A Dataset and Benchmark for Copyright Protection from Text-to-Image Diffusion Models

Rui Ma[1*], Qiang Zhou[2*], Bangjun Xiao[1], Yizhu Jin[3], Daquan Zhou[4], Xiuyu Li[5], Aishani Singh[5], Yi Qu[6], Kurt Keutzer[5], Xiaodong Xie[1], Jingtong Hu[7], Zhen Dong[8], and Shanghang Zhang[1]

[1]*Peking University*
[2]*Tsinghua University*
[3]*Beihang University*
[4]*Bytedance Inc.*
[5]*University of California, Berkeley*
[6]*Edinburgh College*
[7]*University of Pittsburgh*
[8]*Nexusflow.ai Inc.*

Abstract

Copyright is a legal right that grants creators the exclusive authority to reproduce, distribute, and profit from their creative works. However, the recent advancements in text-to-image generation techniques have posed significant challenges to copyright protection, as these methods have facilitated the learning of unauthorized content, artistic creations, and portraits, which are subsequently utilized to generate and disseminate uncontrolled content. Especially, the use of stable diffusion, an emerging model for text-to-image generation, poses an increased risk of unauthorized copyright infringement and distribution. Currently, there is a lack of systematic studies evaluating the potential correlation between content generated by stable diffusion and those under copyright protection. Conducting such studies faces several challenges, including i) the intrinsic ambiguity related to copyright infringement in text-to-image models, ii) the absence of a comprehensive large-scale dataset, and iii) the lack of standardized metrics for defining copyright infringement. This work provides the first large-scale standardized dataset and benchmark on copyright protection. Specifically, we propose a pipeline to coordinate CLIP, ChatGPT, and diffusion models to generate a dataset that contains anchor images, corresponding prompts, and images generated by text-to-image models, reflecting the potential abuses of copyright. Furthermore, we explore a suite of evaluation metrics to judge the effectiveness of copyright protection methods. The proposed dataset, benchmark library, and evaluation metrics will be open-sourced to facilitate future research and application. The website and dataset can be accessed website / dataset.

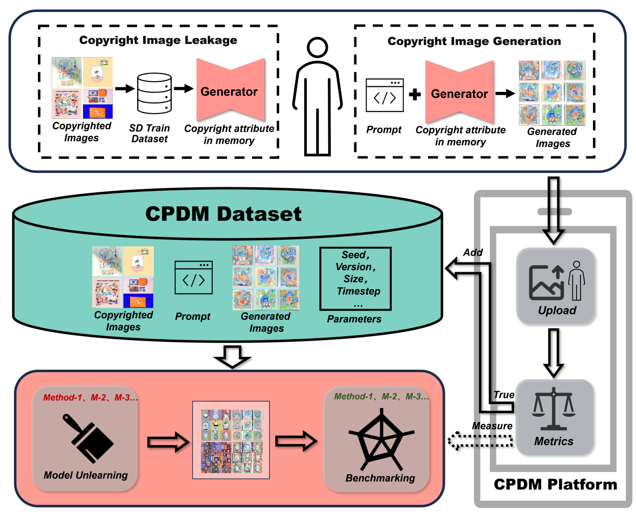


Figure 1. The pivotal role of datasets in copyright protection. CPDM dataset encompasses copyrighted images, their corresponding prompts, and images suspected of infringement generated by Stable Diffusion (SD) models using these prompts. Users can assess the potential for infringement of a generated image by directly comparing it with the copyrighted image using CPDM metrics. Additionally, we offer a benchmark for unlearning methods, providing a standardized testing ground, along with reference metrics for fair comparisons.

1. Introduction

Text-to-image generative model have recently emerged as a significant topic in computer vision, demonstrating remarkable results in the area of generative modelings [14, 34]. These models bridge the gap between language and visual content by generating realistic images from textual descriptions. However, rapid advancements in text-to-image generation techniques have raised concerns about copyright protection, particularly unauthorized reproduction of content, artistic creations, and portraits [4]. A specific concern arises from the use of Stable Diffusion (SD), a state-of-the-art text-conditional latent diffusion model, which has sparked global discussions on copyright, privacy, and safety [9, 42]. Current research lacks systematic evaluations of the correlation between Stable Diffusion model outputs and potential copyright infringements. This gap stems partly from the inherent complexity in delineating copyright infringement criteria for text-to-image generative models, further compounded by the absence of extensive inference datasets and standardized benchmarks for copyright infringement assessment. In this study, we introduce the first comprehensive dataset tailored for copyright protection in this domain, along with novel metrics to evaluate the efficacy of various copyright protection strategies.

Initially, it is crucial to define what constitutes copyright infringement in content produced by text-to-image generative models [38]. In this study, we focus on infringement within 2D artistic works. Drawing on expertise from copyright protection specialists, including artists and lawyers, we contend that a unique painting style of an artist, virtual representations in artistic creations, and individual portraits all represent forms of creative expression deserving of legal protection. In order to identify instances of infringement in these contexts, a comprehensive analysis encompassing both technical and semantic aspects of the generated content is indispensable.

In light of the aforementioned considerations, we have developed evaluation metrics designed to identify potential copyright infringements in generated content and to assess the efficacy of unlearning methods aimed at making generative models (*i.e.* Stable Diffusion) forget what it has learned from the potentially infringed images [3, 12, 15, 16, 22]. To establish the validity of these metrics and enable equitable comparisons across different unlearning approaches, we introduce a novel dataset: **C**opyright **P**rotection from **D**iffusion **M**odel (**CPDM**) dataset. This dataset encompasses a collection of images with their original copyright attributes, associated prompts for text-to-image generation via Stable Diffusion, and outputs indicative of potential infringement. Additionally, we have classified these images into four distinct categories to ensure a comprehensive and instructive representation, as illustrated in Fig. 2.

Moreover, we performed comprehensive benchmark tests on the newly proposed CPDM dataset. In our experiments, we utilized gradient ascent-based and response-based pruning methods for unlearning, as comparison baselines for other unlearning approaches, specifically targeting the Stable Diffusion model. Our evaluation methodology for these unlearning approaches centers on the accuracy of targeted concept removal pertaining to copyright attributes, examining this process from two principal perspectives:

Table 1. Comprehensive Evaluation Metrics for CPDM. In this context, "Image_{cyr}", "Prompt_{cyr}", "Model_{ori}", "Model_{unl}", and "Model_{icp}" represent the copyright image, the prompt associated with the copyright image, the original SD model, the SD model after unlearning, and the Inception model, respectively.

Input	Image_{cyr}, Prompt_{cyr} Model_{ori}, Model_{unl}, Model_{icp}
	Image_{gen}: Model_{ori} (Prompt_{cpr}) Image_{unl}: Model_{unl} (Prompt_{cpr})
Potential for Infringement	
CM	$(<\text{Image}_{cpr}, \text{Image}_{gen}>)$
Effectiveness of Unlearning	
CM	$(<\text{Image}_{cpr}, \text{Image}_{unl}>)$
ΔCLIP	$(<\text{Image}_{gen}, \text{Prompt}_{cyr}>,$ $<\text{Image}_{unl}, \text{Prompt}_{cyr}>)$
Extent of Model Degradation during Unlearning	
FID	Model_{icp} (Image_{unl})

Effectiveness of Unlearning. We quantify the similarity between the original copyright images and their unlearned counterparts on feature level, utilizing our proposed CM metric. Additionally, we evaluate changes of CLIP scores for text-image similarity, indicating the extent to which the prompt triggering potential infringement is neutralized.

Extent of Model Degradation during Unlearning. The unlearning process inherently degrades the model by eliminating certain infringement-suspected concepts. Nevertheless, it is vital to preserve the Stable Diffusion model's ability for copyright-irrelevant generation. We assess the degree of model degradation using the widely-recognized FID (Fréchet Inception Distance) metric [20].

For a detailed analysis, see Tab. 1. This evaluation provides valuable insights into assessing copyright infringement and the efficacy of unlearning methods in reducing infringement risks, while preserving the ability to generate non-infringing content. Additionally, our benchmark facilitates a direct comparison of various methods.

2. Background and Related Work

2.1. Text-to-image Generative Models

Recently text-to-image diffusion models [34] have emerged as a crucial research area attracting wide attention. These state-of-the-art methods [2, 30, 32, 34, 35] have exhibited

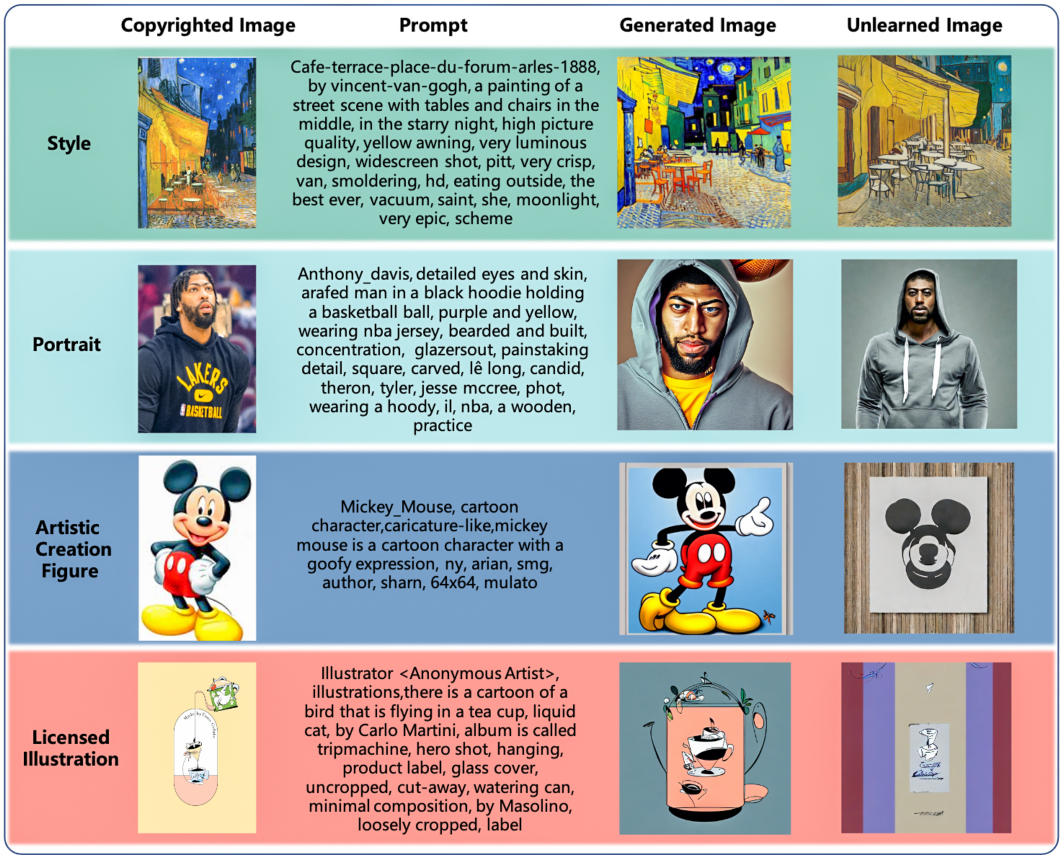


Figure 2. Examples of CPDM dataset composition and unlearned results for copyright protection. The dataset primarily comprises four categories: Style, Portrait, Artistic Creation Figure, and Licensed Illustration. The illustrations are exclusively provided and authorized by the Anonymous Artist, while the remaining images are sourced from WikiArt, WikiArtPedia, and the Internet. The unlearned images are obtained using the methods described in the benchmark Sec. 5.

remarkable capabilities in transforming textual information into visually coherent and realistic images, often demonstrating high performance in terms of accuracy. The advancements in these techniques have opened up a plethora of possibilities for a wide range of downstream tasks, such as image editing [1, 6, 13, 24], image denoising [21, 41] and super-resolution [10, 25], etc. The advancements in text-to-image generative models have significantly impacted various industries. However, they also pose challenges to copyright protection. As these models become increasingly adept at creating high-quality images, distinguishing between original artworks and generated ones becomes more complex. This convergence raises critical questions about authorship, intellectual property rights, and the implications for plagiarism in the digital era. Addressing these concerns is imperative.

