ZeroFake: Zero-Shot Detection of Fake Images Generated and Edited by Text-to-Image Generation Models

Zeyang Sha
CISPA Helmholtz Center for Information Security
Saarbrücken, Germany
zeyang.sha@cispa.de

Yicong Tan
CISPA Helmholtz Center for Information Security
Saarbrücken, Germany
yicong.tan@cispa.de

Mingjie Li
CISPA Helmholtz Center for Information Security
Saarbrücken, Germany
mingjie.li@cispa.de

Michael Backes
CISPA Helmholtz Center for Information Security
Saarbrücken, Germany
director@cispa.de

Yang Zhang*
CISPA Helmholtz Center for Information Security
Saarbrücken, Germany
zhang@cispa.de

ABSTRACT

The text-to-image generation model has attracted significant interest from both academic and industrial communities. These models can generate the images based on the given prompt descriptions. Their potent capabilities, while beneficial, also present risks. Previous efforts relied on the approach of training binary classifiers to detect the generated fake images, which is inefficient, lacking in generalizability, and non-robust. In this paper, we propose the novel zero-shot detection method, called ZeroFake, to distinguish fake images apart from real ones by utilizing a perturbation-based DDIM inversion technique. ZeroFake is inspired by the findings that fake images are more robust than real images during the process of DDIM inversion and reconstruction. Specifically, for a given image, ZeroFake first generates noise with DDIM inversion guided by adversary prompts. Then, ZeroFake reconstructs the image from the generated noise. Subsequently, it compares the reconstructed image with the original image to determine whether it is fake or real. By exploiting the differential response of fake and real images to the adversary prompts during the inversion and reconstruction process, our model offers a more robust and efficient method to detect fake images without the extensive data and training costs. Extensive results demonstrate that the proposed ZeroFake can achieve great performance in fake image detection, fake artwork detection, and fake edited image detection. We further illustrate the robustness of the proposed ZeroFake by showcasing its resilience against potential adversary attacks. We hope that our solution can better assist the community in achieving the arrival of a more efficient and fair AGI.[1]

*Corresponding author.

[1]Our code is available at https://github.com/TrustAIRLab/ZeroFake.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

CCS '24, October 14–18, 2024, Salt Lake City, UT, USA

© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 979-8-4007-0636-3/24/10
https://doi.org/10.1145/3658644.3690297

CCS CONCEPTS

• **Security and privacy** → **Social aspects of security and privacy**.

KEYWORDS

Text-To-Image Models; Deepfake Detection; Image Editing

ACM Reference Format:
Zeyang Sha, Yicong Tan, Mingjie Li, Michael Backes, and Yang Zhang. 2024. ZeroFake: Zero-Shot Detection of Fake Images Generated and Edited by Text-to-Image Generation Models. In *Proceedings of the 2024 ACM SIGSAC Conference on Computer and Communications Security (CCS '24), October 14–18, 2024, Salt Lake City, UT, USA.* ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3658644.3690297

1 INTRODUCTION

Text-to-image generation models [34, 36, 37, 49, 51] have recently marked a significant milestone. These kinds of models can generate authentic images according to the description of the given prompts. An increasing number of platforms, like Midjourney [1], are utilizing text-to-image generation models to enable users to create or alter images according to their specific intentions. These advancements have revolutionized the digital art and creative industries, enabling artists and designers to explore new realms of creativity and innovation with unprecedented ease and flexibility. Additionally, they open up possibilities for personalized content creation, allowing individuals to bring their unique visions to life with just a prompt.

However, these powerful models can also be leveraged by the adversary to generate fake images, which can potentially be propagated to spread misinformation, incurring ethical concerns, causing identity theft, and so on. For instance, even though Dutch politician Frans Timmermans is traveling on an ordinary plane, text-to-image generation models can create the fake, misleading images that Frans Timmermans is sitting in a luxurious private jet, which may be used to influence the results of the election. Moreover, the text-to-image generation models are also used to generate the AI porn of Taylor Swift, which is widely spread. Such incidents highlight a broader trend where these fake images pose increasing threats to both social harmony and individual privacy.

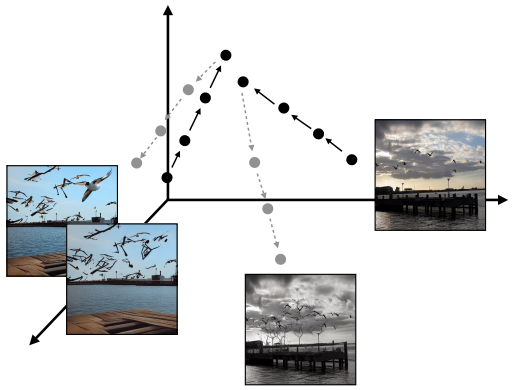


Figure 1: An illustration of ZeroFake. Fake images share more common features with the reconstructed images under the DDIM inversion than real images.

Previous works have already paid a lot of attention to how to detect these fake images. Normally, these works treat the task as a binary classification problem. Wang et al. [43] just trained a classifier to detect fake images generated by GAN. The performance is great due to the poor capability of traditional GAN models. Based on that, more works [5, 12, 29, 43, 50, 53] try to figure out how to enhance the performance of the binary classifier from the perspective of datasets [43, 50], training process [12, 53], model architecture [5, 29], and so on. Among these works, DeFake [38] is the first work to detect text-to-image-based fake images. It leverages the prompt and image together to train a hybrid detector. During the inference phase, given the images, DeFake first generates the prompts using BLIP. Then, the detector can distinguish whether the given images are fake or real by feeding both images and the generated prompts. Besides DeFake, DIRE [44] is another state-of-the-art detection method proposed against diffusion models. It also needs to train a binary classifier to detect fake images. However, these training-based methods need a large number of fake images to train a classifier, which is impractical in domains like artwork or cartoons, where the number of real images is limited. Moreover, binary classifiers are vulnerable to potential adversary example attacks. The adversary can easily create fake images with noise to mislead detectors into making wrong judgments.

1.1 Our Contribution

In this paper, to address the problems faced by current detection methods, we introduce ZeroFake, a novel zero-shot method for detecting fake images. Before ZeroFake, previous papers [6, 42, 44] on image editing showed that real images are hard to reconstruct via DDIM inversion and lead to worse performance on editing real images compared with the generated images. Tov et al. [41] proposed a possible reason called "distortion-editability tradeoff," where input prompts can influence a lot. ZeroFake leverages the unique behaviors of real and fake images during the DDIM inversion [15] and reconstruction processes. The core intuition behind ZeroFake is that fake images, due to their synthetic nature, are inherently more robust to perturbations introduced by adversary prompts during the DDIM inversion process than real images. Consequently, when these images are subjected to noise addition and subsequent noise reduction under the guidance of an adversary prompt, they tend to retain more of their original characteristics compared to real images. We show the overview of ZeroFake in Figure 1.

Methodology. Based on the above intuition, we propose ZeroFake to distinguish fake images apart from real ones. The process begins by feeding a given image into the BLIP model [19] to generate a reversed prompt. ZeroFake then constructs an adversary prompt by altering the reversed prompt with words from an adversarial text list, such as "big tree" or "small dog." Specifically, the first noun of the reversed prompt is replaced with these adversarial texts to produce multiple adversary prompts. ZeroFake evaluates these adversary prompts by calculating the cosine distance from the original reversed prompt, selecting the most divergent prompt as the final adversary prompt. This adversary prompt then guides the DDIM inversion process, which aims to perturb the original image. The perturbation is achieved by incrementally adding noise generated by a UNet that has been pre-trained on diffusion models. After the inversion, the estimated initial noise, which is supposed to contain features of the original images, is obtained. Then, guided by the embeddings from the adversary prompt again, the estimated initial noise is methodically reconstructed back to its original form, which can be considered as the reconstructed image. In the end, ZeroFake compares the similarity between the reconstructed image and the initial given image to determine the image's authenticity.

Fake Image Detection. We conduct experiments using two benchmark datasets, MSCOCO [26] and Flickr [48], selecting 500 prompt-image pairs randomly due to computational constraints. We generated fake images using five state-of-the-art models, including Stable Diffusion [36], SDXL [25], XLBase [2], DALL·E 2 [33], and GLIDE [32]. The extensive results show that our proposed ZeroFake performs exceptionally well across all scenarios. For example, when detecting fake images generated by Stable Diffusion on MSCOCO, DIRE and DeFake can achieve 0.590 and 0.853 accuracy, while our proposed zero-shot ZeroFake can achieve 0.952 accuracy. These results confirm that fake images are more robust to the perturbation of the adversary prompts than real images.

Fake Artwork Detection. In the realm of digital art, distinguishing genuine creations from AI-generated fakes is also critical, especially as text-to-image generation models increasingly win art competitions. Moreover, artworks are considered harder to detect than fake images for traditional detection methods because human-based paintings and machine-based paintings are more similar to each other than common fake photos. Therefore, instead of only focusing on the authentic prompt-image datasets, we also test the performance of ZeroFake on the collected artworks. Our ZeroFake model was tested across various artwork types, including cartoons and oil paintings, demonstrates remarkable efficacy in identifying fake artworks. The tests were conducted on a curated set of 120 image-prompt pairs derived from diverse artistic styles, ensuring broad validation. The results underscore ZeroFake's superior performance compared to conventional models like DIRE and DeFake, which falter particularly in the nuanced domain of artwork authenticity.

