Design and Implementation of Decentralized Data Protection Protocol for Generative AI

Bingxue Zhang[1][0000−0002−9227−2649], Yang Shi[1][0009−0000−1360−6827] ✉, Feida Zhu[2][0000−0001−6077−4356], and Wang-Chien Lee[3][0000−0002−8949−489X]

[1] University of Shanghai for Science and Technology, Shanghai, China
[2] Singapore Management University, Singapore
[3] Pennsylvania State University, Pennsylvania, USA
shiyang1121@st.usst.edu.cn

Abstract. In recent years, generative AI has made significant advancements. Specifically, in the image domain, one major breakthrough is the adoption of diffusion model-based customized image generation technology, allowing users to generate personalized content based on their own copyrighted images. However, along with this development, the misuse of copyrighted data has become a growing concern. To tackle this issue, this paper first focuses on protecting data from being maliciously trained or inferred by generative AI models in image domain. Given the limitations of current adversarial example methods, we propose the EMCF (Enhanced Mist, Color space and Frequency) data protection method with stronger robustness and better visual quality. Furthermore, from the standpoint of AI governance, merely preventing data misuse without considering its circulation and creative reuse demands will fail to unleash its full potential. To bridge this gap, beyond EMCF data protection method, this paper further incorporates blockchain technology to design a decentralized data protection protocol, ensuring the scalable data circulation within authorized boundaries. Experimental results verify its effectiveness in preventing AI-generated image models from using the protected data for training, providing new idea for AI governance in image domain.

Keywords: Generative AI · diffusion model · adversarial example · blockchain · AI governance

1 Introduction

Generative AI has made rapid strides in recent years, revolutionizing the industries with its innovative capabilities. Specifically in image field, one of the key breakthroughs is the advent of Diffusion models [1]. These models enable users to create high-quality images directly from text descriptions, leading to widespread applications such as advertising poster creation, artwork design, etc. Beyond this, customized image generation techniques like fine-tuning [2] or image-to-image synthesis [3] enable these models to create tailored contents just based on small amounts of user's copyrighted images. However, since image data directly

reflects personal identity traits or an artist's unique painting style, any abuse of these copyrighted image data can pose serious risks, e.g., digital sexual crime [4], artwork misuse [5], etc. This also brings new challenges to AI governance.

To address copyrighted image data abuse for customized generation, the adversarial example technology [6] is a well-received data protection method. It embeds subtle perturbations to the protected images, so that generative AI models are unable to extract useful features from them, thus making these models difficult to generate qualified content. However, existing adversarial example methods still have shortcomings in terms of robustness and visual quality. In addition, from the perspective of AI governance, data protection should not hinder the demands for data circulation and creative reuse. For this reason, current approaches have attempted to achieve data circulation within authorized boundary. Nevertheless, they are primarily centralized methods and show weakness in issues such as the difficulty of tracing data ownership and the interoperability across different applications, making the scalability unattainable. In contrast, blockchain technology, with its decentralization, immutability and interoperability features, offers a new solution. However, applying blockchain also brings potential risks such as sensitive data leakage, malicious behavior of computational nodes and so on. Therefore, how to achieve data security while ensuring efficient data circulation scalably, thereby fostering more effective AI governance, remains a pressing challenge.

Facing above issues, our main innovations and contributions are as follows:

1) Proposing EMCF data protection method for generative AI in image field. Based on recent advancements in adversarial examples, this paper proposes a dynamic target image generation strategy to improve the robustness of perturbations. Additionally, it integrates frequency domain and color space theory to enhance the perturbation imperceptibility, while ensuring that image data cannot be abused by Diffusion models.

2) Proposing a decentralized data protection protocol for copyrighted data circulation and creative reuse demands in the image domain of generative AI. By incorporating blockchain technology and tailoring the protocol design for AI-generated image context, it enhances data ownership traceability and ensures cross-application interoperability, offering a fresh perspective on decentralized data protection.

2 State of Art

This section provides an overview of the frontier research for customized image generation, adversarial example methods, and the data circulation strategies within authorized boundary.

2.1 Research on Customized Image Generation

In image domain of generative AI, emerging Diffusion models [1] have already demonstrated their unparalleled capabilities in high-quality image generation

and editing tasks. Recently, with the development of customized image generation techniques, users can create personalized images using their private images through open-source Diffusion models. These customized generation techniques can be divided into two stages: inference and training stage.

Inference-stage image customization technique operate without the need for fine-tuning on the source image, mainly involving Textual Inversion [9] and Image-to-Image synthesis [3]. Of these, the Textual Inversion [9] generates personalized images by optimizing word embeddings without adjusting the model weights, resulting in lower image quality. Image-to-Image synthesis [3] directly inputs both the original image and a text prompt into a pre-trained model for inference to generate customized image, ideal for style transfer and image inpainting.

Training-stage image customization techniques needs fine-tune on the source images to extract the desired information and generate new images, primarily includes DreamBooth [11] and LoRA [12]. In particular, DreamBooth [11] finetunes all Diffusion model parameters for high-quality images but requires high computational cost. LoRA [12], on the other hand, freezes the pre-trained weights and adds trainable layers in the attention layer, balancing image quality and computational cost. These two methods are most commonly used methods in image customization field [13].

