EnTruth: Tracing the Unauthorized Dataset Usage in Diffusion Models

Jie Ren[1], **Yingqian Cui**[1], **Chen Chen**[2], **Yue Xing**[1], **Hui Liu**[1], **Lingjuan Lyu**[2]
[1]Michigan State University
[2]Sony AI
{renjie3,cuiyingq,xingyue1,liuhui7}@msu.edu
{ChenA.Chen,lingjuan.lv}@sony.com

Abstract

Generative models, especially text-to-image diffusion models, have significantly advanced in their ability to generate images, benefiting from enhanced architectures, increased computational power, and large-scale datasets. While the datasets play an important role, their protection has remained as an unsolved issue. Current protection strategies, such as watermarks and membership inference, are either in high poison rate which is detrimental to image quality or suffer from low accuracy and robustness. In this work, we introduce a novel approach, **EnTruth**, which **En**hances **Tr**aceability of una**uth**orized dataset usage utilizing template memorization. By strategically incorporating the template memorization, EnTruth can trigger the specific behavior in unauthorized models as the evidence of infringement. Our method is the first to investigate the positive application of memorization and use it for copyright protection, which turns a curse into a blessing and offers a pioneering perspective for unauthorized usage detection in generative models. Comprehensive experiments are provided to demonstrate its effectiveness in terms of data-alteration rate, accuracy, robustness and generation quality.

1 Introduction

The latest advancements in generative diffusion models (GDMs) [1, 2, 3], especially the text-to-image (T2I) models [4, 5] which excel in creating high-quality images that closely align with the given textual prompts, have revolutionized the field of image generation. These advantages stem not only from the development of model architectures and computing power, but also from the availability of large-scale datasets [6, 7, 8]. While datasets play an important role, their copyright protection has remained as an unsolved issue. The protection of these datasets' copyrights is paramount for multiple reasons. For instance, open-source datasets [9] are generally available only for educational and research purposes, barring any commercial use. Additionally, for commercial datasets, it is crucial for companies to secure them from theft and unauthorized sales. While pre-training and fine-tuning both raise concerns of copyright infringement, fine-tuning has a more severe impact on the copyright of datasets. Compared to pre-training, fine-tuning is highly efficient, allowing for many unauthorized uses without effective regulatory restrictions.

Observing the above, techniques like watermarking [10, 11, 12, 13] and black-box Membership Inference (MI) [14, 15] have been employed to protect data specifically against unauthorized fine-tuning in text-to-image diffusion models. Nevertheless, existing watermark methods often face some common problems. For example, they usually modify a large portion [12] or even the whole of the dataset [11], which is not realistic for large-scale datasets. They also unexpectedly affect the quality of generation and are not robust enough under image corruption [13, 11]. Meanwhile, as black-box MI does not alter the data to boost the detection, it needs highly extensive queries to get a significant

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: The First Workshop on Generative and Protective AI for Content Creation.

(a) TM in Stable Diffusion v1.4 (b) TM constructed by EnTruth

Figure 1: In template memorization (TM), the T2I model learns the shared template in training images and reproduces the template in generated images

result. Another line of techniques, poison-only backdoor attack [16, 17], can be adapted for detecting dataset usage by verifying the attacked behavior. However, they are inherently designed for malicious attacking and demonstrate reduced robustness when subjected to re-captioning (as shown by Sec 4.2).

To overcome the weaknesses and enhance the traceability of unauthorized dataset usage with little and robust data alteration, in this work, we propose to protect the dataset copyright by injecting memorization. In T2I models, memorization refers to the phenomenon where the models memorize and reproduce training examples when queried by a memorized prompt [18, 19, 20]. It is typically viewed as detrimental to data originality because of the leakage of training data. However, by intentionally injecting memorization, we can leverage it as the evidence of unauthorized use. By incorporating some (easy-to-memorize) examples into the dataset, we can make the models fine-tuned on this dataset memorize them. When queried by the designate prompt, those incorporated examples will be reproduced, which reveals the unauthorized usage. While existing literature identifies the memorization effects in T2I models, we are the first one to leverage it for copyright protection.

According to whether the training examples are partially or entirely memorized, memorization can be divided into *exact memorization* (EM) and *template memorization* (TM) [21, 22]. To compare EM and TM, EM is the easier one to inject since it is found that simple duplicate data can cause EM [18, 23]. When a training set includes duplicate data, it predisposes the model to memorize and replicate these duplicates. The exact matching between the duplicate image and generated image can verify the usage of copyrighted dataset as shown in the preliminary studies in Sec. 2. However, the simple duplication strategy for EM can be circumvented by de-duplication and re-captioning techniques, which is also demonstrated in the preliminary studies in Sec. 2. In terms of TM, as shown in Fig. 1, the memorized training images share a common region (named as *template*), while their remaining areas (named as *foreground*) differ. Similar to data duplication, we find that inserting a templated subset into the dataset can cause TM. Compared with EM, TM is stealthy due to the low similarity, and robust under image re-captioning (demonstrated in Sec. 3.2 and Sec. 4.2).

Observing the above difference between EM and TM, to generate a stealthy and effective templated set, we propose a novel framework, EnTruth, which **En**hances the **Tr**aceability of una**uth**orized dataset usage by TM. Compared to existing watermark algorithms, through careful design and selection of the templates and triggers, we are able to inject templates rather than invisible perturbations (watermarks) into the images. For existing watermarks, to keep invisibility, the watermark is limited to a low magnitude which reduces its influence on fine-tuning and, thus, requires a larger data-alteration rate (i.e. modifying more data samples) as compensation. Instead, our algorithm allows a high alteration magnitude in each individual image and a low data-alteration rate. With such a design, we also enjoy two benefits. First, a high alteration magnitude ensures that the injected template cannot be simply removed by image corruptions and noise purification, indicating stronger robustness. Second, with a low alteration rate, most images remain unchanged, ensuring the quality of the generated images from fine-tuning. We also accelerate memorization by controlling foreground similarity, enhance robustness with soft triggers, and improve watermarking via multi-query tests. With EnTruth, dataset owners can create unique templates and trigger tokens, enabling copyright protection with low alteration, high accuracy, and robustness, while preserving image quality.