Image Editing Detection. Besides detecting fake images that are totally generated by text-to-image generation models, it is also

critical to detect the images that are edited by text-to-image generation models. In this paper, we regard the edited images as the fake images. Therefore, ZeroFake is assessed using a custom dataset designed to mimic realistic editing scenarios in political and social contexts. This dataset includes examples of text-driven edits that substantially edit the appearance and context of images by three state-of-the-art image editing methods, including prompt-to-prompt [44], Ledit [42], and Instruct Pixel [6]. We show that ZeroFake also performs well in image editing detection. For instance, when the real images are edited by prompt-to-prompt, one of the state-of-the-art image editing methods, ZeroFake can achieve 0.966 accuracy, while DeFake can only achieve 0.533. Our findings reveal that ZeroFake outperforms existing methods, effectively discerning between edited and original images, thereby demonstrating robustness against sophisticated image manipulation techniques.

Robustness. Moreover, the main drawback of the previous methods lies in the fact that they are all binary classifiers, which are proven vulnerable to potential image transformations. We test ZeroFake against various types of image transformations, including blurring, sharpening, gaussian noise, and adversary example attacks [14, 16], which are designed to alter image properties to mislead detection algorithms subtly. Our results confirm that ZeroFake maintains high accuracy and reliability in detecting fake images, artworks, and edited images even when subjected to these sophisticated attacks. For instance, even if the added noise is 0.01, which is large enough to be noticed by human eyes, ZeroFake can still achieve 0.891 accuracy. This demonstrates that ZeroFake not only outperforms traditional detection methods in standard scenarios but also holds strong potential in adversary environments, ensuring its practical applicability in real-world settings where robustness against evasion techniques is paramount.

Implications. In conclusion, we present a novel approach to addressing the challenges posed by fake images generated by text-to-image generation models. The performance of our proposed ZeroFake method in detecting fake images and accurately attributing them to their originating models is promising. These findings indicate that our approach could be crucial in counteracting the risks associated with these fake images. To advance research in this domain, we plan to make our source code publicly available, thereby supporting the broader academic community in exploring and enhancing the efficacy of fake image detection.

2 PRELIMINARIES

2.1 Text-to-Image Generation Models

Text-to-image generation models try to reverse the diffusion process starting from a random noise vector x_t to an output image x_0 through a denoising process based on the textual prompt $\mathcal{P}$, which is related to the given prompt representing users' requirements. During generation progress, the image is gradually denoised with the predicted noise via noise estimator ϵ_θ in the diffusion models, which is trained based on the following objective to predict the artificial noise in different steps in Equation 1.

$$\min_\theta E_{x_0,\epsilon\sim\mathcal{N}(0,I),t\sim\text{Uniform}(1,T)}\|\epsilon-\epsilon_\theta(x_t,t,\text{emb})\|^2 \quad (1)$$

where $\text{emb}=\psi(P)$ here denotes the embedding of the text condition and x_t is a noised sample generated by adding t stemp noise to the sampled images x_0. After training, the diffusion model can generate images x_0 by de-noising a random sampled x_T with its noise predictor ϵ_θ. Besides operating directly on the pixel space, the diffusion model can also be applied to the latent space. In this scenario, z_0 is a latent embedding encoded from an image encoder E with $z_0=E(x_0)$, where x_0 is a sample of the real image. After obtaining generated latent code z_0 via the diffusion model, users can also convert to a real image with an image decoder E with $x_0=D(z_0)$, where x_0 is a sample of real images. Due to the efficiency and flexibility of this kind of diffusion model, most text-to-image models like Stable-Diffusion use this kind of diffusion model, which is also called the Latent Diffusion Model. Thus, we mainly consider this kind of diffusion model in our paper.

A primary obstacle in text-guided generation lies in enhancing the influence of the provided text. Addressing this, Song et al. [39] introduced a novel approach known as classifier-free guidance. This technique involves unconditional prediction, which is subsequently combined with conditioned prediction to amplify its impact. Formally, let $\varnothing=\psi(\text{“”})$ represent the embedding of an empty text and denote w as the guidance scale parameter. Then, the classifier-free guidance prediction is formulated in Equation 2.

$$\tilde{\epsilon}_\theta(z_t,t,\text{emb},\varnothing)=w\cdot\epsilon_\theta(z_t,t,\text{emb})+(1-w)\ \cdot\epsilon_\theta(z_t,t,\varnothing) \quad (2)$$

where w is a constant scalar representing the strength of the classifier-free guidance. DDIM process is the opposite direction of the generation process, which aims to add noise to the original images. The process can be formulated in Equation 3.

$$z_t=\sqrt{\frac{\alpha_t}{\alpha_{t-1}}}z_{t-1}+\sqrt{\alpha_t}\left(\sqrt{\frac{1}{\alpha_t}-1}-\sqrt{\frac{1}{\alpha_{t-1}}-1}\right)\cdot\epsilon_\theta(z_t,t,\text{emb}), \quad (3)$$

However, the challenge is that $\epsilon_\theta(z_t,t,\text{emb})$ is impossible to obtain. We will explain it later in Section 4.

2.2 Fake Image Detection

Several fake image detection methods have been proposed previously for text-to-image generation models. As we have mentioned in the introduction phase, all of the previous methods are developed based on training a binary classifier. To the best of our knowledge, there are two state-of-the-art detection methods that can be regarded as the baseline in our paper.

- **DeFake** [38] Instead of just training a classifier on the images, DeFake was proposed to train the classifier on the combination of images and the prompts. Note that the prompts are generated by the BLIP model, as mentioned in the DeFake paper. DeFake argues that the involvement of the prompt can help the classifier better identify the fake images since fake images are always closer to the prompt than real images. However, it takes too much computational resources to train the classifier. We also show later that DeFake has poor performance in domains like artwork and image editing.
- **DIRE** [44] DIRE is another fake image detection method that also needs to train a binary classifier. DIRE first adds the noise to the given images and then reconstructs the

images from the noise. Then, the reconstructed images, together with the original images, are fed into the classifier for detection. However, the prompt will not be considered in the whole detection process. In this paper, we show that the proposed DIRE cannot work well in the text-to-image generation models. Our ZeroFake approach utilizes prompt guidance during the DDIM reversion process, demonstrating significantly improved performance over DIRE without requiring any training.

3 THREAT MODEL

To better demonstrate how the proposed ZeroFake works, we introduce the threat model of our proposed detection process in this section.

3.1 Detector's Goals

- **Distinguishing Fake Images Apart From Real Images.** The primary goal of our proposed ZeroFake is to distinguish fake images generated by text-to-image generation models apart from real images. Note that in this paper, we refer to the images that are generated or edited by text-to-image generation models as fake images. Therefore, the detector is supposed to be able to recognize not only generated images but also modified images. Moreover, instead of detecting authentic images, our detector should also be able to distinguish fake artworks from real artworks.
- **Agnostic to Models and Datasets.** The rapid development of text-to-image generation models highlights the challenge of developing a comprehensive detector due to the impracticality of including all models and the high resource consumption involved. Thus, it's vital to determine if a detector, based on a few models, can generalize across unknown models. Additionally, without knowing the exact prompts used for generating fake images, our detector must also effectively identify fakes from diverse prompt-image datasets.
- **Low/Zero Training Cost.** Nowadays, most of the current fake image detection frameworks rely on a large amount of training data to achieve great performance. However, as the text-to-image generation model continues to update, the detector also needs to continuously update training data, which will cost a lot of waste of computational resources. Therefore, in this paper, we aim to propose a training-free detection method, which can be considered zero cost while achieving super great detection performance.
- **Robustness to Potential Attacks.** As we have mentioned before, one major drawback of the binary classifiers is that they are all vulnerable to potential adversary example attacks. Therefore, in this paper, we aim to develop a robust detector that would not be disturbed by any form of malicious noise.

3.2 Detector's Capabilities

In this paper, we assume that the detector has access to one text-to-image generation model, which is a reasonable assumption as most text-to-image generation models are open-source models. Note that we conduct experiments to show that one text-to-image generation model can be used to detect other fake images generated by unknown and totally different models. We also assume the detector does not know the original prompts that generate the fake images. Therefore, in the first stage of our detection, we need to generate the prompt. We will detail the process of detection in Section 4.

4 ZEROFAKE

In this section, we introduce the detailed pipeline for our proposed ZeroFake. We start by introducing the overview of how ZeroFake works. Then, we introduce three critical parts of ZeroFake to describe the proposed mechanism in detail.

Algorithm 1 Overview Pipeline of ZeroFake.

Ensure: Given image, x; Threshold, τ
Require: Verification result, y (Real or Fake)

$ReversedPrompt \leftarrow \textsc{Reverse}(x)$
$AdvSets \leftarrow$ ["tree", "dog", "cat", ...]
for $word$ in $AdvSets$ **do**
    $ReversedPrompt \leftarrow \textsc{RepFirstNoun}(ReversedPrompt, word)$
end for
$LeastSimilar \leftarrow \textsc{FindLeastSimilar}(ReversedPrompt, AdvSets)$
$ImageLatent \leftarrow \text{DDIM}(x, LeastSimilar)$
$x' \leftarrow \textsc{Forward}(ImageLatent, LeastSimilar)$
$Sim \leftarrow \textsc{Similarity}(x', x)$
if $Sim \leq \tau$ **then**
    $y \leftarrow 0$
else
    $y \leftarrow 1$
end if

4.1 Overview

In general, the fake image detection problem typically revolves around identifying differences between real and fake images. Previous methods have typically relied on training classifiers to implicitly discern these visual distinctions. However, such approaches are often impractical due to the high costs associated with data collection and model training, particularly given the rapid evolution of diffusion models, where visual differences can vary significantly across different generative models.

