \text{Concat}(\hat{Z}_1^{(0)}, \hat{Z}_2^{(0)})$
return x_{rec}

C. Trajectory Calibration Module

In the recovery stage, the stego image x_{stego} is inverted back to the latent representation without semantic conditioning. From Eq. (9), the inversion at time t_i can be written as follows:

$$x'_{t_i} = \frac{\sigma_{t_i}}{\sigma_{t_{i-1}}} x'_{t_{i-1}} - \alpha_{t_i}(e^{-h_i} - 1)\epsilon_\theta(x'_{t_{i-1}}, t_{i-1}). \tag{10}$$

Since $x'_{t_{i-1}}$ is unknown during the current step, the prediction term $\epsilon_\theta(x'_{t_{i-1}}, t_{i-1})$ cannot be directly obtained. In naïve DDIM inversion, x'_{t_i} is used to approximate the unknown term $x'_{t_{i-1}}$, which inevitably introduces numerical errors. To address this issue, we introduce the TCM that predicts the next state and then calibrates the recovery trajectory through single-step discrepancy feedback. The TCM first computes an estimate state $x'_{t_{i-1}}$ from the current state x'_{t_i} using naïve DDIM inversion:

$$x'_{t_{i-1}} = \frac{\sigma_{t_{i-1}}}{\sigma_{t_i}}\left(x'_{t_i} + \alpha_{t_i}(e^{-h_i} - 1)\epsilon_\theta(x'_{t_i}, t_i)\right). \tag{11}$$

The estimated state $x'_{t_{i-1}}$ is used to construct a simulated forward state x''_{t_i}:

$$x''_{t_i} \leftarrow \frac{\sigma_{t_i}}{\sigma_{t_{i-1}}} x'_{t_{i-1}} - \alpha_{t_i}(e^{-h_i} - 1)\epsilon_\theta(x'_{t_{i-1}}, t_{i-1}). \tag{12}$$

The mismatch between the simulated state x''_{t_i} and the observed current state x'_{t_i} is measured by:

$$\mathcal{R}_i = \|x''_{t_i} - x'_{t_i}\|_2^2. \tag{13}$$

The estimated state $x'_{t_{i-1}}$ is then refined by using the gradient of $\mathcal{R}_i$:

$$x'_{t_{i-1}} \leftarrow x'_{t_{i-1}} - \vartheta \nabla_{x'_{t_{i-1}}} \mathcal{R}_i, \tag{14}$$

where ϑ denotes the update step size for the gradient trajectory calibration. This operation performs gradient correction based on the mismatch between the forward-projected state and the current state. In this way, TCM not only compensates for the approximation error of DDIM inversion but also corrects reverse trajectory deviations caused by distortions in the transmission channel. The recovered latent representation $\hat{x}_T$ is then fed into CACM for accurate reconstruction of the secret image x_{rec}. Algorithm 2 provides a detailed outline of the comprehensive steps involved in the proposed method for recovering the secret image.

D. Loss Functions

We combine the loss terms introduced to form the total loss function $\mathcal{L}_{\text{total}}$. This objective is composed of a recovery loss $\mathcal{L}_{\text{rec}}$ and a trajectory consistency loss $\mathcal{L}_{\text{tc}}$, which are defined as follows:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{rec}}\mathcal{L}_{\text{rec}} + \lambda_{\text{tc}}\mathcal{L}_{\text{tc}}, \tag{15}$$

where λ_{rec} and λ_{tc} are weighted coefficients for training.

a) Recovery Loss: To ensure accurate reconstruction of the secret image during recovery, we introduce a recovery loss $\mathcal{L}_{\text{rec}}$, which minimizes the discrepancy between the recovery image x_{rec} and the secret image x_{secret}. The loss is defined as:

$$\mathcal{L}_{\text{rec}} = \|x_{\text{rec}} - x_{\text{secret}}\|_1, \tag{16}$$

where $\|\cdot\|_1$ denotes the L_1 norm regularization.

b) Trajectory Consistency Loss: To ensure accurate recovery, we introduce a trajectory consistency loss that measures the discrepancy between states. The loss computes the squared L_2 distance between the original state and the simulated state $\hat{x}_T$ obtained through reverse inversion:

$$\mathcal{L}_{\text{tc}} = \|x_T - \hat{x}_T\|_2^2. \tag{17}$$

IV. Experiment

In this section, we first introduce the experimental settings. Then, we provide a quantitative and qualitative evaluation on the visual appearance of the generated stego images. The security performance against some steganalyzers is reported also. Additionally, we assess its robustness against different degradations. Finally, ablation experiments are conducted to illustrate the necessity of each component in our framework.

A. Implementation Details

a) Experimental Settings: In our experiments, Stable Diffusion v2 [13] provided by HuggingFace is selected as the baseline. The guidance scale of the diffusion model is set to 3.0. During inference, DPMSolver++ [47] is used for sampling with 50 steps.