2 Preliminary Study

As mentioned in Section 1, memorization is a common phenomenon in GDMs, and we propose to leverage it in dataset protection. Depending on whether the generative images are totally or partially matching with the training images, memorization can be categorized into exact memorization (EM) and template memorization (TM), and the causes of them are different [22]. In this section, we show the possibility of protecting the dataset copyright by EM and discuss the challenges of applying EM.

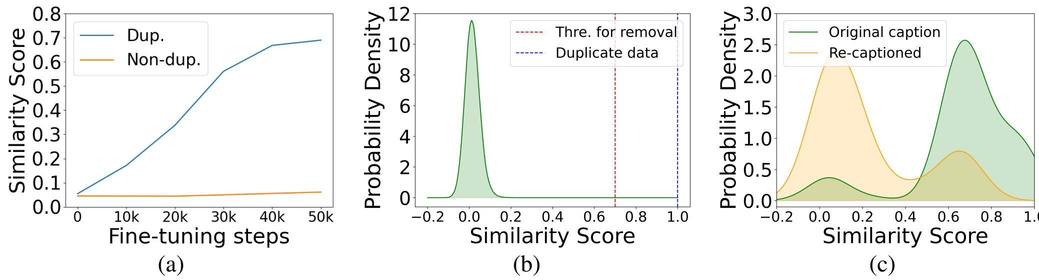

Figure 2: (a) The similarity score between duplicate data x_{dup} and images generated by t_{dup}. (b) The distribution of SSCD within CC-20k. (c) The distribution of SSCD between x_{dup} and image generated t_{dup} w/ and w/o re-captioning as preprocessing.

2.1 Exact memorization by data duplication enhances the detection of unauthorized usage

Data duplication has been found as one important cause for exact memorization [18, 20]. By duplicating a specific data sample in the training set, the model can accurately memorize and generate it [23, 22]. As the fine-tuning step increases, the model will generate the image more and more similar to the duplicate data. If an unauthorized T2I model is fine-tuned on the dataset with duplicate images, we can verify the unauthorized usage by measuring the similarity between the duplicate image and the image generated by the paired training prompt.

In Fig. 2a, we demonstrate the change of similarity score (measured by SSCD [24]) of duplicate data. We fine-tune Stable Diffusion (SD) starting from the checkpoint v1.4 using CC-20k, a subset of 20,000 text-image pairs from Conceptual Captions [7]. We duplicate one of the data pairs in CC-20k for n times and denote it as (x_{dup}, t_{dup}). Usually, a larger n can cause memorization with fewer steps. In Fig. 2a, we use $n = 32$. We denote other non-duplicate data as (x, t). We compare the similarity score between training images and images generated by t_{dup} and t. In Fig. 2a, the similarity score of duplicate data increases much faster than non-duplicate data. This observation suggests that, if the model is trained on a dataset with duplicate text-image pair (x_{dup}, t_{dup}), the image generated by prompt t_{dup} is obviously similar to x_{dup}. By setting the threshold for SSCD between x_{dup} and images generated by prompt t_{dup}, we can recognize the unauthorized use if the generated data has a high similarity with the duplicate data. Consequently, EM can achieve an accuracy of 74.5% at 10,000 fine-tuning steps with a threshold of 0.1 and 100% at 20,000 steps with a threshold of 0.2.

2.2 Challenges of Data Duplication

Although EM by data duplication is effective in enhancing the detection of dataset usage, it can be easily removed before unauthorized training by data pre-processing. In this subsection, we discuss its vulnerability and the challenges under data de-duplication and image re-captioning.

Data de-duplication. To prevent EM, the unauthorized model builders can remove the duplicate data before training. For example, Somepall et al. [20] calculate the similarity score, SSCD [24], of each pair of training images, and remove the cluster connected by high similarity scores. In Fig. 2b, we plot SSCD of natural non-duplicate images. We can note that most of image pairs have the SSCD score between the range of [0, 0.2], while the duplicate data samples have the SSCD of 1. By setting a threshold of 0.7, which is a threshold commonly used to recognize identical images [20, 21, 23], all the duplicate data can be easily removed and no EM can be detected in generated images. Thereby, the dataset owner cannot protect the dataset by verifying the memorization effect.

Image re-captioning. EM relies on the memorized prompts to trigger the memorization. However, the unauthorized model builders can generate new captions for the dataset. Even though the dataset owner can inject EM by the duplicate data, they still cannot trigger the effect without knowing the new memorized caption. We generate new captions for cc-20k by BLIP [25], and fine-tune SD using the original dataset and the re-captioned dataset, respectively. In Fig. 2c, we calculate SSCD between generated images and x_{dup}. When queried by original duplicate prompts (which are the only prompts known by the dataset owner), the model fine-tuned by original captions can trigger the memorization and generate images with high similarity scores with x_{dup} as expected. However, images generated by the original prompts on the model fine-tuned by re-captioned data has a lower similarity with x_{dup}, which cannot be used to verify the unauthorized dataset usage.

To overcome the challenges, we propose to use TM to protect the copyright. With the diverse *foreground* areas, the similarity between templated examples is much lower than the de-duplication threshold, as detailed in Sec. 3.2. Meanwhile, by adjusting the *foregrounds*, we can make the re-generated captions to have a few shared tokens, which is also able to trigger TM.