In contrast to classifier-based approaches, we propose a novel method that capitalizes on a fundamental trait of text-to-image generation models to differentiate between fake and real images. Specifically, we observe an obvious disparity in the stability of final generated images when starting from initial states obtained via DDIM inversion on fake versus real images. We show some examples in Figure 2 and Figure 3.

According to the above findings, the basic idea of the proposed ZeroFake is that fake images are more robust than real images during the DDIM inversion process. Specifically, we find that when adding some perturbation, fake images are easier to reconstruct than real images. To leverage the difference above, we adopt the following steps to conduct the fake image detection. We first conduct the DDIM inversion under the guidance of our adversary prompts.

(a) Real (b) Reconstructed Real

Figure 2: Real image V.S. reconstructed real image.

(a) Fake (b) Reconstructed Fake

Figure 3: Fake image V.S. reconstructed fake image.

After the DDIM inversion, we can get the Gaussian noise that is supposed to be able to generate the original images. Then, we leverage the de-noise process to de-noise the Gaussian noise to reconstruct the images. We then compute the similarity between the reconstructed images and the original images. If the similarity is smaller than the predefined threshold, then we classify the given images as real or fake images. We draw the overall pipeline in Figure 4. We also show the pseudo code to better demonstrate the process of ZeroFake in Algorithm 1.

4.2 Prompt Generation

As illustrated above, we leverage the robustness difference between fake and real images' inversion under perturbed prompts to distinguish fake images apart from real ones. Therefore, we first need to obtain the prompt for each given image, as we only take images for the classification. To achieve that goal, we adopt the BLIP model, one of the state-of-the-art image captioning models that can generate a proper description $\mathcal{P}_{oracle}$ of the given image, denoted in Equation 4.

$$\mathcal{P}_{\text{oracle}} = \text{BLIP}(\text{x}). \tag{4}$$

In this paper, we refer to the output of BLIP as the reverse prompt.

Then, we use the reversed prompts as oracle prompts to obtain the adversary ones using our one-word replacement method in the following steps. Firstly, we change the first noun words in the reversed prompt with an alternative noun set $\mathcal{N} = \{n_1, n_2, \ldots, n_k\}$ consisting of many unrelated noun terms n_i like dog, tree, car, etc. We manually craft this adversary noun list. We show how we construct the adversary noun list and why the list can be considered high quality in Section A.1. Since noun words usually contribute a lot to the whole meaning of sentences, such a method can easily perturb the meaning of the prompts. Then, we can obtain an alternative prompt set $\mathcal{S}$ containing different adversary prompts. The process is denoted in Equation 5:

$$\mathcal{P}_{\text{perturbed},i} = \text{ReplaceNoun}(\mathcal{P}_{\text{oracle}}, n_i), \tag{5}$$

where n_i belongs to $\mathcal{N} = \{n_1, n_2, \ldots, n_k\}$, which is an alternative noun set.

For example, suppose the reversed prompt is "A motorcycle parked in the parking space next to another motorcycle". In that case, the alternative prompt set $\mathcal{S}$ will contain "A tree parked in a parking space next to another motorcycle", "A dog parked in a parking space next to another motorcycle" and others formed by the first noun replacement methods. Then we compute the cosine similarity of the perturbation prompts $\mathcal{P}_{\text{perturbed},i}$ with the reversed prompt $\mathcal{P}_{\text{oracle}}$ and find the least similar one, which can be considered as the most effective perturbation prompts. The process is formulated in Equation 6 and Equation 7:

$$\text{similarity}_i = \text{CosineSimilarity}(\mathcal{P}_{\text{oracle}}, \mathcal{P}_{\text{perturbed},i}) \tag{6}$$

$$\mathcal{P}_{\text{effective}} = \underset{\mathcal{P}_{\text{perturbed},i}}{\text{argmin}} \ \text{similarity}_i \tag{7}$$

4.3 DDIM Inversion and Reconstruction

After obtaining the adversary prompts corresponding to the given images, in this stage, we conduct the DDIM inversion, leveraging both the original images and the derived adversary prompts. DDIM inversion inherently involves the strategic incorporation of noise into the images, followed by a systematic denoising procedure to recover the original images. Specifically, we first need to feed the given image to a pre-trained Variational Autoencoder (VAE) encoder, thereby yielding the initial latent representation of the image, denoted in Equation 8:

$$z_0 = \text{VAE\_Encoder}(\text{x}) \tag{8}$$

Also, we need to get the prompt embedding, which will serve as guidance when we conduct the add-noise and de-noise process in Equation 9.

$$emb = \text{TEXT\_Encoder}(\mathcal{P}_{\text{effective}}) \tag{9}$$

After getting z_0 and emb, we need to add the noise to the initial latent representation of the image, employing the DDIM inversion. According to DDIM's sampling procedure, the exact procedure is mathematically represented in Equation 10:

$$z_t = \sqrt{\frac{\alpha_t}{\alpha_{t-1}}} z_{t-1} + \sqrt{\alpha_t} \left(\sqrt{\frac{1}{\alpha_t} - 1} - \sqrt{\frac{1}{\alpha_{t-1}} - 1} \right) \cdot \epsilon_\theta(z_t, t, \text{emb}), \tag{10}$$

where:

- z_t symbolizes the de-noised latent state at time step t.

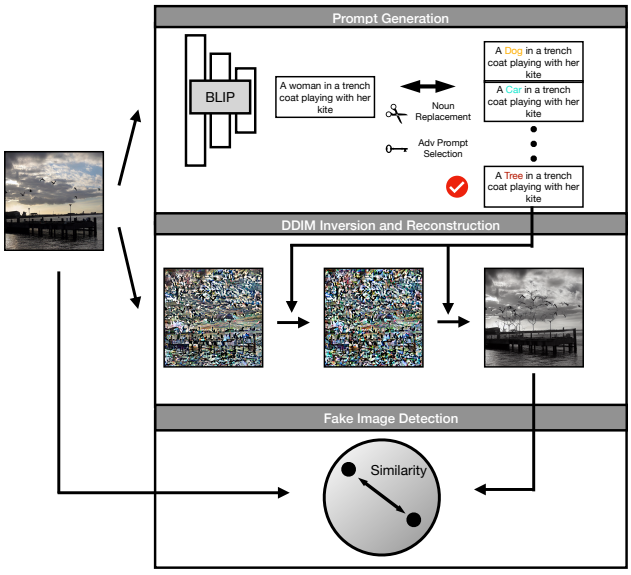


Figure 4: An illustration of fake image detection. Specifically, ZeroFake first generates the adversary prompt via prompt reversion and selection. Then, ZeroFake conducts the DDIM inversion and reconstruction on the given images to obtain the reconstructed images. ZeroFake finally computes the similarity between the given image and the reconstructed image to determine whether it is a fake or real image.

- α_t and α_{t+1} signify the noise scales at consecutive time steps t and $t+1$, respectively. These parameters are integral to the variance schedule that dictates the progressive modulation of noise throughout the diffusion steps, maintaining a delicate balance between noise and original data fidelity.
- $\epsilon_\theta(z_t, t, \text{emb})$ denotes the UNet, tasked with estimating the noise component.

However, we cannot get z_t in practice to estimate the noise in Equation 10. Thus, we use z_{t-1} to estimate the noise ϵ in practice shown in Equation 11,

$$z_t \approx \sqrt{\frac{\alpha_t}{\alpha_{t-1}}} z_{t-1} + \sqrt{\alpha_t}\left(\sqrt{\frac{1}{\alpha_t} - 1} - \sqrt{\frac{1}{\alpha_{t-1}} - 1}\right) \cdot \epsilon_\theta(z_{t-1}, t-1, \text{emb}). \tag{11}$$

Integrating z_0 into the step-by-step process we've described and moved through a set of T time steps helps us slowly polish the hidden features of the image. By repeating this process, which ends at z_t, we finally get the approximate original noise, which can be used to generate the given images. In this paper, we set forward step t as 999. However, the process is not accurate since the strict inversion requires estimating noise at step 't', but the actual inversion can only use noise from step 't-1', whereas the restoration process uses noise from step 't'. Also, the guidance of the given prompt, which is *emb*, will disturb the direction of the noise-adding process. Therefore, there is a gap between the original noise latent and the reversed noise latent. Moreover, based on our findings, the perturbation of the prompts is stronger in real images than in fake images, which will contribute to the fact that the reversed noise latent of fake images is closer than the original noise latent. This leads to the reconstructed images being more similar to the original images.

The reconstruction process is the reverse function of the DDIM inversion process, which can be formulated in Equation 12.

$$z_{t-1} = \sqrt{\frac{\alpha_{t-1}}{\alpha_t}} z_t - \sqrt{\alpha_{t-1}}\left(\sqrt{\frac{1}{\alpha_t} - 1} - \sqrt{\frac{1}{\alpha_{t-1}} - 1}\right) \cdot \epsilon_\theta(z_t, t, \text{emb}) \tag{12}$$

However, as we have stated before, the process will also introduce bias as the initial noise z_0 is not the approximated one. Moreover, the adversary prompt embedding also disturbs the process in this phase. Therefore, after the de-noise process, we will get the new initial latent code z_0^T. In practice, we reconstruct the image in 20 steps.

