DiffStega: Towards Universal Training-Free Coverless Image Steganography with Diffusion Models

Yiwei Yang[1] , **Zheyuan Liu**[1] , **Jun Jia**[1*] , **Zhongpai Gao**[2] , **Yunhao Li**[1] , **Wei Sun**[1] ,
Xiaohong Liu[1] and **Guangtao Zhai**[1*]

[1]Shanghai Jiao Tong University
[2]United Imaging Intelligence
{evyoung, lzy233, jiajun0302}@sjtu.edu.cn, gaozhongpai@gmail.com,
{lyhsjtu, sunguwei, xiaohongliu, zhaiguangtao}@sjtu.edu.cn

Abstract

Traditional image steganography focuses on concealing one image within another, aiming to avoid steganalysis by unauthorized entities. Coverless image steganography (CIS) enhances imperceptibility by not using any cover image. Recent works have utilized text prompts as keys in CIS through diffusion models. However, this approach faces three challenges: invalidated when private prompt is guessed, crafting public prompts for semantic diversity, and the risk of prompt leakage during frequent transmission. To address these issues, we propose DiffStega, an innovative training-free diffusion-based CIS strategy for universal application. DiffStega uses a password-dependent reference image as an image prompt alongside the text, ensuring that only authorized parties can retrieve the hidden information. Furthermore, we develop Noise Flip technique to further secure the steganography against unauthorized decryption. To comprehensively assess our method across general CIS tasks, we create a dataset comprising various image steganography instances. Experiments indicate substantial improvements in our method over existing ones, particularly in aspects of versatility, password sensitivity, and recovery quality. Codes are available at https://github.com/evtricks/DiffStega.
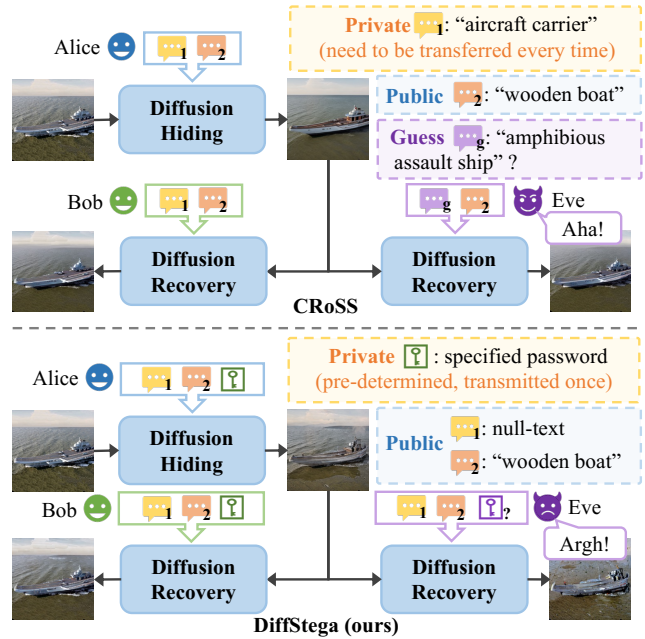
Figure 1: In this scenario, Alice represents a military organization that Eve regards as a target for espionage. Instead of using text prompt 1 as private key for diffusion-based CIS like previous work (CRoSS), DiffStega uses pre-determined password as private key, and null-text as prompt 1. DiffStega has no risk of text prompt leakage, and can encrypt the original image with arbitrary prompts.

1 Introduction

The internet revolution has significantly facilitated communication, yet posing challenges in securing messages transmitted over the Internet [Mandal *et al.*, 2022]. Steganography is a popular technique to hide information into the container in an imperceptible manner [Yu *et al.*, 2024]. As a result, only trusted receivers are able to recover the information from the steganographic content. As a subset of this field, image steganography specializes in disguising the secret message as an image, offering a high degree of security and privacy. It has applications in diverse areas, including

image compression [Jafari *et al.*, 2013], secure communication [Duluta *et al.*, 2017], and cloud computing [AlKhamese *et al.*, 2019]. Traditional cover-based image steganography schemes hide the secret message in a cover image by altering its statistical properties [Meng *et al.*, 2023]. Once the cover image is leaked, the hidden message can be easily detected by steganalysis [Karampidis *et al.*, 2018]. In contrast, coverless image steganography (CIS) [Zhou *et al.*, 2015] aims to encode or map the secret message into a stego image rather than modifying a cover image. Thus, it has greater imperceptibility compared to cover-based techniques.

CRoSS [Yu *et al.*, 2024] has demonstrated the potential of using diffusion models for CIS. This approach consists of two

---

stages: hiding and recovery. In the hiding stage, text prompt 1, serving as a private key, guides the Denoising Diffusion Implicit Model (DDIM) [Song *et al.*, 2021] inversion process to convert an original image to an initial noisy latent code. Subsequently, text prompt 2, acting as a public key, directs the DDIM denoising process to produce a stego image from this latent code. However, the reliance on text prompts as private keys is vulnerable. As illustrated in Figure 1, if Eve knows about what Alice might encrypt through a background investigation, he could easily infer the correct private prompt. In practice, even a similar private prompt may be sufficient to breach the encryption. Furthermore, since the private key is contingent on the content of the original image, it must be transmitted for each use, creating a potential risk of leakage.

We posit that a general CIS task should offer substantial flexibility, meaning that the images intended for concealment could be encrypted into various contents or styles. To this end, we categorize the general CIS task into two distinct types: content-based steganography and style-based steganography. Content-based steganography alters the primary object categories within an image while preserving its inherent structure. In contrast, style-based steganography achieves near-total imperceptibility by translating images into artistic renditions. This approach further obscures the detection and identification of hidden objects. Meanwhile, diffusion-based CIS should be able to produce stego images with arbitrary target text, especially general or similar texts.