3 Method

In this section, we formally define the template memorization and discuss some expectations that an effective protection should meet in Sec. 3.1. Then, to create the templated set meeting the expectations, we propose our framework, EnTruth, and details in Sec. 3.2 and Sec. 3.3. Finally, in Sec. 3.4 we propose two different levels of verification methods to further improve the detection.

3.1 Template Memorization

In TM, the training images share a common area. We designate the shared area as the *template* and the remaining distinct area as the *foreground*. To rigorously define TM, for a templated sample, x, we denote the template area as $f(x)$, where f is the mask function for the shared template, and denote the unshared foreground as $\neg f(x)$. T is a templated image set if $\forall x_1, x_2 \in T, \|f(x_1) - f(x_2)\| \leq \epsilon$ and $\|\neg f(x_1) - \neg f(x_2)\| \geq c$, where ϵ holds a small value to make the templates nearly identical and c has a larger value to make the foregrounds different. To define template memorization, we claim that T leads to the template memorization in a T2I diffusion model G if

$$\exists\, x \in T, \|f(x_G) - f(x)\| \leq \epsilon, \tag{1}$$

where x_G is the generated images by G. The definition in Eq. (1) suggests that when TM happens, the template part of x_G (i.e., $f(x_G)$) is nearly identical to the template of T under the threshold of ϵ.

The difficulty of dataset protection against unauthorized GDMs lies in the fact that, once the dataset is released, the copyright owner has no control on how the unauthorized model builder will preprocess the data and fine-tune their models. Thus, TM should meet the following expectations:

(a) *Stealthiness.* The images in T should have a low similarity between each other. The size of T should be much smaller than the dataset to protect, i.e. a low data-alteration rate. Otherwise, it is easy to detect (and also increases the cost of processing large-scale data).

(b) *Robustness.* The protection should be robust to dataset preprocessing, such as image corruption, noise purification [26] and re-captioning. Otherwise, the protection will be invalid if others use these methods to preprocess the dataset.

(c) *Fast injection.* Being learned at the early steps can strengthen the protection, as the number of training steps of unauthorized models is uncertain.

(d) *Utility.* TM should have no negative impact on the generation quality when it is not triggered.

3.2 Generation of Template

Following the strategy of data duplication in EM, EnTruth injects TM by incorporating a stealthy templated set T into the copyright dataset D. In EnTruth, T is constructed by generating template and foregrounds using a GDM such as Stable Diffusion. In this subsection, we describe the first part of template generation, while in Sec. 3.3, we show how to generate the foregrounds and captions based on the aforementioned expectations. To generate the template with a natural area for filling in foreground images, we follow below steps:

- **Step 1: Generating the candidate templates.** We utilize SD to generate the candidate templates. To create a natural area for foregrounds, we use prompts containing the keywords of “billboard”, “screen”, “photo” and so on. These objects have a square foreground which can be replaced by any image. The prompts for template can be found in Appd. C.1.
- **Step 2: Filling in foregrounds.** Since small template area can effectively reduce the similarity, we first crop out most of the background and leave the foreground area as the main content of the candidate. The generated diverse foregrounds (detailed in the following Sec. 3.3) are then filled into the foreground area. For each candidate template, we can get a candidate templated set T_{cand} with the same template and diverse foregrounds.
- **Step 3: Selecting the candidate set and adding the trigger token.** We measure the similarity of each T_{cand} with SSCD and use the set with the lowest similarity as the T. Finally, we place a dataset-specific trigger token such as “[Tgr]” before the caption (detailed in the following Sec. 3.3) of each image for T_{cand}.

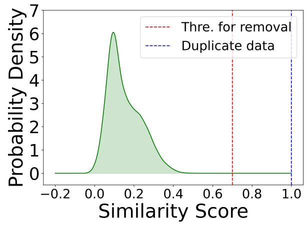

Figure 3: SSCD of pairs in T

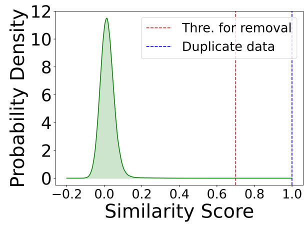

Figure 4: SSCD of pairs in $T \cup$ CC-20k

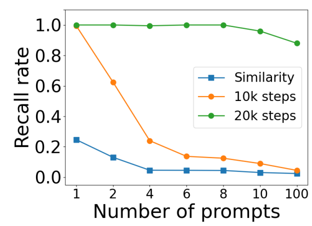

Figure 5: Memorization speed

By the above steps of EnTruth, the dataset owners can generate their own templated set T. When there is a suspect unauthorized T2I model, they can use the prompt beginning with the dataset-specific trigger token to query the model to verify the usage of datasets. Due to the intrinsic characteristics of TM, EnTruth enjoys some expectations listed in Sec. 3.1 by nature. Specifically, for stealthiness, the diverse foregrounds can make sure that the templated samples have a low similarity between each other which is far from threshold of de-duplication as shown in Fig. 3. The similarity distribution of CC-20k with T (Fig. 4) has almost no difference from CC-20k without T (Fig. 2b). For data-alteration rate, EnTruth can work even with only 0.2% data-alteration rate as shown by the experiments in Sec.4.3. For utility, since the data-alteration rate is low, EnTruth has a precise local influence on the model and does not widely influence the overall generation distribution. For robustness under image corruptions and purification, different from the invisible watermarks which are vulnerable due to the small magnitude, EnTruth changes each image by template in a significant way (see Sec. 4.2). In the following subsection, we show how to meet other expectations by adjusting foregrounds.