2.2 Research on Adversarial Example

Although the above customized image generation techniques meet the user's personalized needs, they also raise concerns about potential misuse of copyrighted data. To protect copyrighted image data, adversarial examples [6] are well-received by injecting subtle perturbations into original images, making it difficult for generative AI models to extract useful features and hence hardly generate qualified images. The adversarial examples for inference and training stages are as follows.

Inference-stage adversarial example methods mainly target preventing data misuse by Textual Inversion [9] and Image-to-Image synthesis [3]. Liang [6] designed adversarial examples for Textual Inversion, making it difficult for Diffusion models to derive suitable word embeddings. In addition, for defensing the image abuse by image-to-image techniques, Salman [10] design the encoder and diffusion attack mechanism to train the perturbation, and let the model generate distorted images.

Training-stage adversarial example methods focus on safeguarding data from misuse by DreamBooth [11] and LoRA [12]. In detail, Anti-DreamBooth [14] designs T-FSMG and T-ASPL methods to train perturbations that can disrupt DreamBooth's diffusion process, but only works on Dreambooth. In contrast, DUAW [13] trains a universal perturbation for both Dreambooth and Lora models. However, the perturbation trained by DUAW lacks robustness, as subtracting the adversarial example from the original image at the pixel level can easily recover the universal perturbation.

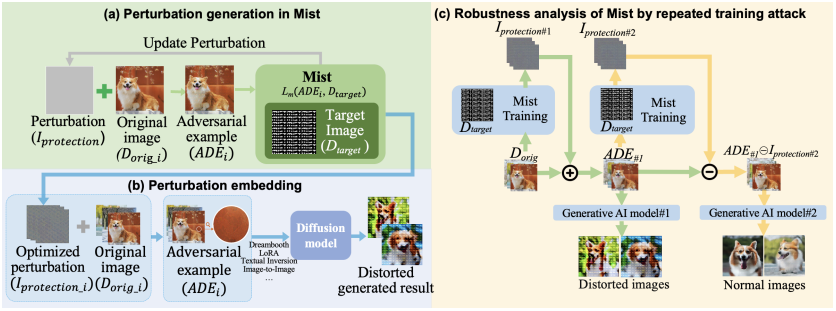

Fig. 1. Mist Framework.

Unlike above phase-specific protection methods, Mist [7] protects both training and inference stages, supporting various Diffusion models' customized technologies such as Textual Inversion [9], Image-to-Image synthesis[3], DreamBooth [11] and LoRA [12], make it the most promising adversarial example method currently. The Mist framework is illustrated in Fig. 1(a).

The idea behind Mist is to shield images by disrupting Diffusion models' training or inference process. For an original image (D_{orig_i}), an initial perturbation ($I_{protection_i}$) is embedded to D_{orig_i} through pixel-bit addition to obtain initial adversarial example (ADE_i). Then, target image (D_{target}) plays a crucial role during Mist adversarial example training process. The Mist aims to make ADE_i resemble to D_{orig_i}, while also misleading Diffusion models to generate the results similar to D_{target}. In Fig.1(b), the Diffusion model trained by ADE_i will generate distorted results. Meanwhile, we can observe that distorted results generated by Diffusion models have similar features to D_{target}.

Despite Mist yielding effective protection effects, Mist's robustness remains a con-cern. As shown in Fig.1(c), Mist fails to resist "repeated training attack". Specifically, through Mist training on original image (D_{orig_i}), $I_{protection_{\#1}}$ is produced. Then, we obtain $ADE_{\#1}$ by embedding $I_{protection_{\#1}}$ into D_{orig_i}. Next, if the malicious user applies same Mist training process to $ADE_{\#1}$ to obtain a new $I_{protection_{\#2}}$, and then perform pixel subtraction between $I_{protection_{\#2}}$ and $ADE_{\#1}$ to get a new image, this new image no longer has the capacity to interfere with Diffusion models. In addition, for imperceptibility, enhancing similarity between adversarial example and original image remains an area for improvement for Mist [7].

2.3 Research on data circulation strategies within authorized boundary

The core of generative AI model's data protection strategy is not only to prevent data misuse, but also facilitate the data circulation within authorized boundary. Platforms like Dawex [15] and Xignite [16] support this by facilitating data authorization and sharing between data owners and data demanders. However,

most existing authorization models often rely on centralized management, which has problems such as single point failure, lack of transparency, and difficulty in tracing data ownership.

With development of blockchain technology, decentralized platforms for data authorization and circulation have emerged. For instance, Michael [17] proposed a blockchain-based Internet of Things (IoT) data authorization scheme using blockchain and smart contracts for decentralized and traceable data use, enabling more IoT data to be effectively used. Similarly, Wang's ArtChain platform [18] also employs blockchain technology to ensure the legitimacy and traceability of digital artwork ownership, while guaranteeing the transparency and fairness of authorization. In addition, the above blockchain-based decentralized platforms offer strong interoperability, making them suitable for cross-application compatibility and demonstrating significant scalability potential.

Despite significant progress in data protection for circulation demands, a decentralized strategy tailored to generative AI in image domain remains lacking, requiring further research.

3 Methodology

Based on the state of art above, we propose the EMCF data protection method in this section. Furthermore, to meet the demands of data circulation and creative reuse, we propose the decentralized data protection protocol.