To achieve general CIS, we introduce a novel diffusion-based pipeline. Akin to traditional cryptography, our method hinges on a specific numerical password, predetermined and known only to trusted parties. Since our password is independent of image contents, the password only needs to be transmitted once. Utilizing public text prompts and the prearranged password, trusted parties can generate a specific image to guide both the hiding stage and the recovery stage. To enhance the critical role of the password, we propose the Noise Flip technique, which encrypts the noisy latent code at the deepest step of the diffusion model while minimally altering its mean and variance. This innovation substantially complicates the task for potential attackers while simultaneously ensuring superior generation quality. To objectively assess our method's efficacy in general CIS tasks, we have curated a specialized dataset. This dataset comprises images alongside their target content, style, and analogous prompts. Our experimental results indicate significant advancements over existing methods, particularly in aspects of versatility, password sensitivity, and the effectiveness of recovery.

Our contributions can be summarized as follows:

- We overcome the limitations of current diffusion-based CIS models by incorporating pre-determined passwords, leveraging existing models without further fine-tuning.

- We introduce a novel CIS pipeline, which uniquely employs pre-determined passwords instead of text prompts alone to ensure security. This eliminates the need to transfer the private key each time the secret image is changed and is adaptable to any text prompt.

- We create a specialized dataset tailored for general CIS tasks. Comprehensive experiments show the superior performance of ours in comparison to existing methods.

2 Related Work

2.1 Image Steganography

Cover-based Methods. Traditional cover-based methods are divided into spatial domain-based methods that directly modify the pixels of the cover image [Yang *et al.*, 2008; Pevnỳ *et al.*, 2010], and transform domain-based methods that embed information into frequency domains [Chen, 2007; McKeon, 2007; Valandar *et al.*, 2017]. Deep learning-based methods use neural networks to hide information in cover images[Baluja, 2017; Jia *et al.*, 2022b; Jia *et al.*, 2020; Jia *et al.*, 2022a]. SteganoGAN [Zhang *et al.*, 2019] uses generative adversarial networks to optimize image quality. HiNet [Jing *et al.*, 2021] introduces invertible neural networks into steganography tasks. Changing the cover image leads to the usual drawbacks of cover-based methods with steganalysis [Karampidis *et al.*, 2018]. Related topics include Image Forgery Detection [Guo *et al.*, 2023] and invisible Watermarking [Fu *et al.*, 2024] have also been studied extensively [Liu *et al.*, 2022; Wu *et al.*, 2023; Gao *et al.*, 2015].

Coverless Methods. Coverless methods hide information without cover images. Early CIS refers to mapping the secret message into another image without modification [Meng *et al.*, 2023]. Therefore, it fundamentally resists steganalysis and significantly improves security. Recently, CRoSS [Yu *et al.*, 2024] uses diffusion models and the inversion technique to achieve CIS. It uses prompts as private and public keys to translate the secret image into another, which is more controllable and robust with high quality. Diffusion-based CIS is promising with the powerful generation ability and rapid development of diffusion models.

2.2 Diffusion Models

Diffusion models are the newly emerged generative models, which synthesize images via progressively denoising Gaussian noise. Among them, DALLE-3 [Betker *et al.*, 2023], Imagen [Saharia *et al.*, 2022] and Stable Diffusion [Rombach *et al.*, 2022] have achieved the state-of-the-art results on many computer vision tasks.

Controlled Generation. Recent advances provide additional control for diffusion process. ControlNet [Zhang *et al.*, 2023] and T2I-Adapter [Mou *et al.*, 2023] leverage adapters to add conditional controls of semantic segmentation, Open-Pose bone image, and Canny edge detection, etc. In addition, BLIP-Diffusion [Li *et al.*, 2023] and IP-Adapter [Ye *et al.*, 2023] handle reference images as additional prompts to control the generation with image embeddings.

Diffusion Inversion. Diffusion inversion deterministically noise the image to the intermediate latent code along the path that the denoising would follow with the same conditioning [Wallace *et al.*, 2023], widely used for image editing tasks. Among them, DDIM [Song *et al.*, 2021] relies on local linearization assumptions, suffering errors with the actual process. To reduce inversion errors, Null-text Inversion [Mokady *et al.*, 2023] introduces additional fine-tuning on optimizing the null embedding, and EDICT [Wallace *et al.*, 2023] uses coupled transformations without fine-tuning.
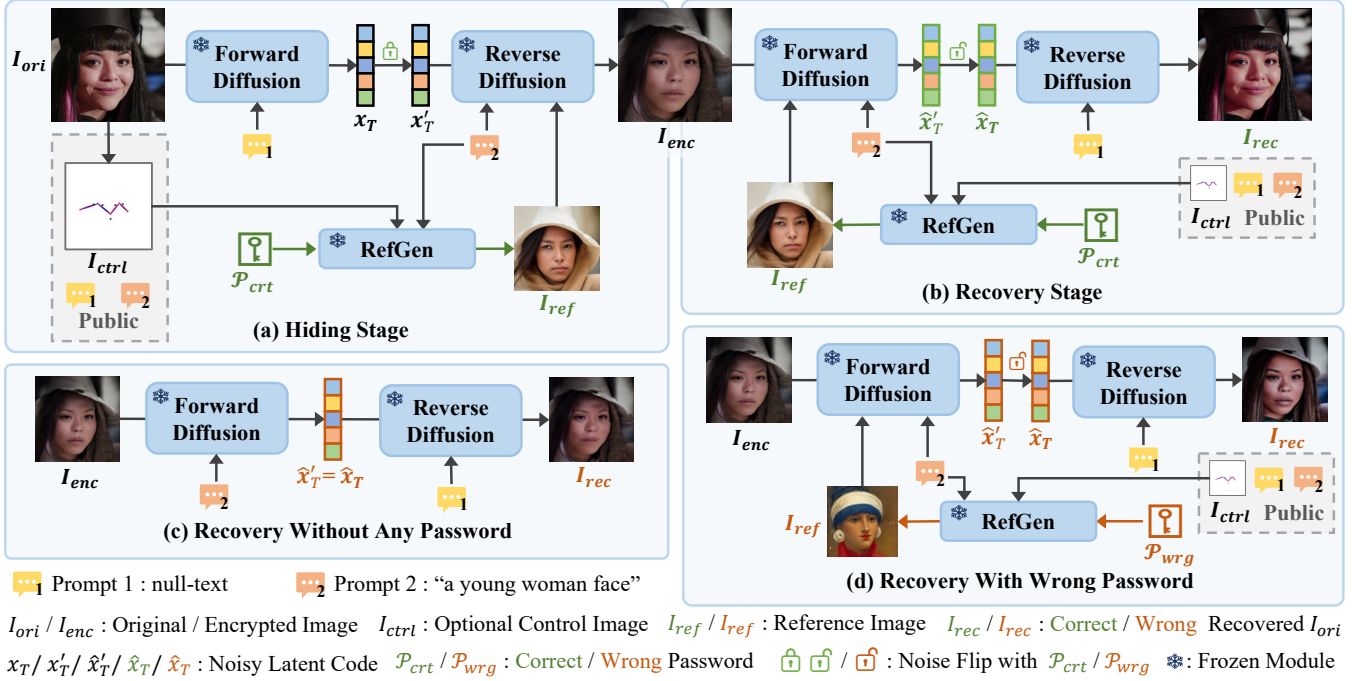
Figure 2: The pipeline of the DiffStega. (a) We use text prompts and I_{ref} generated by RefGen with password $\mathcal{P}_{crt}$ to guide the diffusion process of hiding stage. The text prompts and the optional control image I_{ctrl} (e.g. OpenPose bone image) is set public. (b) With public resources, authenticated parties could reproduce the same I_{ref} with $\mathcal{P}_{crt}$ to guide the diffusion process of recovery stage with text prompts. (c) It illustrates the scenario where attackers attempt to directly recover the image without any password. (c) Wrong password $\mathcal{P}_{wrg}$ would result in wrong I_{ref}, which is distinct from the correct reference image, resulting in misleading the recovery diffusion process. **Green** / **Red** denotes the **correct** / **wrong** decrypted items. For brevity, we omit the encoder and decoder of VAE for latent diffusion models.