Then, we input the new initial latent code to the VAE-Decoder and get the new reconstructed image as Equation 13:

$$x' = \text{VAE\_Decoder}(z_0^T) \tag{13}$$

4.4 Fake Image Detection

The process of Fake image detection is quite simple. The above steps can help us get the reversed image based on the given images. We find that the real images are less robust to the prompt perturbations, which means that the reversed real images may share less similarity of original real images than fake images. Therefore, in this step, we will compute the SSIM [45] similarity between the reversed images and the original images in Equation 14.

$$Distance = \text{SSIM}(x, x') \tag{14}$$

SSIM is a metric used to measure the similarity between two images. SSIM considers the visual impact of three characteristics of an image: luminance, contrast, and structure. By comparing these elements between a reference image and a test image, SSIM quantifies their perceptual differences, offering a value between 0 and 1, where 1 indicates perfect similarity. If the SSIM is smaller than the predefined threshold, then we can conclude that the given images are real; otherwise, they are fake. To decide the threshold, we first need to generate 100 fake images together with real images. Then, we leverage SVM to find the best threshold based on the 200 samples. We test the threshold in other unseen samples.

5 FAKE IMAGE DETECTION

5.1 Experimental Setup

Text-to-Image Generation Models. Nowadays, text-to-image generation models have attracted an increasing amount of attention. Typically, these models use a text prompt along with random noise as inputs. The process involves denoising the image while aligning it with the guidance provided by the text prompt, ensuring the final image corresponds accurately to the given prompt. In this work, we focus on the following four state-of-the-art text-to-image generation models, which are considered the most powerful models in hugging faces,

- **Stable Diffusion 1.4 [36].** Stable Diffusion is a latent text-to-image generation model noted for its capability of generating photo-realistic given textual prompts. It optimizes

the image generation process by conducting the diffusion process in the latent space with a conditioning mechanism that enables data from other modalities to control the synthesis process, improving the training efficiency and achieving competitive performance across a range of image synthesis tasks.

- **SDXL-Lightning [25].** SDXL-Lightning is a state-of-the-art one-step/few-step text-to-image generation model, which is trained through a progressive, adversarial diffusion distillation approach. It can produce high-quality images in 2, 4, or 8 steps. The number of steps is set to 2 in our text-to-image generation.
- **Stable-diffusion-xl-base [2].** In the case of Stable-diffusion-xl-base, the two-stage pipeline is adopted. First, the noisy latent of the desired output size is produced from the base model, and then the specialized high-resolution refiner with the technique SDEdit [30] is applied to generate the output fake images of size 1024x1024. The number of steps and high noise fraction are set to 40 and 0.8, respectively.
- **DALL·E 2 [33].** DALL·E-2 is a text-to-image generation model developed by OpenAI that generates images given textual prompts. It utilizes a two-stage model for text-to-image generation, initially employing a Diffusion prior model to generate image embeddings from CLIP text embeddings, followed by a classifier-free guidance diffusion decoder that inverts these image embeddings back into images, demonstrating the capability of synthesizing complex and realistic images with text conditions. For DALL·E 2, we obtain 1024×1024 images by making requests from the OpenAI API.
- **GLIDE [32].** GLIDE, a text-to-image generation model developed by OpenAI, is accessible via its GitHub page[2]. This model has been trained on a curated dataset consisting of hundreds of millions of prompt-image pairs. Furthermore, GLIDE exhibits limitations in processing prompts involving "person" topics, as images of this nature were excluded from its training dataset to address ethical considerations.

Note that based on the description, Stable Diffusion and Stable-diffusion-xl-base are totally different model structures trained on different datasets. Therefore, they can be regarded as different generation models. Also, the DALL·E 2 we used in this paper is the official DALL·E 2 model from Open AI. Therefore, we just query the API and get the fake images.

Datasets. In this paper, we take advantage of the following benchmark datasets as the evaluation datasets. Both of them are prompt-image pair datasets.

- **MSCOCO [26].** The MSCOCO dataset, developed by Microsoft, stands as a comprehensive resource for object detection, segmentation, key-point detection, and image captioning tasks. This large-scale dataset consists of 328,000 images, facilitating a range of visual tasks by providing a wide variety of annotated data, i.e., prompts.
- **Flickr30k [48].** The Flickr30k dataset comprises 31,000 images sourced from Flickr, each accompanied by five reference sentences crafted by human annotators. This collection offers images alongside detailed textual descriptions, supporting a wide range of visual tasks, including image retrieval, image captioning, image-to-text generation, and multi-modal understanding.
- **AGFD-20K [3].** Besides MSCOCO and Flickr30k, we also consider the face dataset AGFD-20k in the evaluation part to show the generalization of the proposed ZeroFake. Note that AGFD-20k only contains the fake face images generated by Stable Diffusion. Therefore, for the real images, we leverage CeleBA [28]. Together with AGFD-20k and CeleBA, we construct our real fake face image pairs.

[2]https://github.com/openai/glide-text2im

Table 1: The text-to-image generation models, datasets, and the number/size of fake images we consider in this work.

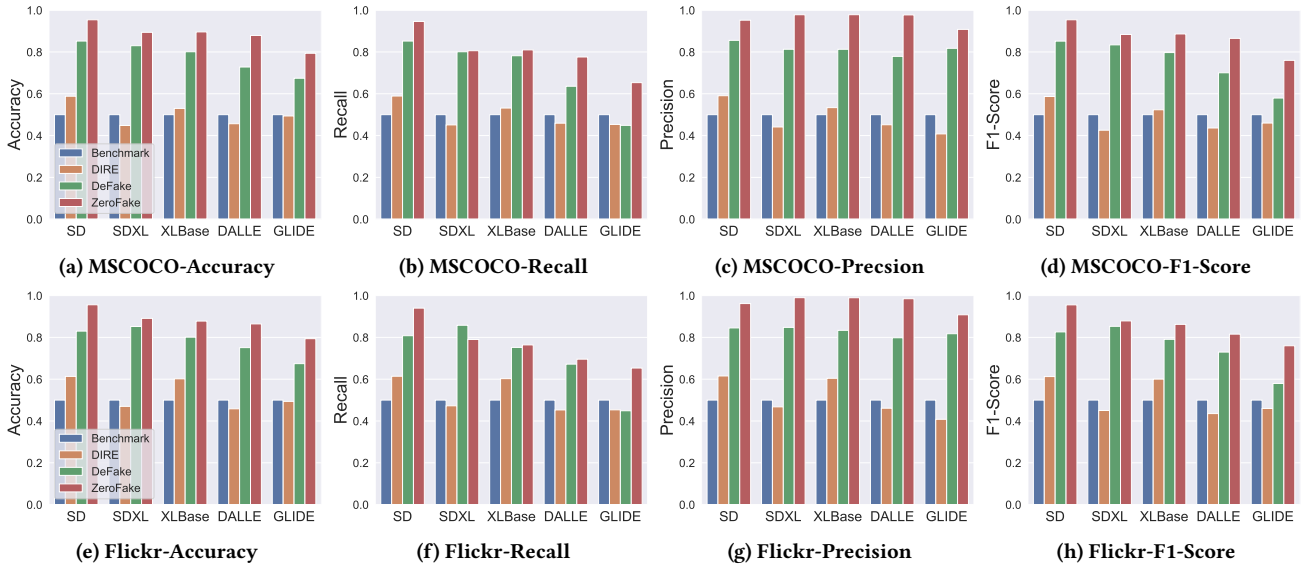
Model	Dataset	Images	Image Size
Stable Diffusion	MSCOCO	500	512×512
	Flickr30k	500	512×512
SDXL-Lighting	MSCOCO	500	512×512
	Flickr30k	500	512×512
XLBase	MSCOCO	500	1024×1024
	Flickr30k	500	1024×1024
GLIDE	MSCOCO	500	256×256
	Flickr30k	500	256×256
DALL·E 2	MSCOCO	500	1024×1024
	Flickr30k	500	1024×1024

Based on the text-to-image generation models and datasets, we generate fake images for the experiments in this paper. We summarize the fake images we used in Table 1. Note that in fake image detection settings, we set the threshold as 0.78.

5.2 Results

We show our results in Figure 5. First of all, we can observe that the previous DIRE fails to consistently perform across various scenarios. For instance, when detecting fake images generated by Stable Diffusion on the prompts from MSCOCO, DIRE can only achieve 0.590 accuracy, which is close to the random guess. This shortfall is attributed to the superior capabilities of modern text-to-image generation models over traditional diffusion models, highlighting DIRE's limitations in adapting to high-quality generative outputs. Consequently, DIRE's performance issues underscore its lack of generality, especially against state-of-the-art text-to-image models capable of producing highly authentic images. Therefore, based on the results of DIRE, we can conclude that instead of just adding noise and conducting the denoise process, the DDIM inversion, together with the prompt guidance we considered in this paper, has significant effects on the performance of fake image detection.

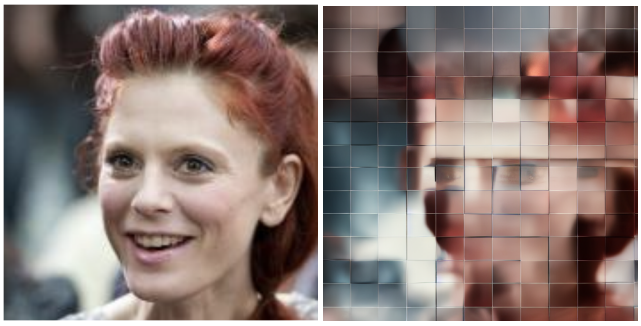
Furthermore, we can see from the figure that while DeFake outperforms DIRE, it still cannot achieve as good a performance as ZeroFake. For instance, DeFake can detect 0.852 percent of fake images generated by SDXL on the prompts from Flickr30k apart from real images. Note that the DeFake we used in this paper is also only trained on 400 fake-real image pairs on SD. We finetune the CLIP to train the DeFake according to the instructions in the DeFake



(a) MSCOCO-Accuracy (b) MSCOCO-Recall (c) MSCOCO-Precsion (d) MSCOCO-F1-Score

(e) Flickr-Accuracy (f) Flickr-Recall (g) Flickr-Precision (h) Flickr-F1-Score

Figure 5: The performance of ZeroFake V.S. DeFake and DIRE.