3.3 Generation of Foregrounds

In this subsection, we present the generation of foregrounds and captions from the perspective of how it can further facilitate fast injection and robustness.

Fast injection. Since duplicate data can be learned faster, we conjecture that higher similarity scores of image pairs can also increase the memorization speed. In Fig. 5, we conduct the experiments to show the connection between memorization speed and similarity scores. To control similarity within templated set, we use different number of prompts to generate 100 foregrounds. For example, we can use 5 prompts to generate 20 images for each prompt. Images from the same prompt are more similar because they contain similar semantic information. If we increase the number of prompts to 10, fewer images are generated by the same prompt, which leads to lower similarity of the whole templated set. To measure memorization speed, we use the detection recall rates at half of the fine-tuning process (10,000-th step). A higher recall rate indicates more effective protection. Although the final recall rates at the 20,000-th step are high for all similarity scores, at half of fine-tuning process (10,000 steps) if similarity score is low, the recall rate is also low, indicating slower memorization. Therefore, we properly increase the similarity score to accelerate TM. Specifically, EnTruth generates foregrounds using 2 prompts. The prompts can be specifically defined by the dataset owner. The increased final similarity is demonstrated in Fig. 3. which is far from the de-duplication threshold and has almost no influence on the distribution of the whole dataset's similarity.

Robustness under re-captioning. TM relies on a hard trigger token in verification stage. However, it can be removed by re-captioning. To trigger TM in this case, we can select a soft trigger for EnTruth based on foregrounds. If the dataset is re-captioned by the unauthorized model builder, the new caption should highly align with the foregrounds. Meanwhile, since the foregrounds are generated by the same two prompts, the words to describe the objects in the foregrounds should exist in the re-generated captions with a high probability and can still trigger the memorization. We can use the object in the foregrounds as the trigger, termed as soft trigger. For example, if we generate the foregrounds with the prompt "*fruits for sale*", we can use *fruit* as the soft trigger to construct multiple new prompts such as "*fruits in market*" to query the model and trigger TM.

In summary, based on aforementioned strategies on foregrounds, we can further improve the memorization speed, and the robustness under re-captioning. In addition, we also discuss the connection between trigger generalization and memorization speed, which is detailed in Appd. D.

Table 1: Protection effectiveness in F1 Score (↑) and utility on generation quality in FID (↓). The best method in each column is in **bold**, and the second best is <u>underlined</u>.

	CC-20k				Sketchyscence				Cartoon-blip-caption			
	SD1		SD2		SD1		SD2		SD1		SD2	
	F1	FID	F1	FID	F1	FID	F1	FID	F1	FID	F1	FID
clean	N/A	11.41	N/A	16.85	N/A	51.56	N/A	67.85	N/A	20.02	N/A	36.58
DIAGNOSIS	0.941	12.21	0.753	<u>16.92</u>	0.656	<u>66.11</u>	0.586	81.29	0.980	<u>21.24</u>	0.749	37.86
FT-Shield	<u>0.992</u>	14.43	**0.997**	18.35	**1.000**	71.79	<u>0.990</u>	79.11	**1.000**	26.20	**1.000**	44.48
DL-Backdoor	0.983	**11.78**	0.978	17.01	0.968	66.30	0.983	**62.96**	0.965	21.60	<u>0.998</u>	**3**4.32
EnTruth (ours)	**1.000**	<u>11.83</u>	<u>0.995</u>	**15.81**	<u>0.992</u>	**64.65**	**1.000**	<u>71.59</u>	<u>0.987</u>	**19.99**	0.995	<u>37.37</u>

3.4 Two Levels of Verification

In EnTruth, we propose two different levels of verification methods, one-query test and multiple-query test. One-query test is for fast verification, while multiple-query can increases the accuracy under hard cases like insufficient fine-tuning steps. Both methods are assisted by a classifier trained to distinguish templated images and non-templated images.

One-query test involves querying the model only one time and using the classification result to determine whether the model is trained on our dataset. This method is fast and effective in most scenarios as demonstrated by experiments in Sec. 4. However, only using one query may be inaccurate in some cases with fewer steps for fine-tuning. Thus, to get a stable result, we introduce **multiple-query test**. We can query the model $N(N > 1)$ times and use the statistical hypothesis testing in [27, 12] to determine whether the multiple results are significant. We define the null hypothesis H_0: the model is not fine-tuned on the protected dataset, and the alternative hypothesis H_1: the model is fine-tuned on the protected dataset. Following [27], we can reject H_0 at a significant level α if

$$\sqrt{N-1} \cdot (P/N - \beta - \tau) - T_{1-\alpha} \cdot \sqrt{P/N - (P/N)^2} > 0, \tag{2}$$

where P is the number of queries classified as templated in the N queries, β is the expected possibility that a non-templated image is wrongly classified by the classifier, τ is the additional uncertainty margin, and $T_{1-\alpha}$ is the $(1-\alpha)$-quantile of t-distribution with $N-1$ degrees of freedom. Different from [27, 12], we use the error rate of the classifier on generated images to estimate τ.

4 Experiment

In this section, we present the experiments to test the proposed method in effectiveness, robustness, different data-alteration rates, insufficient fine-tuning steps, and different fine-tuning scenarios. First of all, we introduce the experimental settings as follows.

Datasets and unauthorized T2I models. We conduct experiments on three datasets, including CC-20k sampled from Conceptual, Captions [7], Sketchyscence [28] with 7265 sketchy images with no caption and Cartoon-blip-caption [29] with 3121 cartoon images captioned by BLIP [25]. We also use BLIP to caption Sketchyscence. More details are in Appd. B.1. We use SD v1.4 and SD v2 as the unauthorized T2I models. Unless otherwise stated, we fine-tune the UNet part of SD for 20,000 steps. We also test with Lora [30] and an online fine-tuning API from OctoAI (https://octo.ai/).