3 Method

3.1 Overview

Our method consists of two stages: hiding and recovery. In hiding stage, shown in Figure 2a, we first generate a reference image I_{ref} with given password $\mathcal{P}_{crt}$, prompt 2 and optional control image I_{ctrl} (e.g. OpenPose bone image). Then we translate original secret image I_{ori} into encrypted image I_{enc} with the guidance of prompt 1, prompt 2, and I_{ref}. We set prompt 1, prompt 2 and I_{ctrl} as public resources. In recovery stage, shown in Figure 2b, the secret image is recovered from I_{enc} with correct password $\mathcal{P}_{crt}$ and public resources in a reverse procedure of hiding stage. If a malicious attacker attempts to directly recover the image without any passwords, or with a wrong password, a distinct recovery would be obtained, shown in Figure 2c and in Figure 2d respectively. To control the diffusion process in the two stages, we propose Reference Generator (RefGen) to generate I_{ref} as image prompts, and Guidance Injection module as detailed in Section 3.2 and Section 3.3 respectively.

Hiding Stage

In hiding stage, DiffStega mainly uses two encryption methods. One is RefGen to generate I_{ref} that guides the reverse diffusion process. Another is Noise Flip to ensure that the original image cannot be recovered with wrong passwords. We introduce the hiding stage as follows.

Preparation. We first input private password $\mathcal{P}_{crt}$, public text prompt 2, and optional control image I_{ctrl} (e.g. Open-Pose bone image, semantic image) into RefGen, to generate a reference image I_{ref}. Note that we only set text prompts and I_{ctrl} public. $\mathcal{P}_{crt}$ and I_{ref} will not be published publicly.

Forward Diffusion. Since we use lattent diffusions [Rombach *et al.*, 2022], I_{ori} is first encoded by VAE into latent codes x_0. Then we use diffusion inversion to convert x_0 to initial noisy latent code x_T, where T is DDIM [Song *et al.*, 2021] steps. Since we choose EDICT [Wallace *et al.*, 2023] as noising inversion method, our method is training-free. This process requires only null-text prompt 1 as guidance.

Noise Flip. To improve the influence of $\mathcal{P}_{crt}$, we use it as random seed to deterministically flipping partial positions in x_T, resulting in a slightly different x'_T. We denote this procedure as Noise Flip. The more positions flipped, the harder for attackers to recover the original image without the correct password. This Noise Flip process follows the formula:

$$x'_T = x_T \odot (1 - M_{rand}(\mathcal{P}_{crt}, \eta)) - x_T \odot M_{rand}(\mathcal{P}_{crt}, \eta)$$

where $\odot$ denotes element-wise multiplication, and M_{rand} is a random binary mask generated with deterministic random seed according to $\mathcal{P}_{crt}$. The proportion of 1 in M_{rand} is controlled through the coefficient $\eta \sim [0, 1]$. Small η is sufficient for significant modification of x_T. After this procedure, the noisy latent code of step T has strong dependence on $\mathcal{P}_{crt}$, while barely altering its mean and variance.

Reverse Diffusion. In this process, we use Guidance Injection to treat I_{ref} as image prompt to guide noise prediction together with text prompt 2. Then, the reverse diffusion process converts x'_T into x'_0 with EDICT [Wallace *et al.*, 2023] denoising. Finally, x'_0 is decoded into encrypted image I_{enc} with VAE. With the aid of I_{ref} and Noise Flip, I_{enc} would be distinct from I_{ori}. Only authenticated users with $\mathcal{P}_{crt}$ could recover the original image. For brevity, we use one latent code to represent coupled latent pairs used in EDICT.

Recovery Stage

As illustrated in Figure 2, the recovery stage is the reverse process of the hiding stage in symmetry. With private $\mathcal{P}_{crt}$, authenticated parties could easily recover the original image, where attackers is hard to predict the correct numerical password. In this stage, we have encrypted image I_{enc} and public resources consisting of prompt 1, prompt 2, and I_{ctrl}. The possible scenarios are that we have the correct password $\mathcal{P}_{crt}$, or the wrong password $\mathcal{P}_{wrg}$, or do not use any password at all. We will discuss the details of these scenarios as follows.