The number of CACM layers is set to $K = 4$. The hyperparameters of the loss function λ_{rec} and λ_{tc} are set to 2.0 and 1.0, respectively. The model is trained for 20 epochs with the Adam optimizer and a learning rate of 1.0×10^{-4}. The batch size is set to 1. All experiments are conducted on an NVIDIA L20 GPU with the PyTorch framework.

b) Baseline Methods: We focus on high-capacity image steganography and select seven classical baseline methods for comparison. These methods include some traditional modification-based image steganography (TS) methods, such as HiNet [31], ISN [30], iSCMIS [33], and MIGI-IHNet [34]. State-of-the-art generative steganography (GS) methods, including CRoSS [12], DiffStega [17], and VQGAN-DiffStega [15]. All these models utilized in this study were pre-trained.

c) Data Preparation: Following the data setting in [15], we use the FFHQ dataset [48] as the secret image set, with 5,000 images for training and 1,000 for testing. For GS methods that require text prompts, we use ChatGPT to extract text prompts from the Bedroom category of the LSUN dataset [49]. For TS methods, the cover images are generated by the same diffusion model. All images are center-cropped to 512×512 pixels and the corresponding latent space has a dimension of $64 \times 64 \times 4$. The embedding capacity for all baselines is set at 24 bpp (bits per pixel).

d) Evaluation metrics: In the experiments, we adopt various metrics to evaluate the performance of diverse steganographic methods. Specifically, we measure extraction accuracy, robustness, security, and visual quality.

Visual quality: The Fréchet Inception Distance (FID) [50] serves as a benchmark for assessing image quality, as it quantifies the Fréchet distance between the feature distributions of stego images and real images. The FID is defined as follows:

$$\text{FID}_{(x,y)} = \|\mu_x - \mu_y\|^2 + \text{Tr}\left(\Sigma_x + \Sigma_y - 2\left(\Sigma_x\Sigma_y\right)^{1/2}\right), \tag{18}$$

where μ_x and μ_y denote the mean feature vectors of the real and stego images, respectively. Σ_x and Σ_y denote their corresponding covariance matrices.

For security, the stego image should be visually authentic and with no content leakage. According to this, we calculate two FIDs, i.e. $\text{FID}_{(\text{ste,real})}$ and $\text{FID}_{(\text{ste,sec})}$. The former measures the distribution distance between stego images and real images (bedroom dataset [49]) to assess authenticity, for which a lower score indicates higher resemblance to real images. For the latter, we use the secret FFHQ dataset as the real image set. A higher score implies less information leakage.

Natural Image Quality Evaluator (NIQE) [51] is a no-reference image quality assessment metric that evaluates the perceptual naturalness of stego images, with lower scores indicating greater statistical similarity to natural images.

Extraction accuracy: We use PSNR and SSIM to measure extraction accuracy by comparing the recovered image with the original secret image, where higher values indicate better reconstruction. PSNR is defined based on the mean squared error (MSE) as follows:

$$\text{MSE} = \frac{1}{W \times H}\sum_{i=1}^{W}\sum_{j=1}^{H}\left[X_{i,j} - Y_{i,j}\right]^2, \tag{19}$$

$$\text{PSNR} = 10 \cdot \log_{10}\frac{\text{MAX}^2}{\text{MSE}}. \tag{20}$$

SSIM is defined as follows:

$$\text{SSIM} = \frac{\left(2\mu_x\mu_y + c_1\right)\left(2\sigma_{xy} + c_2\right)}{\left(\mu_x^2 + \mu_y^2 + c_1\right)\left(\sigma_x^2 + \sigma_y^2 + c_2\right)}. \tag{21}$$

Security: The security refers the detection resistance of stego image against the steganalysis tool. It is commonly evaluated by the detection error P_E of the steganalyzer, which is defined as follows:

$$P_E = \frac{P_{\text{FA}} + P_{\text{MD}}}{2}, \tag{22}$$

where P_{FA} is the false alarm rate and P_{MD} is the missed detection rate. The value of P_E ranges from 0 to 1. $P_E = 0.5$ corresponds to random guessing (equal likelihood of misclassifying cover or stego images), indicating optimal resistance to steganalysis. To measure P_E, we adopt three steganalysis tools: StegExpose [52], XuNet [35] and KeNet [36]. StegExpose is a widely used steganalysis tool that integrates multiple statistical methods, including RS analysis and the Chi-square attack. While XuNet and KeNet are representative deep learning-based steganalyzers.

Robustness: To evaluate robustness against unknown real-world distortions, we test extraction accuracy under several attacks, including JPEG compression, Gaussian noise, Salt-and-pepper noise, Gaussian filtering, Median filtering, rotation, and resizing.

B. Visual Quality Comparison

Since traditional methods(TS) preserve visual quality by modifying a cover image, visual comparison with GS methods on visual quality is unfair. Therefore, we focus on the visual quality comparison among GS methods.

TABLE I
THE VISUAL COMPARISON OF STEGO IMAGES GENERATED BY DIFFERENT GENERATIVE-BASED SCHEMES. THE BEST RESULTS ARE IN BOLD AND THE SECOND-BEST RESULTS ARE UNDERLINED.