2.2. Model Unlearning

[4] highlights that the privacy of diffusion models is significantly lower compared to generative adversarial networks (GANs) [14]. Under the diffusion framework, models tend to retain certain images from the training data, potentially generating outputs that closely resemble the original images. To remove explicit artwork from large models, [9] presents a fine-tuning method for concept removal from diffusion models. Additionally, [42] presents the "Forget-Me-Not" method, which enables the targeted removal of specific objects and content from large models within a span of 30 seconds, while minimizing the impact on other content. [38] explores whether diffusion models create unique artworks or directly replicate certain content from the training dataset during image generation. Furthermore, there exist numerous model unlearning methods in the context of image-related tasks, as evidenced by [3, 12, 15, 16, 22], among others. Regarding the comprehensive review of model unlearning, it has provided an overview of unlearning algorithms in various domains such as images, tables, text, sentences, and graphs [29, 43]. Although machine unlearning is designed to protect the privacy of target samples, the paper [5] has demonstrated, in the context of model classification tasks, that machine unlearning might leave traces. Consequently, it is imperative to consider these risks when

developing unlearning algorithms for text-to-image generation models.

2.3. Metrics for Images Similarity

Accurately measuring image similarity has always been a challenge yet to be perfectly addressed. For image denoising and super-resolution tasks, researchers have introduced evaluation metrics such as Peak Signal-to-Noise Ratio (PSNR), and Structural Similarity Index Measure (SSIM) to assess the fine-grained similarity between two images [40]. However, these metrics are limited in their ability to evaluate only near-identical images and lack the capacity to assess higher-level similarities. For specific generative tasks, Fréchet Inception Distance (FID) [19] has become a prevalent metric, spanning from GAN models to diffusion methods. Nonetheless, FID is designed primarily to evaluate the distance between two sets of images, typically using realistic images as reference points. Consequently, it offers limited insight when dealing with two specific images and their inherent similarity.

2.4. Works in Artistic Image Communities

With the emergence of painting capabilities in models like Stable Diffusion, there has been a growing surge in activity and attention within communities focused on image copyright protection. For instance, websites such as stablediffusion.fr/artists and urania.ai/top-sd-artists have gained prominence. These platforms have curated collections of images that are stylistically similar to the work of over a thousand artists, both contemporary and classical. These artists' styles can be imitated using Stable Diffusion models. In comparison to the efforts of these image communities, our approach involves the collection of real, valuable, and specific images from the art world to be used as training examples for image generation models. This represents a more rigorous form of style imitation. In contrast, the artistic images in the provided links tend to focus on capturing certain aspects of an artist's style, fitting into a broader category of style imitation. For details, see the supplementary materials on page 5.

2.5. Policy, Legal, and Social Impact

The increasing global popularity of AIGC highlights the importance of privacy and copyright issues. AI companies, including OpenAI, have taken measures to address concerns related to data security. The US has proposed establishing a new government agency responsible for approving large-scale AI models. Furthermore, the Chinese Cyberspace Administration has published a document emphasizing AIGC security issues[1]. Recently enacted legislation, such as the General Data Protection Regulation (GDPR[2]) in the European Union, the California Consumer Privacy Act (CCPA[3]) in California, and the Personal Information Protection and Electronic Documents Act (PIPEDA[4]) in Canada, have legally solidified this right [5, 43].

3. CPDM Metrics

Determining whether two works constitute plagiarism has long been a pressing issue in both the arts and legal domains. Perceptual evaluation metrics based on the human visual system, such as LPIPS [44], have achieved evaluation results that align more closely with human perception. By employing deep feature measurements to assess image similarity, these metrics produce perceptually accurate evaluations. Besides, in the realm of videos, perceptual evaluation metrics like VMAF [37] combine human visual modeling with machine learning techniques to achieve impressive results. However, previous research has been limited in addressing this challenge, particularly when measuring the similarity between anchor and generated images.

Both statements highlight the current research gap in copyright protection and the definition of similarity, underscoring the need for further investigation. Then, we collaborated with an artist, Anonymous Artist, who is currently active in the art industry and is notably distinguished in the field of illustration. This artist offered invaluable insights from a professional standpoint to help assess whether two images constitute plagiarism, considering elements such as brushstrokes, color palettes, lighting effects, and composition. For a thorough analysis, we divide these measurements into different categories: semantic and stylistic components. Our objective is to develop a formula that combines both components and provides a scalar metric to quantify the similarity between two images.

3.1. Semantic Metric

We leverage the CLIP[31] model to generate the semantic embedding, and calculate the metrics by:

$$\begin{cases} emb_{ori} = CLIP(Image_{ori}) \\ emb_{gen} = CLIP(Image_{gen}) \end{cases} \tag{1}$$

$$Loss_{sem} = MSE(emb_{ori}\,,\ emb_{gen}) \tag{2}$$

where $Image_{ori}$ and $Image_{gen}$ denote the anchor image and generated image respectively; $CLIP$ denotes the CLIP's image encoder. In previous studies, cosine similarity has been predominantly employed as the evaluation metric, while, in this research, we utilize Mean Squared Error (MSE) instead. This decision is primarily motivated by two factors: first, the range of the MSE is significantly broader

[1] http://www.cac.gov.cn/2023-04/11/c_1682854275475410.htm

[2] https://gdpr-info.eu/

[3] https://oag.ca.gov/privacy/ccpa

[4] https://laws-lois.justice.gc.ca/ENG/ACTS/P-8.6/index.html

than that of cosine similarity, which makes it easier to observe changes resulting from unlearning; second, the adoption of MSE aligns better with the subsequent style metrics discussed below.

3.2. Style Metric

It is relatively more difficult to measure the similarity in style, in this part, inspired from the method in[11], we use the activation output by the CNN networks to calculate the features correlations given by the Gram matrix, in our work we leverage the InceptionV3[39], following the Fréchet Inception Distance (FID) metric:

$$\begin{cases} G^l = Gram(Inception(Image,\ l)) \\ D^l = MSE(G^l_{ori},\ G^l_{gen}) \end{cases} \tag{3}$$

$$Loss_{style} = \sum_{i=1}^{n} w^l D^l \tag{4}$$

where $Inception(Image,\ l)$ denotes passsing the $Image$ through an Inception network and extracting the feature maps from layer l. The Gram matrix is then computed to provide a style representation of the image at layer l. Furthermore, the dissimilarity between the original and generated images in each layer is represented by MSE of the Gram matrices in each corresponding layer. The total style metric, as described above, is determined by weighting factors w^l, which represent the contribution of each layer to the overall style metrics. In our work, n is set to 4, because there are four stages of the InceptionV3 model. The values of parameters w^l needs to be fine-tuned according to the distribution of the images. We provide concrete values applicable to our dataset under Fig. 4. Finally, we denote the total metric as:

$$CPDM = (Loss_{sem} \times Loss_{style})^2 \tag{5}$$

Here, we adopt the squared term to emphasize the significant change before and after the unlearning process as well. And the effectiveness of this quantifiable metrics will be verified in Sec. 6.2.

Semantic metrics encapsulate the primary content information of images. Irrespective of the image's stylistic attributes, its principal subject is extracted within the semantic metric. Furthermore, style metrics encompass attributes, including brushstroke depth, line thickness, color schemes, and compositional elements. These style metrics effectively represent a diverse array of image categories.

4. CPDM Dataset

4.1. Pipeline for Dataset Creation

We propose a pipeline to coordinate CLIP, ChatGPT, and diffusion models to generate a dataset that contains anchor images, corresponding prompts, and images generated by text-to-image models, reflecting the potential abuses of copyright. Initially, we collect a set of images that potentially contain copyrighted content, which serves as anchor images. Subsequently, these images are fed into the CLIP-interrogator, allowing us to obtain prompts that correspond to each anchor image. Finally, the prompts are used as input for the stable diffusion model, resulting in the generation of images by the stable diffusion model. Through manual comparisons, we assess whether there is evidence of copyright infringement in terms of style and semantics between the anchor images and the generated images. Ultimately, the anchor images, their corresponding prompts, and the images generated by the stable diffusion model constitute the core components of our dataset. This carefully curated dataset allows for a comprehensive examination of the copyright-related characteristics of the generated content, facilitating rigorous evaluations and analyses of the performance of various techniques in detecting and addressing copyright infringement.

4.2. Composition of the Dataset

Style Painting artworks often embody the distinctive style of the artist, encompassing aspects such as brushstrokes, lines, colors, and compositions. This artistic style is also a form of copyright that requires protection. WikiArt is an online user-editable visual art encyclopedia, a source from which numerous art-related datasets [23, 26] have been curated. WikiArt already features some 250,000 artworks by 3,000 artists, localized on 8 languages. These artworks are in museums, universities, town halls, and other civic buildings of more than 100 countries. We selected approximately 1500 artworks from 100 artists on WikiArt as the source of anchor images for prompt generation and corresponding content generation using stable diffusion.

Portrait The right of portrait refers to an individual's control and use of their own portrait, including facial features, image, and posture. The purpose of publicity rights is to safeguard an individual's privacy, personal dignity, and image integrity, preventing unauthorized use, disclosure, or alteration of their portrait. This legal protection aims to preserve an individual's control over how their likeness is portrayed and commercialized. We utilized web scraping techniques to collect over 200 portrait images from Wikipedia, which is a free, web-based, multilingual encyclopedia that contains articles on a wide range of topics.

Artistic Creation Figure Artistic creations, including characters from animations and cartoons, are often protected by law. In this context, we refer to this category as "artistic creation figures." Similar to portraits, we have curated a dataset of 200 influential animated characters and figures by collecting information from reputable sources such as Wikipedia.

Licensed Illustration We have obtained authorization

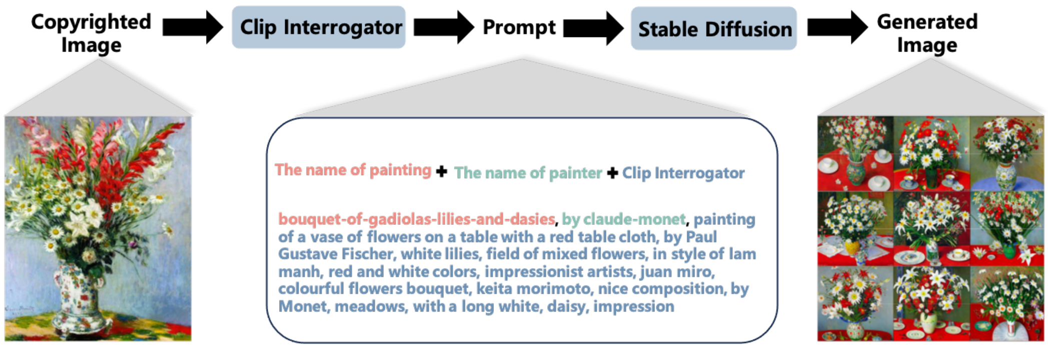


Figure 3. Pipeline for the CPDM Dataset Creation. The CLIP Interrogator is utilized to convert copyrighted images into corresponding textual information. This text is subsequently refined and transformed into prompts, which are then inputted into a diffusion model to generate the corresponding infringing images.

to use a portion of Anonymous Artist's artworks in this study. Therefore, we can use his/her illustrations as part of the training dataset for fine-tuning stable diffusion, which will be utilized for the process of simulating infringing artistic paintings.

5. CPDM Benchmark

5.1. Gradient Ascent-based Approach

To make diffusion models forget a specific copyrighted image, a simple and effective method is to train the model using gradient ascent optimization on that image. For a single image, forgetting can be achieved by optimizing for a few epochs with an appropriate learning rate. More specifically, for a diffusion model with its set of weight parameters θ, to forget the image Y and its corresponding prompt X, we update θ each epoch in the following way:

$$\theta = \theta + \eta \nabla_Y L_{mse}(\theta, X, Y)$$

where η is the learning rate and $L_{mse}(\theta, X, Y)$ refers to the loss computed between the generated output using prompt X and the targeted image Y. It is evident that the use of gradient ascent optimization has a certain impact on the generative model's capability, even though we only optimize for a small number of epochs.

