DE-FAKE: Detection and Attribution of Fake Images Generated by Text-to-Image Diffusion Models

Zeyang Sha[1] Zheng Li[1] Ning Yu[2] Yang Zhang[1]

[1]*CISPA Helmholtz Center for Information Security* [2]*Salesforce Research*

Abstract

Diffusion models emerge to establish the new state of the art in the visual generation. In particular, text-to-image diffusion models that generate images based on caption descriptions have attracted increasing attention, impressed by their user controllability. Despite encouraging performance, they exaggerate concerns of fake image misuse and cast new pressures on fake image detection. In this work, we pioneer a systematic study of the authenticity of fake images generated by text-to-image diffusion models. In particular, we conduct comprehensive studies from two perspectives unique to the text-to-image model, namely, visual modality and linguistic modality. For visual modality, we propose *universal detection* that demonstrates fake images of these text-to-image diffusion models share common cues, which enable us to distinguish them apart from real images. We then propose *source attribution* that reveals the uniqueness of the fingerprints held by each diffusion model, which can be used to attribute each fake image to its model source. A variety of ablation and analysis studies further interpret the improvements from each of our proposed methods. For linguistic modality, we delve deeper to comprehensively analyze the impacts of text captions (called *prompt analysis*) on the image authenticity of text-to-image diffusion models, and reason the impacts to the detection and attribution performance of fake images. All findings contribute to the community's insight into the natural properties of text-to-image diffusion models, and we appeal to our community's consideration on the counterpart solutions, like ours, against the rapidly-evolving fake image generators.

1 Introduction

With the superpower of harmonizing imaginative and controllable content creation, text-to-image generation that takes a natural language description as input and generates an image to match that description, has become one of the most exciting topics in recent years. Especially in the past few months, the outputs of state-of-the-art text-to-image models, such as OpenAI's DALL-E2 [31], Google Brain's Imagen [37], and StabilityAI's Stable Diffusion [35] are approaching the authenticity of real photographs and professional the hand-drawn art.

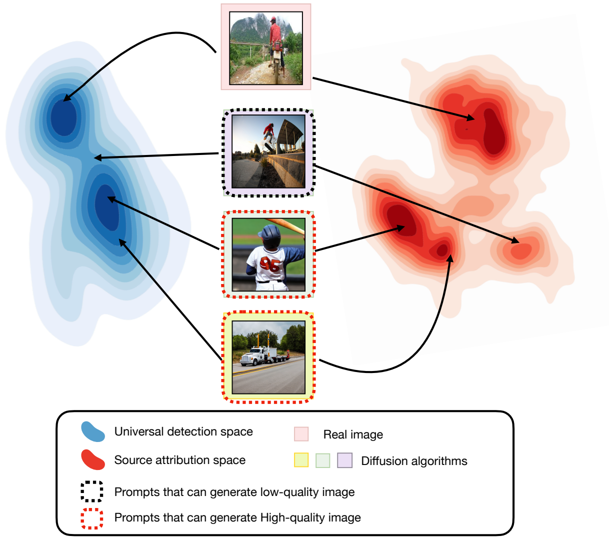


Figure 1: An illustration of universal detection, source attribution and prompt analysis, as well as a high level overview of DE-FAKE.

Text-to-image generative models, despite their extraordinary creativity, can be easily misused for malicious purposes, such as the misinformation and disinformation of their generated contents [3, 9, 34]. For instance, fake images can be maliciously used in legal cases (falsifying electronic evidence) or politics (misleading the public). As recent news VICE[1] reported, AI-generated artworks impersonated the styles of famous artists and even won the first place in an art competition, which raised serious complaints from the involved artists. Such fake images pose a huge threat to the copyrights of existing artworks, leading to widespread of unfair competition.

To mitigate the aforementioned concerns, researchers establish fake detection models to distinguish generated images from real ones. To the best of our knowledge, existing works [21, 24, 36, 51] solely focus on traditional generative models (e.g., GANs [28]), leaving text-to-image generation models unexplored.

In this paper, we pioneer to conduct the first systematical study on visual forgeries from text-to-image generation models. In particular, our goals are in two thrusts as shown in Figure 1. First, we aim to distinguish diffusion-based fake im-

[1] https://www.vice.com/

ages apart from real images. Second, we aim to further track the source model of each fake image. The two goals jointly interpret the common cues of fake images and evidence the security risks from the rapidly-developing diffusion models.

To this end, we conduct sophisticated studies from two perspectives specific to text-to-image diffusion models, namely, visual modality and linguistic modality. For the visual modality, we investigate the existence and universality of artifacts shared between different models, as well as the uniqueness and visualization of fingerprints held by diffusion models. For the linguistic modality, we study the impact of its semantics and structures on the authenticity of generated images. We therefore formulate to investigate the following research questions (**RQ**s).

- **RQ1: Existence and Generality.** Do there exist common features or artifacts shared across different diffusion models?
- **RQ2: Uniqueness and Visualization.** Does each diffusion model enjoy its unique fingerprint [4, 19, 24] that differentiates itself from other diffusion models? And how to expose these unique fingerprints?
- **RQ3: Prompt Analysis.** Which kind of semantics and structures of prompts can generate more real\fake images?

Methodologies. To answer the above three research questions, we first leverage the existing methods as baselines and then propose a novel method to compensate for the disadvantages of the baselines.

Technically, to respond to **RQ1**, we propose *universal detection* to expose common artifacts of fake images. Our *universal detection* learns to distinguish apart fake and real images, and generalize towards open-world unseen fake images. We adopt two different methods: image-only detection and hybrid detection. Image-only detection shares the spirit with previous work [44] that validates the efficacy of a simple CNN to distinguish apart GAN's fake images and real images. In this work, we consider the most challenging and realistic scenario where we train a simple CNN on fake images generated by only *one* model and then evaluate it on fake images generated by *many* unknown models. Hybrid detection is a brand-new detection method where text information is non-trivially incorporated into learning. We take advantage of CLIP image and text encoders [30] and concatenate the two embeddings to train a real/fake binary classifier. In the evaluation phase, if the natural texts of images are not available, we leverage BLIP [18] to generate captions for us and then feed the images and texts to the hybrid detection.

What CLIP is used for?

To respond to **RQ2**, we propose *source attribution* by adopting two methods similar to **RQ1** to validate that fake images from different models contains distinct fingerprints. Different from **RQ1**, in this stage, our multi-label classifiers are trained on fake images from multiple diffusion models, each model of which corresponds to one label. The image-only attribution takes solely images as input, whereas the hybrid attribution concatenates image embeddings with text embeddings.

Different from **RQ1** and **RQ2**, **RQ3** focuses on the linguistic modality. To respond to these questions, we reuse the detectors we have pretrained above. More concretely, we conduct a comprehensive evaluation from two angles of linguistics, namely semantic analysis and structure analysis. In the former, we leverage two different topic extraction methods to analyze the impact of topic on the authenticity of fake images. Both two topic extraction methods facilitate us to reach the conclusion that prompts describing "person" generate more realistic images. For sentence analysis, we study the impact of prompt captions based on sentence length and the proportion of nouns in the sentence, respectively. Empirically results demonstrate that 25 to 75 is the best length for generating high-authenticity fake images while the proportion of nouns in the sentences has no impact on the authenticity of fake images. Our findings can help the developers and researchers who use the text-to-image diffusion model to generate better fake images, thereby further helping the community understand the natural properties of the fake images generated by the text-to-image diffusion model and design methods with higher detection capabilities.

Main Discoveries. The general discoveries and the answers to the research questions can be summarized as follows.