Methods	NIQE ↓	$\text{FID}_{(\text{ste,real})}$ ↓	$\text{FID}_{(\text{ste,sec})}$ ↑
CRoSS [12]	4.42	49.052	71.608
DiffStega [17]	4.966	48.698	74.055
VQGAN-DiffStega [15]	<u>3.25</u>	<u>17.709</u>	<u>292.933</u>
Ours	**3.15**	**9.254**	**342.506**

Figure 4 shows stego images generated by CRoSS and the proposed method under four semantic conditions: 'Flower', 'Bedroom', 'Animal', and 'Human'. Each column shares the same semantic condition. The first two rows show stego images from CRoSS, guided by label names as text prompts. Since text prompts lack the ability to capture fine-grained details, the stego images suffer from reduced visual authenticity and content leakage. The third and fourth rows present our method guided by semantic priors from reference images. It is obvious that the proposed method produce a diverse set of images from the same secret image. The obtained stego images exhibit clearer textures, and a better semantic alignment compared to the obtained from the other approaches.

For quantitative comparison, we further adopt NIQE and $\text{FID}_{(\text{ste,real})}$ to evaluate the visual quality of our method. The results are given in the first two columns of Table I. Compared with CRoSS [12] and DiffStega [17], VQGAN-DiffStega [15] shows relatively competitive performance ob both two metrics. The proposed method achieves the lowest NIQE and $\text{FID}_{(\text{ste,real})}$ scores, which further confirms its superior visual quality.

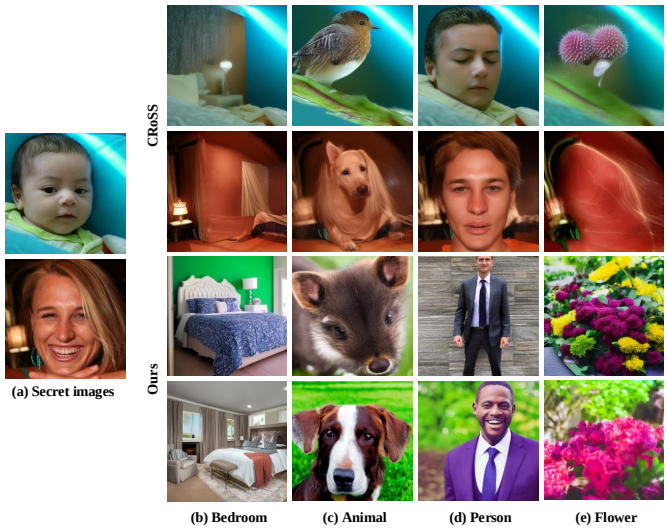


Fig. 4. Stego images generated by CRoSS and our proposed method with the same semantic conditions.

C. Security Analysis

We compare the proposed framework with several state-of-the-art (SOTA) methods to evaluate its security in terms of imperceptibility and undetectability. Imperceptibility is determined by how much secret-related content remains in the stego image. Following [15], we quantify this effect by measuring the distribution distance between the secret and stego image sets. As shown in the third column of Table I, CRoSS [12] and DiffStega [17] obtain lower $\text{FID}_{(\text{ste,sec})}$ scores. This indicates a higher resemblance between the secret and stego image sets, which implies a severe risk of semantic leakage. By contrast, the proposed framework overcomes this issue by aligning the generation with the target semantics. As expected, the proposed method achieves the highest $\text{FID}_{(\text{ste,sec})}$ score, which suggests the lowest risk of content leakage.

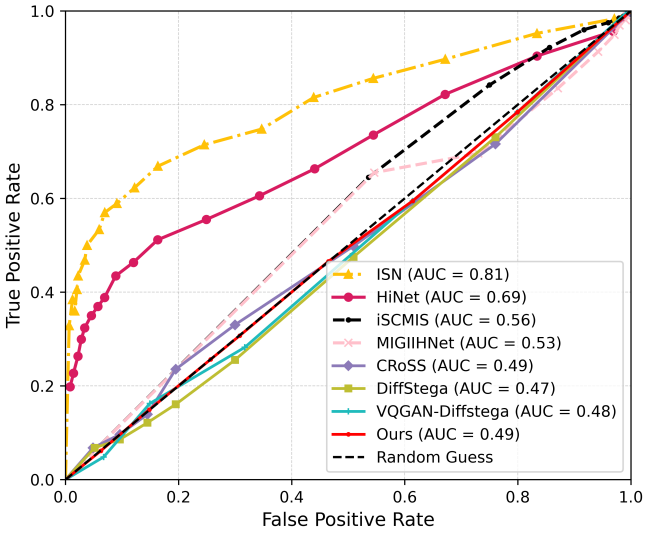


Fig. 5. ROC curves of StegExpose for different schemes.

To evaluate the undetectability of the proposed method, 2,000 cover/stego image pairs generated by state-of-the-art (SOTA) methods were utilized for steganalysis. For fair comparison, the cover images are generated by the same diffusion model without embedding any secret. These cover images serve both as carriers for TS methods and as a reference for GS-generated stego images.

TABLE II
QUANTITATIVE COMPARISON OF UNDETECTABILITY AMONG DIFFERENT METHODS. THE BEST RESULTS ARE IN BOLD AND THE SECOND-BEST RESULTS ARE UNDERLINED.