Recovery with Correct Password. With $\mathcal{P}_{crt}$, public prompt 2 and I_{ctrl}, we first use RefGen to reproduce the correct I_{ref}, identical to that used in hiding stage. Then we use I_{ref} and prompt 2 to guide noise prediction together via Guidance Injection. VAE encodes I_{enc} into $\hat{x}'_0$. The forward diffusion process convert $\hat{x}'_0$ into $\hat{x}'_T$ with EDICT noising inversion. Following the same procedure of Noise Flip in hiding stage, we could just flip back the previous reversed positions in $\hat{x}'_T$, resulting in $\hat{x}_T \approx x_T$. Then we use reverse denoising process via EDICT to convert $\hat{x}_T$ to $\hat{x}_0$. After VAE decoder, we finally get correctly recovered image I_{rec}.

Recovery without Any Password. For malicious attackers, the simplest way they would try is directly conducting recovery with only public text prompts, as shown in Figure 2c. It removes the guidance of I_{ref} and Noise Flip from the correct recovery procedure. However, without the guidance of I_{ref}, there is a gap between $\hat{x}'_T$ and x'_T. Moreover, since flipped positions in $\hat{x}'_T$ remain, $\hat{x}_0$ is distinct from x_0, leading to awfully wrong recovery.

Recovery with Wrong Password. As shown in Figure 2d, wrong password $\mathcal{P}_{wrg}$ would produce wrong I_{ref}, which is distinct from that used in hiding stage. It results in wrong $\hat{x}'_T$ from $\hat{x}'_0$ in the forward diffusion process. After Noise Flip with a wrong random seed, $\hat{x}_T$ would be further far different from the original x_T. Since prompt 1 is null-text, it is nearly impossible for the final I_{rec} to present the content with the same semantics as the original image.

3.2 Reference Generator

In this section, we will introduce the inside procedure of RefGen. It aims to produce deterministic images according to the password and I_{ctrl}. As shown in Figure 3, RefGen first generates a deterministic Gaussian noise with the random seed according to given password $\mathcal{P}$. The Gaussian noise serves as the initial noisy latent code of generation. We use ControlNet [Zhang *et al.*, 2023] to add additional control with I_{ctrl} (e.g. OpenPose bone image) to diffusion models. Then the reference image I_{ref} is obtained from pretrained diffusion models with the guidance of I_{ctrl} and prompt 2.
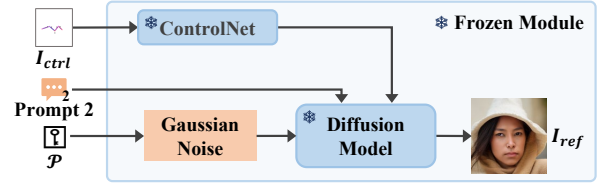

Figure 3: Details of Reference Generator. It first generates a deterministic initial Gaussian noise according to given password. I_{ref} is generated from pretrained diffusion models with the guidance of I_{ctrl} with ControlNet and prompt 2.

3.3 Guidance Injection

This procedure aims to inject image features of I_{ref} to the diffusion models. In order to use images to guide the diffusion process in the same way as the text, we adapt IP-Adapter [Ye *et al.*, 2023] to inject image features into U-Net on the low dimensional latents. Furthermore, we notice that if we use optional control image I_{ctrl} to force I_{ref} have the similar structure with I_{ori}, the encrypted image I_{enc} usually have the similar structure with the aid of Guidance Injection.

3.4 Security Guarantee

Previous work uses prompt 1 as private key and prompt 2 as public key [Yu *et al.*, 2024]. They believe that attackers can only guess the prompt 1 by exhaustive method, being unable to judge which is the true I_{ori} from candidate recovered images. However, we believe that potential attackers usually conduct many investigations on the target. For example, if the target is a military organization, the private key is likely to be related to weapons and equipment. Moreover, frequently transferring private prompt 1 faces risks. In order to make the steganography and recovery only depend on the specified password, DiffStega encrypts the original image with the guidance of images generated with passwords.

In the whole pipeline, $\mathcal{P}_{crt}$ only needs to be transmitted once and is independent of the original image contents. For trusted parties, I_{ref} can be losslessly reproduced through $\mathcal{P}_{crt}$ and guides recovery stage to restore the original image. Null-text prompt 1 makes it impossible for an attacker to speculate on the original content of the image, which makes the wrong decryption less likely to be the original content. And our pipeline can resist steganalysis and many kinds of distortion because of the inherent advantages of diffusion models, which has been explained in [Yu *et al.*, 2024].

4 Experiment

4.1 Implementation Details

For all experiments in this paper, we use the pre-trained SD v1.5[1] in the forward diffuion of hiding stage and the reverse diffusion of the recovery stage. And we use PicX_real[2] in RefGen. We use SD v1.5 for experiments on style prompts in the reverse diffuion of hiding stage and the forward diffusion of the recovery stage, but PicX_real for other experiments. We set $T = 50$, and the mixing coefficient of EDICT is 0.93. We