- **Discovery I.** There indeed exist common artifacts across different diffusion models. Furthermore, using text information in the proposed hybrid detection better benefits the detector to locate these common artifacts and achieves better performance, especially encountering open-world fake images.
- **Discovery II.** Different diffusion models indeed generate distinctive fingerprints together with their images, which allow us to attribute fake images to their sources. We further visualize these unique fingerprints using frequency spectra inspired by Zhang et al. [51].
- **Discovery III.** For regular sentences, their structures does not affect the authenticity of the generated image, whereas extreme structures tend to generate low-authenticity images. Orthogonally, semantics matter the image authenticity. For example, descriptions about "person" often produce higher-authenticity images.

Contributions. By answering the three general research questions above, we cast the following contributions:

- We pioneer systematically assessing the authenticity of fake images generated by text-to-image diffusion models from two perspectives, i.e., visual modality (*universal detection* and *source attribution*) and linguistic modality (*prompt analysis*).
- For visual modality, we design two types of detectors, i.e., the image-only detector that only utilizes fake images and the hybrid detector that utilizes both the fake image and its corresponding text caption. Extensive evaluation on four state-of-the-art text-to-image diffusion models and two benchmark datasets show that the fake images generated by these models share common visual artifacts that allow us to distinguish them from

real images, and further reveal the uniqueness of the fingerprints held by each diffusion model that can be used to attribute each fake image to its model source.

- For linguistic modality, we delve deeper to comprehensively analyze the impacts of text captions on the image authenticity of text-to-image diffusion models and reason the impacts on the performance of fake image detection. Our findings thereby can help the community understand the natural properties of the fake images generated and design methods with higher detection capabilities.

2 Preliminaries

2.1 Text-to-Image Diffusion Models

In this work, we focus on text-to-image diffusion models. Text-to-image diffusion models always takes a prompt and a random noise as the input and then denoise the image under the guidance of prompt information to generate the desired image.

To collect fake images from different diffusion models, we leverage the following pretrained models available online.

- **Stable Diffusion [35]:** Stable Diffusion is a latent text-to-image diffusion model. The available models are pre-trained on 512x512 images from a subset of the LAION-5B [39] dataset. CLIP's [30] text encoder is used to condition the model on text prompts.

- **Latent Diffusion [35]:** Latent Diffusion is also a latent text-to-image diffusion model. It is pre-trained on LAION-400M, a smaller dataset compared to LAION-5B. It also takes advantage of CLIP's text encoder to guide the direction of fake images.

- **GLIDE [29]:** GLIDE is text-to-image diffusion model released by google. The available GLIDE model is trained on a filtered version of a dataset comprised of several hundred million text-image pairs. The online model is not good at understanding texts containing people because such images are removed from the train dataset for ethical issues.

- **DALLE2-PyTorch [31]:** DALLE2-PyTorch is an unofficial implementation of DALL-E 2. It follows the DALL-E 2 paper's algorithm to use an extra layer of indirection with the prior network. It is trained on a partial LAION dataset.

2.2 Datasets

We use the following two text-image datasets to conduct our experiments. Text from the datasets will be used to generate fake images. Then we compare fake images and real images from the same text to study whether we can tell them apart.

- **MSCOCO [23]:** MSCOCO dataset is a large-scale objective, segmentation, and captioning dataset. We will leverage its image-caption pair to construct fake images. MSCOCO's images contain 80 classes.

Table 1: Fake images settings. We generate fake images using four available pre-trained models and two datasets.

Model	Dataset	Images	Image size
Stable Diffusion	MSCOCO	59247	512*512
	Flickr30k	13231	512*512
Latent Diffusion	MSCOCO	29276	256*256
	Flickr30k	17969	256*256
GLIDE	MSCOCO	41685	256*256
	Flickr30k	27210	256*256
DALL-E2	MSCOCO	1028	256*256
	Flickr30k	300	256*256

- **Flickr30k [46]:** Flickr30k dataset contains 31,783 images and 158,915 English captions on various of scenarios.

In summary, the fake images we have collected are listed in Table 1. Note that since different available models are trained on different size images, and in real-world settings, fake images always appear in different resolutions, we generate fake images of different sizes based on the used model's default setting. Also, DALLE2-Pytorch only has 1328 fake images due to its generating process being much slower than the above three.

3 Fingerprint Learning for Visual Modality

We present our work addressing these two **RQ**s respectively by learning the fingerprints of diffusion model for image attribution:

- (**RQ1**) We introduce diffusion model artifacts and exploit them to differentiate generated images of diffusion models from real ones. We denote this task as *universal detection*.

- (**RQ2**) For generated images from diffusion models, we further attribute their sources by learning fingerprints unique to each model. We denote this task as *source attribution*.

For both *universal detection* and *source attribution*, we approach them by constructing a machine learning classifier and predicting the source of an image. More specifically, for *universal detection*, we build a binary classifier to predict whether the image is real or generated. For *source attribution*, we build a multi-classifier to predict from which diffusion model the image comes.

3.1 Universal Detection

Here, we present the first study, i.e., *universal detection*, addressing **RQ1** as aforementioned, we start by introducing our design intuition and goals. Then we describe how to build the binary classifier. Finally, we present the evaluation results.

3.1.1 Design Intuition

We derive our intuition behind the universal detection from the common features or artifacts that are (probably) shared among different diffusion models. Since some features or artifacts shared across different GAN models have been verified in previous work [10, 13, 44, 50, 51]. We believe that diffusion models also hold the same or similar property as GAN models, there are also some shared features or artifacts among different diffusion models. To verify it, we perform a simple yet promising research approach where we investigate whether it is possible to perform forensic detection on images from the *one* diffusion model that generalizes to images from *many* unseen models.

3.1.2 Design Goals

As we aim to differentiate the images from the *one* diffusion model that generalizes to images from *many* unseen models, we should achieve the following goals:

- **Differentiate Between Generated and Real.** The primary goal of *universal detection* is to differentiate the images generated by diffusion models from real ones. This can verify whether there exist fingerprints in diffusion models, and such fingerprints can be reflected in their corresponding generated images, which are different from real images.

- **Agnostic to Algorithms and Datasets.** Since there are and will be more and more diffusion algorithms and datasets to generate fake images, it is not possible to construct a detector for each diffusion algorithm and dataset to distinguish the generated from the real ones. Therefore, it is crucial to explore whether we can train universal detectors that are independent of diffusion algorithms and datasets, i.e., the detector can distinguish between real images and images generated by unknown diffusion algorithms trained on unknown datasets. Note that this is a more realistic and challenging scenario. Furthermore, this can verify whether there are some common features and artifacts shared across different diffusion models.

3.1.3 Fingerprint Learning

In this section, we present two different methods based on different background knowledge exploited, i.e., image-only detection that only exploits visual output by diffusion models, as well as hybrid detection that exploit both visual output and linguistic prompts. See Figure 2 for an illustration of how to conduct the fingerprint learning.

Image-Only Detection. Following the previous work [44], we first explore whether we can train a simple binary ML classifier to differentiate generated images from real images. As we mentioned that the universal detection should be agnostic to diffusion algorithms and datasets. Thus, we only choose the generated images of one diffusion model trained on one certain dataset, as well as the corresponding real images to build the binary classifier.

As shown in the pink part of Figure 2, our image-only detection consists of the following stages.

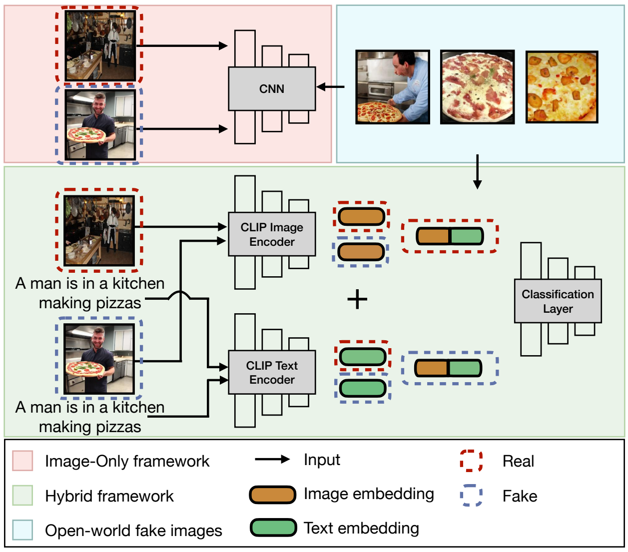


Figure 2: Illustration of universal detection framework. The red part is the image-only framework. The green part is the hybrid framework. The blue part is open-world fake images.