Methods	Type	XuNet [35]	KeNet [36]
HiNet [31]	TS	0.5591	0.5833
ISN [30]		0.5512	0.5827
iSCMIS [33]		0.4755	0.5483
MIGIIHNet [34]		0.4910	0.5163
CRoSS [12]	GS	0.5132	0.5211
DiffStega [17]		0.5078	-
VQGAN-DiffStega [15]		0.5198	-
Ours		**0.5005**	**0.5049**

Fig. 5 shows the receiver operating characteristic (ROC) curves for StegExpose [52] for different methods. The proposed method achieves an area under the ROC curve (AUC) of 0.49, which is close to the level of a random classifier. This indicates the StegExpose cannot reliably distinguish the

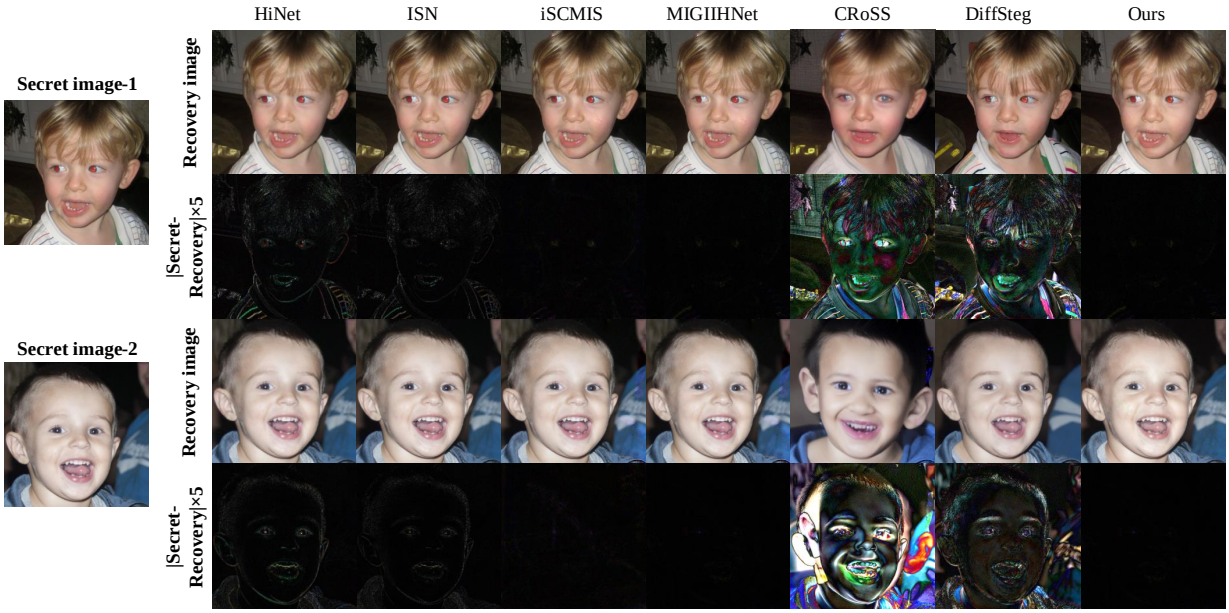


Fig. 6. Demonstrations of the recovery images obtained by different schemes.

TABLE III
PSNR/SSIM OF RECOVERED IMAGES FROM DEGRADED STEGO IMAGES. THE BEST RESULTS ARE IN BOLD AND THE SECOND-BEST RESULTS ARE UNDERLINED.

Methods	Type	Clean	Gaussian Noise				JPEG Compression			
			$\sigma = 0.035$	$\sigma = 0.025$	$\sigma = 0.015$	$\sigma = 0.005$	QF=30	QF=50	QF=70	QF=90
HiNet [31]	TS	37.05/0.93	10.64/0.05	12.72/0.09	16.46/0.18	19.65/0.28	11.16/0.32	11.23/0.32	11.25/0.33	11.30/0.33
ISN [30]		36.77/0.96	13.42/0.12	16.04/0.19	17.85/0.24	20.19/0.33	8.30/0.38	8.61/0.40	8.75/0.40	8.82/0.42
iSCMIS [33]		44.71/0.98	13.45/0.13	19.88/0.25	21.30/0.34	26.83/0.57	11.34/0.44	11.30/0.38	11.27/0.34	11.34/0.27
MIGIIHNet [34]		**46.76/0.99**	15.44/0.20	18.47/0.25	20.63/0.27	28.27/0.33	11.51/0.38	11.47/0.37	11.48/0.35	11.47/0.32
CRoSS [12]	GS	22.75/0.74	20.98/0.51	21.55/0.58	22.28/0.65	22.77/0.68	20.36/0.58	21.01/0.60	22.01/0.67	22.54/0.70
DiffStega [17]		23.13/0.74	21.08/0.52	21.66/0.59	22.34/0.65	22.67/0.68	20.57/0.60	21.23/0.63	22.13/0.69	22.64/0.71
VQGAN-DS [15]		23.65/0.72	21.20/0.63	21.74/0.64	22.47/0.66	23.09/0.69	20.86/0.62	21.39/0.64	22.24/0.67	22.88/0.69
Ours		46.50/0.99	**33.29/0.97**	**33.33/0.97**	**33.23/0.97**	**33.14/0.97**	**37.34/0.99**	**38.24/0.99**	**39.35/0.99**	**40.75/0.99**