(a) Real Face (b) Reconstructed Real Face

Figure 6: Real face V.S. reconstructed real face.

(a) Fake (b) Reconstructed Fake Face

Figure 7: Fake face V.S. reconstructed fake face.

paper. The performance of DeFake argues that the prompt does have a significant influence when conducting fake image detection. However, how to leverage prompts to maximize the performance of fake image detection was still unexplored.

The proposed ZeroFake can achieve better performance than the previous DeFake and DIRE. For instance, when detecting fake images generated by Stable Diffusion on the prompts from MSCOCO, our proposed ZeroFake can achieve 0.957 accuracy, 0.948 recall, 0.965 precision, and 0.956 F1 score, which demonstrates that the proposed methods can easily detect fake images apart from real ones. Moreover, the proposed ZeroFake can also achieve great generability, as shown in the results. For instance, even if we do not know the exact architecture of DALL·E 2, our ZeroFake can still achieve great performance. As we have shown in Figure 2 and Figure 3, it can be seen that with the adversary prompt guidance, the real images can be reconstructed into totally different images, while the fake images can always keep the original features.

We also show that the proposed ZeroFake can have great performance not only in complex scene datasets but also in single scene datasets like AGFD-20k, a fake face dataset. Specifically, the proposed ZeroFake can achieve 0.894 accuracy when detecting fake face images and real face images. We show some examples of the reconstructed fake and real images in Figure 6 and Figure 7. It can be seen from the figure that even if the fake image only contains a single object, such as a human face, there is still a very obvious gap between the real and fake images during the reconstruction process.

6 FAKE ARTWORK DETECTION

Beyond traditional fake image detection, one of the primary applications of these text-to-image generation models is in the field of artwork. For instance, BBC reported that some fake artworks generated by text-to-image generation models had won first place in an art competition, disappointing lots of participating artworks. However, as we have tested, both DIRE and DeFake cannot achieve

Celebrate Mickey Mouse's birthday with some classic Heathcliff cat vs mouse antics _ Cartoon

Woman in Black _ Oil Painting

Figure 8: The collected artwork examples.

great performance on the artworks. In this section, we will show that our proposed ZeroFake can perform well when dealing with fake artwork. We will start the section by introducing the process of collecting artwork. Then, we will show the results of ZeroFake.

6.1 Artwork Collection

We collected 120 prompt-image pairs of artwork from the artwork's online web pages and galleries. As the original artworks do not have prompts, we derive them from the titles or descriptions provided for the artwork together with the art style of the artworks. We show two examples in Figure 8. We collect artworks covering a wide range of artistic expressions, including Cartoons, Chinese Paintings, Oil Paintings, Quick Sketches, Sketches, and Watercolor Paintings. Note that to ensure the collected artworks are not used in the training phase of text-to-image generation models, we test the similarity between the generated fake images and real images according to [47]. All artworks used in this section are not contained in the training dataset of the models. Note that in fake artwork detection settings, we set the threshold as 0.75, which is chosen from 40 real fake image pairs.

6.2 Results

We show the results of the detection performance in Figure 9. It can be concluded from the figure that the proposed ZeroFake can achieve much better performance than the previous DIRE and DeFake. For instance, When detecting fake Chinese Paintings, ZeroFake can achieve 0.881 accuracy while DIRE and DeFake can only achieve 0.410 and 0.690, which can be regarded as a random guess. To better demonstrate the effectiveness of the proposed ZeroFake, we also show some examples of real and fake artworks in Figure 10. It can be easily observed that the reversed fake image obviously shares more common features than the real image. This observation again demonstrates that prompt-guided image reversion can help us better distinguish fake images from real images. Besides the above observations, we can also conclude from the results that different types of artworks have different detection performances. For instance, Chinese painting can achieve 0.881 accuracy, while quick sketches can achieve 0.973 accuracy. We assume it is because of the different abilities of the UNet that we use in different art styles. However, even though there exist gaps between different types of artworks, our proposed ZeroFake can always have the best performance. Therefore, we can conclude that ZeroFake is better at finding the difference between real artworks and fake artworks.

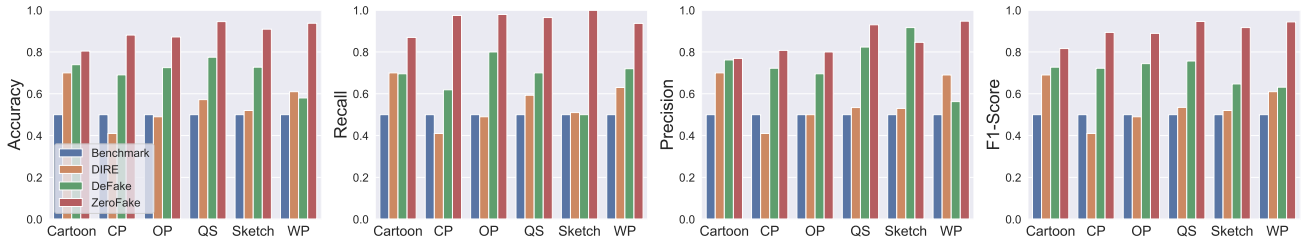
7 IMAGE EDITING DETECTION

Despite the detection of fake artwork, one of the most important and broad applications of text-to-image generation models is real image editing. For instance, a politician can be edited into any adversary's desired appearance. This kind of edition can cause serious social damage, like affecting the voting situation, misleading public perceptions, and fueling misinformation campaigns. Therefore, it is also very important to distinguish edited images apart from real images. However, previous works have not considered edited images to be fake images. In this paper, we demonstrate that previous state-of-the-art methods cannot work well in the image editing domain, while ZeroFake can easily detect edited images apart from real ones. We will start by introducing the experimental setup of the image editing detection. Then, we will introduce our results.

7.1 Experimental Setup

Image Editing. Image editing entails the alteration of images through additional inputs such as text prompts, masks, or reference images. This process allows for adjustments in characteristics such as color and style, as well as the addition or removal of elements to achieve the desired visual outcome. In our study, we consider the following three state-of-the-art image editing techniques within text-to-image generation models.

- **Prompt-to-Prompt (p2p) [44].** Prompt-to-Prompt is a text-only, cross-attention-based image editing method designed to maintain the structure of the input image while altering attributes as specified by the revised prompt. The key idea behind p2p is that the spatial layout and geometry of an image depend on cross-attention maps, which are formed between image pixel features and prompt tokens. Therefore, p2p injects cross-attention elements derived from the original prompt P into a new cross-attention map generated with the modified prompt P*.
- **Ledit [42].** Unlike the p2p approach, which utilizes semantic information from cross-attention maps between image pixel features and prompt tokens, Ledit performs image-editing while maintaining the original image structure and content by incorporating the idea of semantic guidance with the Denoising Diffusion Probabilistic Models(DDPM) [15] inversion process. DDPM utilizes a series of white Gaussian noise samples to reconstruct an image. Those noise maps could be considered latent codes that contain structure and content information. Leveraging this principle, Ledit complements the iterative denoising process for the edited image by incorporating the original image's noise maps obtained from DDPM inversion, thereby preserving the image structures irrelevant to edit concepts.
- **InstructPix2Pixl (Pixel) [6].** InstructPix2Pix proposed a method for image editing based on human instructions by



(a) Accuracy (b) Recall (c) Precision (d) F1-Score

Figure 9: The performance of ZeroFake on Artworks.