1. Randomly selecting 20,000 prompt captions from MSCOCO dataset to query the stable diffusion model to generate fake images.

2. Labelling all fake images as 0 and the real images of the same size from MSCOCO dataset as 1, i.e., 40,000 in total.

3. Training the binary ML classifier with 0-fake and 1-real images in conjunction with classical training techniques.

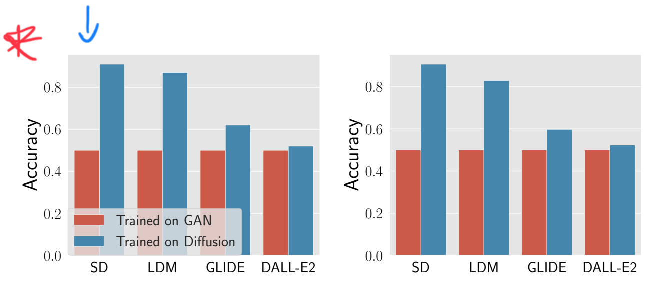
4. Evaluating trained classifiers to distinguish unseen fake images generated by unknown diffusion models (e.g., LDM and DALL-E2) from real images.

The reason why we choose stable diffusion is that stable diffusion can generate more realistic pictures, compared to other models we found in the open-source community. This is actually more challenging for us to train a good classifier. Note that we also choose other diffusion models and datasets to train our detector and then conduct the evaluation.

Hybrid Detection. In addition to the visual output of the diffusion model that we utilize, there are also linguistic prompts unique to the diffusion model, which are different from the traditional domain. Usually, it is not possible to obtain cue captions that faithfully and fully reflect the real image content. In other words, given the prompts input to the diffusion model, the visual output may reflect the content of the prompts more faithfully than the real caption-image pairs in the dataset. Such disparity constitutes our intuition for hybrid detection, which exploits visual output and linguistic input to build ML classifiers.

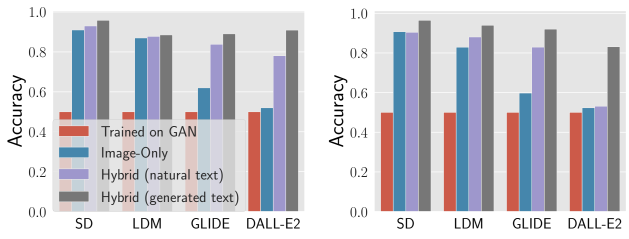
The green part of Figure 2 shows the work pipeline of our hybrid detection. The general pipeline is same as the that of image-only detection. The only one difference is in step 3:

the detector is only trained here



(a) MSCOCO (b) Flickr30k

Figure 3: The performance of previous GAN-based detectors and image-only detectors. Detectors are tested on MSCOCO(a) and Flickr30k(b).



(a) MSCOCO (b) Flickr30k

Figure 4: The performance of image-only detectors and hybrid detectors. Detectors are tested on MSCOCO(a) and Flickr30k(b).

we leverage CLIP's image encoder and text encoder as feature extractors to get embeddings for both images and text. Then, we concat these two embeddings together as a new embedding and use this embedding to train our linear classifier. In this case, our classifiers can learn from image information as well as text information. Note that we leverage CLIP trained on ViT-B/32 as our feature extractors.

To evaluate the well-trained hybrid detector, we also need both images and captions to query it. In a real-world scenario, users are likely to attach a caption to the image they post on the Internet. Therefore, for this scenario where the detector can acquire the captions of the images during the evaluation phase, we call it hybrid detection ("natural text"). In our experiments, we adopt the real prompt caption from datasets to conduct the evaluation.

In a more realistic and challenging scenario where the detectors cannot get the natural prompts, we propose a simple yet effective method to generate the prompt caption, which is named hybrid detection ("generated text"). Concretely, we leverage the BLIP [18] model to generate captions for us and then use this caption as the natural caption to train our hybrid detectors. BLIP is a captioner which can generate synthetic captions. In this way, we can still get the text information and our further experiments demonstrate that detectors trained on natural captions can also work well on the generated captions. In our work, we leverage BLIP trained on ViT-B and CapFilt-L as our captioner.

In a word, we consider two scenarios in the evaluation phase based on whether the detector can get the natural texts. We show in the result section that our detector works well in both scenarios.

3.1.4 Results

In this section, we present the detection performance of our proposed image-only detection and hybrid detection.

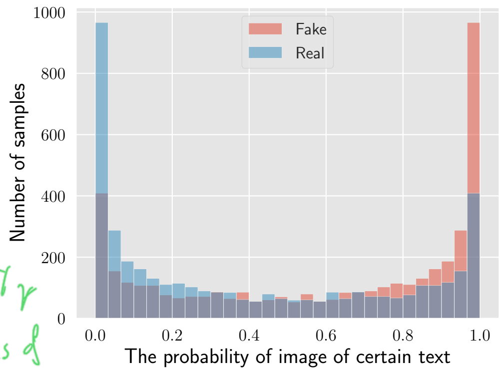
Image-Only Detection. For a convincing evaluation, we adopt existing work [44] that detects visual forgeries in the GAN domain as a comparison. Note that the GAN-based detector is trained on GAN-generated images and achieve close 0.98 performance on other unknown GAN algorithm [5, 16, 17, 36, 52]. Figure 3 depicts the performance of GAN-based detector on fake images generated by diffusion models. We can observe that most diffusion-generated fake images cannot be detected by the GAN-based detectors. In all cases, the detection can only achieve an accuracy score of 0.5, which means all fake images are classified as real images. Therefore, it can be concluded that there is a huge gap between the GAN model and the diffusion model, leading to the complete inability of the GAN-based detector to handle the fake images generated by the diffusion model. These results also indicate that it is crucial to explore the properties of diffusion models for image attribution.

Furthermore, we can see that the image-only detector achieves better performance on all models and datasets compared to the GAN-based detector. For example, the detector can achieve accuracy score of 0.597 in detecting fake images from the GLIDE trained by Flickr30k. Note that the detector here, i.e., the binary ML classifier, is only trained on fake images generated by stable diffusion trained on MSCOCO. We further build detectors on other models and datasets in Section 3.1.6. These results show that the fake images generated by the diffusion model do have some common properties, which enable detectors based on one diffusion model to distinguish between fake images generated by an unknown diffusion model and real images. These results also verify the existence and generality of the fingerprint of diffusion models.

Lastly, another interesting finding is the much larger variation in detection performance under the effect of the model compared to the effect of the datasets. For example, in Figure 3a, the detector scores 0.913 accuracy on stable diffusion, compared to 0.526 on DALL-E2. In contrast, the detector achieves very close accuracy across different datasets on all diffusion models. The results indicate that the diffusion model indeed plays a key role in leaving fingerprints in the generated images.

However, though the image-only detector achieves better performance on all models and datasets compared to the GAN-based detector, we acknowledge that the current detection performance is far from the design goal due to the poor performance on GLIDE and DALL-E2. As mentioned earlier, text-to-image generation with text as input may provide new signals to boost detection performance, which motivates us to perform the following hybrid detection.

Hybrid Detection. We report the accuracy score of our hybrid detection in Figure 4. We report our exact numbers in Section 7.1. We can find that the linguistic information, i.e.,



The probability is computed using distances of embeddings

Figure 5: The distribution of closeness of real\fake images to the prompts. The X-axis represents the probability that the image sample matches the given text description. The Y-axis represents the number of samples.