generated stego images from cover images, which confirms the superior undetectability of our approach. Furthermore, the detection error rates obtained by XuNet and KeNet are also reported in Table II. As shown in the table, stego images generated by TS methods, such as HiNet [31] and ISN [30], are more easily detectable, as modifying cover images inevitably introduces statistical artifacts. Unlike TS methods, GS methods synthesize stego images directly without any cover modification, which minimizes detectable artifacts and offers stronger resistance against steganalysis. Notably, the detection error rate of the proposed method remains consistently closer to 0.5 than the compared methods, indicating its superior undetectability.

D. Robustness

To evaluate robustness, we simulate stego image degradation using Gaussian noise and JPEG compression at different levels. Table III reports the PSNR and SSIM scores of the recovered images. The first column of the table shows that TS methods outperform standard GS methods in recovering secret images from clean stego images. Our method achieves competitive performance, a conclusion further supported by the visual comparison of recovered images in Fig. 6.

However, for degraded stego images, TS performance degrades significantly while GS methods including CRoSS, DiffStega and VQGAN-DS demonstrate stronger resilience and yield relatively consistent recovery accuracy. The proposed method achieves the best and most stable recovery scores across all degradation levels. Fig. 7 provides a visual comparison, including the original secret images and their corresponding recovered versions from stego images degraded at different levels. Our scheme demonstrates remarkable fidelity in recovering the secret content, whereas images recovered by other baseline schemes exhibit noticeable color and texture distortions.

E. Ablation Studies

To assess the contribution of each component in our framework, we conduct ablation experiments by adopting CRoSS [12] as the baseline. The CRoSS result is taken from the reported result in [15]. The proposed modules are then progressively integrated into this baseline, and the resulting variants are evaluated under the same experimental setting.

a) Effectiveness of the CACM: CACM is employed to embed secret image into the latent representation. Rows 1

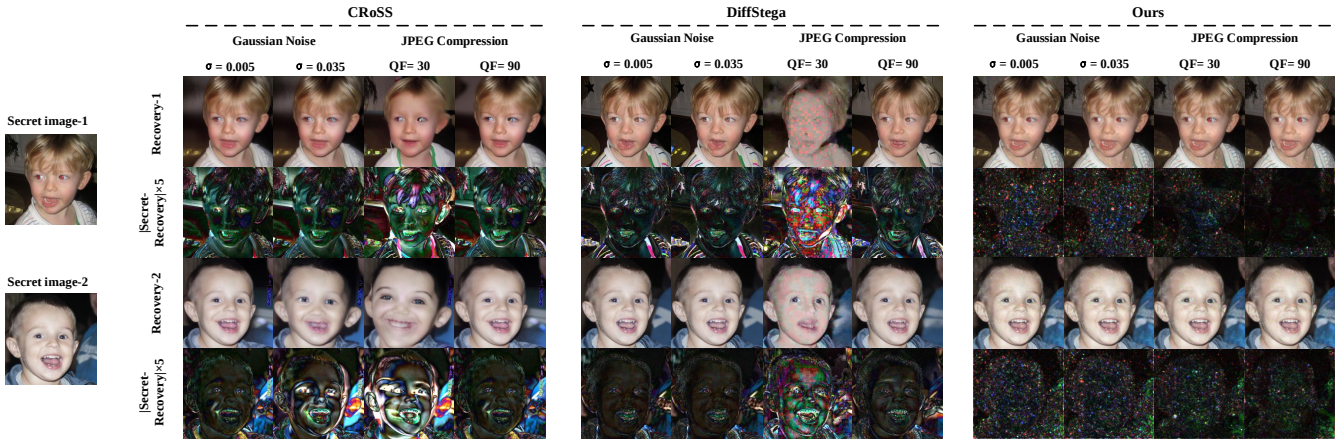


Fig. 7. Visual comparison of the recovered image under different levels of degradation.

TABLE IV
QUANTITATIVE RESULTS OF THE ABLATION STUDY ON DIFFERENT CONFIGURATIONS.

Setting	NIQE ↓	$\text{FID}_{(\text{ste,real})}$ ↓	Secret/Recovery image pairs		Detection error(%)	$\text{FID}_{(\text{ste,sec})}$ ↑
			PSNR ↑	SSIM ↑		
Baseline	4.42	49.052	22.75	0.74	0.52	71.608
Baseline + CACM	4.38	48.292	32.89	0.96	0.52	75.698
Baseline + TCM	4.40	48.885	34.65	0.97	0.47	72.328
Baseline + SADM + TCM	3.58	18.712	35.12	0.97	0.48	278.57
Ours	3.07	8.670	46.50	0.99	0.51	342.51