In response to the challenges analyzed in Section 2, we focus on addressing the following two research questions in this section.

- RQ 1. How to improve the robustness of adversarial example while also improving its imperceptibility?
- RQ 2. How to develop a decentralized data protection protocol tailored to circulation demands, enabling data authorization under specified conditions?

3.1 EMCF Data Protection Method

To address the RQ1, based on Mist [7] (shown in Fig. 1), we propose a new EMCF data protection method. As illustrated in Fig. 2(a), first, through an in-depth analysis of Mist method, we identify the key factors affecting robustness of adversarial example, and propose Enhanced Mist (EM) module by designing the random target image generation mechanism. Beyond this, to enhance the visual quality of adversarial examples while remaining their interference capacity to generative AI models, we further propose the EMCF data protection method by designing Color space (C) and Frequency domain (F) constraint optimization in perturbation training stage. Then, as shown in Fig. 2(b), during perturbation embedding phase, the optimized perturbation trained by EMCF is embedded into D_{orig_i} to produce adversarial example ADE_i. The ADE_i can effectively disrupt Diffusion models' training and inference process. The details are as follows.

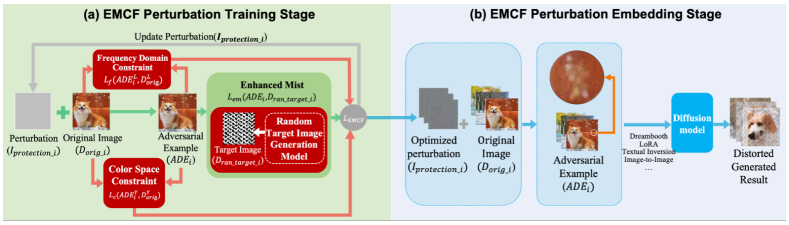

Fig. 2. EMCF data protection method

Enhanced-Mist with random target image generation mechanism As shown in Fig. 1(c), Mist is vulnerable to the "repeated training attack", i.e., after pixel-bit subtraction of $I_{protection_{\#2}}$ and $ADE_{\#1}$, its interference ability to Diffusion models is greatly diminished. We further analyzed the Mist-generated perturbations using the FID [23] metric, and calculated the FID of just 39.862 between 50 pairs of $I_{protection_{\#1}}$ and $I_{protection_{\#2}}$, which indicates that their distributions are closely aligned.

We analyzed that this vulnerability arises because Mist uses same target image in all its trainings. To address the robustness issue of Mist, we proposed the Enhanced-Mist, featuring a random target image generation module within Mist, this module ensures that a unique target image can be generated every time. As shown in Fig. 3(a), we detail its design and implementation.

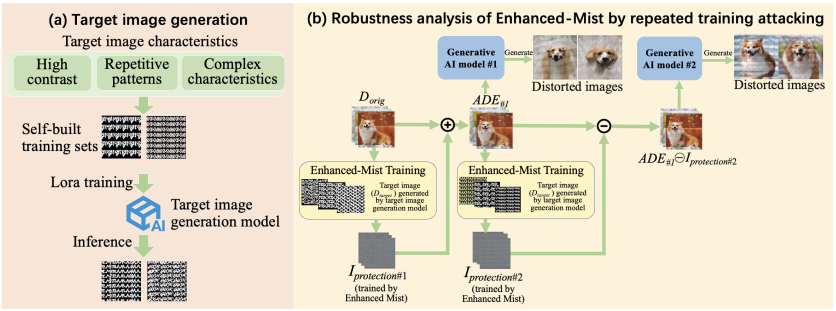

Fig. 3. Random Target Image Generation Module

Note that, to train the target image (D_{target}) generation module, it is imperative to create a training set consists of high-quality target images. Research by Liang [7] indicates that target images with high contrast and repetitive patterns can yield superior interference on Diffusion models. Thus, we utilize PhotoShop to create 20 target images characterized by high contrast, certain repetitive patterns, and complex features to serve as a training set for target image generation model. Next, we use LoRA [12] for customized training on our dataset. Compared to full-model fine-tuning methods like DreamBooth [11], the LoRA offers

similar performance while being more efficient and resource-friendly in training and inference. Therefore, by fine-tuning LoRA model, we trained a target image generation model that consistently generate distinct and effective target images for each perturbation training process.

Based on the joint loss of Mist [7], denoted as L_m, we further propose the Enhanced-Mist module, L_{em}. The L_{em} incorporates the random target image generation mechanism, as shown below.

$$L_{em} = L_m(ADE_i, D_{ran\_target_i}) \tag{1}$$

In Equation (1), ADE_i refers to the adversarial example, $D_{ran\_target_i}$ represents the random target image generated based on the design approach outlined above. This ensures that the target image in Enhanced-Mist is infinitely diverse and non-repetitive.

As shown in Fig.3(b), through the "repeated training attack", even if $I_{protection_{\#2}}$ and $ADE_{\#1}$ are used to perform pixel subtraction and further train generative AI model$_{\#2}$, the generative AI model$_{\#2}$ can still generate distorted results. Furthermore, by calculating the FID to measure the differences between $I_{protection_{\#1}}$ and $I_{protection_{\#2}}$, we found an FID value of 60.127, which is higher than that of Mist-generated perturbations. This indicates that the perturbation distribution in Enhanced-Mist is more diverse, demonstrating better robustness to "repeated training attacks".