(a) Real Painting **(b) Fake Painting**

Figure 6: Real and fake images examples on "the scream."

prompt caption, significantly improves the detection performance on all diffusion models and datasets. For instance, hybrid detectors with real captions achieves accuracy score of 0.909 on DALL-E2, which is much higher than the accuracy score of 0.522 achieve by image-only detector. Moreover, we can find that the detection performance is less affected by different models than image-only detection. We also conduct an evaluation by relaxing the assumption that we can get the real captions of the images. As shown in the green bars, we can observe that our hybrid detector still can achieve remarkable performance compared to image-only detector.

Our extensive evaluation convincingly demonstrates that our proposed hybrid detection can achieve the design goals of universal detection. All these above results verify the existence and generality of fingerprints of diffusion models.

3.1.5 Discussions

The above results fully demonstrate the effectiveness of our universal detection. Here we delve more deeply into the reasons for successfully distinguishing fake images from real ones and why additional linguistic information can improve universal detection performance.

Why Does Linguistic Information Enhance Detection Performance. We conduct studies here to further investigate why linguistic information, i.e., prompt caption, enhance helps us distinguish between real and fake images. For each prompt caption, we collect its corresponding real image and fake image generated by Stable Diffusion. Then, we use CLIP's text encoder and image encoder to compute the similarity between the prompt caption and the real/fake respectively. In our experiments, we adopt 2,000 texts by randomly sampling from MSCOCO dataset. Figure 5 show the similarity distribution. We can see that the distance between the fake image and the text information is closer than the real image, leading to a clear difference between their loss distribution. This verifies our intuition aforementioned, given the prompts input to the diffusion model, the visual output reflects the content of the prompts more faithfully than the real caption-image pairs in the dataset. Furthermore, we can also conclude that it is not the linguistic information itself that enhances the performance of the detector, but the linguistic information can be exploited as an extra "anchor" to provide a new measure to distinguish between true and false images.

Case Study of Artwork. All the above experiments are performed on distinguishing real fake images from real photos. Our experimental results show that we can train hybrid detectors to achieve excellent performance on these datasets. However, as we stated in the introduction, the misuse of diffusion models in art competitions already pose serious security concerns. Therefore, it is crucial for us to evaluate whether our detector is able to distinguish between real and fake paintings.

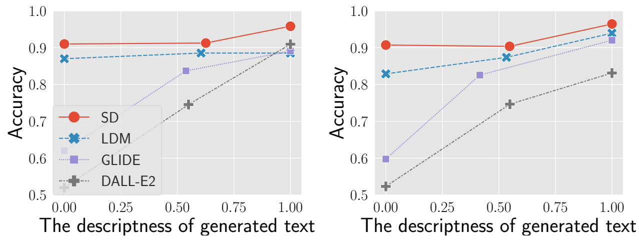
Since there is no existing dataset on paintings and their captions, we collect only 50 real artworks from the Internet and 50 works generated by the stable Stable Diffusion, which are then fed to the previous detector. In addition, paintings are always caption-less. Therefore, we evaluate these artworks using only an image-only detection and a hybrid detection with generated text. Our experiments show that the image-only detector can achieve an accuracy of 0.710 and the hybrid detector can achieve 0.690. The results show that even with different topics and different styles (e.g., artworks and photos), fake images still have some common properties.

As an additional discovery, we also found that stable diffusion is able to fully replicate some paintings knowing only the name of the painting, which we show in Figure 6, which also proves the feasibility of membership inference attacks [15, 20, 22, 41] and model stealing attacks [40, 42] on the diffusion model.

3.1.6 Ablation Study

Impact of Generated Prompt Authenticity. In hybrid detection(generated text), when the natural captions of the image are not available to the detector, we utilize the BLIP as a caption generator to generate the prompt caption and then use these generated prompt caption to build the detector.

Here, we explore whether the authenticity of generated prompts affects the performance of detector. In particular, we leverage a new term called prompt descriptiveness [8, 11, 26, 38] to evaluate the authenticity of prompt by computing the similarity between prompt embeddings and image embeddings. The descriptiveness of one dataset is the mean



(a) MSCOCO **(b) Flickr30k**

Figure 7: The impact of generated texts' descriptiveness on detectors' performance.

of the descriptiveness of each text. Figure 7 depicts the detection performance in terms of the authenticity of generated prompts. We can see that the authenticity of the prompts greatly affects the performance of the detector. For instance, when the descriptiveness is 0, i.e., no prompts are used as the input, the detector can only achieve the accuracy score of 0.621 on GLIDE MSCOCO, while the performance can be improved to 0.838 with the generated text with descriptiveness of 0.538. In addition, it also can be concluded that we can easily obtain the generated prompts by BLIP that have good descriptiveness among all datasets and algorithms.

Impact of Training Dataset Size. In this section, we explore the impact of the training dataset's size on the performance of detector. More concretely, for each diffusion model, we build the detector based on Stable Diffusion by varying the size of training dataset from 500 to 40000 (half is real, half is fake). Note that the defaults size we adopt in previous evaluation is 40000.

We report the detection performance in terms of the size of training dataset in Figure 8. As expected, the performance of different detectors indeed is affected by the size the training dataset, and the general trend is that all the detectors performs better with the increase of the training dataset's size. For instance, as the performance of hybrid detection shown in Figure 8b, when the training dataset size is 1000, the detector can achieve 0.792 while the accuracy can be improved to 0.885 when the training dataset size is 40000. More encouragingly, we can also find that the hybrid detector achieves significant performance even with a small number of images of only 500 images, which is much fewer compared to 40,000 images. For example, in Figure 8a, the detection performance is 0.830 with only 500 training data, while the detection performance is 0.958 with 40,000 training data. Lastly, we again find that the hybrid detection achieves much better performance than image-only detection, as shown previously. For example, in Figure 8c, the hybrid detector achieves 0.714 accuracy with 500 training data, while the vision-only detector achieves only 0.523 with 40,000 training data. This again demonstrates the advantages of introducing linguistic information in the detection process.

Impact of the Diffusion Models for Building Detector. As aforementioned, we chose stable diffusion to build all our detectors, and empirical results show that the detector can generalize to unknown diffusion models. In this section, we explore whether the detection performance is affected when the detector is built on different diffusion models.

Concretely, we adopt the fake images generated by GLIDE as training data to build the detectors and compare the detection performance with detectors from Stable Diffusion. Our results are shown in Figure 9. We can observe that both image-only detector and hybrid detector can perform well on fake images from the same algorithm as the training data source, but fail on the other algorithms. The overall performance is worse than that shown in Figure 4, where the detector is built on the stable diffusion model. Especially, the detectors built on GLIDE work poorly on stable diffusion model. These results suggest that the higher-authenticity fake images generated by stable diffusion model can boost the hybrid detector to achieve better generalization over other models, especially for hybrid detection.

3.1.7 Takeaways

In summary, to answer **RQ1**, we propose both image-only and hybrid detection to differentiate diffusion-based fake images apart from real ones. Extensive evaluations show that there indeed exists some common features and artifacts among different diffusion models. Besides, with the linguistic information, the hybrid detector can achieve great performance compared to image-only detector. Overall, these empirical results verify the existence and generality of fingerprints of diffusion models.

3.2 Source Attribution

After demonstrating the existence and generality of fingerprint of diffusion models, Here, we present the second study i.e., *source attribution*, addressing the **RQ2**. We start by introducing our design intuition and goals. Then we describe how to build the binary classifier. Finally, we present the evaluation results.