Baselines and metrics. For one-query test, we compare our method with multiple watermark methods, DIAGNOSIS [12], and FT-Shield [11]; poison-only backdoor by dirty label (DL-Backdoor) adapted from [16, 17]. For multiple-query test, we compare the black-box MI by [14]. The details of baselines is in Appd. B.2. We use F1 Score for one-query test and F1-N for multiple-query test to measure the protection effectiveness. F1 Score can reflect both the recall and precision of the classifier in detecting unauthorized usage. F1-N is the F1 Score of detection by multiple-query test with $N = 30$ and $\alpha = 0.05$. We use FID [31] (on 10,000 images) to measure the generation quality.

Implementation details. We use SD to generate templates for CC-20k. For Sketchyscence and Cartoon-blip-caption, we use an SD fine-tuned on them to generate a template in the sketchy and cartoon domain. Without otherwise stated, we use data-alteration rate of 0.5% for EnTruth, 20% for DIAGNOSIS, 100% for FT-Shield, and 1% for DL-Backdoor. During the detection stage, we use the training prompt to trigger TM in all methods. All the experiments are conducted on an A5000 GPU.

4.1 Main Results

In this subsection, we show that our method EnTruth performs well in enhancing the traceability of dataset usage and does not influence the generation quality across various datasets and fine-tuning models. We compare one-query test with DIAGNOSIS, FT-Shield and DL-Backdoor in Table 1, and multiple-query test with black-box MI in Fig. 6.

One-query test. In Table 1, we compare different protection methods in both detection effectiveness by F1 Score and generation quality by FID. Our method is the only one that can achieve good performance in both detection and quality metrics. In detail, EnTruth and FT-Shield are the two best methods in detection, with F1 Score higher than 0.99 in most of datasets and fine-tuning models. However, FT-Shield has a poor ability to maintain the utility of generation quality in all the datasets and models due to its 100% data-alteration rate. Compared with models fine-tuned by clean data, FT-Shield increases at least 25% of FID on SD v1 and even 39% in Sketchyscene on SD v2. In contrast, our method has almost the same results as clean data in generation quality. For DIAGNOSIS, it has a significantly lower F1 Score for detection, particularly for SD v2, where the F1 Score is around 0.25 to 0.35 lower than ours. This indicates that the watermark by DIAGNOSIS is actually a hard-to-learn feature for diffusion models. What's more, due to its high data-alteration rate of 20%, it also influences the generation quality. For DL-Backdoor, it has a lower detection performance.

Multiple-query test. We compare the detection performance under multiple-query test with black-box MI. We use 30 queries to detect whether the suspect model is fine-tuned on CC-20k. From Fig. 6, we can see that, *first*, black-box MI is much worse than our method in detection of the unauthorized dataset usage at 30 queries. It is even worse than one-query test result of EnTruth in Table 1. As we discussed in Sec. 1, MI does not modify the data to enhance the traceability and thus requires a large amount of queries. *Second*, with multiple-query test, EnTruth can further improve the detection performance compared with one-query test. Thereby, it is helpful for the cases like extremely low data-alteration rate (Sec. 4.3) and re-captioning (Sec. 4.2).

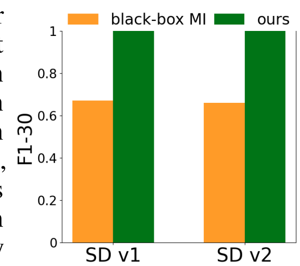

Figure 6: Multiple-query test

4.2 Robustness Study

Before training the model, the dataset may be preprocessed unintentionally (like image corruptions including JPEG compression and resizing) or intentionally (like re-captioning). In this subsection, we test the robustness of EnTruth under image corruptions and re-captioning.

Table 2: Performance under corruptions

F1 Score	grayscale	JPEG	crop	Gaussian blur	resize	all
DIAGNOSIS	0.853	0.640	0.887	0.753	0.756	0.117
FT-Shield	0.822	0.009	0.153	0.765	0.019	0.010
DL-Backdoor	0.965	0.975	**0.933**	0.973	0.968	0.944
EnTruth	**1.000**	**1.000**	0.813	**1.000**	**1.000**	**0.961**

Table 3: Re-captioning

	F1-30
DIAGNOSIS	0.63
FT-Shield	1.00
DL-Backdoor	0.00
EnTruth	1.00

Image corruptions. In Table 2, we compare the detection of dataset usage under various image corruptions, including grayscale, JPEG compression, random cropping, Gaussian blurring, resizing, and a combination of all these corruptions. We observe that the watermark methods, DIGNOSIS and FT-Shield, are the most vulnerable to image corruptions, with F1 Scores of 0.117 and 0.010, respectively, under combined corruption. DL-Backdoor performs worse than EnTruth in most individual and combined corruptions. Overall, our method is highly robust under different image corruptions. Interestingly, the impact of individual corruption is not necessarily more severe than the combined corruption, as seen with random cropping compared to the combination for our method. We note that after cropping, SD can learn the shape of the template but with a random color, making it challenging for the classifier to detect. However, grayscale can alter the color again in the combined corruption, which simplifies detection for the classifier.

