CustomMark: Customization of Diffusion Models for Proactive Attribution

Vishal Asnani[1,2] John Collomosse[1,3] Xiaoming Liu[2] Shruti Agarwal[1]
[1]Adobe Research, [2]Michigan State University, [3]University of Surrey
{vasnani, collomos, shragarw}@adobe.com liuxm@msu.edu

Abstract

Generative AI (GenAI) presents challenges in attributing synthesized content to its original training data, particularly for artists whose styles are replicated by these models. We introduce CustomMark, a novel technique for customizing pre-trained text-to-image GenAI models to enable attribution. With CustomMark, text prompts can be modified to embed a watermark in generated images, linking them to training concepts such as an artist's style, specific objects, or the GenAI model itself. Our approach supports sequential customization, allowing new concepts to be attributed efficiently and scalably without retraining from scratch. We demonstrate that CustomMark can robustly watermark hundreds of individual concepts and support multiple attributions within a single image while preserving the high visual quality of the generation.

1. Introduction

Given GenAI's potential to democratize creativity, ethical concerns have emerged among artists regarding the unauthorized use of their works. Many seek recognition or compensation for the derivative use of their styles in generated images [44]. In the past, such creative recognition has relied on collaborations between technology, legal frameworks, and artistic practices [9]. GenAI currently lacks such mechanisms, leading to artist discontent and prompting adversarial strategies like "Glaze" [51], "Anti-DreamBooth" [60], and others [21, 24, 80] to protect their works.

To address this discontent, it is needed that GenAI models provide attribution when generated images are derived from artists' works in training data. Such attribution could potentially unlock new revenue streams in the creator economy, rewarding creative opt-in to GenAI training [15]. A decentralized framework to compensate creators based on visual similarities between generated and training images was proposed in [8]. Several similarity embeddings have been explored [8, 48, 64] to determine the subset of training images that influenced the generation. While intuitive, these visual correlation-based attribution methods [8, 48, 64] often fail to provide definitive explanations and can also incorrectly attribute works not present in the training set.

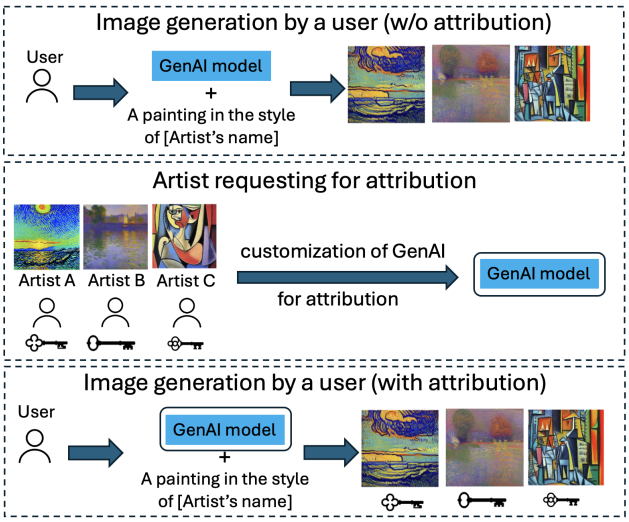

Figure 1. **Overview of concept attribution by GenAI models.** (a) A user generates images of various artists' styles using artists' tokens in the prompt (w/o attribution). (b) Artists request to the companies to provide attribution for their work. Using CustomMark, companies customize their models to enable attribution only for the artists who have requested the same. (c) A user generates the images using the improved GenAI model with an artist's specific watermark for attribution to the artists.

Alternative approaches attempt to establish direct causal relationships using techniques like proactive watermarking [6] or influence estimation via data removal [65]. However, these methods require modifications to training data or inference paradigms, making them computationally heavy.

In response, we propose CustomMark, an efficient technique for attribution in pre-trained GenAI models. Similar to [6], we use concept-specific watermarking but without requiring predefined concepts before training. CustomMark enables selective attribution of specific concepts in a pre-trained model, supporting sequential learning for newly emerging seen or unseen concepts. This approach avoids exhaustive retraining and allows attribution only for relevant concepts.