---

[1] https://huggingface.co/runwayml/stable-diffusion-v1-5
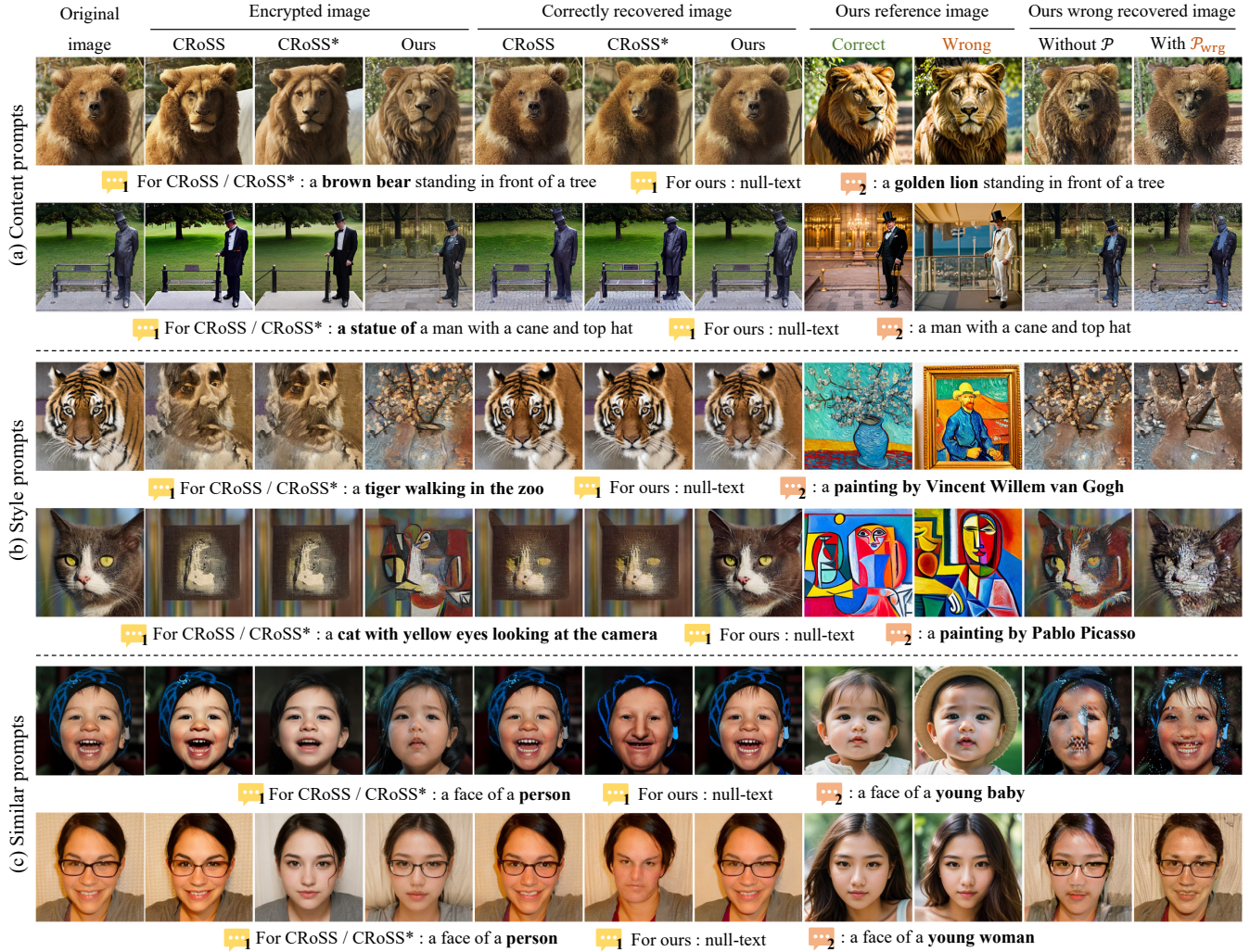[2] https://huggingface.co/GraydientPlatformAPI/picx-real

Figure 4: The visual comparison of DiffStega and CRoSS families on UniStega dataset with different prompts. For DiffStega (ours), we categorize the recovery it into three possible scenarios. Besides recovering with correct private key, malicious attackers may attempt recovery without any password, ignoring the use of Guidance Injection and Noise Flip, or recovery with the wrong password $\mathcal{P}_{wrg}$. Note that although we display prompt 1 bellow images, DiffStega still uses null-text as prompt 1 instead. CRoSS* uses two diffusion models consistent with DiffStega rather than a single model in CRoSS.

use IP-Adapter-plus [Ye *et al.*, 2023] in Guidance Injection, and its weight factor is 1. The guidance scale of diffusion models is 1. $\eta = 0.05$ in Noise Flip. The diffusion process for ours is executed over steps $[0, \xi T]$. We set $\xi = 0.7$ for experiments on style prompts and $\xi = 0.6$ for other prompts. DiffStega uses ControlNet [Zhang *et al.*, 2023] with additional control images in RefGen for all experiments except for style prompts. For additional control images, DiffStega uses semantic segmentations from OneFormer [Jain *et al.*, 2023] as control images in general, but OpenPose bone images for face images. Experiments on style prompts use no control image. All experiments are conducted on single Nvidia RTX 3090 GPU, requiring no additional training or fine-tuning.

Since there is only one diffusion-based CIS model, we mainly compare DiffStega with CRoSS [Yu *et al.*, 2024]. For fair comparison, CRoSS* is the modefied version which uses the same two models used in DiffStega, rather than only SD v1.5 in the original CRoSS. And all models use EDICT inversion. Also, CRoSS families use ControlNet with the same control images in the reverse diffuion of hiding stage and the forward diffusion of the recovery stage when DiffStega uses them. For sufficient inversion steps, ξ is set to 1 for CRoSS families. Meaningful description texts are only used as prompt 1, i.e. private keys, for CRoSS families. On the contrary, DiffStega uses null-text as prompt 1. All models use target texts as prompt 2. We also conduct security and robustness experiments compared with HiNet [Jing *et al.*, 2021].

4.2 Datasets and Metrics

Datasets. Since our method supports universal CIS under arbitrary text prompts and can be applied to versatile images, we build a dataset UniStega, consisting of 3 subsets with a

Method	Encryption with steganography					Recovery using correct private key				Recovery without any password				Recovery using wrong password			
	PSNR↓	SSIM↓	LPIPS↑	ID Sim↓	CLIP Score↑	PSNR↑	SSIM↑	LPIPS↓	ID Sim↑	PSNR↓	SSIM↓	LPIPS↑	ID Sim↓	PSNR↓	SSIM↓	LPIPS↑	ID Sim↓
CRoSS	19.034	0.657	0.365	0.892	26.912	21.248	0.711	0.320	0.877	-	-	-	-	-	-	-	-
CRoSS*	**17.139**	0.606	0.435	0.516	28.251	19.553	0.673	0.348	0.750	-	-	-	-	-	-	-	-
DiffStega	18.611	_0.590_	_0.461_	_0.343_	**29.446**	23.290	0.769	0.266	0.893	**18.491**	**0.595**	**0.457**	_0.331_	**17.530**	**0.540**	**0.476**	**0.477**
DiffStega†	_18.529_	**0.586**	**0.462**	**0.340**	29.203	_23.365_	_0.773_	_0.261_	_0.902_	_18.840_	_0.614_	_0.429_	**0.315**	_17.739_	_0.555_	_0.457_	_0.486_
DiffStega‡	19.734	0.653	0.421	0.413	29.037	**23.920**	**0.785**	**0.252**	**0.924**	20.228	0.682	0.390	0.497	20.680	0.703	0.350	0.814