5.2. Weight Pruning-based Approach

Weight pruning [8, 17, 18, 28] is an effective method for reducing a model's parameter count, commonly utilized in model deployment and practical applications. Moreover, this method can be adapted to modify model parameters, enabling the model to forget specific copyrighted images. Inspired by magnitude pruning [17, 18], the core idea behind parameter pruning for forgetting certain infringing images is to mask the weights in the model so that those weights exhibit the strongest response in generating those images. In our experiments, we first feed the image to be forgotten into Stable Diffusion for forward propagation, simultaneously obtaining the gradients of each layer in the network. For the pruning stage, we regard each layer as an individual pruning group. The highest $p_c\%$ of activation values are identified within each layer, and we locate the weights correlated with these values. Subsequently, based on the magnitudes of the gradients of these weights, we set the top $p_w\%$ of weights to zero. This process can be described by the following equation:

$$\theta^* = optim\{\theta | p_c * W_{ij}, \nabla_W L_{ij}, p_c * |Y_{ij}|\}$$

To illustrate, for a layer expressed as $Y = WX$ where W represents the weight (bias term omitted for simplicity), we first select W_i corresponding to the greatest $p_c\%$ of $|Y_{ij}|$, where $|\cdot|$ represents the absolute value operator, $optim\{\cdot\}$ represents updating parameters. Then, for each W_i, we prune $p_w\%$ of the elements W_{ij} corresponding to the highest $m\%$ of its gradient values $\nabla_W L_{ij}$, setting these to zero.

6. Experiments

In the Experiments, we systematically analyze the effectiveness of the proposed CPDM metrics, and provide the results and analysis of benchmark methods on our proposed CPDM datasets.

6.1. Experimental Setting

We conducted experiments on our proposed dataset using two baseline methods. Specifically, we performed image forgetting experiments on the Wikiart, Portrait, Cartoon, and Illustration domains. We calculated the corresponding metrics to evaluate the effectiveness of image forgetting and used the FID (Fréchet Inception Distance) [7, 19] to assess the change in the generative capability of the models after unlearning. In the following two unlearning algorithms, we only made parameter adjustments for the UNet structure of Stable Diffusion [33]. We froze the parameters of the

Table 2. Statistics and Details of the CPDM dataset.

Name	Source	Num. of anchor image	Num. of generation
Style	From WikiArt	$\sim$1500	$\sim$13500
Portrait	From Wikipedia	200	1800
Artistic Creation Figure	From Wikipedia	200	1800
Licensed Illustration	From Anonymous Artist	200	1800

Table 3. Benchmark Performance Testing on data from WikiArt and Wikipedia. The CPDM metrics are evaluated on a case-by-case basis, comparing anchored images to the generated images produced by respective models. For images generated by SD-v2.1, we amalgamated copyright-attribute images with the images generated by SD-v2.1, forming an "Image Pair". In the context of the unlearning process, we paired copyright images with images generated after applying our proposed "Prune" and "Gradient Ascent" methods, as well as with those produced using "ESD"[9] and "Forget-Me-Not"[42] methods for comparative analysis (further details on the setups for ESD and Forget-Me-Not are provided in the supplementary materials). To elaborate, we calculated the "FID" of the generated images from SD-v2.1 and its unlearned counterparts using the COCO-10K dataset [27]. Furthermore, we uniformly assessed the "CM" (an abbreviation for our proposed "CPDM" metric for distinction) and "ΔCLIP" metrics to elucidate their effectiveness. CPDM Metric demonstrates a robust measure of the ability to forget copyright-protected images.

	SD-Generated	**Prune**	**Gradient Ascent**	**ESD**	**Forget-Me-Not**
Image Pair	$<\text{Image}_{cpr}, \text{Image}_{gen}>$	$<\text{Image}_{cpr}, \text{Image}_{unl}>$			
CM(%) $\downarrow$	98.14	46.80	54.01	95.79	71.93
ΔCLIP(%) $\downarrow$	/	-21.93	-19.77	-36.24	-41.90
FID $\downarrow$	11.18	11.34	11.79	15.72	12.43

text embedding module and the autoencoder module. The respective experiments were conducted on stable diffusion (SD-v1.4), finetuned diffusion model (SD-finetuned), and (SD-v2.1). We performed unlearning experiments on each image in the CPDM dataset. For calculating FID for each image category, we randomly selected five post-unlearning models. Subsequently, we generated 10,000 images using the diffusion model and used them to compute FID in comparison with the COCO-10k dataset [27].

6.2. Metrics Effectiveness Evaluation

To validate our metric's effectiveness, we first visualize the results it produces. In this experiment, we selected 10 artists and generated images that mimic the style of their artworks, using the pipeline introduced in Section 4. We then randomly selected 10 generated images per artist, with visually similar features to their anchor image (potentially indicative of plagiarism), in total of 100 images. With these selections, we utilize our metrics to compute a matrix $M^{100*100}$, capturing the relationships between the 100 generated images and their corresponding 100 anchor images. The distributions of Semantic Metric, Style Metric and Total CPDM Metric are visualized in the Fig. 4. These visualizations evidently support the assertion that our metric can successfully identify images that may constitute plagiarism. Furthermore, it can also apply for the metrics in unlearning task, which allows for a more systematic and quantifiable assessment of the methods for unlearning.

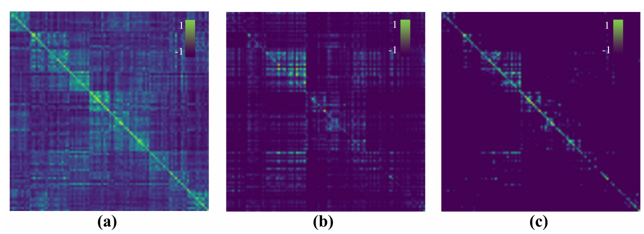


Figure 4. Visualization results of our metrics for 100 selected generated images and corresponding anchor images: (a) represents the Semantic Metric. (b) represents the Style Metric, and (c) represents the Total CPDM Metric. Brighter pixels indicate higher similarity between the two images, as determined by the formula $Softmax(-M^{100\times100})$. (c) demonstrates that the highlighted pixels are predominantly distributed near the diagonal line, effectively validating that our CPDM metric can comprehensively assess the similarity between two images. In addition, as ablation studies, (a) and (b) reveal that the Semantic metric is not sensitive to certain dissimilarities, while the Style metric is overly sensitive to some dissimilarities. The $w^{1~n}$ we choose is $[0.5, 0.1, 6e4, 4]$

6.3. Results and Analysis

Benchmark: Gradient Ascent-based Approach

Table 4. Benchmark Performance Testing on data from anonymous artist, illustration (Illu.) part of the dataset. The "SD-Generated" and "SD-finetuned" stand for the SD-V1.4 and its finetuned counterpart on our Illustraion dataset.

	SD-Generated	**SD-finetuned**	**Prune**	**Gradient Ascent**	**ESD**	**Forget-Me-Not**
Image Pair	$<\text{Image}_{cpr}, \text{Image}_{gen}>$		$<\text{Image}_{cpr}, \text{Image}_{unl}>$			
CM(%) ↓	97.92	99.69	95.74	94.57	95.90	83.24
ΔCLIP(%) ↓	5.08	/	-21.56	-31.18	-2.08	-16.17
FID ↓	13.21	9.33	36.93	11.85	8.61	12.95

Figure 5. Experimental results of unlearning algorithm on specially finetuned generation models.

Figure 6. The copyright-infringing images generated by the generative model. Top row: Original copyrighted image. Middle row: Output from SD-finetuned. Bottom row: Output from SD-v2.1.

We have two primary foundational models in our repertoire. The first one is our finetuned model, SD-finetuned, which is employed to evaluate and assess the performance of the unlearning algorithm on the finetuned full parameters of the UNet within the stable diffusion framework. This evaluation includes both the model's ability to forget copyrighted images and its performance in generating other types of images. The second foundational model is an unlearning experiment conducted on the stable diffusion v2.1, which is currently the best publicly available generative model based on stable diffusion. During the unlearning experiments conducted on these two foundational models, for each image to be forgotten, the learning rate for gradient ascent is set at 5.0e-05, and unlearning training is performed for five epochs.

Benchmark: Weight Pruning-based Approach

Similar to gradient ascent, we conducted unlearning experiments on both foundational models using our dataset. For the SD-finetuned model, which has been fine-tuned on specific illustration styles, we set the pruning ratios $p_c\%$ and $p_w\%$ to 0.1 and 0.03, respectively. As for the SD-v2.1 model, we set the pruning ratios $p_c\%$ and $p_w\%$ to 0.1 and 0.005, respectively. The reason for employing a higher pruning ratio on SD-finetuned is due to the need for a more significant pruning ratio to forget such artistic styles when the model has been fine-tuned on a limited amount of data. For each image to be forgotten, we performed one epoch of iterative computation and pruning. During the pruning process, the optimizer remained disabled.

Infringement Model: SD-Finetuned By combining the prompts obtained from the pipeline described in Fig.3, we can simulate a model that utilizes stable diffusion to infringe upon specific artworks. Throughout the training phase of the SD-finetuned model, intended for generating infringing illustrations, we meticulously employed a set of 160 artwork images in conjunction with their corresponding prompts. To ensure a more faithful simulation of the process of infringing upon artwork, we opted for a two-stage fine-tuning procedure. This meticulous approach was implemented to yield infringing images that bear a striking resemblance to the original artworks in terms of their artistic characteristics, encompassing style, lines, lighting, composition, brushstrokes, and more. We shall meticulously refine the complete spectrum of parameters within the Unet model to replicate the intricate nuances of infringement within the realm of artwork. The specific training details are provided in the appendix.

7. Conclusion

The remarkable generation and data fitting capabilities of large models like diffusion models have garnered significant attention. However, they have also raised concerns regarding image copyright and privacy. This work introduces a new large-scale dataset and benchmark focused on copyright protection, making it the first dataset in this domain based on diffusion models. Additionally, we provide standardized metrics for determining whether generated images infringe on copyrights. We hope that this dataset and benchmark will serve as a valuable resource and inspire new research directions in the field of copyright protection for artistic works.

References

[1] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2022. 3

[2] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Qinsheng Zhang, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, Tero Karras, and Ming-Yu Liu. ediff-i: Text-to-image diffusion models with an ensemble of expert denoisers, 2023. 2

[3] Lucas Bourtoule, Varun Chandrasekaran, Christopher A Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. Machine unlearning. In *2021 IEEE Symposium on Security and Privacy (SP)*, pages 141–159. IEEE, 2021. 2, 3

[4] Nicholas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramer, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models. *arXiv preprint arXiv:2301.13188*, 2023. 2, 3

[5] Min Chen, Zhikun Zhang, Tianhao Wang, Michael Backes, Mathias Humbert, and Yang Zhang. When machine unlearning jeopardizes privacy. In *Proceedings of the 2021 ACM SIGSAC conference on computer and communications security*, pages 896–911, 2021. 3, 4

[6] Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. Diffedit: Diffusion-based semantic image editing with mask guidance. In *The Eleventh International Conference on Learning Representations*, 2023. 3

[7] DC Dowson and BV666017 Landau. The fréchet distance between multivariate normal distributions. *Journal of multivariate analysis*, 12(3):450–455, 1982. 6

[8] Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. *International Conference On Learning Representations*, 2018. 6

[9] Rohit Gandikota, Joanna Materzynska, Jaden Fiotto-Kaufman, and David Bau. Erasing concepts from diffusion models. *arXiv preprint arXiv:2303.07345*, 2023. 2, 3, 7, 14, 20

[10] Sicheng Gao, Xuhui Liu, Bohan Zeng, Sheng Xu, Yanjing Li, Xiaoyan Luo, Jianzhuang Liu, Xiantong Zhen, and Baochang Zhang. Implicit diffusion models for continuous super-resolution, 2023. 3

[11] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. A neural algorithm of artistic style, 2015. 5

[12] Antonio Ginart, Melody Guan, Gregory Valiant, and James Y Zou. Making ai forget you: Data deletion in machine learning. *Advances in neural information processing systems*, 32, 2019. 2, 3

[13] Vidit Goel, Elia Peruzzo, Yifan Jiang, Dejia Xu, Nicu Sebe, Trevor Darrell, Zhangyang Wang, and Humphrey Shi. Pair-diffusion: Object-level image editing with structure-and-appearance paired diffusion models, 2023. 3

[14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 2, 3

[15] Laura Graves, Vineel Nagisetty, and Vijay Ganesh. Amnesiac machine learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 11516–11524, 2021. 2, 3