As shown in Fig. 1, we focus on attribution in text-to-image Latent Diffusion Models (LDMs), where attributable concepts appear in prompts, such as "A painting in the style of V*" or "An image of V*." If the owner of concept V* requests attribution, CustomMark embeds a concept-specific watermark into generated images while preserving visual quality. Unlike [6], which attributes to a subset of training images, CustomMark directly attributes the concept itself. The watermark remains robust against non-editorial modifications, ensuring traceability to the original concept and the GenAI model as the image circulates online. Since CustomMark embeds watermarks in a concept-specific manner without requiring exhaustive retraining, it effectively functions as a form of model customization.

Current customization methods [19, 22, 28, 32, 36, 47, 52, 69, 72, 75] struggle to scale across many distinct concepts, often compromising generation quality. To address this, we propose a novel architecture that customizes pretrained LDMs for large-scale watermarking. Building on [21], we use a concept encoder to map a bit-secret to token-embedding perturbations, but find it insufficient for scalability. Thus, we introduce a mapper network that perturbs input Gaussian noise, we fine-tune the LDM's attention layers, and leverage CSD [54] loss for faster training and improved image quality. CustomMark enables fine-tuned LDMs to generate watermarked images aligned with text prompts while embedding corresponding watermarks. Its sequential learning capability allows new attributions with just 10% additional finetuning, preserving visual quality while protecting artist styles. Our contributions are:

1. An efficient, scalable technique to customize LDMs for imperceptibly watermarking single or multiple seen/unseen concepts in a generated image, enabling robust concept attribution in pre-trained text-to-image LDMs.
2. Sequential attribution capability, allowing fine-tuning for new concepts dynamically without retraining the model, ensuring selective attribution of relevant seen and unseen concepts.
3. Demonstration that diffusion models can attribute 100s of artists' styles and 1000 ImageNet classes while maintaining high visual quality of watermarked concepts.

2. Related Works

Proactive Schemes. Proactive methods enhance various tasks by embedding signals or perturbations into input images, providing benefits to deepfake tagging [63], detection of manipulated content [2], localization of manipulations [4], object detection [3, 21], and concept attribution [6]. Some approaches focus on altering the training data to disrupt the output of generative models [46, 70]. Meanwhile, Alexandre *et al*. [49] introduce a fixed signal method to enable attribution of training datasets. Recently, a survey by Asnani *et al*. [7] discuss various proactive approaches, encryption schemes, learning process, and their applications, such as vision model defense [58, 67], LLM defense [37, 66, 77], privacy protection [43, 57, 68, 81], improving GenAI models [30, 33, 35, 38, 53, 74], 3D domain [26, 27, 29, 59, 71, 76], *etc*. In CustomMark, we use proactive techniques to do concept attribution in an efficient and scalable manner, with a focus on practical application to real-world scenarios.

IP Protection and Concept Attribution. For IP protection of AI-generated models and content, watermarking techniques embed signals into outputs via model fine-tuning [23], prompt verification [34, 79], and token-level adjustments [31]. Copyright-focused tools like DiffusionShield [16] and detection watermarking [39] prevent misuse, while latent fingerprinting [5] and audio watermarking [12] extend protection across media. Additional model security is provided by DeepSigns [17], DeepMarks [14], and network embedding [62], as well as deep spatial encryption [73], backdoor triggers [1], and dynamic defenses like DAWN [55].

Concept attribution identifies which training data influenced a generated output, distinct from model [11] or camera attribution [13]. Traditional methods passively assess visual similarities between generated and training images using predefined criteria. For instance, Wang *et al*. [64] propose Attribution by Customization (AbC), modifying embeddings like CLIP and DINO with customized diffusion models. Style-specific attribution methods such as ALADIN [48] and EKILA [8] employ perceptual hashing for patch-based matching. MONTRAGE [10] monitors weight updates to attribute pre-trained concepts, while Asnani *et al*. [6] embed concept-specific watermarks in training images for direct attribution. In contrast, we introduce a proactive watermarking technique that requires no training data modifications and enables selective, sequential attribution after training.