Table 1: Quantitative assessment and ablation results of encrypted images and recovered images on UniStega dataset. CRoSS* is the revised version of CRoSS for fair comparation, which uses two diffusion models consistent with DiffStega rather than a single model in CRoSS. DiffStega† is the ablation comparation when prompt 1 is not null-text but meaningful text. DiffStega‡ is the version without Noise Flip.

total of 100 images under different scenarios: (1) UniStega-Content comprises of 42 images with corresponding prompts and target content prompts. It applies to the most common CIS scenario with similar shape but different content; (2) UniStega-Style comprises of 28 images and prompt pairs. The target prompts refer to artworks in the styles of famous artists. It applies to CIS with higher difficulty to guess what the original images are; (3) UniStega-Similar comprises of 30 images and analogous prompt pairs. It applies to CIS with similar or overly general prompts. All images are from public dataset COCO [Lin *et al.*, 2014], AFHQ [Choi *et al.*, 2020], FFHQ [Karras *et al.*, 2019], CelebA-HQ [Karras *et al.*, 2018] and Internet, center cropped and resized to 512×512. We use BLIP [Li *et al.*, 2022] to generate description prompts, and Llama2 [Touvron *et al.*, 2023] to generate target content prompts with semantic modifications, or artificial adjustment to generate other prompts, following CRoSS [Yu *et al.*, 2024].

Metrics. We use PSNR, SSIM [Wang *et al.*, 2004], LPIPS [Zhang *et al.*, 2018], and ID Cosine Similarity from Facenet [Schroff *et al.*, 2015] for face images to assess the quality of hiding and recovery. We use CLIP Score [Radford *et al.*, 2021] to evaluate whether the encrypted image matches the target text promt or not. We use NIQE [Mittal *et al.*, 2012] to blindly assess the naturalness of encrypted images.

4.3 Experimental Results

Qualitative Results. Figure 4 shows the visual comparision of DiffStega and CRoSS families on UniStega dataset. We categorize it into three possible scenarios. The first is that authenticated users recover the original images with correct private key. Although DiffStega makes more modifications to the original images, it performs more accurate recovery. The others are that attackers attempt to recover without any password, i.e. not using RefGen and Noise Flip, or recovering with a wrong password. Attackers even hardly recover the correct category of original objects. For style prompts, the encrypted image of CRoSS loses too many details of original images, resulting in a significant difference between recovered images and the original. For similar prompts, using a single diffusion model, the encrypted images of CRoSS are almost the same as the original. CROSS* fails in recovery because of inversion errors and general prompt 1. However, with the guidance of reference image and fewer diffusion steps, DiffStega could still achieves satisfactory performance.

Quantitative Results. For encryption with steganography shown in Table 1. The smaller the similarity to the original
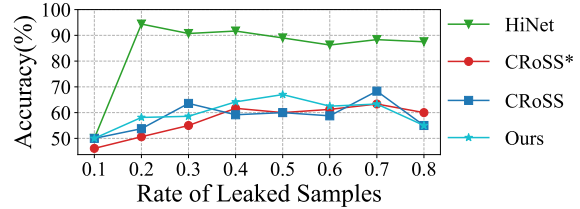

Figure 5: Deep steganalysis accuracy by XuNet. As the rate of leaked samples increases, the closer the curve approximates 50%, the more secure the method is. Diffusion-based methods are similar.

(a) Different controls. (b) Different checkpoints.

Figure 6: The encrypted and recovered images of DiffStega with different controls (bottom right) and checkpoints of diffusion models.

image, the better the hiding performance. DiffStega makes the encrypted images distinct from the original ones. Due to the lack of more effective guidance, CRoSS has the worst encryption performance. Meanwhile, the CLIP Scores between encrypted images and prompt 2 shows that DiffStega has better consistency with target prompts. Moreover, DiffStega has stronger identity hiding capability of face images than CRoSS families. For recovery of different scenarios shown in Table 1. The greater the similarity to the original image, the better the recovery performance with correct private key. For scenarios of wrong recovery, the opposite is true. DiffStega shows better recovery performance than CRoSS families.

Imperceptibility and Security. Table 2 shows that ours has similar NIQE score to the original images, hardly suspected by people. For anti-analysis security, We use XuNet [Xu *et al.*, 2016] to distinguish the encrypted images to general images without steganography. As shown in Figure 5, the closer the detection accuracy approximates 50%, the more secure the method is. DiffStega shares the similar performance with CRoSS families, much better than HiNet.

(a) Null-text and meaningful prompt.

(b) Different Noise Flip scales η.

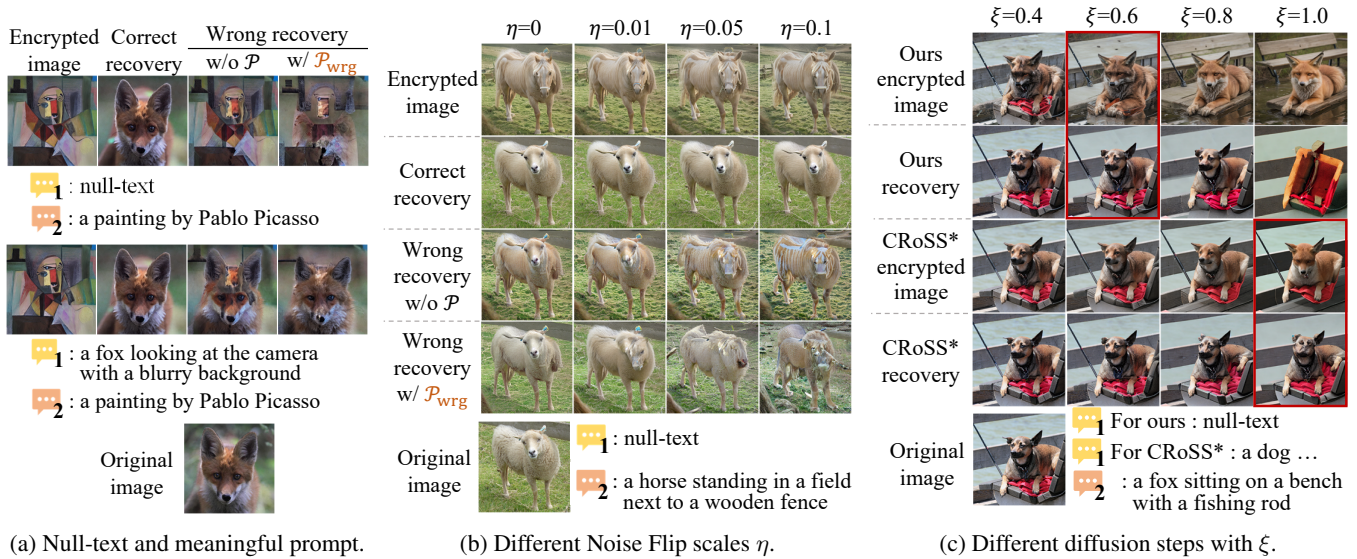(c) Different diffusion steps with ξ.