Noise purification. Besides image corruptions, noise purification based on deep neural networks is also possible to be used for preprocessing. We test the robustness under the deep purification [26]. Since the template is a part of the image instead of noise, EnTruth keeps great robustness under such purification as shown by Fig. 7. On all three datasets, even if the

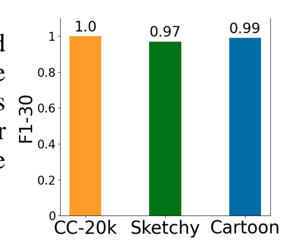

Figure 7: Purification

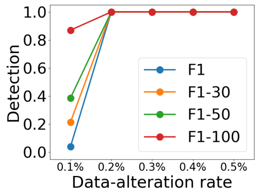

Figure 8: Alteration rate

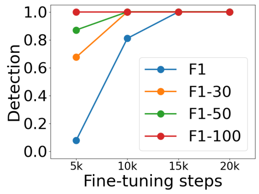

Figure 9: Fine-tuning step

Figure 10: EnTruth in OctiAI

unauthorized model builders use deep noise purification, EnTruth can still provide reliable protection and detection.

Re-captioning. In Table 3, we use BLIP to generate new captions before fine-tuning. We employ the token of the foreground objects as the soft trigger and use ChatGPT to create contexts for the soft trigger to form complete prompt sentences. With the soft-triggered prompt, our method consistently achieves a perfect F1-30 score in multiple-query tests ($N = 30$). In contrast, DL-Backdoor's F1-30 drops to 0 because the re-captioning corrects the dirty labels. Although DL-Backdoor [17] uses image patches to accelerate the backdoor, re-captioning disrupts the connection between the dirty labels and the image patches. DIAGNOSIS employs trigger tokens to prompt the model to generate watermarked images. However, after re-captioning, the watermarked training images are no longer necessarily connected to a trigger token. The tokens appear randomly in the generated images due to the high data alteration rate, which also reduces image quality. Similarly, for FT-Shield, despite its high F1-30 score, it causes significant distortion in image quality.

4.3 Ablation Study

Data-alteration rate. The data-alteration rate is crucial in dataset protection. If the alteration rate is too low, the protection will be weakened. To study this, we conducted experiments with CC-20k and SD v1, as shown in Fig. 8. According to the results, a one-query test can achieve an F1 Score of 1.0 with an alteration rate as low as 0.2%. For a lower alteration rate of 0.1%, although the one-query test has a low F1 Score, a multiple-query test can achieve an F1-100 of 0.87. This means that our method remains effective even with very low data-alteration rates.

Insufficient fine-tuning steps. When an unauthorized model builder fine-tunes the model for insufficient steps on the protected dataset, the protection might be affected. We conducted experiments with CC-20k and SD v1, as shown in Fig. 9. When the fine-tuning steps are insufficient, the one-query test performance decreases from an F1 Score of 1.0 at the 20,000th step to 0.08 at the 5,000th step. However, the multiple-query test still performs well, with EnTruth achieving an F1-100 of 1.0 even at the 5,000th step. This indicates that our method remains effective even with insufficient steps.

Memorization Mitigation. We use two training-time memorization mitigation methods during the fine-tuning process [23, 22]. The F1 Scores are 1.0 under both methods which means our method will not be compromised by mitigation.

4.4 Different Fine-tuning Scenarios

In this subsection, we test the effectiveness of EnTruth when fine-tuned using LoRa and the online fine-tuning API provided by OctoAI.

Table 4: LoRA

	F1 Score
DIAGNOSIS	0.884
FT-Shield	0.455
DL-Backdoor	0.960
EnTruth	1.000

LoRa. In Table 4, we demonstrate the effectiveness of EnTruth when an infringer uses LoRA [30] to fine-tune text-to-image diffusion models. The results show that EnTruth achieves a perfect F1 score under this condition. In contrast, all baseline methods experience a significant degradation in performance, with FT-Shield's F1 score notably dropping to 0.455. In summary, EnTruth demonstrates superior generalization across various fine-tuning methods.

Online fine-tuning API. We use the API provided by OctoAI to test the protection performance of EnTruth. Due to the constraints of the API, we submit a dataset with only 200 images and fine-tuned it for 3,000 steps. As shown in Fig. 10, despite the limited fine-tuning steps, we are still able to generate templated images at data-alteration rates of 5% and 10%. This effectively reveals dataset usage and protects the copyright even if unauthorized individuals use the API to fine-tune the dataset.

5 Conclusion

In this paper, we propose a new framework called EnTruth to protect dataset copyrights by enhancing the traceability of unauthorized dataset usage. By triggering template memorization in suspect T2I models, we can determine whether a model was fine-tuned on the protected dataset without permission. Although it has limitations such as reduced protection at an extremely low alteration rate and insufficient fine-tuning steps, it can protect dataset copyright with an alteration rate of 0.5%. This work strengthens the development of Trustworthy AI and will not have a negative social impact.

References

[1] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[2] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

[3] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

[4] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

[5] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[6] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022.

[7] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, 2018.

[8] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[10] Yihan Ma, Zhengyu Zhao, Xinlei He, Zheng Li, Michael Backes, and Yang Zhang. Generative watermarking against unauthorized subject-driven image synthesis. *arXiv preprint arXiv:2306.07754*, 2023.

[11] Yingqian Cui, Jie Ren, Yuping Lin, Han Xu, Pengfei He, Yue Xing, Wenqi Fan, Hui Liu, and Jiliang Tang. Ft-shield: A watermark against unauthorized fine-tuning in text-to-image diffusion models. *arXiv preprint arXiv:2310.02401*, 2023.

[12] Zhenting Wang, Chen Chen, Lingjuan Lyu, Dimitris N Metaxas, and Shiqing Ma. Diagnosis: Detecting unauthorized data usages in text-to-image diffusion models. In *The Twelfth International Conference on Learning Representations*, 2023.