GenAI Customization. Advances in GenAI customization leverage techniques like Video Motion Customization [28], Custom Diffusion [32], and CustomNet [72] to adapt models to specific concepts and motions, while approaches like Modular Customization [41] and CIDM [20] enhance scalability and prevent catastrophic forgetting. Efficiency-focused methods [19] and LoRA-Composer [69] optimize customization with minimal parameter adjustments, while AquaLoRA [22] provides watermarking for unauthorized use protection, and textual inversion [36, 47, 75] enables precise text-based editing. Privacy-oriented anti-customization [61] offers additional security by adapting adversarial strategies. We propose a proactive concept attribution technique using model customization, which hasn't been explored before.

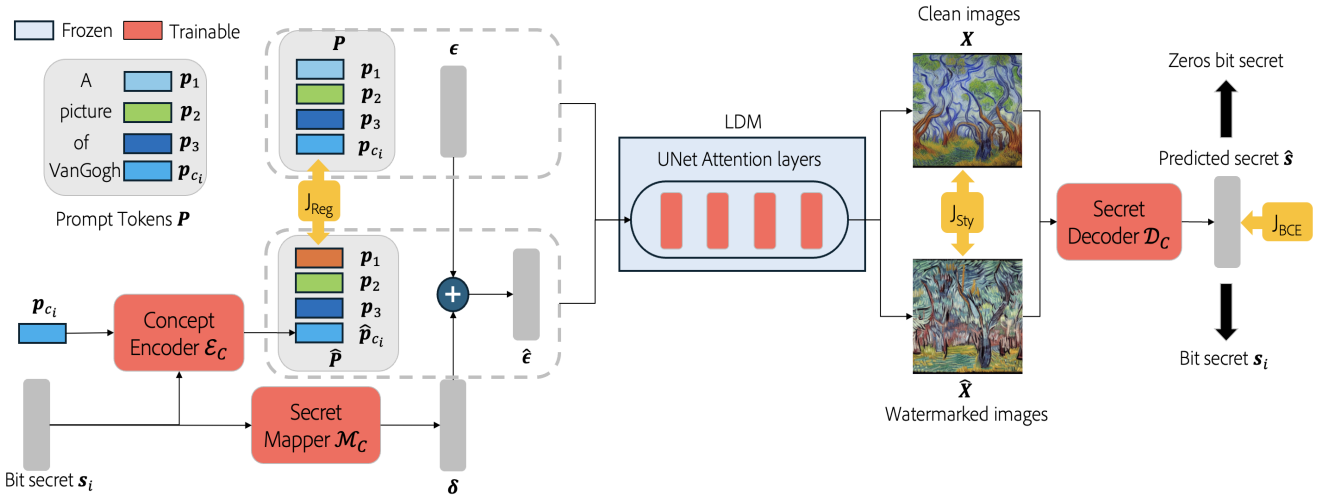

Figure 2. **Overview of CustomMark.** Illustrating the training workflow for CustomMark. A concept token $\boldsymbol{p}_{c_i}$ is encoded through the Concept Encoder $\mathcal{E}_C$ to generate a modified prompt $\hat{\boldsymbol{p}}_{c_i}$ with embedded watermark information. The Secret Mapper $\mathcal{M}_C$ maps a bit secret $\boldsymbol{s}_i$ to perturb the concept token, producing δ, which is added to the Gaussian noise ϵ. The LDM using the prompt tokens and perturbed Gaussian noise, producing watermarked images $\hat{\boldsymbol{X}}$ that carry the bit secret in visual form. During inference, the Secret Decoder $\mathcal{D}_C$ extracts the bit secret from watermarked image $\hat{\boldsymbol{X}}$ and the clean image $\boldsymbol{X}$ to extract the bit secret. CustomMark is guided by various constraints, namely regularization loss J_{Reg} to make the artist token embedding similar, style loss J_{Sty} to maintain style consistency between clean and watermarked images, and the bit secret loss J_{BCE} to predict the added bit secret. Best viewed in color.

.