Figure 7: Ablations. (a) The comparison of encrypted and recovered images when prompt 1 is null-text and meaningful text for DiffStega. (b) Visual comparisons of DiffStega with different Noise Flip scales. The first row shows the original image and the encrypted images. (c) The encrypted and recovered images with different diffusion steps according to ξ. CRoSS* is the revised version of CRoSS for fair comparation.

	HiNet	CRoSS	CRoSS*	Ours	Original image
NIQE ↓	**3.125**	3.601	3.795	3.408	3.083

Table 2: NIQE scores that indicate the quality of encrypted images. All methods is relatively as natural as the original images.

Robustness and Controllability. Real-ESRGAN [Wang *et al.*, 2021] is used to perform nolinear image enhancement for degradations, denoted as Gaussian deblur and JPEG enhancer. As shown in Table 3, DiffStega has similar robustness to CRoSS, while HiNet suffers significant drops in PSNR. Figure 6 shows that our pipeline is flexible and applicable to different controls and checkpoints[3].

4.4 Ablation Study

Influence of Null-text Prompt. As shown in Figure 7a, meaningful prompt 1 narrows the similarity between the wrong decrypted image and the original image. Considering that publicly disclosing meaningful prompt 1 is equivalent to letting everyone know what is encrypted, but only brings minimal performance improvement as shown in Table 1, we prefer to set null-text as prompt 1 for better security.

Influence of Noise Flip. As shown in Table 1 and Table 1 for DiffStega[‡], Noise Flip increases the reliance on the private password. Without Noise Flip, DiffStega has slight improvement in the quality of recovery, but double performance drop on encryption and wrong decryption. As η grows in Figure 7b, although wrong decrypted images are getting further away from the original, the quality of the encrypted images is getting worse and worse. Therefore $\eta = 0.05$ is sufficient. The second row demonstrates that different Noise Flip scales have limited influence of the correct recovery.

[3]https://huggingface.co/stablediffusionapi/majicmix-fantasy

Method	Clean	Gaussian blur	Gaussian deblur	JPEG compression	JPEG enhancer
HiNet	46.152	10.262	10.992	10.946	10.856
CRoSS	21.248	20.080	**19.354**	20.195	19.267
CRoSS*	19.553	16.371	16.696	16.697	16.032
DiffStega	23.290	**20.849**	18.620	**21.161**	**20.154**

Table 3: PSNR(dB) results of recovered images on UniStega dataset when encrypted images suffer degradations. CRoSS* uses two diffusion models consistent with DiffStega. For Gaussian blur and deblur, the kernel size is 7 and σ is 10 . Q is 40 for JPEG degradations.

Influence of ξ. As shown in Figure 7c, DiffStega achieves distinct change and meets the target description when $\xi \geq 0.6$. However, only when ξ is close to 1 can CRoSS families make the encrypted image distinct from the original. The red box indicates the values we recommended in the experiments.

5 Limitation

Because there is inevitable error in diffusion inversion methods, DiffStega suffers the poor recoverability if ξ is too large. Meanwhile, DiffStega introduces additional processes of generating reference images, which means more computational overhead. However, the additional cost will gradually be negligible with fast sampling methods [Luo *et al.*, 2023].

6 Conclusion

This paper proposes DiffStega, cleverly designed for general coverless image steganography with diffusion models. We uses the pre-determined password as private key, and propose Noise Flip to achieve high quality steganography and nearly undistorted recovery of the original images. Extensive experiments show our superiority compared with previous methods. How to directly influence the image generation process with passwords is a promising research topic in the future.

Acknowledgments

Contribution Statement

Yiwei Yang and Zheyuan Liu contributed equally.

References

[AlKhamese *et al.*, 2019] Aya Y AlKhamese, Wafaa R Shabana, and Ibrahim M Hanafy. Data security in cloud computing using steganography: a review. In *International Conference on Innovative Trends in Computer Engineering*, pages 549–558, 2019.

[Baluja, 2017] Shumeet Baluja. Hiding images in plain sight: Deep steganography. In *Advances in Neural Information Processing Systems*, pages 2069–2079, 2017.

[Betker *et al.*, 2023] James Betker, Gabriel Goh, Li Jing, Tim Brooks, et al. Improving image generation with better captions. https://cdn.openai.com/papers/dall-e-3.pdf, 2023. Accessed: January 2, 2024.

[Chen, 2007] Wen-Yuan Chen. Color image steganography scheme using set partitioning in hierarchical trees coding, digital fourier transform and adaptive phase modulation. *Applied Mathematics and Computation*, 185(1):432–448, 2007.

[Choi *et al.*, 2020] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 8185–8194, 2020.

[Duluta *et al.*, 2017] Andrei Duluta, Stefan Mocanu, Radu Pietraru, et al. Secure communication method based on encryption and steganography. In *International Conference on Control Systems and Computer Science*, pages 453–458, 2017.

[Fu *et al.*, 2024] Kang Fu, Xiaohong Liu, Jun Jia, Zicheng Zhang, et al. Rawiw: Raw image watermarking robust to isp pipeline. *Displays*, 82:102637, 2024.

[Gao *et al.*, 2015] Zhongpai Gao, Guangtao Zhai, and Chunjia Hu. The invisible qr code. In *ACM international conference on Multimedia*, pages 1047–1050, 2015.