Color space and frequency domain constraint optimization To enhance perturbation's imperceptibility while keeping interference with Diffusion model's training or inference, we propose color space and frequency domain constrained optimizations.

Color space constrained optimization Perturbations within chroma channels in color space are almost undetectable to human vision but can still effectively affect Diffusion models. Therefore, we propose a color space constraint strategy limiting the perturbation to chroma channels. This is achieved by maintaining the luminance information's similarity between D_{orig_i} and ADE_i. Since luminance channel Y and chroma channel UV are distinctly separated in YUV color space, we convert image from RGB to YUV color space according to BT.470 system [19]. Next, we can extract the luminance information, denoted as D^Y_{orig}. The specific method is outlined below.

$$D^Y_{orig} = \alpha D^R_{orig} + \beta D^G_{orig} + \gamma D^B_{orig} \tag{2}$$

In Equation (2), D^R_{orig},D^G_{orig}, D^B_{orig} respectively represent R, G and B channels of the RGB color space image. For the weight coefficients α, β and γ, we set them to 0.299, 0.589 and 0.114 respectively, as specified in BT.470 system [19].

To limit the perturbation to chroma channel UV, we design a color space constraint strategy. It aims to maximize the similarity between the D^Y_{orig} (luminance information of original image) and ADE^Y_i (luminance information of

adversarial example) to improve similarity between D_{orig} and ADE_i. The loss function of color space constraint is shown in Equation (3):

$$L_c = \sum_{i}^{K} (1 - MS\_SSIM(D^{Y}_{orig_i}, ADE^{Y}_{i})) \tag{3}$$

Where $D^{Y}_{orig_i}$ is luminance information of the image in YUV color space. Here, K represents the number of images in the training set, and i denotes the current image. We use MS-SSIM [22] to calculate the multi-dimensional similarity of two images in brightness, contrast and structure, with values ranging from 0 to 1. The training process is complete when L_c approaches 0.

Frequency domain constrained optimization According to human visual theory, the human eye has a limited capacity to discern fine details in images. This insight led us to restrict $I_{protection_i}$ to image details that are more easily overlooked by humans, thereby improving the similarity between the D_{orig_i} and ADE_i. From a frequency domain standpoint, high-frequency components reflect details like noise and texture, while low-frequency embodies the fundamental structure. Therefore, perturbations in high-frequency information are less noticeable than those in low-frequency. In light of this, we propose frequency domain constraint, aiming to limit perturbation to high-frequency information. We first use discrete cosine transform (DCT) [20] to convert image from spatial domain to frequency domain, and then decompose it to obtain low-frequency and high-frequency components. Next, we reconstruct image by maximizing the similarity between $D^{L}_{orig_i}$(low-frequency image from D_{orig_i}) and ADE^{L}_{i} (low-frequency image from ADE_i). This allows the perturbation to be better hidden in the high-frequency areas of the image. The formula for obtaining low-frequency information is as follows.

$$x^L(k_1, k_2) = \begin{cases} X^L(k_1, k_2) & , 0 \leq k_1 + k_2 \leq 2Mr \\ 0 & , otherwise \end{cases} \tag{4}$$

In Equation (4), $x^L(k_1, k_2)$ represents the low-frequency information of image $x^{M \times M}$ in frequency domain, where $2Mr$ mainly controls low-frequency component image, and r is usually taken as 0.08. To obtain low-frequency information from image $x^{M \times M}$, we reconstruct it through inverse DCT. The formula is as follows.

$$x^L(m_1, m_2) = \sum_{n_1=0}^{M-1} \sum_{n_2=0}^{M-1} X^L(k_1, k_2) C_1(n_1, k_1) C_2(n_2, k_2) \tag{5}$$

Furthermore, this paper proposes a frequency domain constrained loss function. The formula is as follows.

$$L_f = \sum_{i}^{K} (1 - MS\_SSIM(D^{L}_{orig_i}, ADE^{L}_{i})) \tag{6}$$

In Equation (6), we maximize the MS-SSIM value between $D^L_{orig_i}$ and ADE^L_i to limit perturbation to high frequency information. The training process is complete when L_f approaches 0.

EMCF Constraint Optimization To improve the similarity between D_{orig_i} and ADE_i, obtaining optimal optimization effect, we integrate L_{em} (Enhanced-Mist loss), L_c (Color space constraint optimization) and L_f (Frequency domain constraint optimization). We propose a unified constraint optimization denoted as L_{emcf} in Equation (7).

$$L_{emcf} = L_{em} + \lambda(L_c + L_f) \tag{7}$$

Where λ is a hyperparameter specific for color space and frequency domain constrained loss function. The training is complete when L_{emcf} converges. Through unified constraint optimization, we both enhance the robustness, but also optimize the imperceptibility and ensure the interference capacity.

3.2 Design of decentralized data protection protocol

In Section 3.1, we design a EMCF data protection method to enhance the performance of adversarial examples from both robustness and imperceptibility perspectives, preventing malicious use of copyrighted data. However, merely protecting data is insufficient to unleash its potential value. How to develop a decentralized data protection protocol tailored to circulation demands, enabling the data authorization under specified conditions, is key to unlocking data value.