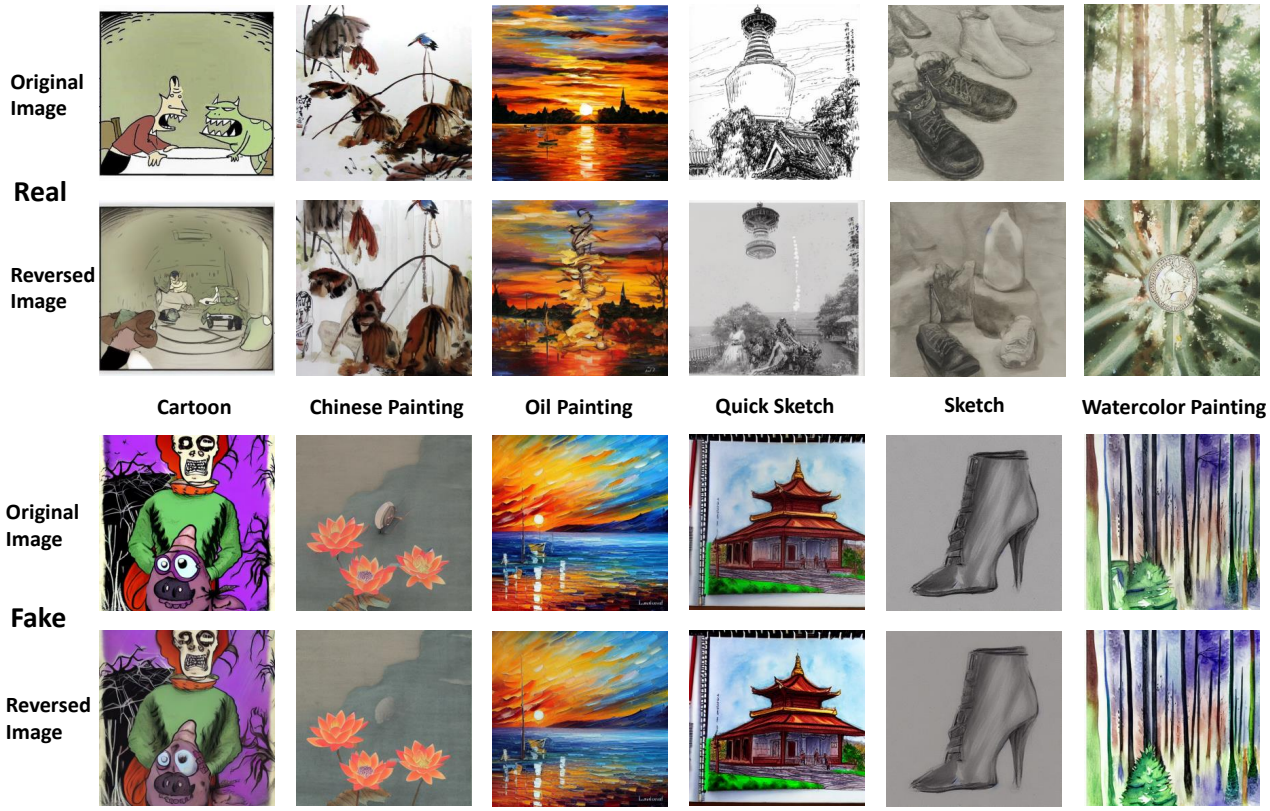
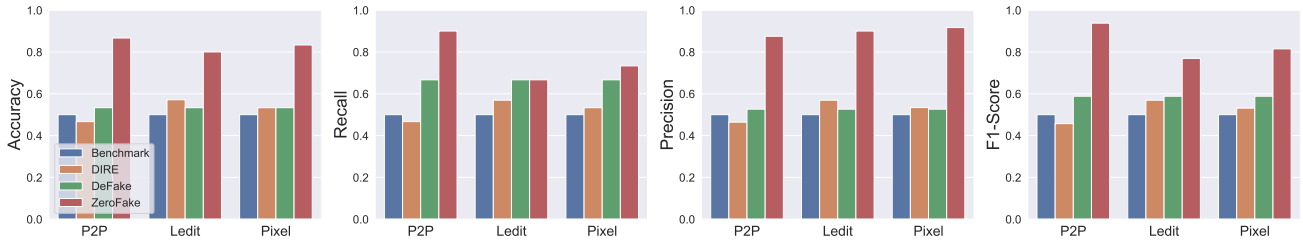


Figure 10: The example of the real and fake artworks from different types and their corresponding reconstructed images.

training a conditional diffusion model on a specially constructed dataset of image pairs with corresponding source captions and instructions. Unlike Prompt-to-prompt and Ledits, InstructPix2Pix facilitates editing directly in the forward pass without necessitating per-example finetuning or inversion, thus significantly accelerating the editing process.

Edited Images. As there is no existing benchmark dataset for image editing, we collect 15 images and write the corresponding editing prompts. We show the collected images together with the editing prompts in Table 2. We first find that current image editing methods have proved to be very unstable. However, as the edited images are all generated and selected by our human efforts, the edited images we used in this paper all perfectly capture the features of the prompt guidance, which can be considered to pose some challenge to the detection methods. However, as we show in Section 7.3, the proposed ZeroFake can still achieve great performance while the DIRE and DeFake failed. Note that in edited image detection settings, we set the threshold as 0.78, which is chosen from 30 real fake image pairs.



(a) Accuracy (b) Recall (c) Precsion (d) F1-Score

Figure 11: The performance of ZeroFake on edited images.

7.2 The Performance of Image Editing

Before we introduce the performance of our proposed ZeroFake on edited images, we first show the performance of the image editing methods we used in this paper. As there are no benchmark metrics for image editing, we have edited some examples here to show the performance of our adapted methods. Our results are shown in Figure 11. It can be seen that image editing methods can normally perform well on real images. For instance, P2P can faithfully turn the goldfish into a shark. However, in some cases, the considered methods may fail. For instance, when asked to turn the cat to the front, only P2P works, while other methods change the cat to another cat. We find that whether good or bad performance of the considered three methods on certain images, our proposed ZeroFake can always easily distinguish fake images, which are the edited images here, apart from real ones. We will show the results in the next section.

7.3 The Performance of ZeroFake

We show the results of our proposed ZeroFake in Figure 11. We can see from the figure that ZeroFake can still achieve the best performance compared to DeFake and DIRE. For instance, when detecting the fake images edited by p2p, ZeroFake can achieve 0.866 accuracy, while DIRE and DeFake can only achieve 0.467 and 0.533. To better demonstrate the effectiveness of the proposed ZeroFake, we show several examples in Table 2. It can be concluded that ZeroFake can always achieve great performance regardless of the performance of image editing methods. For instance, even if the Ledit cannot effectively turn the cat towards the camera, ZeroFake can still precisely reconstruct the edited image while the reconstruction of real images fails. Therefore, it can be concluded that the proposed ZeroFake can perform not only on the generated fake images but also on the edited images. This great performance enables ZeroFake to have a boarder application in the field of fake image detection.

8 ROBUSTNESS EVALUATION

We show the excellent performance of the proposed ZeroFake on fake images, fake artworks, and edited images above. One major advantage of ZeroFake is that it is robust enough to detect any potential attacks on images. In this section, we will introduce the robustness evaluation of the ZeroFake and previous detection methods. We consider several different perturbations on the fake images to evaluate the robustness of the proposed ZeroFake, including blurring, gaussian noise, sharpening, and adversary example attacks. Note that for adversary example attacks, as ZeroFake is the first work that leverages the prompt-guided DDIM inversion to conduct the fake image detection, there are no existing adversary example attacks against it. Therefore, we adapt the transfer attacks for FGSM and BIM.

8.1 Image Transformations

We consider three traditional image transformations in this section including blurring, sharpening, and gaussian noise. We also consider two state-of-the-art adversary attacks in this section, including FGSM and BIM.

Blurring. Blurring is a common image processing technique used to reduce noise and detail in an image. This is typically achieved by averaging the pixels within a local neighborhood, which results in a smooth, less detailed version of the original image. Blurring can help to obscure small imperfections and details, making it useful for various applications in image processing and computer vision.

Sharpening. Sharpening is an image processing technique used to enhance the edges and fine details within an image. This process increases the contrast between different regions, making features more distinct and the overall image crisper. In this paper, we set the factor for sharpening as 2.

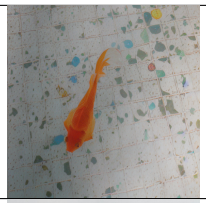
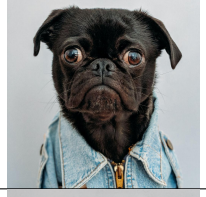
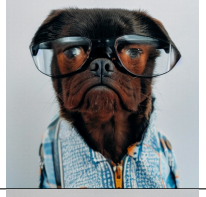
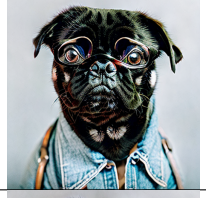
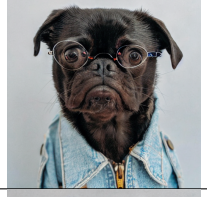
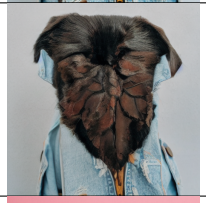
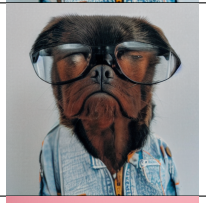
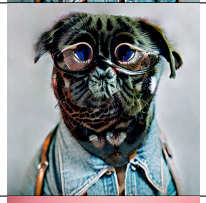
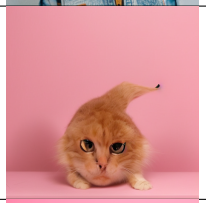
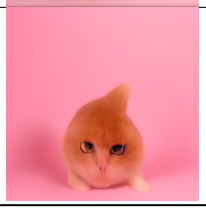
Gaussian Noise. Gaussian noise involves adding random noise with a Gaussian distribution to an image. This type of noise has a probability density function equal to that of the normal distribution, which is characterized by its mean and variance. In this paper, we generate Gaussian noise with a mean of 0 and a variance of 2.

FGSM [14]. The fast gradient sign method (FGSM) is one of the most representative attacks in the adversary example domain. FGSM can be formulated in Equation 15.

$$x' = x + \epsilon \cdot \text{sign}(\nabla_x J(\theta, x, y)) \tag{15}$$

x is the input, y is the true label of x, θ is our detector's parameters, and $J(\theta, x, y)$ is its cost function. The main idea of FGSM is to maximize the loss for the given model by adding a small and scaled version of the sign of the gradient to the original image. For our proposed ZeroFake, as no classifier was used in the detection process, no gradient can be calculated to generate the noise. Therefore, we adapt the transfer attacks based on FGSM. Specifically, we compute the adversary noise based on the binary fake detectors' gradients. Then the noise will be transferred to the fake images to

Table 2: The example of the real and edited images by different state-of-the-art methods and their corresponding reconstructed images.

Classes	Original Image	Edited Prompt	P2P	Ledit	Instruct Pixel
Images		A Shark swims in the water.			
Reversed Images					
Images		The dog is wearing glasses.			
Reversed Images					
Images		The cat moved towards me.			
Reversed Images					

test whether the proposed ZeroFake can detect them. BIM takes a similar process.