[Guo *et al.*, 2023] Xiao Guo, Xiaohong Liu, Zhiyuan Ren, Steven Grosz, et al. Hierarchical fine-grained image forgery detection and localization. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3155–3165, 2023.

[Jafari *et al.*, 2013] Reza Jafari, Djemel Ziou, and Mohammad Mehdi Rashidi. Increasing image compression rate using steganography. *Expert Systems with Applications*, 40(17):6918–6927, 2013.

[Jain *et al.*, 2023] Jitesh Jain, Jiachen Li, Mang Tik Chiu, Ali Hassani, et al. Oneformer: One transformer to rule universal image segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2989–2998, 2023.

[Jia *et al.*, 2020] Jun Jia, Zhongpai Gao, Kang Chen, Menghan Hu, et al. Rihoop: Robust invisible hyperlinks in offline and online photographs. *IEEE Transactions on Cybernetics*, 52(7):7094–7106, 2020.

[Jia *et al.*, 2022a] Jun Jia, Zhongpai Gao, Dandan Zhu, et al. Rivie: Robust inherent video information embedding. *IEEE Transactions on Multimedia*, 2022.

[Jia *et al.*, 2022b] Jun Jia, Zhongpai Gao, Dandan Zhu, Xiongkuo Min, Guangtao Zhai, and Xiaokang Yang. Learning invisible markers for hidden codes in offline-to-online photography. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2263–2272, 2022.

[Jing *et al.*, 2021] Junpeng Jing, Xin Deng, Mai Xu, Jianyi Wang, and Zhenyu Guan. Hinet: Deep image hiding by invertible network. In *International Conference on Computer Vision*, pages 4713–4722, 2021.

[Karampidis *et al.*, 2018] Konstantinos Karampidis, Ergina Kavallieratou, et al. A review of image steganalysis techniques for digital forensics. *Journal of Information Security and Applications*, 40:217–235, 2018.

[Karras *et al.*, 2018] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *International Conference on Learning Representations*, 2018.

[Karras *et al.*, 2019] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019.

[Li *et al.*, 2022] Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 12888–12900, 2022.

[Li *et al.*, 2023] Dongxu Li, Junnan Li, and Steven CH Hoi. Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. *ArXiv preprint*, abs/2305.14720, 2023.

[Lin *et al.*, 2014] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, et al. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pages 740–755, 2014.

[Liu *et al.*, 2022] Xiaohong Liu, Yaojie Liu, Jun Chen, and Xiaoming Liu. Pscc-net: Progressive spatio-channel correlation network for image manipulation detection and localization. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(11):7505–7517, 2022.

[Luo *et al.*, 2023] Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference. *ArXiv preprint*, abs/2310.04378, 2023.

[Mandal *et al.*, 2022] Pratap Chandra Mandal, Imon Mukherjee, Goutam Paul, and BN Chatterji. Digital image steganography: A literature survey. *Information sciences*, 2022.

[McKeon, 2007] Robert T McKeon. Strange fourier steganography in movies. In *IEEE International Conference on Electro/Information Technology*, pages 178–182, 2007.

[Meng *et al.*, 2023] Laijin Meng, Xinghao Jiang, and Tanfeng Sun. A review of coverless steganography. *Neurocomputing*, page 126945, 2023.

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

[Mokady *et al.*, 2023] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 6038–6047, 2023.

[Mou *et al.*, 2023] Chong Mou, Xintao Wang, Liangbin Xie, Jian Zhang, et al. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *ArXiv preprint*, abs/2302.08453, 2023.

[Pevnỳ *et al.*, 2010] Tomáš Pevný, Tomáš Filler, and Patrick Bas. Using high-dimensional image models to perform highly undetectable steganography. In *Information Hiding: 12th International Conference*, pages 161–177, 2010.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763, 2021.

[Rombach *et al.*, 2022] Robin Rombach, Andreas Blattmann, Dominik Lorenz, et al. High-resolution image synthesis with latent diffusion models. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 10674–10685, 2022.

[Saharia *et al.*, 2022] Chitwan Saharia, William Chan, Saurabh Saxena, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022.

[Schroff *et al.*, 2015] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 815–823, 2015.

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

[Touvron *et al.*, 2023] Hugo Touvron, Louis Martin, Kevin Stone, et al. Llama 2: Open foundation and fine-tuned chat models, 2023. *ArXiv preprint*, abs/2307.09288, 2023.

[Valandar *et al.*, 2017] Milad Yousefi Valandar, Peyman Ayubi, and Milad Jafari Barani. A new transform domain steganography based on modified logistic chaotic map for color images. *Journal of Information Security and Applications*, 34:142–151, 2017.

[Wallace *et al.*, 2023] Bram Wallace, Akash Gokul, and Nikhil Naik. Edict: Exact diffusion inversion via coupled transformations. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 22532–22541, 2023.

[Wang *et al.*, 2004] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

[Wang *et al.*, 2021] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *International Conference on Computer Vision Workshops*, pages 1905–1914, 2021.

[Wu *et al.*, 2023] Guangyang Wu, Weijie Wu, Xiaohong Liu, et al. Cheap-fake detection with llm using prompt engineering. In *International Conference on Multimedia and Expo Workshops*, pages 105–109, 2023.

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

[Yang *et al.*, 2008] Cheng-Hsing Yang, Chi-Yao Weng, Shiuh-Jeng Wang, et al. Adaptive data hiding in edge areas of images with spatial lsb domain systems. *IEEE Transactions on Information Forensics and Security*, 3(3):488–497, 2008.

[Ye *et al.*, 2023] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *ArXiv preprint*, abs/2308.06721, 2023.

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

[Zhang *et al.*, 2018] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018.

[Zhang *et al.*, 2019] Kevin Alex Zhang, Alfredo Cuesta-Infante, et al. Steganogan: High capacity image steganography with gans. *ArXiv preprint*, abs/1901.03892, 2019.

[Zhang *et al.*, 2023] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *International Conference on Computer Vision*, pages 3836–3847, 2023.

[Zhou *et al.*, 2015] Zhili Zhou, Huiyu Sun, Rohan Harit, et al. Coverless image steganography without embedding. In *Cloud Computing and Security: First International Conference*, pages 123–132, 2015.