As analyzed in Section 2.3, current data circulation strategies within authorized boundary often adopt centralized mode, which lacks transparency and makes tracing data ownership difficultly. Blockchain technology offers new approaches to solve these issues. First, the blockchain's decentralization, non-tampering and traceability enhance trust in data circulation. Then, smart contracts automate the transactions based on preset conditions, reducing risks related to centralized intervention. Besides, tokenization enables data assetization and make data managed like physical assets, clarifying ownership and use rights.

Based on the above design idea, we propose a decentralized data protection protocol, aiming to protecting data owners' original data in a trustworthy way while promoting data reuse with authorized boundaries, ultimately unleashing data value.

As shown in Fig. 4, the overview of decentralized data protection protocol is as follows. Next, we will provide a detailed introduction to roles and smart contracts involved in the protocol, and finally introduce overall execution process.

Roles The protocol involves three key roles: data owner, data user and worker.

1) Data owners intend to protect their data from unauthorized access. They can set authorization conditions such as fees and duration to decide the authorized data users.

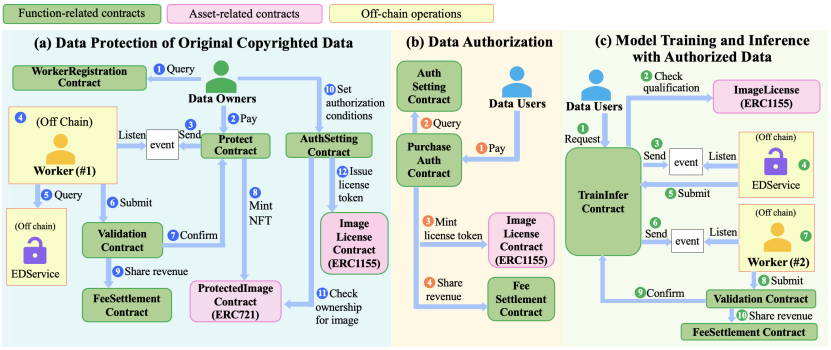

Fig. 4. Overview of decentralized data protection protocol

2) Data users request authorization for model training or inference. To prevent their potential malicious behavior, our protocol ensures they access only AI-generated results, not the original data.

3) Workers provide computational support, using verifiable computing technology for off-chain training tasks, including performing perturbation training on data owner's images, and Diffusion model training or inference for data users.

Contracts In addition, we design a collection of smart contracts to achieve efficient collaboration in data protection and authorization management. They can be primarily divided into function-related and asset-related contracts.

Function-related contracts:

1) WorkerRegistration contract: Records the worker's availability status.

2) Protect contract: Creates training tasks, assigning workers for perturbation training, and triggers the minting process for image owner's ownership.

3) Validation contract: Verifies the results of EMCF perturbation training and Diffusion model training or inference through verifiable computing [26], ensuring computing reliability and security.

4) AuthSetting contract: Configures the authorization conditions set by data owner, including authorization fees, usage time limits, etc.

5) PurchaseAuth contract: Verifies whether the data users meet the authorization conditions and triggers FeeSettlement contract for fee allocation.

6) FeeSettlement contract: Allocates fees revenue among participants, e.g., the computing workers.

7) TrainInfer contract: Assigns off-chain workers to perform training or inference tasks.

Asset-related contracts:

1) ProtectedImage contract: Based on ERC721 [21], it links the image owner's address, adversarial examples, encrypted perturbations and encrypted keys for

decrypting the perturbations. Its non-fungible nature each image is uniquely managed as an independent asset.

2) ImageLicense contract: Serves as a credential for authorized users. Leveraging ERC1155 [21], it supports batch operations, and reduces gas consumption.

Execution process The entire process can be divided into three main stages: (a) protection of original data, (b) authorization management, and (c) model training and inference with authorized data.

Protection of original data The process for the data owners to protect their original data is shown in Fig. 4(a). First, the data owner queries an available worker (e.g., $\text{Worker}_{\#1}$) via WorkerRegistration contract. To ensure the secure transmission of the original data, the data owner encrypts the original image (D_{orig_i}) using $\text{Worker}_{\#1}$'s public key and uploads it to IPFS (Step 1). Following this, data owner proceeds to pay for image protection (Step 2). Then, the Protect contract creates a protection task, and assign $\text{Worker}_{\#1}$ to perform perturbation training (Step 3). For balancing the security and efficiency during the protection process, we design a hybrid encryption strategy for transmission. In detail, $\text{Worker}_{\#1}$ decrypts the original image with $\text{Worker}_{\#1}$'s private key and applies EMCF data protection. This process generates both the perturbation ($I_{protection_i}$) and adversarial example (ADE_i). $\text{Worker}_{\#1}$ then generates a random symmetric key ($PertKey_i$) and encrypt the $I_{protection_i}$ with $PertKey_i$, resulting in $EI_{protection_i}$ (Step 4). Then, $\text{Worker}_{\#1}$ queries the EDService's public key and uses it to encrypt $PertKey_i$ to get $E_{PertKey\_EDservicePub}$ (Step 5). The ADE_i, $EI_{protection_i}$, $E_{PertKey\_EDservicePub}$ are stored in IPFS. Once these actions are completed, $\text{Worker}_{\#1}$ submits the results to Validation contract for verification (Step 6). This checks that $\text{Worker}_{\#1}$ has strictly followed the requirements in training the perturbations, and has not engaged in any malicious behavior, such as providing incorrect results or data theft (Step 7). If the verification is successful, Protect contract calls ProtectedImage contract to mint an NFT (ERC721) for the adversarial example ADE_i (Step 8). At this point, FeeSettlement contract distributes the payment from data owner to $\text{Worker}_{\#1}$ and other participants such as the platform (Step 9). Finally, data owner configures authorization conditions for the protected image using AuthSetting contract (Step 10). AuthSetting contract first validates data owner's NFT ownership and issues an ImageLicense contract (ERC1155) as the authorization license for protected image (Step 11-12).