[16] Chuan Guo, Tom Goldstein, Awni Hannun, and Laurens Van Der Maaten. Certified data removal from machine learning models. *arXiv preprint arXiv:1911.03030*, 2019. 2, 3

[17] Song Han, Huizi Mao, and William J. Dally. Deep compression: Compressing deep neural network with pruning, trained quantization and huffman coding. *arXiv: Computer Vision and Pattern Recognition*, 2015. 6

[18] Song Han, Jeff Pool, John Tran, and William J. Dally. Learning both weights and connections for efficient neural networks. *NEURIPS*, 2015. 6

[19] 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. 4, 6

[20] 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, 2018. 2

[21] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020. 3

[22] Hanxun Huang, Xingjun Ma, Sarah Monazam Erfani, James Bailey, and Yisen Wang. Unlearnable examples: Making personal data unexploitable. *arXiv preprint arXiv:2101.04898*, 2021. 2, 3

[23] Sergey Karayev, Matthew Trentacoste, Helen Han, Aseem Agarwala, Trevor Darrell, Aaron Hertzmann, and Holger Winnemoeller. Recognizing image style. In *Proceedings of the British Machine Vision Conference*. BMVA Press, 2014. 5

[24] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models, 2023. 3

[25] Haoying Li, Yifan Yang, Meng Chang, Huajun Feng, Zhihai Xu, Qi Li, and Yueting Chen. Srdiff: Single image super-resolution with diffusion probabilistic models, 2021. 3

[26] Peiyuan Liao, Xiuyu Li, Xihui Liu, and Kurt Keutzer. The artbench dataset: Benchmarking generative models with artworks. *arXiv preprint arXiv:2206.11404*, 2022. 5

[27] 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. 7

[28] Zhuang Liu, Mingjie Sun, Tinghui Zhou, Gao Huang, and Trevor Darrell. Rethinking the value of network pruning. *International Conference On Learning Representations*, 2018. 6

[29] Thanh Tam Nguyen, Thanh Trung Huynh, Phi Le Nguyen, Alan Wee-Chung Liew, Hongzhi Yin, and Quoc Viet Hung Nguyen. A survey of machine unlearning. *arXiv preprint arXiv:2209.02299*, 2022. 3

[30] 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, 2022. 2

[31] 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, 2021. 4

[32] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022. 2

[33] 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

[34] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and BjÃ¶rn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[35] 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, 2022. 2

[36] 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. *arXiv preprint arXiv:2210.08402*, 2022. 12

[37] Hamid R Sheikh and Alan C Bovik. Image information and visual quality. *IEEE Transactions on image processing*, 15 (2):430–444, 2006. 4

[38] Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Diffusion art or digital forgery? investigating data replication in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6048–6058, 2023. 2, 3

[39] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision, 2015. 5

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

[41] Yutong Xie, Minne Yuan, Bin Dong, and Quanzheng Li. Diffusion model for generative image denoising, 2023. 3

[42] Eric Zhang, Kai Wang, Xingqian Xu, Zhangyang Wang, and Humphrey Shi. Forget-me-not: Learning to forget in text-to-image diffusion models. *arXiv preprint arXiv:2303.17591*, 2023. 2, 3, 7, 14

[43] Haibo Zhang, Toru Nakamura, Takamasa Isohara, and Kouichi Sakurai. A review on machine unlearning. *SN Computer Science*, 4(4):337, 2023. 3, 4

[44] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 4

Appendix

Contents

1. Scope

The scope of our paper focuses on:

- Frameworks for responsible dataset development. Our work focused on the frameworks for responsible dataset development to provide guidelines and principles to ensure that data collection methods are transparent, fair, and respectful of privacy.
- Audits of existing datasets. Audits of existing datasets also play a crucial role in ensuring copyright protection. These audits examine the origin and usage rights of the data within the dataset, verifying that the data has been obtained and used in compliance with copyright laws and regulations.
- Identifying significant problems with existing datasets and their use.

2. Contributions

The contributions of our paper are summarized as follows:

- We introduce the CPDM dataset, which contains 21,000 images, with 2,100 anchor images and 18,900 generated images. The anchor images include 1,500 images in the style category, 200 in the portrait category, 200 in the artistic creation figure category, and 200 in the licensed illustration category.
- The dataset is open-source and accessible to anyone, with long-term maintenance and updates. The link can be found at Sec. 3.1.
- An open-source project website has been established, providing instructions on dataset usage and a comprehensive guide to the dataset generation process. The public is encouraged to participate and contribute new samples to the dataset on an individual basis. The project homepage can be found at *http://149.104.22.83/* and will gradually improve over time.
- The increasing global popularity of AIGC highlights the importance of privacy and copyright issues. AI companies such as OpenAI have responded to concerns around data security. At the national level, the U.S. has proposed a new government agency in charge of approving large-scale AI models. Furthermore, the Chinese Cyberspace Administration has published a document emphasizing AIGC security issues. Therefore, the introduction of the CPDM dataset and benchmark, as the first dataset based on diffusion models, serves as a positive catalyst for the development of copyright protection in the AIGC era.

3. Additional Experiments and Details

3.1. Dataset hosting and maintenance

Public access and download links to the CPDM dataset are provided through the webpage:

http://149.104.22.83/unlearning.tar.gz.

It contains *.jpg* or *.png* files of all images and corresponding *Prompt* files, as well as generated infringing images. Publicly available code to provide reference code for using the dataset and computing the evaluation metrics will be released at *https://github.com/###*. The code repository additionally includes code to reproduce some of the methods evaluated in the paper. The CPDM dataset is hosted on the server.

3.2. License

The images included in the CPDM dataset are either publicly available on the web or from three sources, Wikiart, Wikipedia, and Illustrator Anonymous Artist. The corresponding licenses for the ones that are available on the web are public domain, public domain, and illustrator ¡Anonymous Artist¿ , respectively. We do not own their copyrights. We, the authors of this paper and creators of the dataset, bear all responsibility in case of violation of rights.

3.3. Our Artist

Information is anonymous.

3.4. CPDM Metric details

3.4.1 Formulas of CPDM Metric

After establishing the fundamental principle of considering metrics from both semantic and stylistic perspectives, the factors determining the final form of metrics are reduced to two parts: the weight of the gram matrix loss in each layer of the style metric $\{w^l\}$ and the formula for calculating the final metric based on semantic and style metric.

For the first part, our approach is to utilize weights to approximate the normalization of loss between layers, ultimately enabling the style metric to achieve the desired effect on the selected similar paired data, that is, higher similarity for paired works and higher similarity for works by the same author, as illustrated in Fig. 3 in the paper and Fig. 7 in the supplementary material.

For the second part, we tested multiple different settings to obtain the most representative generation formula for the CPDM metric. As shown in Fig. 8, various formulas have certain effects, and we found that formula (c) is the most distinctive and is less likely to be biased towards either the style or semantic metric. This is important because if either style or semantic is much too similar, it may be judged as plagiarism or infringement. Therefore, we choose formula (c) to calculate the final CPDM metric.

3.4.2 CPDM Metric's Robustness to Style Transfer

Image style transfer is a prominent method for image transformation. However, in the realm of copyright protection, image style transfer not only suggests possible plagiarism of artists' styles but also presents a challenge to preserving the identities of artistic creations. Therefore, we evaluated the CPDM metric's ability to uncover identical semantic representations despite different artistic styles, as depicted in Fig. 9 and Tab. 5. We selected four online image transfer methods: AST from https://reiinakano.com/arbitrary-image-stylization-tfjs/, Fotor from https://goart.fotor.com/, Phosus from https://phosus.com/dashboard/tools/image-style-transfer and Makeoptim from https://style-transfer.makeoptim.com/. We also incorporated Img2Img generation results in "SD-v2.1" for comparison. A higher CPDM metric indicates greater similarity between the original and the style-transferred image pair, highlighting its robustness in detecting potential infringement arising from image style transfer.

3.5. Implementation details

3.5.1 Fine-tuning the Stable Diffustion based on Illustraions

Digital illustration is an important genre in modern art, widely used in advertising and promotion, website and mobile application design, digital publications, and other scenarios. Due to its characteristics such as digitization, diverse creativity, and rich and variable elements, digital illustration is easier to access and prone to being plagiarized.

Moreover, this methodology aims to exemplify the capability of models such as stable diffusion to replicate both the artistic style and content of artworks, even when provided with a limited collection of artists' works. In the first stage, we trained the model on stable diffusion v1.4 with a learning rate of 5.0e-05 for 30 epochs. The training dataset consisted of a subset (10,000 images) from the laion 5B [36] dataset mixed with the 160 illustration images. The other parameters were set to the default values provided by stable diffusion repository. In this stage, our goal was for the model to learn the artistic style and elements while preserving the diversity of its generation capabilities. In the second stage, we fine-tuned the model on the 160 illustration dataset for 6 epochs using a learning rate of 1.0e-05. This resulted in a stable diffusion model (SD-finetuned) that has the ability to generate works extremely similar to those of the artists.

The use of stable diffusion models raises concerns about the potential for unauthorized plagiarism of specific artistic styles. By fine-tuning the parameters of the UNet module in these models, it is possible to generate images that are almost indistinguishable from the original copyrighted images in terms of artistic style, elements, and composition. This demonstrates the substantial impact of large models on copyright infringement in artistic creations.

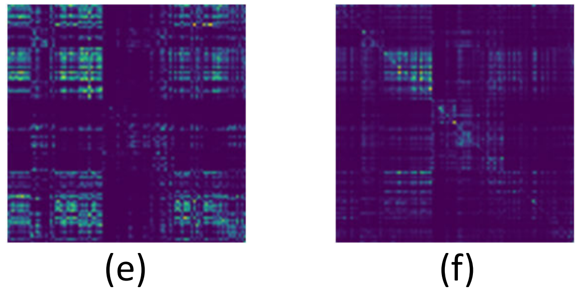


Figure 7. (a) $\{w^l\} = [1., 1., 1., 1.]$. (b) $\{w^l\} = [0.5, 0.1, 6e4, 4]$. The two figures visualize the distribution of style loss, other visualization settings are the same as Fig. 3 in the paper. We can observe that, by adopting the selected weights as (b), the visualization results demonstrate superior performance. Specifically, the brightness of the pixels near the diagonal line is significantly higher, which indicates a higher similarity. Additionally, there is a tendency to form bright clusters according to different authors.

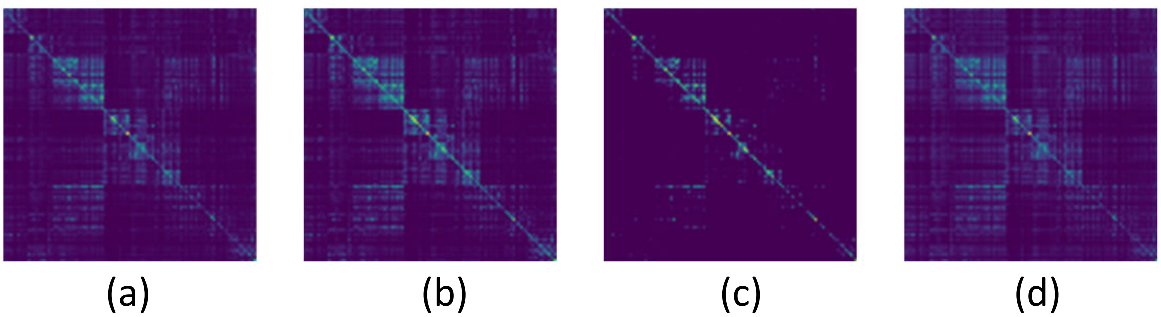


Figure 8. (a) $= (Loss_{sem} + Loss_{style})$, (b) $= (Loss_{sem}^2 + Loss_{style})$, (c) $= (Loss_{sem} * Loss_{style})^2$, (d) $= (Loss_{sem} + Loss_{style})$ The visualization settings are the same as Fig. 3 in the paper.