TABLE V
ABLATION STUDY ON THE INVERSION STRATEGY UNDER DIFFERENT DISTORTION LEVELS

Method	Clean	Gaussian Noise				JPEG compression			
		$\sigma = 0.035$	$\sigma = 0.025$	$\sigma = 0.015$	$\sigma = 0.005$	QF = 30	QF = 50	QF = 70	QF = 90
Baseline	23.65/**0.72**	20.98/0.51	21.55/0.58	22.28/**0.65**	22.77/**0.68**	20.36/0.58	21.01/0.60	22.01/**0.67**	22.54/**0.70**
Baseline + TCM	**29.24**/0.66	**28.81/0.60**	**28.68/0.61**	**28.84**/0.61	**28.85**/0.61	**29.28/0.60**	**29.27**/0.60	**28.75**/0.61	**29.06**/0.58

and 2 of the Table IV report the results of baseline and CACM-enabled configuration, respectively. Compared to the baseline, the integration of CACM substantially boosts reconstruction fidelity, which increases the PSNR from 22.75 dB to 32.89 dB and the SSIM from 0.74 to 0.96. These gains indicate that CACM establishes a stable and deterministic mapping between the secret image and the latent representation, which is a critical property for reliable information embedding.

b) Effectiveness of the SADM: The proposed SADM is used to guide stego image generation with style priors. The efficiency of SADM is demonstrated by comparing the configurations shown in Rows 3 and 4 of Table IV. With SADM, the visual quality of the stego images is significantly improved, whose $\text{FID}_{(\text{ste,real})}$ drops from 41.885 to 18.712. This improvement is attributed to the semantic priors introduced by SADM, which enhance the undetectability of the generated stego images.

c) Effectiveness of the TCM: TCM is an essential module for trajectory correction during the recovery process in the prompt-free framework. The effectiveness of this module is demonstrated by comparing Rows 1 and 3 in Table IV.The introduction of TCM substantially improves reconstruction fidelity, which improves PSNR from 22.75 dB to 34.65 dB, and SSIM increase from 0.74 to 0.97. These results indicate that TCM reduces the numerical errors accumulated and improves latent recovery accuracy.

To further compare the robustness of inversion strategies, text prompt-guided inversion and TCM are evaluated under different distortion. As shown in Table V, text-prompt-based inversion becomes less reliable under distortions. This is because text prompts cannot capture the latent deviation caused by image degradation, which limits their ability to stabilize the inversion trajectory. In constant, TCM addresses this limitation by correcting trajectories and stabilizing the reconstruction process to preserve high fidelity under noise.

TABLE VI
ABLATION ON THE NUMBER OF AFFINE COUPLING BLOCKS.

Number of CACM	NIQE↓	$\text{FID}_{(\text{ste,real})}$↓	Secret/ Recovery image pairs		Detection error(%)
			PSNR↑	SSIM ↑	
1	4.86	49.61	30.49	0.97	0.56
2	4.18	48.81	31.02	0.97	0.50
4	3.07	8.670	47.08	0.99	0.49
6	3.89	12.11	37.19	0.99	0.48

d) Ablation on the Number of CACM Blocks: We further evaluate the impact of using different numbers of CACM blocks, as reported in Table IV. The results reveal that the performance of the framework is sensitive to the choice of K. With a small value of K (e.g., $K < 4$), the mapping from the secret image to the latent space remains insufficient, and the

reconstruction fidelity tends to be limited. Conversely, with an excessively large K, the overall performance tends to decrease. This degradation stems from the progressively accumulated information loss across successive invertible transformations.The system get the best overall performance with $K = 4$. The above results suggest that an appropriate number of coupling blocks enables effective latent transformation and reduces excessive information loss, which leads to high visual quality of reconstruction.

V. Conclusion

This paper presents a prompt-free generative image steganography framework to improve the security and robustness of the steganography system. The proposed framework removes the dependence on transmitted text prompts and instead uses semantic priors extracted from reference datasets to enable user-controllable stego image generation. Within this framework, the hiding stage maps the secret image to the latent representation through CACM and uses SADM to guide controllable stego image generation with semantic priors. During recovery, the stego image is inverted to the latent representation through TCM, which uses single-step feedback to improve recovery stability under transmission distortions. Extensive experiments demonstrate that the proposed framework achieves competitive visual quality, stronger controllability, and improved robustness compared with existing methods. Future work will focus on increasing embedding capacity and enhancing recovery fidelity under complex real-world degradations.

Acknowledgments

This research was supported by the National Natural Science Foundation of China (Grant Nos. 62376238, 62372170, 12571591), the Scientific Research Fund of Hunan Provincial Education Department (Grant No.2023JGSZ032), and the Postgraduate Scientific Research Innovation Project of Hunan Province (Grant No.CX20250999).

References

[1] J. Lv, C. Fu, L. Chen, M. Liu, S. He, S. Jiang, and L. Han, "Dopsteg: Program steganography using data-oriented programming," *Science of Computer Programming*, vol. 245, p. 103311, 2025.

[2] H. Zeng, Y. Xing, S.-T. Kim, and X. Li, "Dualmodal and multifunctional steganography of three-dimensional integral imaging for the internet of medical things," *Signal Processing*, p. 110217, 2025.

[3] X. Ni, Z. Wu, L. Liu, and S. Song, "Face video steganography for privacy-protection automatic depression assessment," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025, pp. 80–86.