Authorization management The procedure for the data user to obtain authorization is illustrated in Fig. 4(b). Initally, data user pays the PurchaseAuth contract for the authorization rights of ADE_i (Step 1). Then, PurchaseAuth contract queries AuthSetting for the authorization price of ADE_i and check the data user's payment (Step 2). Following this, PurchaseAuth mints an image license (ERC1155) for data user and distributes the payment to the data owner and other participants like the platform via FeeSettlement contract (Step 3-4).

Model training and inference with authorized data The process for the data user to train or perform inference with Diffusion model based on authorized images is shown in Fig. 4(c). Data user submits image license (ERC1155) to the TrainInfer contract for Diffusion model training or inference (Step 1). At this stage, TrainInfer contract verifies data user's authorization via ImageLicense contract and sets up the corresponding training or inference task (Step 2). To continue, TrainInfer contract selects an available worker (e.g.Worker$_{\#2}$) and notifies EDService for further processing. In detail, EDService then decrypts the received $E_{PertKey\_EDservicePub}$ using EDService's private key, thereby obtaining $PertKey_i$ for recovering $I_{protection_i}$. Following this, EDService re-encrypts the $PertKey_i$ using Worker$_{\#2}$'s public key, thus getting a new encrypted key ($E_{PertKey\_W2}$) that only Worker$_{\#2}$'s private key can decrypt. Then, EDService send the $E_{PertKey\_W2}$ to TrainInfer contract (Step 3-5). Once the TrainInfer contract receives this, it allocates Worker$_{\#2}$ to decrypt $E_{PertKey\_W2}$ with Worker$_{\#2}$'s private key, and then obtain $PertKey_i$. Next, Worker$_{\#2}$ downloads the $EI_{protection_i}$ from IPFS and use $PertKey_i$ to restore $I_{protection_i}$. By performing pixel-bit subtraction of $I_{protection_i}$ and ADE_i, Worker$_{\#2}$ finally obtains D_{orig_i}, and further performs the Diffusion model's training or inference task (Step 6-7). After the task is completed, Worker$_{\#2}$ submits results for Validation contract for verification to ensure that the training and inference were conducted as agreed, without any malicious behavior such as data theft or submitting incorrect results. Upon successful validation, TrainInfer contract is notified (Step 8-9). Lastly, Validation contract calls FeeSettlement contract to distribute a portion of the authorization fee to Worker$_{\#2}$ (Step 10).

Therefore, our proposed decentralized data protection protocol, integrated with blockchain technology, can effectively protect data owners' original data in a trustworthy manner, while ensuring interoperability across different platforms. At the same time, it effectively maintains the separation of data ownership and usage rights during the authorization process, ultimately unlocking the value of the data.

4 Experiment and Evaluation

In this section, we will evaluate the effectiveness of EMCF data protection method, and assess the perturbation trained by EMCF from three aspects: interference performance, imperceptibility, and robustness.

4.1 Experiment Configuration

The sampling step size of training perturbation is 100, and perturbation step is set to 1/255. In addition, by testing multiple values of λ and comparing their effects on adversarial examples, the hyperparameter λ for color space and frequency domain constraints is set to 2×10. For evaluation metric, MS-SSIM [22] measures the similarity between D_{orig_i} and ADE_i. To assess $I_{protection_i}$'s effect, we uses FID [23] to evaluate the similarity between AI-generated results and

original images, while IL-NIQE [25] evaluates image's perceptual quality. Higher FID and IL-NIQE indicates better interference effects on Diffusion models.

4.2 Evaluation of interference performance

We assess $I_{protection}$'s interference effect on Diffusion models by comparing Inf_m (Mist method), Inf_{em} (Enhanced-Mist method), Inf_{emcf} (EMCF method), where Inf represents the inference result of the Diffusion model trained with respective ADE using Lora and Dreambooth. We use the same experimental setting as Mist [7], with each method generating 50 inference results for experiment, and the FID and IL-NIQE values are calculated as the average of the 50 inference results. The experimental results are shown in Fig. 5. Compared with the inference result of the Diffusion model trained on original images, the FID of Inf_{emcf} in both LoRA and Dreambooth is the highest, indicating the best interference performance. From a qualitative standpoint, through the Inf_{emcf}, the generated image and background are more distorted, and covered with obvious watermarks. It demonstrates that our proposed color space and frequency domain constraints have yielded favorable results in interference with the training or inference process of Diffusion models.