3.5.2 Utilization of ChatGPT

The primary application of ChatGPT was during the collection of anchor image data, where it facilitated the creation of a list containing the most renowned and potentially infringing images. This approach streamlined the process of efficiently gathering the most probable anchor images. Furthermore, we experimented with employing ChatGPT to refine prompts generated by the CLIP Interrogator. Leveraging the advanced textual processing capabilities of large language models, we aimed to optimize prompts by addressing issues such as semantic ambiguity, lack of coherence, and incomplete phrasing. The goal was to produce prompts that exhibited a closer alignment with the input image. Regrettably, the experimental outcomes demonstrated that the integration of ChatGPT-optimized prompts did not result in substantial improvements. Specifically, the enhanced prompts failed to significantly aid in identifying more closely related infringing images to the input image. Instead, this integration led to an escalation in pipeline complexity and necessitated consideration of response time and stability concerns associated with the large language model. Consequently, in the final version of our work, we restricted the invocation of ChatGPT solely to the process of collecting anchor images.

3.5.3 Data Collection Method and Data Quality

During the process of data collection, we employed a combined approach involving manual validation, ChatGPT, and CPDM metric to ensure the quality of the collected data. Specifically, we initiated the process by utilizing ChatGPT to compile a list of the most renowned and likely infringing image names. This list was curated to encompass a diverse range of image types. Subsequently, a manual review was undertaken to eliminate anchor images lacking in representativeness. For images generated by the model, we employed a dual strategy involving the CPDM metric and manual assessment. This combination served to guarantee the accuracy and representativeness of the collected images. Moreover, it is crucial to recognize that the method of data collection has a direct impact on the overall quality of the dataset. The CPDM metric played a crucial role as an initial screening tool, while the subsequent manual review ensured the ultimate quality of the dataset. The synthesis of these approaches not only ensured the data's integrity but also contributed to a dataset that better reflects the diversity and accuracy required for the research.

Table 5. Testing CPDM metric's robustness for image style transfer.

	SD-v2.1	AST	Fotor	Phosus	Makeoptim
CM (%)	97.65	91.17	88.83	98.19	92.66
ΔCLIP (%)	/	-2.59	-3.42	3.17	-9.52

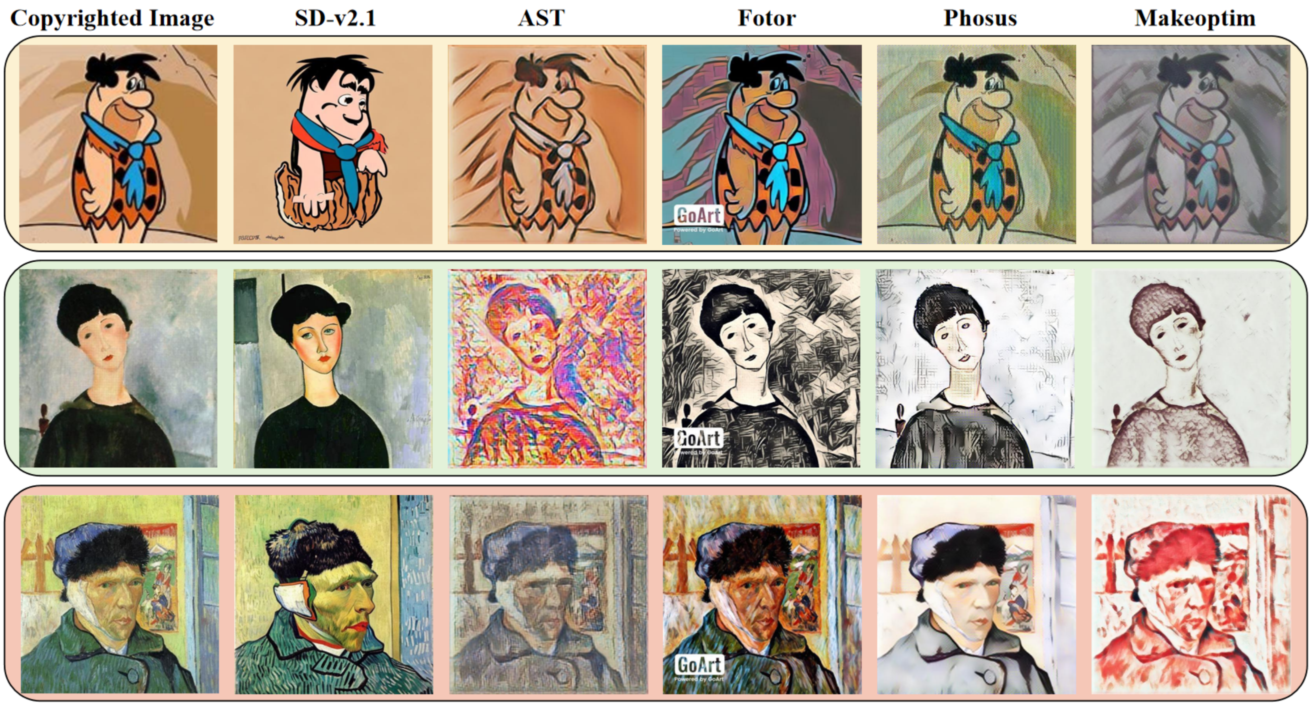


Figure 9. Testing CPDM metric's robustness for image style transfer.

3.5.4 Baseline for Unlearning Algorithm Design

We have proposed two fundamental unlearning methods for text-to-image generation models: the Gradient Ascent-based Approach and the Weight Pruning-based Approach. During the unlearning experiments, we have kept the parameters of the text embedding module and the autoencoder module frozen, focusing solely on adjusting the model's UNet structure. Taking the example of stable diffusion, the model's text embedding module employs either "openai/clip-vit-large-patch14" or "ViT-H-14," both of which are pretrained modules. Consequently, during the fine-tuning of the diffusion model, it's a common practice to freeze these parameters to maintain the semantic information of the text module and the feature information of the image encoder-decoder module. Similarly, in conducting unlearning experiments with copyrighted images, based on insights from the ESD [9] and Forgot-Me-Not [42] papers, we've found that the cross-attention structure of the UNet module greatly influences the connection between prompts and the semantic and stylistic information of generated images. Adjusting these parameters through unlearning enables the model to forget copyrighted images without excessively compromising its generative capability. Thus, we've introduced two fundamental unlearning methods specifically targeting the latent space of the UNet module.

3.5.5 Forgetting Copyrighted Images

For text-to-image generation models, when we identify that the model has generated copyrighted images based on a specific prompt and we intend to forget those corresponding copyrighted images, a two-step process is employed. Firstly, the relevant unlearning algorithm is applied to the generation model. Subsequently, the unlearned model is employed with the same prompt to regenerate the corresponding image. Following this, the CPDM metric is utilized to measure the similarity between the generated images before and after the unlearning process. This assessment helps determine whether the model has successfully forgotten the corresponding copyrighted images.

3.5.6 Experimental Setups for ESD and Forget-Me-Not

In the evaluation of both unlearning algorithms, the open-source code for ESD is tailored for Stable Diffusion v1.4. For exploratory experiments, we made appropriate code adjustments to accommodate Stable Diffusion v2.1, while maintaining other parameter configurations consistent with those specified in Tab. 2 of the paper. We utilized 100 iterations for ESD due to observed limitations in image quality and excessive decline in FID when using the official 1000 iterations. This could be attributed to the disparities in model versions between Stable Diffusion v1 and Stable Diffusion v2. Regarding the Forget-Me-Not experiment, we introduced modifications to the official parameter settings. The default image size for the official experiment is 512, employing the "stable-diffusion-2-1-base" model. We adapted this to an image size of 768 and employed the "v2-1-768-ema-pruned" model to ensure consistency with the baseline methods used in several experiments outlined in Tab. 2 of the paper. It's worth noting that, in both experiments, the code and parameter adjustments were not meticulously fine-tuned, potentially affecting the representativeness of the experimental outcomes for the optimal efficacy of the proposed approach.

3.6. Relevant Work in Artistic Image Communities

There has been considerable attention towards community efforts that employ Stable Diffusion models to imitate artistic styles. Notable examples include websites like https://stablediffusion.fr/artists and https://www.urania.ai/top-sd-artists. These platforms curate collections of images that bear stylistic resemblance to the works of over a thousand artists, spanning both contemporary and classical art. In the workflow of these image communities, artistic style images are generated by manually providing a prompt, where the artist's name is incorporated as a style cue in the prompt, such as "A woman + [artist]." This approach emphasizes capturing the unique artistic traits of an artist's style. In our framework, our approach begins by utilizing existing, specific, and valuable artistic images as anchor images. Unlike arbitrarily determining a subject or content for a painting, we derive corresponding prompts through multimodal analysis of these anchor images. These prompts are then combined with the artist's name and fed into a text-to-image model to generate images that are both content-wise and stylistically similar to the anchor image. For instance, we might create a prompt like "fritillaries-in-a-copper-vase-1887, by vincent-van-gogh, painting of a vase with fritillaria flowers in it, in the starry night, orange flowers, pine, luscious brushstrokes, prizewinning, cone, high details, hips, tyler, mesmerizing, description, pot, visually stunning, unlit." This prompt includes a description of the painting's content, its title, and the artist's name, enabling the identification of corresponding infringing images. In essence, our method involves collecting authentic, valuable, and specific images from the art world to be used as training examples for image generation models. This represents a more rigorous form of style imitation. In contrast, the artistic images in the provided links tend to focus on capturing certain aspects of an artist's style, falling into the category of broader style imitation. Furthermore, certain images from these links closely resemble those in our dataset, such as the character image generated from the prompt "A woman, [Vincent Van Gogh]."

3.7. Metrics and Human Perception Alignment

We conducted the following experiment to validate the correspondence between the CPDM metric and human perceptual evaluation. Specifically, we randomly selected 10 anchor images from each category in the CPDM dataset, resulting in a specialized dataset of 40 images for this experiment. We then divided the prompts corresponding to each image into three different lengths: short, medium, and long. These prompt lengths capture different levels of completeness in describing the anchor images. Using these prompts, we generated corresponding counterfeit images and obtained metric values. In the final step, we had human evaluators manually rate the similarity between the images generated from prompts of varying lengths and the corresponding anchor images. We then compared these human perceptual ratings with the metric values and conducted a correlation analysis. The experimental results demonstrated a significant alignment between the benchmark's metric values and human assessments of image counterfeiting trends. Experimental results refer to Tab. 6 and Fig. 12.

3.8. Balancing Among Different Users

Balancing the interests between users of art and imitators with the original creators of art is a practical challenge. For example, imitators argue that their derivative works incorporate many of their own design elements, making them significantly distinct from the original image. On the other hand, original creators believe that imitated images still contain their own creativity and artistic achievements. The CPDM dataset currently encompasses four major categories of images: art paintings, cartoons, portraits, and illustrations. Our goal is to protect a broader range of image types. Through the demo, we aim to establish a community that can accommodate various image categories and styles, facilitated by extensive voluntary searches by users and artists. This necessitates the provision of evaluation metrics to determine whether infringement has occurred. These metrics should be reliable, consistent, widely applicable to diverse image categories and styles, and generally accepted by the majority. For instance, an image that has been legally determined to be infringing due to copy-

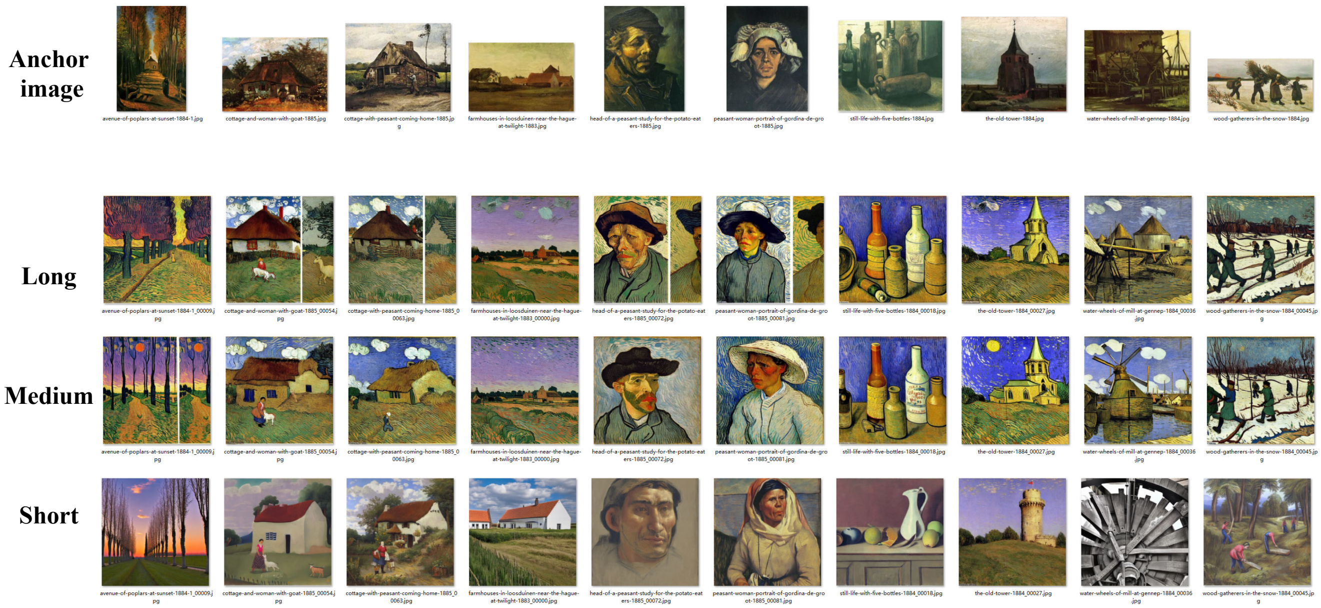


Figure 10. Style: figures generated under prompts of different lengths.