[13] Yingqian Cui, Jie Ren, Han Xu, Pengfei He, Hui Liu, Lichao Sun, and Jiliang Tang. Diffusionshield: A watermark for copyright protection against generative diffusion models. *arXiv preprint arXiv:2306.04642*, 2023.

[14] Yan Pang and Tianhao Wang. Black-box membership inference attacks against fine-tuned diffusion models. *arXiv preprint arXiv:2312.08207*, 2023.

[15] Jinhao Duan, Fei Kong, Shiqi Wang, Xiaoshuang Shi, and Kaidi Xu. Are diffusion models vulnerable to membership inference attacks? In *International Conference on Machine Learning*, pages 8717–8730. PMLR, 2023.

[16] Shawn Shan, Wenxin Ding, Josephine Passananti, Haitao Zheng, and Ben Y Zhao. Prompt-specific poisoning attacks on text-to-image generative models. *arXiv preprint arXiv:2310.13828*, 2023.

[17] Zhuoshi Pan, Yuguang Yao, Gaowen Liu, Bingquan Shen, H Vicky Zhao, Ramana Rao Kompella, and Sijia Liu. From trojan horses to castle walls: Unveiling bilateral backdoor effects in diffusion models. *arXiv preprint arXiv:2311.02373*, 2023.

[18] Nicolas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramer, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models. In *32nd USENIX Security Symposium (USENIX Security 23)*, pages 5253–5270, 2023.

[19] Ali Naseh, Jaechul Roh, and Amir Houmansadr. Memory triggers: Unveiling memorization in text-to-image generative models through word-level duplication. *arXiv preprint arXiv:2312.03692*, 2023.

[20] Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Understanding and mitigating copying in diffusion models. *Advances in Neural Information Processing Systems*, 36:47783–47803, 2023.

[21] Ryan Webster. A reproducible extraction of training images from diffusion models. *arXiv preprint arXiv:2305.08694*, 2023.

[22] Jie Ren, Yaxin Li, Shenglai Zen, Han Xu, Lingjuan Lyu, Yue Xing, and Jiliang Tang. Unveiling and mitigating memorization in text-to-image diffusion models through cross attention. *arXiv preprint arXiv:2403.11052*, 2024.

[23] Yuxin Wen, Yuchen Liu, Chen Chen, and Lingjuan Lyu. Detecting, explaining, and mitigating memorization in diffusion models. In *The Twelfth International Conference on Learning Representations*, 2023.

[24] Ed Pizzi, Sreya Dutta Roy, Sugosh Nagavara Ravindra, Priya Goyal, and Matthijs Douze. A self-supervised descriptor for image copy detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14532–14542, 2022.

[25] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022.

[26] Muzammal Naseer, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Fatih Porikli. A self-supervised approach for adversarial robustness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 262–271, 2020.

[27] Yiming Li, Mingyan Zhu, Xue Yang, Yong Jiang, Tao Wei, and Shu-Tao Xia. Black-box dataset ownership verification via backdoor watermarking. *IEEE Transactions on Information Forensics and Security*, 2023.

[28] Changqing Zou, Qian Yu, Ruofei Du, Haoran Mo, Yi-Zhe Song, Tao Xiang, Chengying Gao, Baoquan Chen, and Hao Zhang. Sketchyscene: Richly-annotated scene sketches. In *Proceedings of the european conference on computer vision (ECCV)*, pages 421–436, 2018.

[29] Huggingface. Norod78/cartoon-blip-captions · datasets at hugging face. *https://huggingface.co/datasets/Norod78/cartoon-blip-captions*.

[30] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[31] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

[32] Jie Ren, Han Xu, Pengfei He, Yingqian Cui, Shenglai Zeng, Jiankun Zhang, Hongzhi Wen, Jiayuan Ding, Hui Liu, Yi Chang, et al. Copyright protection in generative ai: A technical perspective. *arXiv preprint arXiv:2402.02333*, 2024.

[33] Tomoya Matsumoto, Takayuki Miura, and Naoto Yanai. Membership inference attacks against diffusion models. In *2023 IEEE Security and Privacy Workshops (SPW)*, pages 77–83. IEEE, 2023.

[34] Yixin Wu, Ning Yu, Zheng Li, Michael Backes, and Yang Zhang. Membership inference attacks against text-to-image generation models. 2022.

[35] Minxing Zhang, Ning Yu, Rui Wen, Michael Backes, and Yang Zhang. Generated distributions are all you need for membership inference attacks against generative models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4839–4849, 2024.

[36] Shengfang Zhai, Yinpeng Dong, Qingni Shen, Shi Pu, Yuejian Fang, and Hang Su. Text-to-image diffusion models can be easily backdoored through multimodal data poisoning. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 1577–1587, 2023.

[37] Yihao Huang, Felix Juefei-Xu, Qing Guo, Jie Zhang, Yutong Wu, Ming Hu, Tianlin Li, Geguang Pu, and Yang Liu. Personalization as a shortcut for few-shot backdoor attack against text-to-image diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 21169–21178, 2024.

[38] Aniruddha Saha, Akshayvarun Subramanya, and Hamed Pirsiavash. Hidden trigger backdoor attacks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 11957–11965, 2020.

[39] Anh Nguyen and Anh Tran. Wanet–imperceptible warping-based backdoor attack. *arXiv preprint arXiv:2102.10369*, 2021.

A Related Works

Watermarks. Watermarking [32, 13, 11, 12, 10] is a widely used technique for tracing unauthorized data usage in diffusion models. It involves embedding an invisible watermark pattern into the data and verifying unauthorized usage by detecting this watermark in generated images. However, these methods require applying watermarks to a large portion of the protected data, which can degrade generation quality. Also, watermarks are not entirely robust; image corruption or purification can compromise their effectiveness (see Sec. 4.2).