BIM [16]. The basic iterative method (BIM) is another popular adversary example attack, which is actually an extension of FGSM. Instead of computing the gradient in one epoch, BIM uses multiple interactions with smaller step sizes to generate the adversarial noise, which can be formulated in Equation 16:

$$x^{(0)} = x, \qquad x'^{(t+1)} = x'^{(t)} + \alpha \cdot \text{sign}(\nabla x' J(\theta, x'^{(t)}, y)) \quad (16)$$

To transfer the attack to the proposed ZeroFake, we take the same actions as FGSM.

8.2 Results

We show the results in Table 3. We can see from the figure that both the detector-based method and the proposed ZeroFake demonstrate satisfying robust performance against traditional perturbations like blurring, sharpening, and Gaussian noise. For instance, ZeroFake

Table 3: The results of robustness evaluation against possible image transformations.

Method	Blurring	Sharpening	Gaussian Noise	FGSM		BIM	
				Noise	Detection	Noise	Detection
DeFake	0.778	0.897	0.783	0.001	0.763	0.001	0.612
				0.005	0.691	0.005	0.574
				0.01	0.517	0.01	0.490
ZeroFake	0.903	0.957	0.894	0.001	0.938	0.001	0.915
				0.005	0.912	0.005	0.908
				0.01	0.891	0.01	0.903

under sharpening can still achieve 0.957 accuracy when detecting fake images. It can also be concluded from the table that the detector-based method is vulnerable to potential adversary example attacks. For instance, when the noise is 0,005, the performance of DeFake will be reduced to 0.691 when faced with FGSM attacks, which is close to a random guess. It demonstrates that the adversary just needs to add small perturbations to the images to confuse the detectors. However, the proposed ZeroFake has proven to be very robust against potential adversary attacks. ZeroFake can maintain good performance against state-of-the-art attacks. For instance, even if the nose is 0.01, ZeroFake can still achieve 0.891 accuracy when detecting fake images against FGSM attacks. Note that the considered FGSM and BIM are both state-of-the-art adversary example attacks nowadays. These results demonstrate that ZeroFake is not easily confused by the adversary, and thus can contribute to a more stable detection performance. Such a robust performance can be attributed to the implicit method used for sampling, as the implicit methods usually show stable and robust performance on different tasks [21–23].

9 RELATED WORK

9.1 Text-to-Image Generation

Text-to-image generation models convert text prompts into corresponding visual representations. Foundational works by Reed et al. [35] and Zhang et al. [52] built on GAN principles introduced by Goodfellow et al. [13]. These systems integrate a prompt embedding with a latent vector. This enables GANs to create images that visually interpret the text prompts. The pioneering studies have inspired a wide range of researchers [4, 10, 11, 17, 17, 18, 24, 27, 31, 40, 46, 51] to explore text-to-image generation further. Despite significant contributions, using GANs sometimes results in suboptimal image generation outcomes [34, 36]. This has spurred the search for more reliable technologies. However, the quality of images generated by GAN is not satisfying.

Recently, text-to-image generation has seen major advancements with the emergence of diffusion models. These models begin with a representation of random noise. They iteratively refine this noise into a detailed and clear image guided by the text prompt. Leading models in this domain include DALL·E [34], Stable Diffusion [36], Imagen [37], GLIDE [32], and DALL·E 2 [33]. Diffusion-based models are now considered state-of-the-art. They demonstrate superior performance in generating high-quality images compared to previous GAN-based methods. The remarkable capabilities of these models are the focus of our current research in text-to-image generation technologies. Moreover, text-to-image generation models can also be used to edit part of images instead of just generating whole fake images [6, 42, 44]. This technology leverages the UNet to control the edition. In this paper, we show that previous training-based methods cannot achieve great detection performance due to the fact that these edited images are not considered in the training set. However, our proposed ZeroFake can have great performance in both fake images and edited images.

9.2 Fake Image Detection and Attribution

Previous research has primarily focused on detecting fake images generated by traditional generation models, such as GANs [13], low-level vision models [7, 9], and perceptual loss models [8, 20]. Wang et al.[43] successfully trained a simple CNN model to differentiate these generated images from authentic ones, capitalizing on common flaws inherent in images from earlier generative technologies. Similarly, Yu et al.[50] demonstrated that it is possible to trace fake images back to the specific traditional models that created them due to distinctive "fingerprints" left by these technologies. Building on this, Girish et al. [12] introduced an attribution method designed for scenarios where the generation model behind the image is unknown.

Recently, there have been some works that focus on detecting fake images generated by text-to-image generation models. the DeFake [38] utilizes both the image and its associated generated prompt for detection. During testing, DeFake generates prompts using the BLIP model to assist the classifier in determining the authenticity of images. Additionally, the DIRE method [44] represents the latest detection techniques against diffusion model-generated images, also employing a binary classifier. We emphasize here that almost all existing works need to train a binary classifier to detect fake images. We show in this paper that instead of the binary classifier, the natural difference between real and fake images can be leveraged to conduct the detection.

10 LIMITATION

In this section, we discuss the possible limitations of the proposed ZeroFake. It should also be noted that, unlike DIRE and DeFake, the proposed ZeroFake is a zero-shot detection method, which takes much less computational resources than the previous methods during the training phase. However, during the inference phase, the DDIM inversion and reconstruction will take more time than

traditional detectors. We show that on our NVIDIA DGX-A100, the inversion and reconstruction will take 30.2 seconds for one image, while DeFake needs 12.3 seconds to load the model and conduct the detection. Moreover, the ZeroFake also takes more computational resources to detect fake images than the traditional detectors. This is due to the fact that ZeroFake leverages the DDIM process to add the noise to the images and then reconstruct them. The whole process is resource-intensive. For one image, DeFake needs 3798 MiB GPU memory while ZeroFake takes 10076 MiB.

11 CONCLUSION

In conclusion, previous methods conducted fake image detection by training a binary classifier, which is shown to be inefficient, lacking in generalizability, and non-robust. In recognition of these limitations, our work introduces a novel zero-shot detection method called ZeroFake to distinguish fake images generated by text-to-image generation models apart from real ones. ZeroFake uniquely leverages the differential features exhibited by fake versus real images during the processes of DDIM inversion and subsequent reconstruction. Central to ZeroFake is the insight that fake images, inherently synthetic in nature, demonstrate great robustness to perturbations during the DDIM inversion process compared to their real counterparts. Our method takes advantage of this robustness by applying a perturbation-based DDIM inversion where each image undergoes strategic noise addition and subsequent noise reduction, guided by carefully crafted adversary prompts. This meticulous process effectively accentuates the resilience differences between fake and real images, enabling accurate classification. Extensive experimental evaluations affirm that ZeroFake surpasses previous state-of-the-art methods in detecting a broad array of fake images, including fake images, fake artwork, and edited images. Furthermore, our results highlight ZeroFake's enhanced robustness against potential adversarial example attacks, outperforming earlier detection techniques. In an era where the authenticity of images is paramount—owing to the ease of creating and disseminating highly realistic fake images—the ability to reliably discern between genuine and fabricated visuals is more crucial than ever. The adoption of ZeroFake could significantly bolster public discourse by shielding it from the influence of deceptive imagery, enhancing the integrity of media content, and providing a robust tool for platforms and regulatory bodies to combat the proliferation of digital misinformation.

12 ACKNOWLEDGMENTS

This work is partially funded by the European Health and Digital Executive Agency (HADEA) within the project "Understanding the individual host response against Hepatitis D Virus to develop a personalized approach for the management of hepatitis D" (DSolve, grant agreement number 101057917) and the BMBF with the project "Repräsentative, synthetische Gesundheitsdaten mit starken Privatsphärengarantien" (PriSyn, 16KISAO29K).