[4] P. Wei, S. Li, X. Zhang, G. Luo, Z. Qian, and Q. Zhou, "Generative steganography network," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 1621–1629.

[5] J. Qin, Y. Luo, X. Xiang, Y. Tan, and H. Huang, "Coverless image steganography: a survey," *IEEE Access*, vol. 7, pp. 171 372–171 394, 2019.

[6] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in Neural Information Processing Systems*, vol. 27, 2014.

[7] D. P. Kingma and P. Dhariwal, "Glow: Generative flow with invertible 1x1 convolutions," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[8] C. Yu, D. Hu, S. Zheng, W. Jiang, M. Li, and Z.-q. Zhao, "An improved steganography without embedding based on attention gan," *Peer-to-Peer Networking and Applications*, vol. 14, no. 3, pp. 1446–1457, 2021.

[9] P. Wei, G. Luo, Q. Song, X. Zhang, Z. Qian, and S. Li, "Generative steganographic flow," in *IEEE International Conference on Multimedia and Expo*, 2022, pp. 1–6.

[10] X. Liu, Z. Ma, J. Ma, J. Zhang, G. Schaefer, and H. Fang, "Image disentanglement autoencoder for steganography without embedding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 2303–2312.

[11] F. Peng, G. Chen, and M. Long, "A robust coverless steganography based on generative adversarial networks and gradient descent approximation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 9, pp. 5817–5829, 2022.

[12] J. Yu, X. Zhang, Y. Xu, and J. Zhang, "Cross: Diffusion model makes controllable, robust and secure image steganography," *Advances in Neural Information Processing Systems*, vol. 36, pp. 80 730–80 743, 2023.

[13] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10 684–10 695.

[14] Y. Xu, X. Zhang, X. Meng, C. Mou, and J. Zhang, "Diffusion-based hierarchical image steganography," in *2025 IEEE International Conference on Multimedia and Expo*, 2025, pp. 1–6.

[15] L. Chen, B. Feng, Z. Xia, W. Lu, and J. Weng, "Robust generative steganography for image hiding using concatenated mappings," *IEEE Transactions on Information Forensics and Security*, 2025.

[16] J. Jiang, Z. Wang, and X. Zhang, "Image-to-image steganography based on multimodal generative model," *Signal Processing*, vol. 238, p. 110106, 2026.

[17] Y. Yang, Z. Liu, J. Jia, Z. Gao, Y. Li, W. Sun, X. Liu, and G. Zhai, "Diffstega: towards universal training-free coverless image steganography with diffusion models," *arXiv preprint arXiv:2407.10459*, 2024.

[18] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 3836–3847.

[19] X. Xu, J. Guo, Z. Wang, G. Huang, I. Essa, and H. Shi, "Prompt-free diffusion: Taking" text" out of text-to-image diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 8682–8692.

[20] H. Ye, J. Zhang, S. Liu, X. Han, and W. Yang, "Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models," *arXiv preprint arXiv:2308.06721*, 2023.

[21] Q. Zhang and F. Huang, "Provably secure generative steganography based on adjustable orthogonal mapping," *IEEE Transactions on Dependable and Secure Computing*, 2025.

[22] R. Mokady, A. Hertz, K. Aberman, Y. Pritch, and D. Cohen-Or, "Null-text inversion for editing real images using guided diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2023, pp. 6038–6047.

[23] B. Wallace, A. Gokul, and N. Naik, "Edict: Exact diffusion inversion via coupled transformations," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2023, pp. 22 532–22 541.

[24] X. Hu, S. Li, Q. Ying, W. Peng, X. Zhang, and Z. Qian, "Establishing robust generative image steganography via popular stable diffusion," *IEEE Transactions on Information Forensics and Security*, vol. 19, pp. 8094–8108, 2024.

[25] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[26] S. Baluja, "Hiding images in plain sight: Deep steganography," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[27] S. Baluja, "Hiding images within images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 7, pp. 1685–1697, 2020.

[28] J. Zhu, R. Kaplan, J. Johnson, and L. Fei-Fei, "Hidden: Hiding data with deep networks," in *Proceedings of the European Conference on Computer Vision*, September 2018.

[29] K. A. Zhang, A. Cuesta-Infante, L. Xu, and K. Veeramachaneni, "Steganogan: High capacity image steganography with gans," *arXiv preprint arXiv:1901.03892*, 2019.

[30] S.-P. Lu, R. Wang, T. Zhong, and P. L. Rosin, "Large-capacity image steganography based on invertible neural networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 10 816–10 825.