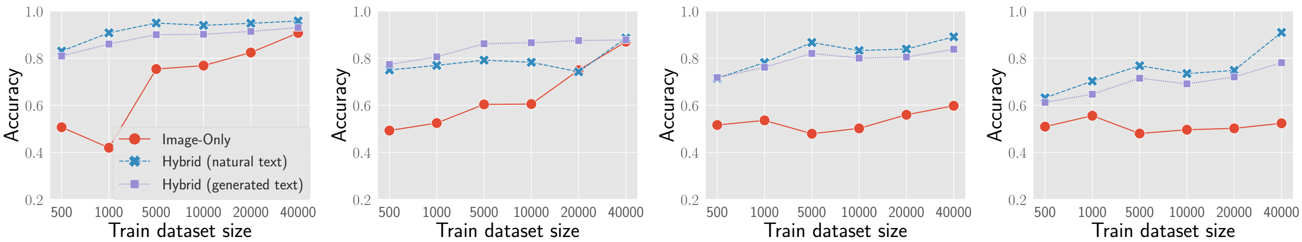
3.2.1 Design Intuition

We derive the intuition behind our *source attribution* from previous works [14,25,48,49] on GANs suggests that GAN-based fake images always carry fingerprints to indicate their algorithms. Therefore, we believe that diffusion-based fake images also have the same behavior, i.e., diffusion models usually carry different fingerprints and leave stable fingerprints in their generated images. To verify it, we propose *source attribution*.

3.2.2 Design Goals

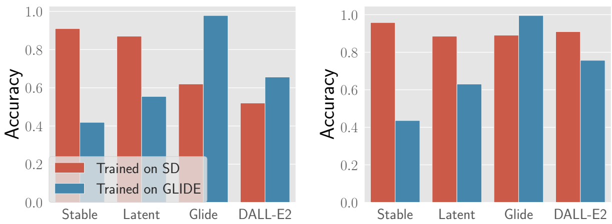
Similar to Section 3.1, we here explain our goal of source attribution.

- **Tracking Original Algorithms.** The primary goal of *source attribution* is to track the fake images generated by which diffusion algorithm. This can verify whether each diffusion model enjoys its unique fingerprint that is different from other diffusion models.



(a) SD **(b) LDM** **(c) GLIDE** **(d) DALL-E2**

Figure 8: The performance of universal detectors under different training dataset sizes on MSCOCO.



(a) Image-Only **(b) Hybrid**

Figure 9: Performance of training fake images from different diffusion algorithms.

- **Agnostic to Datasets.** In real-world settings, fake images are always generated by prompts from different distributions. To be more universal, our designed detector should be agnostic to datasets.

- **Tracking Unknown Algorithms.** Furthermore, it is not possible to collect all the diffusion algorithms to generate fake images to build our detector, thus, a more realistic yet challenging scenario is that a detector built on a limited scope of diffusion can track an unknown diffusion algorithm and easily add it by retraining.

3.2.3 Fingerprint Learning

Here, we present two methods designed to conduct *source attribution*: image-only attribution and hybrid attribution. Similar to *universal detection*, the former only exploits visual output by diffusion models, and the latter exploits both visual output and linguistic prompts. See Figure 2 for an illustration of how to conduct fingerprint learning.

Image-Only Attribution. We first explore whether we can train a simple ML multi-classifier to track the source of generated images. More concretely, we adopt a limited range of diffusion models to generate fake images to build our detector. Besides, as we mentioned that the *source attribution* should be agnostic to datasets, these diffusion models we adopted are only trained on one dataset.

Our image-only tracking contains the following steps.

1. Randomly selecting 20,000 prompts from MSCOCO to query SD, LDM, and GLIDE, respectively. We also collected 20,000 real images from MSCOCO corresponding to the above prompts.

2. Labelling fake images according to their source algorithms and real images.

3. Training the multi-class classifier with the labeled data.

4. Evaluating trained multi-classifier to predict the source of a given image, which may belong to SD, LDM, and real.

Note that in this section, DALL-E2 is considered an unknown algorithm because it is too slow to generate fake images. We adopt ResNet18 to build our detector.

Hybrid Attribution. Previous evaluations in *universal detection* have shown that the linguistic information does greatly enhance the distinction between real and fake images, therefore we also investigate whether the linguistic information improves the performance of *source attribution*, i.e., hybrid tracking.

The hybrid attribution is quite similar to the above hybrid detection framework. In the training phase, we use the natural captions and their corresponding fake images to train a classifier containing the clip's image and text encoder as well as a classification layer. Then, in the testing phase, we consider two possible scenarios. In hybrid(natural text), we assume the tracker can get the natural text, which is always posted together with images by users. In hybrid(generated text), we assume the tracker only has images. In this scenario, we leverage BLIP to generate text for us and then use the generated texts and images to test the given images.

3.2.4 Results

In this section, we present the performance of our proposed two types of *source attribution*.

Image-Only Attribution. We report the performance of image-only source attribution in Table 2. We can find that our proposed image-only source attribution achieves remarkable performance. For example, when these evaluated diffusion models are trained on MSCOCO, i.e., the evaluated diffusion models are trained on the same dataset as the trained diffusion models to build our image-only detector, our detector can achieve an accuracy score of 0.815, a very high detection performance. These results verify that the different diffusion models indeed enjoy their unique fingerprint that is different from other diffusion models, leading to remarkable performance of our image-only detector.

Table 2: **Performance of image-only attribution and hybrid attribution.**

	MSCOCO	Flicker30k
Image-only	0.815	0.727
Hybrid (natural text)	0.880	0.834
Hybrid (generated text)	0.850	0.765

Further, we can also find that when these evaluated diffusion models are trained on Flicker30k, i.e., the evaluated diffusion models are trained on the different dataset as the trained diffusion models to build our image-only detector, our detector still achieve an accuracy score of 0.727, which is a fairly high detection performance. Note that the random guess for the 4-class classification task is only 0.25. These results indicate that our image-only detector is agnostic to datasets.

Hybrid Attribution. Table 2 shows the performance of our proposed hybrid detection. We can clearly see that the hybrid detector achieves much better performance than the image-only detector, regardless of datasets, which again indicates that the linguistic information indeed can enhance the detection performance.

Moreover, these empirical results demonstrate once again that, in addition to sharing some common properties, these fake images from different algorithms also have their own special properties, which allow us to keep track of their own algorithms, i.e., verifying the fingerprint unique to the diffusion model.

3.2.5 Discussions

The above evaluations convincingly demonstrate the effectiveness of our source attribution. Here, we explore how to expose the fingerprints of diffusion models and evaluate the adaptation to unknown diffusion algorithms.

Fingerprint Visualization. In the previous experiments, we quantitatively demonstrate that each diffusion model enjoys its unique fingerprint and such fingerprint can be reflected in their generated images. Here, we aim to verify the uniqueness of fingerprints qualitatively.

Inspired by Zhang et al. [51], we draw the frequency spectra of different diffusion models built on MSCOCO in Figure 10. For each diffusion algorithm, we randomly select 2,000 fake images and then calculate the average of their Fourier transform outputs. The reason we adopt Fourier transform is that Fourier transforms typically provide more invisible information about images.

As shown in Figure 10, we can clearly observe that there are distinct patterns visible in images generated by different diffusion models, especially in GLIDE and DALL-E2. We can also find that the frequency spectra of stable diffusion is similar to that of latent diffusion, which can explain why detectors built on stable diffusion can also achieve very high performance on latent diffusion. The reason behind this is that they follow similar algorithms, although trained on different datasets. Their qualitative evaluations verify again that each diffusion model enjoys its unique fingerprint.

Adaptation to Unknown Algorithms. As aforementioned, one of our design goals is to identify fake images generated by unknown diffusion algorithms. Therefore, we explore how to adapt our detector to the unknown algorithm.

To this end, we propose a simple yet effective approach named confidence-based prediction. The core idea is to classify the unconfident samples from the detector prediction, i.e., lower than pre-defined confidence, as unknown algorithms. In our experiment, we set the threshold as 0.9 and Table 3 shows the results. As we can see, the performance of our detector drops but stays in a stable range when the test data contains unknown algorithms. Moreover, we can still conclude that hybrid detection can always achieve better performance than the image-only detector in both settings. These results indicate that with a simple modification, our detection can adapt to unknown algorithms easily and can achieve not-so-bad performance.