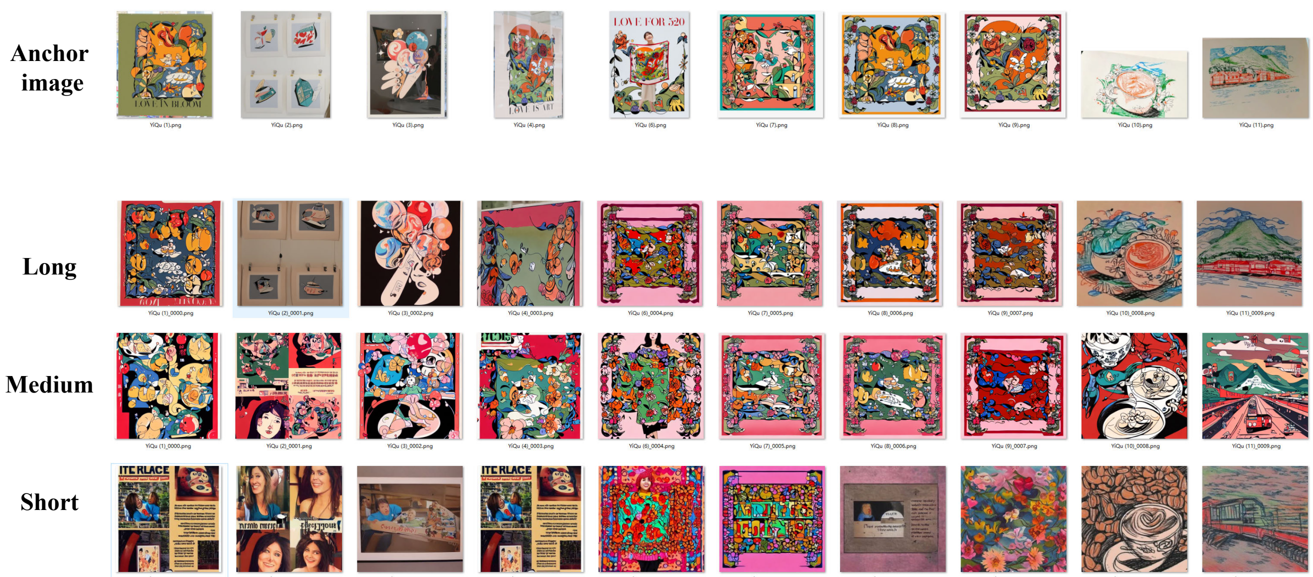


Figure 11. Illustration: figures generated under prompts of different lengths.

right disputes might not be perceived as such by its imitator, who believes their imitation includes their knowledge and creativity. However, according to prevailing universal standards (which could be legal), the imitation indeed constitutes infringement. Regarding the category of art paintings, artists often consider inspiration and creativity as the most valuable aspects of a piece, while the style, lines, and other elements are expressions of their thoughts, techniques, and habits. Imitators, through simple modifications or style transfers, might perceive the modified image to be significantly different from the original, leading them to believe it doesn't infringe copyright. However, by widely accepted standards, the modified image might still be deemed an infringement. Therefore, we aim to provide specific evaluation metrics rather than ambiguous or multiple indicators. Despite the fact that copyright disputes involve varying perspectives from creators and imitators and have some gray areas, we aspire to propose metrics that align as closely as

Table 6. Correlation Analysis of Image Metrics and Human Perception. The table lists the specific CPDM metric values(%), corresponding to Fig. 12.

Licensed Illustration

Sample	1	2	3	4	5	6	7	8	9	10
Long	99.96	99.17	99.34	97.33	99.25	99.33	99.97	98.01	99.33	99.78
Medium	99.71	87.45	88.53	92.46	77.3	52.24	99.88	97.14	71.24	99.84
Short	95.55	96.99	89.75	57.71	96.58	1.41	31.39	92.7	97.19	94.72

Portrait

Sample	1	2	3	4	5	6	7	8	9	10
Long	68.44	98.77	98.77	97.46	97.43	98.59	93.68	97.06	68.2	97.73
Medium	83.68	98.19	99.51	96.09	96.16	88.9	90.03	93.67	60.81	84.05
Short	81.48	98.12	98.02	96.79	96.97	98.61	91.39	97.13	82.93	95.81

Style

Sample	1	2	3	4	5	6	7	8	9	10
Long	95.81	98.09	92.36	97.35	89.47	97.16	96.26	96.79	96.94	97.08
Medium	89.48	96.49	94.11	96.61	63.39	97.21	97.85	95.01	96.79	98.27
Short	76.1	95.47	99.0	93.02	87.59	99.17	88.38	88.58	98.31	82.78

Artistic Creation Figure

Sample	1	2	3	4	5	6	7	8	9	10
Long	92.24	99.36	0	99.41	82.15	97.63	90.67	99.01	98.71	92.33
Medium	98.39	94.55	85.91	98.0	82.43	39.13	86.84	98.45	93.05	92.57
Short	0	74.59	0	0	79.08	87.77	93.52	91.33	90.17	88.88

possible with universally accepted evaluation criteria. We maintain close discussions with the artists within our team, continuously improving and refining our evaluation methods. Our aim is for the outcomes of our evaluation criteria to closely align with the assessments made by artists.

3.9. Usage of Copyrighted Images

The current dataset includes copyrighted content obtained from Wikipedia. The images provided by this community have shared copyrights and can be accessed and utilized for non-commercial purposes and educational research (https://www.wikiart.org/en/terms-of-use). Additionally, a portion of illustrative images comes from the Anonymous Artist . These illustrative images are also available for non-commercial and educational research purposes. For commercial usage, please reach out to Anonymous Artist at https://###/ to request authorization. After the demo link is launched, the dataset will gradually expand to encompass various other categories of images. Through the engagement and usage of a wide range of users, we will curate and collect images that may have been subject to copyright infringement. The collection of such images is carried out with the authorization of the image providers, adhering to non-commercial use and scientific research purposes.

3.10. Prospects for Dataset Scale and Diversity

Currently, our research focus is restricted to the domain of two-dimensional image copyright to analyze and explore the infringement scenarios where copyright images are used in training datasets for text-to-image generation models. This is a highly relevant issue that has attracted significant attention. The data collection and preparation required for the development of large models have a profound impact on image copyright issues. For instance, training sets for image generation models like Stable Diffusion include massive amounts of natural images, some of which unavoidably include copyrighted content. As a result, issues related to copyright infringement in the context of two-dimensional content and text-to-image generation models have gained widespread attention recently. We acknowledge that real-world copyright infringement encompasses a broader spectrum, including three-dimensional content and multimedia presentations. However, attempting to cover all potential infringement domains is a complex and challenging endeavor. We have chosen to focus on the copyright issues of two-dimensional image content and text-to-image generation models, with the aim of laying the groundwork for future expansion into other dimensions of image infringement. We hope that future work can expand the scope to encompass three-dimensional and multimedia content, thus conducting a comprehensive analysis of copyright infringement issues across various dimensions. Meanwhile, the demo for our proposed method will be launched soon and continuously

	Sample	CPDM Metric				Human Perception		
Prompt Lentth		Long	Midium	Short		Long	Midium	Short
Licensed Illustration	1	1	2	3		1	2	3
	2	1	3	2		1	2	3
	3	1	3	2		1	2	3
	4	1	2	3		1	2	3
	5	1	3	2		1	2	3
	6	1	2	3		1	2	3
	7	1	2	3		1	2	3
	8	1	2	3		1	2	3
	9	1	3	2		1	2	3
	10	1	2	3		1	3	2
Portrait	11	2	3	1		2	1	3
	12	1	2	3		1	2	3
	13	2	2	4		3	1	2
	14	1	3	2		1	3	2
	15	1	3	2		1	3	2
	16	1	3	2		1	2	3
	17	1	3	2		1	2	3
	18	1	3	2		1	2	3
	19	3	1	2		1	2	3
	20	1	3	2		1	2	3
Style	21	1	2	3		1	2	2
	22	1	2	3		1	2	3
	23	3	2	1		1	2	3
	24	1	2	3		1	2	3
	25	1	3	2		1	2	3
	26	3	1	2		1	2	3
	27	2	1	3		1	2	3
	28	1	2	3		1	2	3
	29	3	1	2		2	1	3
	30	2	1	3		1	2	3
Artistic Creation Figure	31	2	1	3		1	2	3
	32	1	2	3		1	2	3
	33	2	3	1		1	2	3
	34	1	2	3		1	2	3
	35	1	2	3		1	2	3
	36	1	3	2		3	1	2
	37	3	2	1		1	2	3
	38	1	2	3		1	2	3
	39	1	2	3		1	2	3
	40	1	2	3		1	2	3

Figure 12. Correlation Analysis of Image Metrics and Human Perception. As the color deepens, the correlation increases. This observation highlights the alignment between the trend of the CPDM metric and human perception

improved. With the demo link, we hope that any artist or individual can upload images they believe to be infringing. If users are willing to share their images with the dataset, we will add infringing images to the CPDM dataset through a dual process of evaluation metrics and human screening. We believe that the dataset's scale will become richer as more people participate, and we will rigorously control the dataset's quality and make timely revisions. Once again, we sincerely appreciate your constructive feedback.

Table 7. The images generated using DALLE-mini are evaluated with the CPDM metric. The rows and columns of the table represent anchor images and generated images, respectively. A higher CPDM metric indicates a higher correlation between images. It can be observed that the images in the diagonal positions exhibit the highest correlation.

CPDM Metric (%)	Style	Portrait	Artistic Creation Figure
Style_g	**97.97**	54.29	79.68
Portrait_g	91.62	**98.15**	88.05
Artistic_Creation_Figure_g	75.14	55.36	**98.13**

Table 8. The prompt for the unlearning experiment of Vincent Van Gogh's art painting in Fig. 4.

Prompts **¡Van Gogh¿**	*"still-life-with-vegetables-and-fruit-1885, by vincent-van-gogh, painting of a ... vegetables, 19 century, 3 heads, early 1 9 century, salad, right side composition, chalk, corner"*
	"head-of-a-peasant-study-for-the-potato-eaters-1885, ... open, tilted head, vert coherent, february), farmer, yellowed with age, grain"
	"cottage-with-peasant-coming-home-1885, by vincent-van-gogh, painting of a man standing in front of a thatched ... milkman, a brick cabin in the woods, poor buildings, pastelle, wandering"
	"cottage-and-woman-with-goat-1885, by vincent-van-gogh, a painting of a woman and a child outside a thatched ... detail, 1852, tonalist style, details"
	"avenue-of-poplars-at-sunset-1884-1, by vincent-van-gogh, a painting of a person walking down a path in the woods, by Van Gogh, church, at sunset in autumn, driveway, green ... siècle, pot, amsterdam, definition"

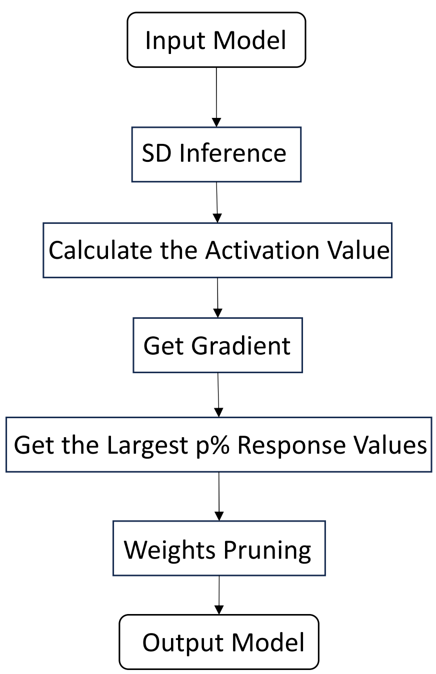


Figure 13. Activation-based pruning method.