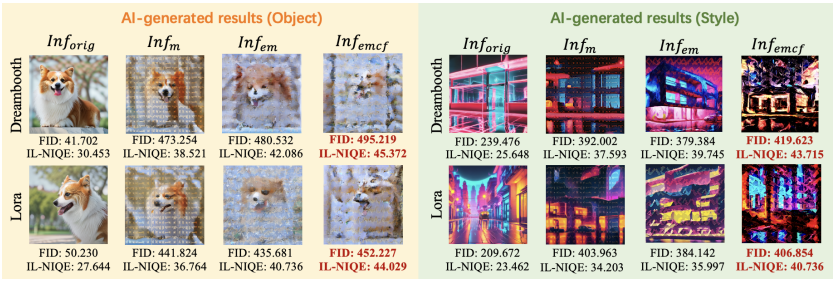

Fig. 5. Interference performance of adversarial examples

4.3 Evaluation of imperceptibility

We assess the similarity between D_{orig_i} and ADE_i. We calculate the MS-SSIM value for D_{orig_i} and ADE_i derived from different methods, including ADE_m, ADE_{em} and ADE_{emcf}, which respectively refer to the adversarial examples produced based on Mist, Enhanced-Mist and EMCF. In this experiment, each MS-SSIM value is calculated as the average value on 20 pair of images. As shown in Fig. 6, EMCF method the highest MS-SSIM values of 0.893 and 0.902 in object and style scenarios respectively, which outperforms Mist methods in imperceptibility. Therefore, it verifies that the adversarial examples generated by EMCF method demonstrate higher imperceptibility than those from other methods.

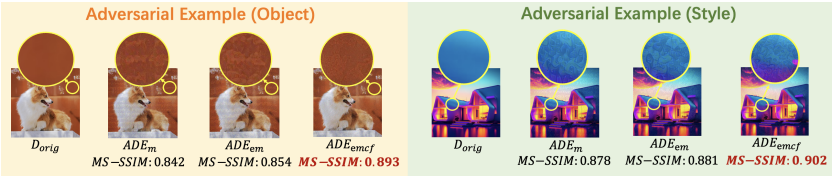

Fig. 6. Similarity comparison between original image and adversarial examples

4.4 Evaluation of robustness

We employ qualitative experiments to assess whether the ADE_i under specific attacks can still interference Diffusion models' training process. Specifically, the attacks include "repeated training attack", "scaling attack", "screenshot attack". The practice of "repeated training attack" is shown as in Fig. 1(c). The "scaling attack" compresses the ADE_i to an image of 256 $\times$ 256 pixels and then upscale it to 512 $\times$ 512 pixels to test clarity loss. The "screenshot attack" involves taking a screenshot of ADE_i and assess the impact of ADE_i under the loss of image detail or decreased resolution caused by the screenshots. Here we use the generation results of the Diffusion models trained with ADE_i under specific attacks (denoted as $aInf$ to evaluate the robustness. As shown in Fig. 7, after "repeated training attacks", "scaling attacks" and "screenshot attacks", the FID and IL-NIQE values of $aInf_{emcf}$ are higher than those of $aInf_m$ and $aInf_{em}$ indicating that EMCF method exhibits superior robustness.

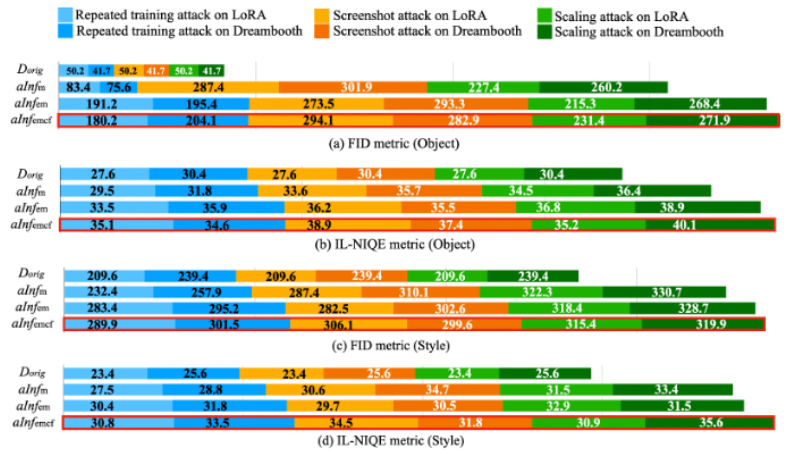

Fig. 7. Evaluation of robustness

5 Conclusion and Future Work

This paper proposes a data protection solution integrating adversarial examples and blockchain technology for generative AI in image field. In view of the shortcomings of existing adversarial example methods in terms of robustness and visual quality, we design an EMCF data protection method. By incorporating a random target image generation module, the color space and frequency domain constraints, this approach significantly enhances the robustness and imperceptibility performance of the adversarial examples. Additionally, considering the demands of data circulation and creative reuse, this paper innovatively integrates blockchain technology to design a decentralized data protection protocol. This protocol not only clarifies data usage rights and data ownership but also enables interoperability across applications, offering new perspectives for large-scale applications in the future. It ensures data sharing and circulation safely within the scope of legitimate authorization, thus ensuring data security while unlocking its potential. Experimental results show that the proposed method achieve positive outcomes in terms of data protection effectiveness.