3. Method

3.1. Background

Prompts and Cross-Attention Mechanism in Diffusion Model. In text-to-image LDMs [45], prompts and cross-attention mechanisms work together to guide image generation. A prompt is processed by a text encoder and converted into a *text embedding*. This embedding conditions the sampling process by capturing the prompt's meaning. Instead of merely producing random images, the cross-attention mechanism allows the model to "attend" to specific parts of the text embedding, guiding the diffusion process to align the output with the input prompt. For key $\boldsymbol{K}$, query $\boldsymbol{Q}$ and value $\boldsymbol{V}$, the scaled dot-product attention is given by:

$$\text{Attention}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \text{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{d_k}}\right)\boldsymbol{V}. \quad (1)$$

Further, multi-head cross attention with respective weight matrices $\boldsymbol{W}_i^*$s is utilized to improve generation quality by processing the prompt with multiple attention heads:

$$\text{MultiHead}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \text{Concat}(\text{H}_1, \ldots, \text{H}_h)\boldsymbol{W}, \quad (2)$$

$$\text{H}_i = \text{Attention}(\boldsymbol{Q}\boldsymbol{W}_i^Q, \boldsymbol{K}\boldsymbol{W}_i^K, \boldsymbol{V}\boldsymbol{W}_i^V). \quad (3)$$

As the multi-head cross-attention in Eq. (3) is the main component to establish a relationship between prompts and the generated image, in CustomMark, we only fine-tune $\boldsymbol{W}_i^*$s. This significantly reduces training time while enhancing critical associations between the concept and its watermarked image.

Concept Attribution. ProMark [6] defines the concept attribution as finding the closest concept in the training dataset for a given generated image. For this purpose, ProMark divides the entire dataset into different concepts and trains with each concept being watermarked. However, this is impractical for the real world, as it is difficult to retrain the GenAI models on the entire watermarked data. Therefore, we redefine the problem of Concept Attribution as follows.

Let $\mathcal{C}$ represent a set of N distinct concepts within the training dataset of a GenAI model. Out of the N concepts, let $\hat{\mathcal{C}} = \{c_1, c_2, \ldots, c_M\}$ be the M concepts that need attribution, whose token embeddings are represented as $\boldsymbol{P}_c = \{\boldsymbol{p}_{c_1}, \boldsymbol{p}_{c_2}, \ldots, \boldsymbol{p}_{c_M}\}$. Given a synthetic image $\boldsymbol{X}$ generated by a GenAI model using $\boldsymbol{p}_{c_i} \in \boldsymbol{P}_c$, along with other prompt token embeddings, forming an input prompt $\boldsymbol{P} = \{\boldsymbol{p}_1, \boldsymbol{p}_2, \ldots, \boldsymbol{p}_{c_i}, \ldots, \boldsymbol{p}_n\}$, the objective of concept attribution is to map $\boldsymbol{X}$ to its corresponding concept c_i. Specifically, we find a mapping function f such that $c_i = f(\boldsymbol{X})$.

3.2. CustomMark

Overview. To add attribution capabilities to a pre-trained LDM, CustomMark perturbs the inputs to the LDM and fine-tunes its attention weights. The input token embed-

ding $\boldsymbol{p}_{c_i}$ and the input Gaussian noise $\boldsymbol{\epsilon}$ are perturbed by the concept encoder $\mathcal{E}_C$ and the secret mapper $\mathcal{M}_C$ networks, which encode a concept-specific bit-secret into the respective inputs. This results in the perturbed embedding $\hat{\boldsymbol{p}}_{c_i}$ and the perturbed Gaussian noise $\hat{\boldsymbol{\epsilon}}$, which are fed into the LDM to sample new images. The synthesized images are then fed to the secret decoder $\mathcal{D}_C$ that outputs the corresponding bit-secret. During training, only the attention weights in Eq. (2), and Eq. (3) of the LDM are fine-tuned. The framework is guided by several constraints that allow for the generation of images with embedded secrets and also maintain the original artistic style. We will now present our method in detail.

Embedding Encryption. In CustomMark we perturb all the concepts in $\boldsymbol{P}_c$ using a single concept encoder $\mathcal{E}_C$. For i^{th} concept, the concept token embedding $\boldsymbol{p}_{c_i}$ is encrypted using $\mathcal{E}_C$ as:

$$\hat{\boldsymbol{p}}_{c_i} = \mathcal{E}_C(\boldsymbol{p}_{c_i}, \boldsymbol{s}_i), \tag{4}$$

where $\boldsymbol{s}_i$ is the concept specific bit-secret of length l, i.e. $\boldsymbol{s}_i = \{b_{i1}, b_{i2}, ..., b_{il}\}$ where $b_{ij} \in \{0, 1\}$.