Table 3: The performance of image-only framework and hybrid framework when encountering unknown data.

	Unknown	MSCOCO	Flicker30k
Image-Only	✗	0.864	0.863
	✓	0.815	0.727
Hybrid (natural text)	✗	0.936	0.933
	✓	0.880	0.834
Hybrid (generated text)	✗	0.903	0.892
	✓	0.850	0.765

3.2.6 Ablation Study

Impact of Training Dataset Size. Both image-only detectors and hybrid detectors achieve remarkable performance in tracking the source algorithm of fake images. Here, we explore the effect of the training data size for constructing the detector on the detection performance.

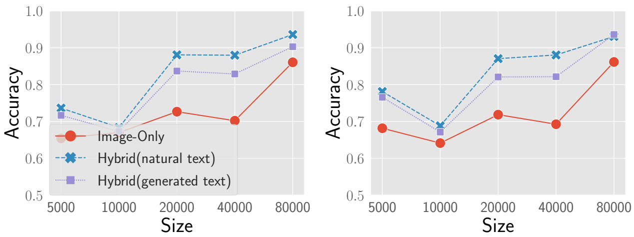
We report the experimental results in Figure 11. We can see that the size of training data indeed has a great influence on detection performance. For example, when training our detector with only 5000 fake images, even the hybrid detector can only achieve 0.736 accuracy, while it can achieve 0.936 accuracy when fed with 80,000 fake images. Besides, we can find that hybrid detectors require less data for convergence compared to image-only methods. For example, hybrid detection achieves a huge performance improvement between 10,000 and 20,000 training size, while this improvement occurs in the 40,000 to 80,000 training size range for the CNN-based approach. From this phenomenon, we can conclude that the hybrid approach requires less data and always achieves good performance even with a small amount of training data.

3.2.7 Takeaways

In summary, to answer **RQ2**, we propose image-only detection and hybrid detection to track the source of fake im-

(a) SD **(b) LDM** **(c) GLIDE** **(d) DALL-E2**

Figure 10: Visualization of frequency analysis on different algorithms' fake images.



(a) MSCOCO **(b) Flickr30k**

Figure 11: The performance of the source detectors under different sizes of training data.

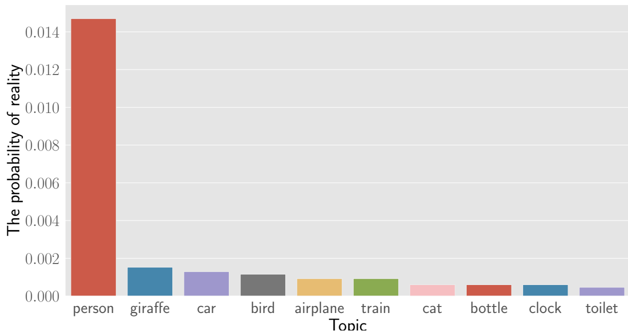


Figure 12: Probability distributions of different topics that can generate high-authenticity images.

ages. Extensive evaluation demonstrates that different diffusion models indeed enjoy unique fingerprint that is different from others, i.e., verifying the uniqueness of the fingerprints held by the diffusion model.

4 Prompt Analysis for Linguistic Modality

One of the most significant differences between text-to-image diffusion models and traditional GANs is that the diffusion models take natural language descriptions as input and generate matching images, i.e., the diffusion models consist of visual and linguistic modalities. The above studies mainly focus on the visual modality; in this section, we shift our attention from the visual modality to the linguistic modality. Here, we focus on addressing our previously mentioned **RQ3**:

- (**RQ3**) Do the semantics and structure of the prompts affect the authenticity of the generated image, thus affecting the detection performance?

To answer this question, we perform a comprehensive study from two perspectives of linguistics, namely semantic analysis, and structural analysis. In this section, we first introduce the definition of images' authenticity. Then we conduct the semantics and structure analysis respectively to explore their influence on images' authenticity.

4.1 Authenticity of Images

Before we start, we will first make the definition of image authenticity from the perspective of fake image detection. Ideally, the image authenticity should reflect whether the given images are more realistic or not. In this work, we adopt the binary detector we designed in *universal detection* as a metric to assess the authenticity of given images. We extract the probability corresponding to the real image as the authenticity metric. Note that, in this section, we aim to analyze the impact of prompts on image authenticity. Based on the principle of control variables, we just use image-only detector as our metric, without text information.

4.2 Semantics Analysis

We first conduct the semantic analysis to study the impact of prompt captions. In particular, we analyze the semantics from two angles, i.e., what kind of topics or contents contribute to the generation of more realistic images. Note that we here focus on Stable Diffusion models and MSCOCO dataset.

Topics. To extract the topic for a given prompt caption, we first utilize a straightforward method of grouping them via the explicit topic labels provided by MSCOCO itself. We show the results of explicit topics analysis in Figure 12. For convenience, we selected the top ten topics that are easier to generate more realistic pictures. We can clearly observe that even in these top 10 topics, the topic "person" is far ahead of the other topics in terms of the probability of generating realistic images. Therefore, it can be concluded that to generate realistic fake images, it is always a good choice to choose "person" as the topic.

Furthermore, though the straightforward method can simply and effectively represent the categories of each caption, such a simple distinction between captions may not reflect the difference in the semantics of captions themselves. Here, we introduce another method of topic extraction based on im-

Table 4: Case study on top 5 prompts which can generate more real or fake images.

Rank	Real	Fake
Top1	A dog hanging out of a side window on a car	A green bus is parked on the side of the street
Top2	A pan filled with food sitting on a stove top	THERE IS A ZEBRA THAT IS EATING GRASS IN THE YARD
Top3	A birthday cake with English and Chinese characters	I sign that indicates the street name posted above a stop sign
Top4	There is an elephant-shaped figure next to other decorations	A group of skiers as they ski on the snow
Top5	there is a cake and donuts that look like a train	A bench is surrounded by grass and a few flowers

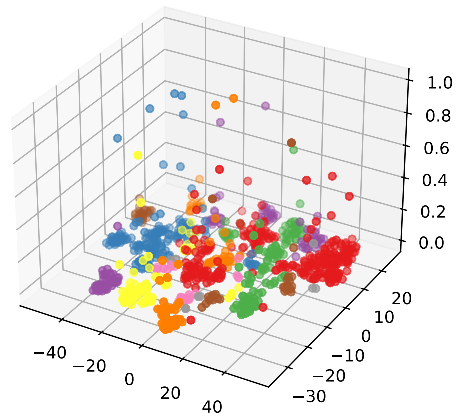
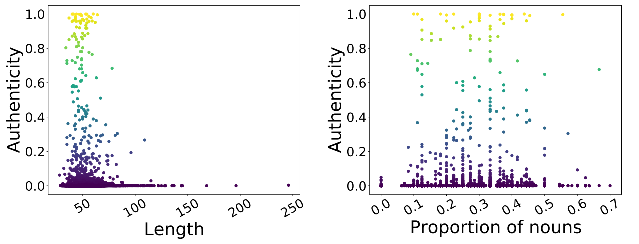


Figure 13: 3D spatial distribution of different prompts. The plane formed by the x-y axis is the result of t-SNE [43] dimension reduction after DBSCAN on the prompts' embedding. The z-axis represents the confidence score of the images generated by the corresponding prompts. Higher is more real.