Future research can be explored from the following aspects: 1) The current data protection method primarily focuses on Diffusion model. In future, extending protection to be applicable to both Diffusion and GAN models represents a promising research direction. 2) For the fee settlement protocol, a more reasonable data revenue distribution mechanism should be designed to better balance the interests of all parties involved. For example, a Shapley value-based revenue sharing strategy for generative AI participants can be explored.

Acknowledgments. The work of Bingxue Zhang is supported by National Natural Science Foundation of China under Grant 62007024. The work of Feida Zhu is supported by the Ministry of Education, Singapore (reference number: 001526-00001, 001508-00001), WEB 3 SECURITY (reference number: 001529-00001), Industry Alignment Fund - Prepositioning (IAF-PP), reference number: 001177-00001. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of Ministry of Education, Singapore. Moreover, we appreciate Xumin Gui and Yang Gao for assistance during experimental process.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Robin, R., Andreas, B., Dominik, L., et al: High-resolution image synthesis with latent diffusion models. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10684–10695. IEEE (2022)
2. Zhang, Q., et al. "Exploring Edge-driven Collaborative Fine-tuning Towards Customized AIGC Services." IEEE Network (2024).
3. Zhu, J., Shen, Y., Zhao, D., et al: In-domain gan inversion for real image editing. In: European Conference on Computer Vision (ECCV), pp. 592–608. Springer (2020)

4. South Korea's Digital Sex Crime Deepfake Crisis. https://www.hrw.org/news/2024/08/29/south-koreas-digital-sex-crime-deepfake-crisis, last accessed 2025/02/20
5. AI art tools targeted with copyright lawsuit, https://www.theverge.com/2023/1/16/23557098/generative-ai-art-copyright-legal-lawsuit-stable-diffusion-midjourney-deviantart, last accessed 2025/02/20
6. Liang, C., Wu, X., et al: Adversarial example does good: Preventing painting imitation from diffusion models via adversarial examples. arXiv preprint arXiv:2302.04578, (2023).
7. Liang, C., Wu, X.: Mist: Towards improved adversarial examples for diffusion models. arXiv preprint arXiv:2305.12683, (2023)
8. Goodfellow, I., Pouget-Abadie, J., et al. "Generative adversarial networks." Communications of the ACM 63(11), 139-144 (2020)
9. Gal, R., Alaluf, Y., Atzmon, Y., et al: An image is worth one word: Personalizing text-to-image generation using textual inversion. arXiv preprint arXiv:2208.01618, (2022).
10. Salman, H., Khaddaj, A., Leclerc, G., et al: Raising the cost of malicious ai-powered image editing. arXiv preprint arXiv:2302.06588, (2023).
11. Ruiz, N., et al: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In: IEEE/CVF conference on computer vision and pattern recognition(CVPR), IEEE, (2023).
12. Edward, J.H., Yelong, S., Phillip, W., et al: Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685, (2021).
13. Xiaoyu, Y., Hao, H., et al: Duaw: Data-free universal adversarial watermark against stable diffusion customization. arXiv preprint arXiv:2308.09889, (2023)
14. Thanh, V.L., Hao, P., Thuan, H.N., et al: Anti-dreambooth: Protecting users from personalized text-to-image synthesis. In: IEEE/CVF International Conference on Computer Vision (CVPR), pp. 2116–2127, IEEE, (2023).
15. Dawex, https://www.dawex.com/en/, last accessed 2025/02/20
16. Xignite, https://www.xignite.com/, last accessed 2025/02/20
17. Sober, M., et al: A blockchain-based IoT data marketplace. Cluster computing, 26(6): 3523-3545 (2023)
18. Wang, Z., et al: ArtChain: Blockchain-enabled platform for art marketplace. 2019 IEEE international conference on blockchain (BLOCKCHAIN). IEEE, (2019)
19. Andrew, Y. "Galaxy Classification Using Transfer Learning and Ensemble of CNNs With Multiple Colour Spaces." arXiv preprint arXiv:2305.00002 (2023)
20. Ahmed, N., Natarajan, T., Rao, K. R.: Discrete cosine transform. IEEE transactions on Computers 100(1), 90-93 (2006)
21. Tan, Y., Wu, Z., Liu, J., et al: Bubble or not: an analysis of ethereum erc721 and erc1155 non-fungible token ecosystem. In: 2024 IEEE international symposium on circuits and systems (ISCAS), pp. 1-5. IEEE, (2024)
22. Zhou, W., Eero, P.S., Alan, C.B. Multiscale structural similarity for image quality assessment. In Thrity-Seventh Asilomar Conference on Signals, Systems and Computers(ACSSC), pp. 1398–1402. IEEE, 2003.
23. Soloveitchik, M., Diskin, T., Morin, E., et al: Conditional frechet inception distance. arXiv preprint arXiv:2103.11521, (2021)
24. Midjourney, https://www.midjourney.com/home, last accessed 2025/02/20
25. Lin, Z., Lei, Z., Alan, C.B.: A feature-enriched completely blind image quality evaluator. IEEE Transactions on Image Processing, 24(8), 2579–2591 (2015).
26. Bontekoe, Tariq, Dimka Karastoyanova, and Fatih Turkmen.: Verifiable privacy-preserving computing. arXiv preprint arXiv:2309.08248 (2023).