[31] J. Jing, X. Deng, M. Xu, J. Wang, and Z. Guan, "Hinet: Deep image hiding by invertible network," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 4733–4742.
[32] Z. Guan, J. Jing, X. Deng, M. Xu, L. Jiang, Z. Zhang, and Y. Li, "Deepmih: Deep invertible network for multiple image hiding," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 1, pp. 372–390, 2022.
[33] F. Li, Y. Sheng, X. Zhang, and C. Qin, "iscmis: Spatial-channel attention based deep invertible network for multi-image steganography," *IEEE Transactions on Multimedia*, vol. 26, pp. 3137–3152, 2023.
[34] K. Zhang, F. Xiao, J. Cai, and X. Gao, "Mutual information guided invertible image hiding network," *Engineering Applications of Artificial Intelligence*, vol. 162, p. 112343, 2025.
[35] G. Xu, H.-Z. Wu, and Y.-Q. Shi, "Structural design of convolutional neural networks for steganalysis," *IEEE Signal Processing Letters*, vol. 23, no. 5, pp. 708–712, 2016.
[36] W. You, H. Zhang, and X. Zhao, "A siamese cnn for image steganalysis," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 291–306, 2020.
[37] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," *arXiv preprint arXiv:2010.02502*, 2020.
[38] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in Neural Information Processing Systems*, vol. 33, pp. 6840–6851, 2020.
[39] E. Hoogeboom, J. Heek, and T. Salimans, "Simple diffusion: End-to-end diffusion for high resolution images," in *International Conference on Machine Learning*, 2023, pp. 13 213–13 232.
[40] B. Xia, Y. Zhang, S. Wang, Y. Wang, X. Wu, Y. Tian, W. Yang, and L. Van Gool, "Diffir: Efficient diffusion model for image restoration," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 13 095–13 105.
[41] B. Fei, Z. Lyu, L. Pan, J. Zhang, W. Yang, T. Luo, B. Zhang, and B. Dai, "Generative diffusion prior for unified image restoration and enhancement," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 9935–9946.
[42] Q. Zhou, P. Wei, Z. Qian, X. Zhang, and S. Li, "Improved generative steganography based on diffusion model," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 35, no. 7, pp. 6494–6507, 2025.
[43] J. Jiang, Z. Wang, Z. Yuan, and X. Zhang, "Generative image steganography based on text-to-image multimodal generative model," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 35, no. 9, pp. 8907–8916, 2025.
[44] X. Li, L. Chen, T. Fu, Z. Fu, and Y. Gao, "Coverless image steganography based on semantic-controlled text-to-image generation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 35, no. 8, pp. 8391–8405, 2025.
[45] Q. Zhang and F. Huang, "Robust generative steganography based on image mapping," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 12, pp. 13 543–13 555, 2024.
[46] Z. Yang, K. Chen, K. Zeng, W. Zhang, and N. Yu, "Provably secure robust image steganography," *IEEE Transactions on Multimedia*, vol. 26, pp. 5040–5053, 2023.
[47] C. Lu, Y. Zhou, F. Bao, J. Chen, C. Li, and J. Zhu, "Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps," *Advances in Neural Information Processing Systems*, vol. 35, pp. 5775–5787, 2022.
[48] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2019.
[49] F. Yu, A. Seff, Y. Zhang, S. Song, T. Funkhouser, and J. Xiao, "Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop," *arXiv preprint arXiv:1506.03365*, 2015.
[50] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," *Advances in Neural Information Processing Systems*, vol. 30, 2017.
[51] L. Zhang, L. Zhang, and A. C. Bovik, "A feature-enriched completely blind image quality evaluator," *IEEE Transactions on Image Processing*, vol. 24, no. 8, pp. 2579–2591, 2015.
[52] B. Boehm, "Stegexpose-a tool for detecting lsb steganography," *arXiv preprint arXiv:1410.6656*, 2014.

Jingwen Cai received the M.E. degree in computer technology in 2022 from the Guilin University of Electronic Technology, Guilin, China, where she is currently working toward the Ph.D. degree in computer technology with the School of Computer Science, Xiangtan University, Xiangtan, China. Her research interests include information hiding, steganography, covert communication, and multimedia security.

Fen Xiao received the B.E. and Ph.D. degrees from Xiangtan University, Xiangtan, China, in 2002 and 2008, respectively. She was a Visiting Scholar at the State Key Laboratory of Pacific Northwest Pacific, Richland, WA, USA. She is currently a Professor with the School of Computer Science, School of Cyberspace Security, Xiangtan University. Her current research interests include visual saliency detection, remote sensing, image analysis, and image description.

Shuhua Deng received the B.S. degree in computer science and the Ph.D. degree in computational mathematics from Xiangtan University, Hunan, China, in 2013 and 2018, respectively. He is currently an Associate Professor at the School of Computer Science, Xiangtan University, China. His current research interests include software-defined networks, network security, and machine learning.

Xieping Gao received the B.S. and M.S. degrees from Xiangtan University, Xiangtan, China, in 1985 and 1988, respectively, and the Ph.D. degree from Hunan University, Changsha, China, in 2003. He was a Visiting Scholar with the National Key Laboratory of Intelligent Technology and Systems, Tsinghua University, Beijing, China, from 1995 to 1996, and the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore, from 2002 to 2003. He is currently a Professor with the Hunan Provincial Key Laboratory of Intelligent Computing and Language Information Processing, Hunan Normal University, Changsha. His research interests include the areas of neural networks, evolution computation, and hyperspectral image processing. Dr. Gao is a regular reviewer for several journals and he has been a member of the technical committees of several scientific conferences.