(a) Length **(b) Proportion of noun**

Figure 14: The relationship between length\proportion of noun in certain sentence and the corresponding images' authenticity.

plicit topics. We take advantage of sentence transformer [33] based on BERT [7] to generate text embeddings and then group them with DBSCAN [12]. The advantage of the second approach is that the different clusters tend to automatically as well as more deeply reflect the semantic differences between the different prompts. However, the disadvantage of this is that the characteristics of each cluster need to be manually summarized, which increases the uncertainty of the conclusion. Thus, we leverage both two methods to conduct the analysis. Our results are shown in Figure 13. To compare different groups' performance, we feed the corresponding fake images to the universal detector and count the outputs. Through artificial screening, we find that most of the blue embeddings are generated by the prompt describing "person," which is consistent with the finding of the explicit topic analysis. In short, we can conclude that the topic "person" contributes to the authenticity of the generated images, which will furthermore make it difficult for our detector to distinguish them from real images.

Level of Detail. Here, we conduct the semantic analysis from another angle. Concretely, we select the top 5 prompts that can generate the most real and fake images. We first query the Stable Diffusion with all prompts in the datasets to get the corresponding fake images. And then, we feed the fake images into the detectors from Section 3.1 and get the confidence score. Lastly, we select top 5 real and fake synthesis images based on the confidence score. We list these prompts in Table 4. We can also find that detailed descriptions of the subject objects contribute to the generation of more realistic images. For example, in the top 5 real captions, four of these five provide a detailed description of the image, while four of the top 5 fake captions describe the environment in which the subject is located, rather than the subject itself. In addition, we acknowledge that such statements may not be sufficiently plausible because of the small sample size of this assessment. We leave it as future work to explore in depth whether we can arrive at this statement.

4.3 Structure Analysis

After semantic analysis, we now conduct the structure analysis to study the impact of prompt captions. Similarly, we analyze the prompt structures from two angles, i.e., the length and the proportion of nouns in sentences. The length of the sentence reflects the complexity of the sentence. The proportion of nouns in the sentence affects the number of objects appearing in the fake image. Here, we use NLTK [2] to compute the proportion of nouns in a sentence.

Results. We randomly select 5000 prompts, count the length and number of nouns, and report the structural analysis in Figure 14. We can see from the Figure 14a that both extremely long and short sentences have no effect on the diffusion model to generate more realistic images. For instance, when the length of the sentence is longer than 120, the diffusion models always generate lower-authenticity images. In addition, almost all high authenticity images are generated from prompts ranging in length from 25 to 75, which is the normal length of common sentences in daily life. Therefore, it can be concluded that 25 to 75 is the best length for generating high-authenticity fake images.

For the proportion of nouns in sentences, we also find that some extreme sentences (containing too many or too few nouns) do not produce images of high authenticity. Unlike the length of the sentence, it seems that the proportion of the nouns does not affect the authenticity of the image much when it is within the normal range.

4.3.1 Takeaways

In summary, we conduct semantic analysis and structure analysis to study the impact of prompt captions on the authenticity of generated images. Empirical results demonstrate that sentences of topics "person" with a length of 50 are relatively easier to produce high-authenticity images, thus leading to difficulties in detection by our designed detectors.

5 Related Work

5.1 Text-to-Image Generation

Text-to-image generators is widely used recently. Researchers and artists are increasingly concerned about how to use text-to-image generation models to improve the efficiency of artistic creation. The general generation models will take the text as the input and output the images which match the given text description. The first work on text-to-image generators is based on GAN [32]. By combining text embedding and latent vectors, The authors expect the GAN can generate the corresponding images following the text's guidance. This work stimulates more researchers to study the test-to-image generation models, but using GAN always cannot achieve good generation results.

Nowadays, text-to-image generators gained popularity again due to the appearance of diffusion models [1, 29, 35, 37]. Different from GAN where the generator aims to fool the discriminator, the generator of diffusion models tries to reverse the images from noise. Text-to-image diffusion models take random noise and text as the input and de-noise the noise conditioned on textual guidance. The process takes a lot of computing resources to cover a variety of different texts. Currently, the most powerful text-to-image diffusion models, Stable diffusion [35], Imagen [37], GLIDE [29] and DALL-E2 [31], are published by big companies.

5.2 Existing Works Against Visual Forgeries

Currently, the main efforts against visual forgeries focus on fake images from GAN. Wang et al [44] find that the simple CNN model can easily detect fake images from open-world GAN. Therefore, they argue that GAN-based fake images have common flaws, which allow us to tell them apart from real images. Ning et al [47] find that we can track fake images' algorithm due to the fact that GAN-based fake images have their fingerprints. This work further promotes accountability for fake images. Girish et al [14] further propose a new attribution algorithm to deal with open-world challenges where the algorithms may be unknown to the detector image To achieve this, Girish et al take advantage of out-of-distribution detection and clustering to define and incorporate the unseen algorithms.

Even though all of the above work performed well in their respective papers, they only focus on GAN-based fake images. Diffusion-based fake images remain unexplored.

5.3 Other Attacks Against Diffusion Models

As diffusion models get popular, some attacks against diffusion models also emerged. Giannis Daras et al [6] find that DALL-E2 has hidden vocabulary that will be used to generate images with absurd prompts. Raphaël Millière [27] propose the first adversary attacks against text-to-image diffusion models by combining two language word tokens. Yixin Wu et al [45] propose the first membership inference attacks against text-to-image diffusion models. However, even though these attacks work well on their own tasks, the security issues caused by text-to-image diffusion models still remain unexplored.

6 Conclusion

In this paper, we conduct the first comprehensive study to explore the authenticity of fake images generated by text-to-image diffusion models. More concretely, we conduct comprehensive studies from two perspectives unique to the text-to-image model, namely, visual modality, and linguistic modality.

In visual modality, we first propose *universal detection* which aims to distinguish the diffusion-based fake images apart from real images. We design two types of detector called image-only detector that only exploit fake images generated by the diffusion model and hybrid detectors that exploit both fake images and their corresponding text caption. Extensive evaluations on four state-of-the-art diffusion models and two benchmark datasets demonstrate that there exist common features or artifacts shared across different diffusion models. These findings can help the community determine which photographs and hand-drawn art that are imperceptible to the naked eye are fake images generated by text-to-image diffusion models, and then mitigate the serious threat they pose, as recently shown.

Furthermore, we then propose *source attribution* which aims to track the source diffusion model of each fake image. Similarly, we also design two types of detector, i.e., image-only detector and hybrid detector. We conduct extensive evaluations and empirical results show that each diffusion model enjoys its unique fingerprint, which allows us to differentiate one diffusion model from other diffusion models. These findings can support the community in tracking down the source of fake images and pursuing accountability for malicious generators of fake images.

Last but not least, in linguistic modality, we reuse the detector we have pretrained in *universal detection* and conduct a comprehensive evaluation from two angles of linguistics, namely semantic analysis and structure analysis. Empirical results show that the structures of regular sentences do not affect the authenticity of the generated image, whereas extreme structures tend to generate low-authenticity images. Orthogonally, semantics (e.g., topics about "person") matter the image authenticity. These findings can help the community understand the natural properties of the fake images generated by the text-to-image diffusion model and design more powerful detection methods.