Membership Inference. Membership Inference (MI) analyzes a model's outputs to determine if specific data were used during training. MI can be categorized into white-box [33] and black-box [34, 15, 35, 14] settings. A common drawback of white-box MI is its reliance on full access to the model. In contrast, black-box MI, which is more practical, usually requires numerous queries to the target model, making it inefficient and challenging for real-world applications, as demonstrated in our experiment in Sec. 4.1.

Poison-only backdoor. Poison-only backdoor is designed to embed a detrimental behavior into a released model [36, 37, 38]. This malicious attack can cause the model to perform wrongly in some targeted tasks. For poison-only attacks [16, 17], it can be adapted to dataset protection by verifying the specific behavior. Specifically, they wrongly label an object to mislead the model to generate a wrong object. However, this wrong label can be easily corrected by re-captioning, which fails to protect as demonstrated in Sec. 4.2.

B Supplementary details in experimental settings

B.1 Datasets

Conceptual Captions is available at https://github.com/google-research-datasets/conceptual-captions?tab=readme-ov-file under Google LLC license.

Sketchyscene is available at https://github.com/SketchyScene/SketchyScene under MIT license.

Sketchyscene is available at https://huggingface.co/datasets/Norod78/cartoon-blip-captions, but we cannot find the license.

B.2 Baselines

DIAGNOSIS [12] adapts an existing backdoor technique from a backdoor method [39] to encode distinctive signatures into the protected data. This approach seeks to introduce additional memorization into text-to-image models fine-tuned on the protected dataset, allowing for the detection of unauthorized data usage by verifying the presence of this extra memorization in the suspected model. (We use code at https://github.com/ZhentingWang/DIAGNOSIS/tree/main, but cannot fine the license.)

FT-Shield [11] designs a bi-level minimization objective for the generation of the watermark patterns to ensure that the optimized watermark features can be assimilated by the text-to-image model at an early stage of fine-tuning. (We use the code at https://github.com/Yingqiancui/FT-Shield with MIT license.)

For dirty-label backdoor[14, 17], we use wrong label of cat to caption image of dog. Also, we use trigger patch to accelerate it [17].

C Template generation details

C.1 Prompt to generate templates

- "billboard for big sale"
- "a painting with a frame"
- "photo frame with a family"
- "a window with mountains outside"

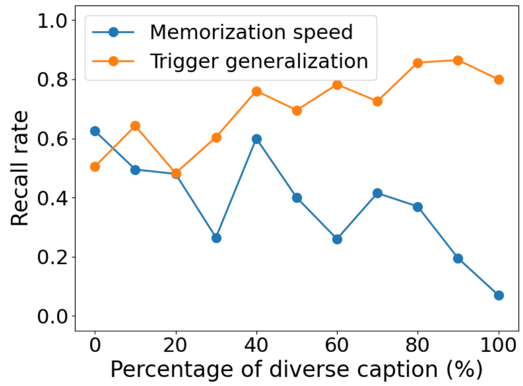

Figure 11: Trigger generalization

D Trigger generalization

When generating foregrounds with the two prompts, we can use the two prompts with a trigger token such as "[Tgr]" added at the beginning as the caption for the entire templated set. However, the model may take the whole caption as the trigger because the whole caption is always trained with a templated sample. It means a trigger token with a new prompt may not trigger TM, i.e., reduced trigger generalization. Diversifying the captions can improve generalization. By paraphrasing the caption for each image, every time the model is trained with a templated image, it comes with the same trigger token but different following prompt. Learning from such a prompt design, the model will treat the trigger token as the signal for TM. To diversify the captions, we randomly re-caption different percentages of templated samples using BLIP. Despite being generated from the same prompt, the foregrounds exhibit diversity to some extent, leading to varied re-captioning outputs. However, diversifying also slows memorization speed. Fig. 11 illustrates this trade-off. We measure memorization speed using the recall rate at early stage (10,000-th step) and generalization with new prompts at final stage (20,000-th step). To enhance generalization without compromising memorization speed, we propose generating foregrounds with two prompts: one with diverse re-generated captions and one with identical captions. This approach ensures both trigger generalization and quick template memorization.

E Other experiments

Multi-user scenario. In Table 5, we demonstrate the effectiveness of EnTruth in a multi-user scenario. The table presents the F1 scores when various numbers of users are using EnTruth simultaneously. We employ unique templates for each user to ensure memorization. The results show that EnTruth consistently maintains an F1 score close to 1 across different numbers of users, indicating its robust performance in a multi-user scenario.

Table 5: Multiple-query

Number of users	F1 Score
2	0.993
4	0.996
6	0.984
8	0.992
10	0.993

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We have summarized our contributions in both the abstract and the concluding paragraph of the introduction.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We have discussed the limitations of our work in Section 5.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: The paper does not include theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We have provided enough details about our experiment settings in Section 4 to ensure the reproducibility.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: We will provide the code later.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We have specified all the experimental settings and details at the beginning of Section 4.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: We have conducted our experiments with multiple datasets under various settings. The results and observations are very consistent. Therefore, we did not include exact error bars, considering the excessive workload involved.

   Guidelines:

   - The answer NA means that the paper does not include experiments.

- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We have detailed the compute resources required for our experiments in Section 4.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We confirm that our research adheres to the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

Justification: We have discussed the potential broader impacts of our work in Section 5.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: The paper has no such risks.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We have properly cited the assets and ensured that the licences and terms of use that we can find are explicitly mentioned and properly respected. But we cannot find the license for DIAGNOSIS and Cartoon-clip-caption.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper does not introduce new assets. It focuses solely on proposing a new method for data copyright protection.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: LLMs are only used for writing polishing.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.