REFERENCES

[1] https://midjourney.com/.
[2] https://huggingface.co/stabilityai/stable-diffusion-xl-base-1.0.
[3] https://github.com/Robin-WZQ/AGFD-20K.
[4] Navaneeth Bodla, Gang Hua, and Rama Chellappa. Semi-supervised FusedGAN for Conditional Image Generation. In *European Conference on Computer Vision (ECCV)*, pages 689–704. Springer, 2018.
[5] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large Scale GAN Training for High Fidelity Natural Image Synthesis. In *International Conference on Learning Representations (ICLR)*, 2019.
[6] Tim Brooks, Aleksander Holynski, and Alexei A. Efros. InstructPix2Pix: Learning to Follow Image Editing Instructions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18392–18402. IEEE, 2023.
[7] Chen Chen, Qifeng Chen, Jia Xu, and Vladlen Koltun. Learning to See in the Dark. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3291–3330. IEEE, 2018.
[8] Qifeng Chen and Vladlen Koltun. Photographic Image Synthesis with Cascaded Refinement Networks. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1520–1529. IEEE, 2017.
[9] Tao Dai, Jianrui Cai, Yongbing Zhang, Shu-Tao Xia, and Lei Zhang. Second-Order Attention Network for Single Image Super-Resolution. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11065–11074. IEEE, 2019.
[10] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, and Jie Tang. CogView: Mastering Text-to-Image Generation via Transformers. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*, pages 19822–19835. NeurIPS, 2021.
[11] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion. *CoRR abs/2208.01618*, 2022.
[12] Sharath Girish, Saksham Suri, Sai Saketh Rambhatla, and Abhinav Shrivastava. Towards Discovery and Attribution of Open-World GAN Generated Images. In *IEEE International Conference on Computer Vision (ICCV)*, pages 14094–14103. IEEE, 2021.
[13] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Nets. In *Annual Conference on Neural Information Processing Systems (NIPS)*, pages 2672–2680. NIPS, 2014.
[14] Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and Harnessing Adversarial Examples. In *International Conference on Learning Representations (ICLR)*, 2015.
[15] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising Diffusion Probabilistic Models. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*. NeurIPS, 2020.
[16] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial Examples in the Physical World. *CoRR abs/1607.02533*, 2016.
[17] Qicheng Lao, Mohammad Havaei, Ahmad Pesaranghader, Francis Dutil, Lisa Di-Jorio, and Thomas Fevens. Dual Adversarial Inference for Text-to-Image Synthesis. In *IEEE International Conference on Computer Vision (ICCV)*, pages 7566–7575. IEEE, 2019.
[18] Ang Li, Yichuan Mo, Mingjie Li, and Yisen Wang. PID: Prompt-Independent Data Protection Against Latent Diffusion Models. In *International Conference on Machine Learning (ICML)*, pages 28421–28447. PMLR, 2024.
[19] Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. *CoRR abs/2201.12086*, 2022.
[20] Ke Li, Tianhao Zhang, and Jitendra Malik. Diverse Image Synthesis From Semantic Layouts via Conditional IMLE. In *IEEE International Conference on Computer Vision (ICCV)*, pages 4219–4228. IEEE, 2019.
[21] Mingjie Li, Lingshen He, and Zhouchen Lin. Implicit Euler Skip Connections: Enhancing Adversarial Robustness via Numerical Stability. In *International Conference on Machine Learning (ICML)*, pages 5874–5883. PMLR, 2020.
[22] Mingjie Li, Yisen Wang, and Zhouchen Lin. Cerdeq: Certifiable deep equilibrium model. In *International Conference on Machine Learning (ICML)*, pages 12998–13013. PMLR, 2022.
[23] Mingjie Li, Yisen Wang, and Zhouchen Lin. GEQ: Gaussian Kernel Inspired Equilibrium models. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*, pages 38767–38785. NeurIPS, 2023.
[24] Wenbo Li, Pengchuan Zhang, Lei Zhang, Qiuyuan Huang, Xiaodong He, Siwei Lyu, and Jianfeng Gao. Object-Driven Text-To-Image Synthesis via Adversarial Training. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1274–12182. IEEE, 2019.
[25] Shanchuan Lin, Anran Wang, and Xiao Yang. SDXL-Lightning: Progressive Adversarial Diffusion Distillation. *CoRR abs/2402.13929*, 2024.
[26] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common Objects in Context. In *European Conference on Computer Vision (ECCV)*, pages 740–755. Springer, 2014.
[27] Vivian Liu and Lydia B. Chilton. Design Guidelines for Prompt Engineering Text-to-Image Generative Models. In *Annual ACM Conference on Human Factors in Computing Systems (CHI)*, pages 384:1–384:23. ACM, 2022.
[28] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep Learning Face Attributes in the Wild. In *IEEE International Conference on Computer Vision (ICCV)*, pages 3730–3738. IEEE, 2015.

[29] Francesco Marra, Diego Gragnaniello, Davide Cozzolino, and Luisa Verdoliva. Detection of GAN-Generated Fake Images over Social Networks. In *IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, pages 384–389. IEEE, 2018.
[30] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: Guided Image Synthesis and Editing with Stochastic Differential Equations. In *International Conference on Learning Representations (ICLR)*, 2022.
[31] Yichuan Mo, Hui Huang, Mingjie Li, Ang Li, and Yisen Wang. TERD: A Unified Framework for Safeguarding Diffusion Models Against Backdoors. In *International Conference on Machine Learning (ICML)*, pages 35892–35909. PMLR, 2024.
[32] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models. *CoRR abs/2112.10741*, 2021.
[33] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical Text-Conditional Image Generation with CLIP Latents. *CoRR abs/2204.06125*, 2022.
[34] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-Shot Text-to-Image Generation. In *International Conference on Machine Learning (ICML)*, pages 8821–8831. JMLR, 2021.
[35] Scott E. Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative Adversarial Text to Image Synthesis. In *International Conference on Machine Learning (ICML)*, pages 1060–1069. JMLR, 2016.
[36] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-Resolution Image Synthesis with Latent Diffusion Models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695. IEEE, 2022.
[37] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. *CoRR abs/2205.11487*, 2022.
[38] Zeyang Sha, Zheng Li, Ning Yu, and Yang Zhang. DE-FAKE: Detection and Attribution of Fake Images Generated by Text-to-Image Diffusion Models. *CoRR abs/2210.06998*, 2022.
[39] Jiaming Song, Chenlin Meng, and Stefano Ermo. Denoising Diffusion Implicit Models. In *International Conference on Learning Representations (ICLR)*, 2021.
[40] Douglas M. Souza, Jonatas Wehrmann, and Duncan D. Ruiz. Efficient Neural Architecture for Text-to-Image Synthesis. In *International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2020.
[41] Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. Designing an encoder for StyleGAN image manipulation. *ACM Transactions on Graphics*, 2021.
[42] Linoy Tsaban and Apolinário Passos. LEDITS: Real Image Editing with DDPM Inversion and Semantic Guidance. *CoRR abs/2307.00522*, 2023.
[43] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A. Efros. CNN-Generated Images Are Surprisingly Easy to Spot... for Now. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8692–8701. IEEE, 2020.
[44] Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, Hezhen Hu, Hong Chen, and Houqiang Li. DIRE for Diffusion-Generated Image Detection. In *IEEE International Conference on Computer Vision (ICCV)*, pages 22388–22398. IEEE, 2023.
[45] 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*, 2004.
[46] Zixu Wang, Zhe Quan, Zhi-Jie Wang, Xinjian Hu, and Yangyang Chen. Text to Image Synthesis With Bidirectional Generative Adversarial Network. In *International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2020.
[47] Yixin Wu, Ning Yu, Zheng Li, Michael Backes, and Yang Zhang. Membership Inference Attacks Against Text-to-image Generation Models. *CoRR abs/2210.00968*, 2022.
[48] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2014.
[49] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, Ben Hutchinson, Wei Han, Zarana Parekh, Xin Li, Han Zhang, Jason Baldridge, and Yonghui Wu. Scaling Autoregressive Models for Content-Rich Text-to-Image Generation. *CoRR abs/2206.10789*, 2022.
[50] Ning Yu, Larry Davis, and Mario Fritz. Attributing Fake Images to GANs: Learning and Analyzing GAN Fingerprints. In *IEEE International Conference on Computer Vision (ICCV)*, pages 7555–7565. IEEE, 2019.
[51] Han Zhang, Jing Yu Koh, Jason Baldridge, Honglak Lee, and Yinfei Yang. Cross-Modal Contrastive Learning for Text-to-Image Generation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 833–842. IEEE, 2021.
[52] Han Zhang, Tao Xu, and Hongsheng Li. StackGAN: Text to Photo-Realistic Image Synthesis with Stacked Generative Adversarial Networks. In *IEEE International Conference on Computer Vision (ICCV)*, pages 5908–5916. IEEE, 2017.
[53] Xu Zhang, Svebor Karaman, and Shih-Fu Chang. Detecting and Simulating Artifacts in GAN Fake Images. In *IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–6. IEEE, 2019.

A APPENDIX

A.1 Quality Assurance of Construction of the Adversary Noun List

In ZeroFake, we need to craft the adversary noun list manually. In this section, we will discuss the quality assurance of the manual construction of the adversary noun list. To make sure there is at least one noun whose textual embedding is far away from the prompts' noun enough to make the prompt semantic change, we craft our adversarial noun list by choosing four nouns that are greatly different from each other, e.g., tree, airplane, cat, and others. Since we can ensure the textual embedding is far away from each other by our manual craft, there is at least one candidate whose textual embedding is far away from any prompts' first noun, as the following proposition states.

Proposition A.1. *Assume textual embedding of N nouns in the adversarial noun list as $e_1, ..., e_N$, and the minimum L_2 distance between each other is d. Then we can conclude that for any word w, there exists a word with embedding e_i in the adversarial noun list, whose L_2 distance between its embedding and w's embedding e_w is no less than $d/2$, shown in Equation 17,*

$$\|e_i - e_w\|_2 \geq d/2, \quad \exists e_i \in \{e_1, ..., e_N\}. \tag{17}$$

Proof. If the L_2 distance between e_w and any embedding e_i in the noun list is always smaller than $d/2$, i.e, e_w is no less than $d/2$, shown in Equation 18,

$$\|e_i - e_w\|_2 < d/2, \quad \forall e_i \in \{e_1, ..., e_N\}. \tag{18}$$

Then we can randomly choose e_{i_1} and e_{i_2} from the noun list, and get the Equation 19,

$$\|e_{i_1} - e_w\|_2 + \|e_{i_2} - e_w\|_2 < d. \tag{19}$$

From the triangle inequality, we have Equation 20,

$$\|e_{i_1} - e_{i_2}\|_2 \leq \|e_{i_1} - e_w\|_2 + \|e_{i_2} - e_w\|_2. \tag{20}$$

And we get Equation 21,

$$\|e_{i_1} - e_{i_2}\|_2 < d, \tag{21}$$

which contradicts the given condition and the proposition is proved. □

From the proposition, one can see that the distance between the user's prompt's embedding and the textual embedding of the candidates in our adversarial noun list is lower bounded by the minimum distance between the textual embedding of two candidates. Thus, in our experiments, we manually select several totally different nouns in the adversarial noun list to ensure the quality of our ZeroFake.