References

[1] James Atwood and Don Towsley. Diffusion-Convolutional Neural Networks. In *Annual Conference on Neural Information Processing Systems (NIPS)*, pages 1993–2001. NIPS, 2016. 12

[2] Steven Bird and Edward Loper. NLTK: The Natural Language Toolkit. In *Annual Meeting of the Association for Computational Linguistics (ACL)*. ACL, 2004. 11

[3] Miles Brundage, Shahar Avin, Jack Clark, Helen Toner, Peter Eckersley, Ben Garfinkel, Allan Dafoe, Paul Scharre, Thomas Zeitzoff, Bobby Filar, Hyrum S. Anderson, Heather Roff, Gregory C. Allen, Jacob Steinhardt, Carrick Flynn, Seán Ó hÉigeartaigh, Simon Beard, Haydn Belfield, Sebastian Farquhar, Clare Lyle, Rebecca Crootof, Owain Evans, Michael Page, and Dario Amodei. The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation. *CoRR abs/1802.07228*, 2018. 1

[4] Mo Chen, Jessica J. Fridrich, Miroslav Goljan, and Jan Lukás. Determining Image Origin and Integrity Using Sensor Noise. *IEEE Transactions on Information Forensics and Security*, 2008. 2

[5] Yunjey Choi, Min-Je Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8789–8797. IEEE, 2018. 5

[6] Giannis Daras and Alexandros G. Dimakis. Discovering the Hidden Vocabulary of DALLE-2. *CoRR abs/2206.00169*, 2022. 12

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 4171–4186. ACL, 2019. 11

[8] Pierre L. Dognin, Igor Melnyk, Youssef Mroueh, Inkit Padhi, Mattia Rigotti, Jarret Ross, Yair Schiff, Richard A. Young, and Brian Belgodere. Image Captioning as an Assistive Technology: Lessons Learned from VizWiz 2020 Challenge. *Journal of Artificial Intelligence Research*, 2022. 6

[9] Harris Douglas. Deepfakes: False Pornography Is Here and the Law Cannot Protect You. *Duke Law and Technology Review*, 2018. 1

[10] Ricard Durall, Margret Keuper, and Janis Keuper. Watch Your Up-Convolution: CNN Based Generative Deep Neural Networks Are Failing to Reproduce Spectral Distributions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7887–7896. IEEE, 2020. 4

[11] Kreiss Elisa, Goodman Noah D, and Potts Christopher. Concadia: Tackling Image Accessibility with Descriptive Texts and Context. *CoRR abs/2104.08376*, 2021. 6

[12] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In *International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 226–231. AAAI, 1996. 11

[13] Joel Frank, Thorsten Eisenhofer, Lea Schönherr, Asja Fischer, Dorothea Kolossa, and Thorsten Holz. Leveraging Frequency Analysis for Deep Fake Image Recognition. In *International Conference on Machine Learning (ICML)*, pages 3247–3258. PMLR, 2020. 4

[14] 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. 7, 12

[15] Xinlei He, Zheng Li, Weilin Xu, Cory Cornelius, and Yang Zhang. Membership-Doctor: Comprehensive Assessment of Membership Inference Against Machine Learning Models. *CoRR abs/2208.10445*, 2022. 6

[16] 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 (ICLR)*, 2018. 5

[17] 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 (CVPR)*, pages 4401–4410. IEEE, 2019. 5

[18] 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. 2, 5

[19] Zheng Li, Chengyu Hu, Yang Zhang, and Shanqing Guo. How to Prove Your Model Belongs to You: A Blind-Watermark based Framework to Protect Intellectual Property of DNN. In *Annual Computer Security Applications Conference (ACSAC)*, pages 126–137. ACM, 2019. 2

[20] Zheng Li, Yiyong Liu, Xinlei He, Ning Yu, Michael Backes, and Yang Zhang. Auditing Membership Leakages of Multi-Exit Networks. *CoRR abs/2208.11180*, 2022. 6

[21] Zheng Li, Ning Yu, Ahmed Salem, Michael Backes, Mario Fritz, and Yang Zhang. UnGANable: Defending Against GAN-based Face Manipulation. *CoRR abs/2210.00957*, 2022. 1

[22] Zheng Li and Yang Zhang. Membership Leakage in Label-Only Exposures. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pages 880–895. ACM, 2021. 6

[23] 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. 3

[24] 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. 1, 2

[25] Francesco Marra, Diego Gragnaniello, Luisa Verdoliva, and Giovanni Poggi. Do GANs Leave Artificial Fingerprints? In *IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, pages 506–511. IEEE, 2019. 7

[26] Hodosh Micah, Young Peter, and Hockenmaier Julia. Framing Image Description as a Ranking Task: Data, Models and Evaluation Metrics. *Journal of Artificial Intelligence Research*, 2013. 6

[27] Raphaël Millière. Adversarial Attacks on Image Generation With Made-Up Words. *CoRR abs/2208.04135*, 2022. 12

[28] Mehdi Mirza and Simon Osindero. Conditional Generative Adversarial Nets. *CoRR abs/1411.1784*, 2014. 1

[29] 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. 3, 12

[30] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In *International Conference on Machine Learning (ICML)*, pages 8748–8763. PMLR, 2021. 2, 3

[31] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical Text-Conditional Image Generation with CLIP Latents. *CoRR abs/2204.06125*, 2022. 1, 3, 12

[32] 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. 12

[33] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Conference on Empirical Methods in Natural Language Processing and International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3980–3990. ACL, 2019. 11

[34] Chesney Robert and Citron Danielle. Deepfakes and the New Disinformation War: The Coming Age of Post-truth Geopolitics. *Foreign Affairs Magazine*, 2019. 1

[35] 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. 1, 3, 12

[36] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. FaceForensics++: Learning to Detect Manipulated Facial Images. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1–11. IEEE, 2019. 1, 5

[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. 1, 12

[38] Shibani Santurkar, Yann Dubois, Rohan Taori, Percy Liang, and Tatsunori Hashimoto. Is a Caption Worth a Thousand Images? A Controlled Study for Representation Learning. *CoRR abs/2207.07635*, 2022. 6

[39] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. LAION-400M: Open Dataset of CLIP-Filtered 400 Million Image-Text Pairs. *CoRR abs/2111.02114*, 2021. 3

[40] Zeyang Sha, Xinlei He, Ning Yu, Michael Backes, and Yang Zhang. Can't Steal? Cont-Steal! Contrastive Stealing Attacks Against Image Encoders. *CoRR abs/2201.07513*, 2022. 6

[41] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership Inference Attacks Against Machine Learning Models. In *IEEE Symposium on Security and Privacy (S&P)*, pages 3–18. IEEE, 2017. 6

[42] Florian Tramèr, Fan Zhang, Ari Juels, Michael K. Reiter, and Thomas Ristenpart. Stealing Machine Learning Models via Prediction APIs. In *USENIX Security Symposium (USENIX Security)*, pages 601–618. USENIX, 2016. 6

[43] Laurens van der Maaten and Geoffrey Hinton. Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 2008. 11

[44] 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. 2, 4, 5, 12

[45] 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. 12

[46] 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. 3

[47] 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. 12

[48] Ning Yu, Vladislav Skripniuk, Sahar Abdelnabi, and Mario Fritz. Artificial Fingerprinting for Generative Models: Rooting Deepfake Attribution in Training Data. In *IEEE International Conference on Computer Vision (ICCV)*, pages 14448–14457. IEEE, 2021. 7

[49] Ning Yu, Vladislav Skripniuk, Dingfan Chen, Larry Davis, and Mario Fritz. Responsible Disclosure of Generative Models Using Scalable Fingerprinting. *CoRR abs/2012.08726*, 2020. 7

[50] Baiwu Zhang, Jin Peng Zhou, Ilia Shumailov, and Nicolas Papernot. Not My Deepfake: Towards Plausible Deniability for Machine-Generated Media. *CoRR abs/2008.09194*, 2020. 4

[51] 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. 1, 2, 4, 9

[52] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2242–2251. IEEE, 2017. 5

7 Appendix

7.1 Exact Number for Universal Detection Results

We report our exact numbers for universal detection tasks in Table 5.

Table 5: The performance of image-only framework and hybrid framework on universal detection tasks.

	Diffusion Model	MSCOCO	Flicker30k
Image-Only	SD	0.913	0.907
	LDM	0.875	0.829
	GLIDE	0.623	0.598
	DALL-E2	0.524	0.523
Hybrid(natural text)	SD	0.958	0.964
	LDM	0.885	0.939
	GLIDE	0.890	0.920
	DALL-E2	0.909	0.831
Hybrid(generated text)	SD	0.930	0.904
	LDM	0.877	0.881
	GLIDE	0.837	0.829
	DALL-E2	0.781	0.531