3.11. Limitations and Negative Societal Impacts

- The scale of the dataset is relatively limited, which may fail to adequately cover a wide range of copyright images and associated cues. This could result in suboptimal performance of the forgetting algorithm when dealing with copyright images not included in the dataset. To enhance the algorithm's accuracy and generalizability, it may be necessary to expand the dataset and incorporate more diverse images and cues.
- The quality of the dataset is paramount to the effectiveness of the forgetting algorithm. If the dataset contains erroneous or inaccurate matches, it can hinder the algorithm's ability to correctly identify and handle copyright images. Therefore, ensuring the quality and accuracy of the dataset is of utmost importance.
- Dataset bias can impact the fairness and accuracy of the forgetting algorithm if there are evident biases in the copyright images and cues within the dataset. To mitigate this issue, it is crucial to ensure that the dataset is broadly representative and diverse.
- Legal and ethical considerations arise when employing copyright images as part of the dataset. When collecting and utilizing such images, it is essential to comply with relevant regulations and respect the rights of the original creators.
- Due to CLIP's inherent limitations including dataset bi-

Figure 14. The experimental results of DALLE-mini are shown in the figure, with 9 generated images on the left and anchor images on the right. The corresponding CPDM metric data is presented in Table 7.

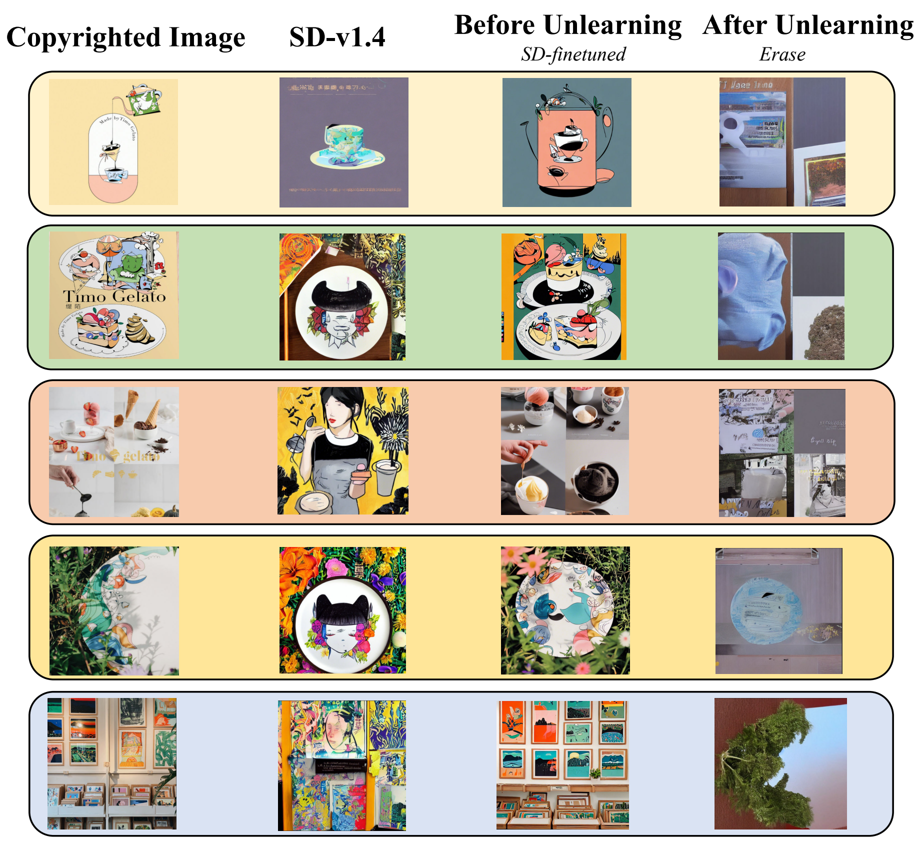


Figure 15. Benchmark experiments with unlearning method Erase[9].

ases, incomplete semantic and texture comprehension, and weak expression of certain image features, it fails to capture all image characteristics and fine details during the image-to-text process. The accuracy of prompts constrains the generation model's ability to produce copyrighted images.

- The use of generic CLIP and Inception as the primary feature extractors for image similarity metrics is subject to limitations stemming from the restricted quantity and variety of pretraining data available to these extractors. Consequently, inherent dataset biases and feature extraction tendencies are present, leading to an inability to capture certain unique or previously unseen image features. This situation may potentially restrict the applicability of evaluation metrics on specific style images.

Table 9. The prompts for illustration image in Fig. 5.

Prompts ¡Anonymous Artist¿	*”Illustrator ¡Anonymous Artist¿ , illustrations,there is a cartoon of a bird that is ... cover, uncropped, cut-away, watering can, minimal composition, by Masolino, loosely cropped, label”*
	”Illustrator ¡Anonymous Artist¿ , illustrations,there is a cartoon picture of a woman with a ... colored accurately, year 2023, a blond, very very happy!, illustratioin,”
	”Illustrator ¡Anonymous Artist¿ , illustrations,someone is holding a drawing of a flower in a ... al fresco, hard morning light, panoramic shot, fuchsia and blue, stipple, 2021, not blurry”
	””Illustrator ¡Anonymous Artist¿ , illustrations,there are a lot of different items that are on ... atmosphere, box, bautiful, star born, michelin restaurant, vivid)”
	”Illustrator ¡Anonymous Artist¿ , illustrations,a close up of a magazine cover with a woman in a dress, amazing ... Bowler, bt21, idyllic, eating, ffffound, inspired by Olive Mudie-Cooke, tummy”
	”Illustrator ¡Anonymous Artist¿ , illustrations,there is a drawing of a cup of coffee with a bird ... promo image, 1956, delightful surroundings, doodles”
	”Illustrator ¡Anonymous Artist¿ , illustrations,a red and gold christmas card with a horse and other holiday ... poster, an engraving, 2019, chocolate, epicurious, cd, album, sk, hello, english, tablecloth, b”

Table 10. The exemplification of illustration prompts. The prompts listed in ascending order within the table align harmoniously with the images sequentially depicted in Fig. 6.

Prompts ¡Anonymous Artist¿	*”Illustrator ¡Anonymous Artist¿ , poster of fruits, created by James Jean and Victo Ngai. Represents love as the beginning of all. ,..., Tom Whalen’s blooming style in 8K resolution.”*
	”Illustrator ¡Anonymous Artist¿ , presents 4 Jetsons-inspired illustrations of a woman, printed on paper and arranged on a wall. ,..., The scenes include icebergs, a carousel, and ships, screen printed as part of an 8-piece portfolio.”
	”Illustrator ¡Anonymous Artist¿ , presents an illustration featuring a person holding balloons, a collaborative work by Tristan Eaton and Greg Rutkowski. ,..., It portrays a slice of life, capturing the essence of caramel.”
	” Illustrator ¡Anonymous Artist¿ presents a poster featuring a woman holding a perfume bottle. ,..., Enclosed in a glass cover, the artwork captures a flower explosion in a snapshot with detailed shots. ”

3.12. Resource Consumption

3.12.1 Human Resources

A team of 13 members, including both computer scientists and artists.

3.12.2 Computational Resources

(4 * 24 * 30) hours (Nvidia A100).

4. CPDM Demo

As shown in Fig. 16, Fig. 17.

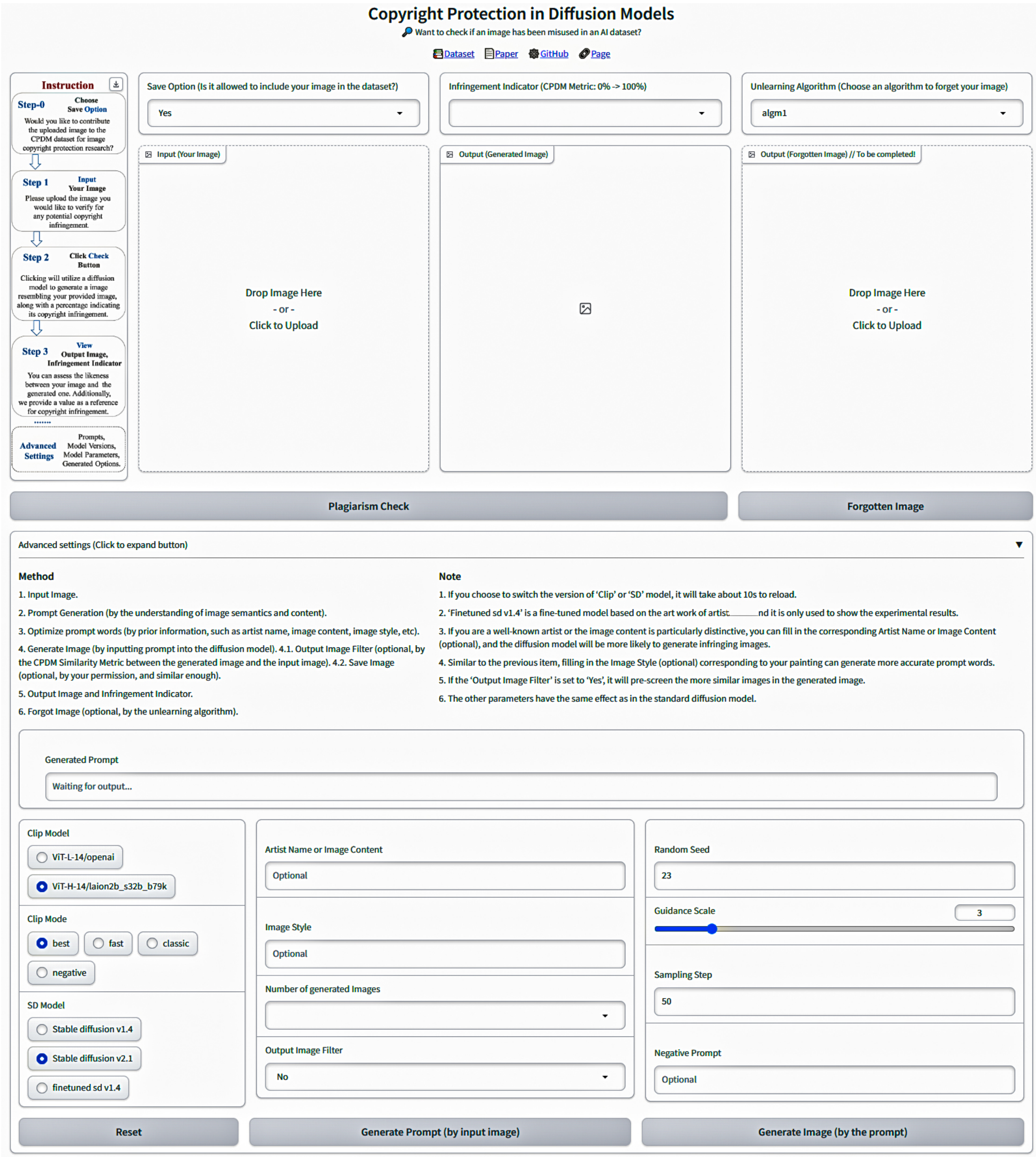
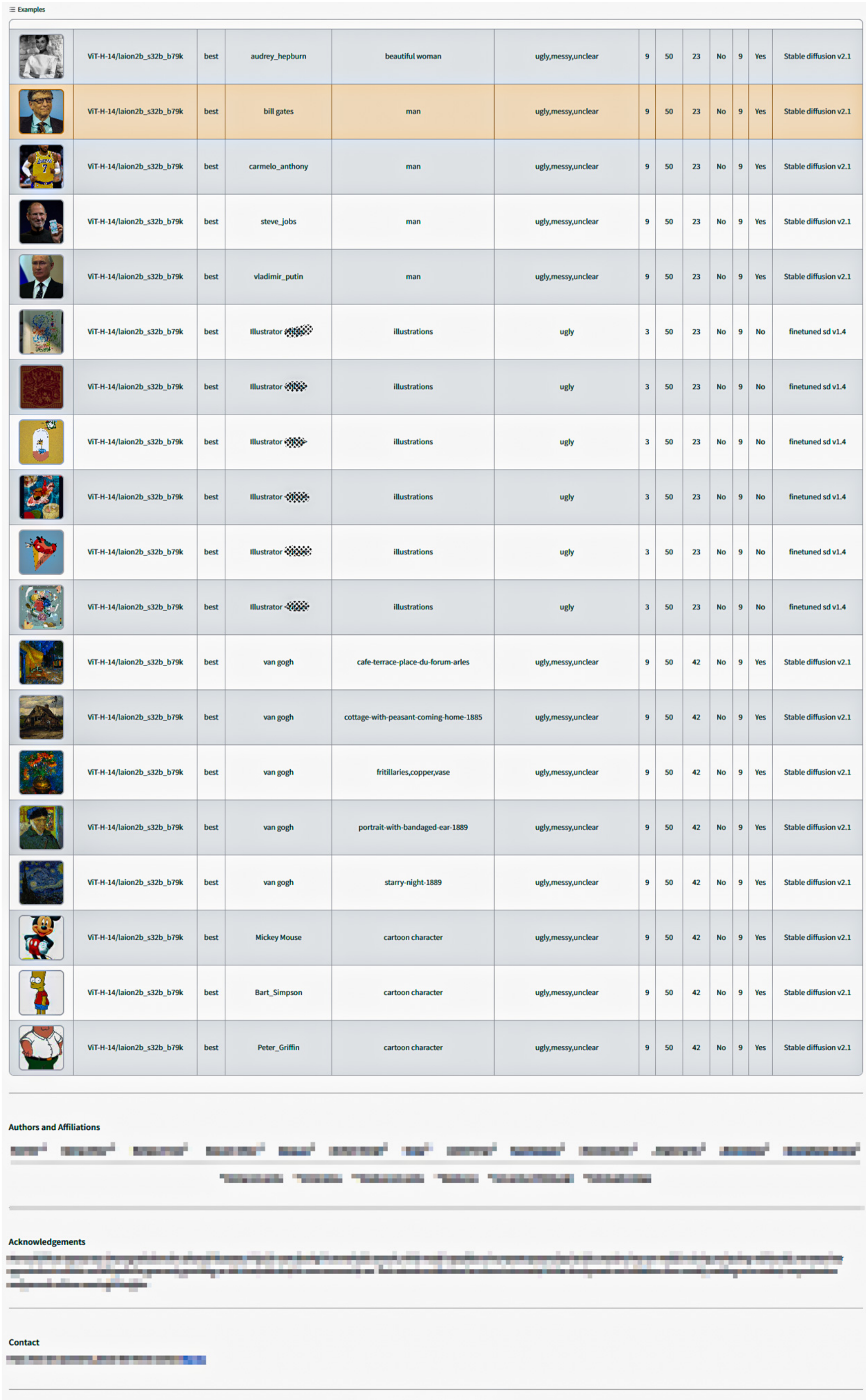


Figure 16. An example of our demo interface. The various functionalities will be continuously improved and updated.



Examples

	ViT-H-14/laion2b_s32b_b79k	best	audrey_hepburn	beautiful woman	ugly,messy,unclear	9	50	23	No	9	Yes	Stable diffusion v2.1
	ViT-H-14/laion2b_s32b_b79k	best	bill gates	man	ugly,messy,unclear	9	50	23	No	9	Yes	Stable diffusion v2.1
	ViT-H-14/laion2b_s32b_b79k	best	carmelo_anthony	man	ugly,messy,unclear	9	50	23	No	9	Yes	Stable diffusion v2.1
	ViT-H-14/laion2b_s32b_b79k	best	steve_jobs	man	ugly,messy,unclear	9	50	23	No	9	Yes	Stable diffusion v2.1
	ViT-H-14/laion2b_s32b_b79k	best	vladimir_putin	man	ugly,messy,unclear	9	50	23	No	9	Yes	Stable diffusion v2.1
	ViT-H-14/laion2b_s32b_b79k	best	Illustrator [redacted]	illustrations	ugly	3	50	23	No	9	No	finetuned sd v1.4
	ViT-H-14/laion2b_s32b_b79k	best	Illustrator [redacted]	illustrations	ugly	3	50	23	No	9	No	finetuned sd v1.4
	ViT-H-14/laion2b_s32b_b79k	best	Illustrator [redacted]	illustrations	ugly	3	50	23	No	9	No	finetuned sd v1.4
	ViT-H-14/laion2b_s32b_b79k	best	Illustrator [redacted]	illustrations	ugly	3	50	23	No	9	No	finetuned sd v1.4
	ViT-H-14/laion2b_s32b_b79k	best	Illustrator [redacted]	illustrations	ugly	3	50	23	No	9	No	finetuned sd v1.4
	ViT-H-14/laion2b_s32b_b79k	best	Illustrator [redacted]	illustrations	ugly	3	50	23	No	9	No	finetuned sd v1.4
	ViT-H-14/laion2b_s32b_b79k	best	van gogh	cafe-terrace-place-du-forum-arles	ugly,messy,unclear	9	50	42	No	9	Yes	Stable diffusion v2.1
	ViT-H-14/laion2b_s32b_b79k	best	van gogh	cottage-with-peasant-coming-home-1885	ugly,messy,unclear	9	50	42	No	9	Yes	Stable diffusion v2.1
	ViT-H-14/laion2b_s32b_b79k	best	van gogh	fritillaries,copper,vase	ugly,messy,unclear	9	50	42	No	9	Yes	Stable diffusion v2.1
	ViT-H-14/laion2b_s32b_b79k	best	van gogh	portrait-with-bandaged-ear-1889	ugly,messy,unclear	9	50	42	No	9	Yes	Stable diffusion v2.1
	ViT-H-14/laion2b_s32b_b79k	best	van gogh	starry-night-1889	ugly,messy,unclear	9	50	42	No	9	Yes	Stable diffusion v2.1
	ViT-H-14/laion2b_s32b_b79k	best	Mickey Mouse	cartoon character	ugly,messy,unclear	9	50	42	No	9	Yes	Stable diffusion v2.1
	ViT-H-14/laion2b_s32b_b79k	best	Bart_Simpson	cartoon character	ugly,messy,unclear	9	50	42	No	9	Yes	Stable diffusion v2.1
	ViT-H-14/laion2b_s32b_b79k	best	Peter_Griffin	cartoon character	ugly,messy,unclear	9	50	42	No	9	Yes	Stable diffusion v2.1

Authors and Affiliations

[redacted]

Acknowledgements

[redacted]

Contact

[redacted]

Use via API · Built with Gradio

Figure 17. An example of our demo interface. The various functionalities will be continuously improved and updated.

5. Datasheets for datasets

Motivation

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

The dataset was created with the purpose of advancing research and development in the field of text-to-image generative models. These models aim to generate realistic images based on textual descriptions, effectively bridging the gap between language and visual content. However, the rapid advancements in text-to-image generation techniques have also raised concerns regarding copyright protection, such as the unauthorized learning of content, artistic creations, and portrait. We aim to develop a dataset and metrics that facilitate the identification of copyright infringement, while enabling a fair comparison of methods for mitigating such infringements.

Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

This dataset was collaboratively created by researchers from Peking University, Tsinghua University, and University of California, Berkeley (UCB), as well as researchers from the industry, specifically ByteDance company.

Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.

No.

Any other comments?

Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

The dataset primarily comprises anchor images, generated images and their corresponding prompts. The anchor images are initially collected as a set of images that potentially contain copyrighted content. These anchor images are then processed using the CLIP-interrogator, which yields prompts associated with each anchor image. Subsequently, the obtained prompts are utilized as inputs for the stable diffusion model, enabling the generation of images by the stable diffusion model.

How many instances are there in total (of each type, if appropriate)?

The dataset consists of 2100 anchor images, each accompanied by a corresponding prompt, resulting in a total of 2100 prompts. Using these prompts as input, a total of 18900 images were generated.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

The dataset encompasses the entirety of all possible instances.

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.

Each instance within the dataset comprises an anchor image, a prompt, and nine corresponding generated images.

Is there a label or target associated with each instance? If so, please provide a description.

Each instance represents an original image, along with its corresponding prompt and nine images generated by the stable diffusion model that potentially exhibit copyright infringement.

Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

No.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.

There is no explicit correlation between individual instances.

Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

No.

Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

No, there are no errors, sources of noise, or redundancies in the dataset.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.
The dataset is self-contained.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals non-public communications)? If so, please provide a description.
No, the dataset does not contain data that might be considered confidential, such as information protected by legal privilege, doctor-patient confidentiality, or the content of individuals' non-public communications.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.
No.

Does the dataset relate to people? If not, you may skip the remaining questions in this section.
Yes

Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.
The dataset primarily comprises four categories: Style, Portrait, Artistic Creation Figure, and Licensed Illustration.

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how.
Yes, portraits are also part of copyright, and therefore, we have included a subset of celebrity portraits in the dataset.

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.
No.

Any other comments?

Collection Process

How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.
The data was directly observable.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated?
We propose a pipeline to coordinate CLIP, ChatGPT, and diffusion models to generate a dataset that contains anchor images, corresponding prompts, and images generated by text-to-image models, reflecting the potential abuses of copyright. Initially, we collect a set of images that potentially contain copyrighted content, which serves as anchor images. Subsequently, these images are fed into the CLIP-interrogator, allowing us to obtain prompts that correspond to each anchor image. Finally, the prompts are used as input for the stable diffusion model, resulting in the generation of images by the stable diffusion model. Through manual comparisons, we assess whether there is evidence of copyright infringement in terms of style and semantics between the anchor images and the generated images. Ultimately, the anchor images, their corresponding prompts, and the images generated by the stable diffusion model constitute the core components of our dataset.

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?
No.

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

The dataset was primarily curated with contributions from the first three authors listed in the author list.

Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.
The data was collected within the past three months.

Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.
No.

Does the dataset relate to people? If not, you may skip the remaining questions in this section.
Yes.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?
We obtained portrait information of public figures from Wikipedia.

Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.
Our images are sourced from Wikipedia, where the images are available for non-commercial or educational use.

Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.
Our images are sourced from Wikipedia, where the images are available for non-commercial or educational use.

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).
We will provide our contact information on the release page of the website. In the event of any potential copyright infringement, we will promptly assess the situation, and if found to be valid, we will take immediate action to remove the corresponding data.

Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.
We assert that our data is unlikely to cause potential negative impacts.

Any other comments?

Preprocessing/cleaning/labeling

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this section.
No.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.

Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.

Any other comments?

Uses

Has the dataset been used for any tasks already? If so, please provide a description.
This dataset is utilized for evaluating the efficacy of unlearning methods applied to stable diffusion.

Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.
No.

What (other) tasks could the dataset be used for?
This dataset can also be utilized to assist in determining whether copyright infringement has occurred.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future

uses? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?
No.

Are there tasks for which the dataset should not be used? If so, please provide a description.
No.

Any other comments?

Distribution

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.
No.

How will the dataset will be distributed (e.g., tarball on website, API, GitHub) Does the dataset have a digital object identifier (DOI)?
We will release this dataset in github.

When will the dataset be distributed?
We plan to release our dataset upon the paper entering the review stage.

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.
The dataset is available for non-commercial or educational use.

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.
The dataset is available for non-commercial or educational use.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.
The dataset is available for non-commercial or educational use.

Any other comments?

Maintenance

Who will be supporting/hosting/maintaining the dataset?
The researcher in this project.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?
We have provided the contact information of the dataset creators on our GitHub website.

Is there an erratum? If so, please provide a link or other access point.
No.

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?
We will release and update our dataset on GitHub, with a monthly update frequency.

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.
If our images infringe upon individuals' portrait rights, we will promptly remove the corresponding data after verification.

Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to users.
With each update, we incorporate the changes based on the original dataset, ensuring that previous versions of the data are preserved.

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.

On the dataset's release page, we have provided corresponding links with the aim of encouraging collaborative expansion of the dataset and fostering the protection of copyright information.