After encryption, the original embedding is replaced by the encrypted text embedding, resulting in encrypted prompt token embeddings $\hat{\boldsymbol{P}} = \{\boldsymbol{p}_1, \boldsymbol{p}_2, \ldots, \hat{\boldsymbol{p}}_{c_i}, \ldots, \boldsymbol{p}_n\}$. To obtain the watermarked image, $\hat{\boldsymbol{P}}$ is fed to the LDM in place of the original token embeddings $\boldsymbol{P}$. Following the architecture of [21], we apply a regularization mean squared error (MSE) loss between $\boldsymbol{P}$ and $\hat{\boldsymbol{P}}$ at initial iterations, so that the encoder $\mathcal{E}_C$ has a good starting point to preserve the style, and support secret learning. The regularization loss is:

$$J_{Reg} = ||\hat{\boldsymbol{P}} - \boldsymbol{P}||_2^2. \tag{5}$$

Secret Learning. We will now discuss the learning of LDM to generate watermarked images given the encrypted token embeddings $\hat{\boldsymbol{P}}$. In addition to $\mathcal{E}_C$, we use a mapper network $\mathcal{M}_C$ to further accelerate the secret learning. Using i^{th} bit-secret $\boldsymbol{s}_i$, we estimate a perturbation $\boldsymbol{\delta} = \mathcal{M}_C(\boldsymbol{s}_i)$ which is added to the initially sampled Gaussian noise ϵ for image generation. Therefore, the perturbed ϵ is given by:

$$\hat{\boldsymbol{\epsilon}} = \boldsymbol{\epsilon} + \alpha \times \mathcal{M}_C(\boldsymbol{s}_i), \tag{6}$$

where α controls the magnitude of $\boldsymbol{\delta}$. The perturbed Gaussian noise $\hat{\boldsymbol{\epsilon}}$ along with $\hat{\boldsymbol{P}}$ is given as input to the LDM to sample an image. Finally, to avoid the complexity of LDM training, we only finetune the attention layers of the LDM while fixing other layers.

During training, we create both clean and watermarked images, $\boldsymbol{X}$ and $\hat{\boldsymbol{X}}$, using the inputs ($\boldsymbol{P}$, $\boldsymbol{\epsilon}$) and ($\hat{\boldsymbol{P}}$, $\hat{\boldsymbol{\epsilon}}$). The style descriptors $\boldsymbol{d}$ and $\hat{\boldsymbol{d}}$ from images $\boldsymbol{X}$ and $\hat{\boldsymbol{X}}$ are extracted using the pretrained Contrastive Style Descriptors (CSD) [54] model. CSD contains concise and effective style information, while being invariant to semantic content and capable of disentangling multiple styles. We maximize the cosine similarity between two descriptors, which ensures that the watermarked images match the style of the original concept. To further support style matching, we apply an MSE loss between the two images, in addition to the CSD loss. Therefore, our style loss is given by:

$$J_{Sty} = 1 - \cos(\hat{\boldsymbol{d}}, \boldsymbol{d}) + ||\boldsymbol{X} - \hat{\boldsymbol{X}}||_2^2. \tag{7}$$

$\boldsymbol{X}$ and $\hat{\boldsymbol{X}}$ are further fed to a secret decoder $\mathcal{D}_C$, which estimates the bit secret in given images. The decoder shall output a zeros secret for $\boldsymbol{X}$, and the secret $\boldsymbol{s}_i$ for $\hat{\boldsymbol{X}}$. To train $\mathcal{D}_C$, we use a binary cross-entropy (BCE) loss between the ground truth bit-sequence $\boldsymbol{s}_i$ and the predicted one $\hat{\boldsymbol{s}}_i$:

$$J_{BCE}(\boldsymbol{s}_i, \hat{\boldsymbol{s}}_i) = -\frac{1}{l}\sum_{j=1}^{l}[b_j \log(\hat{b}_j) + (1-b_j)\log(1-\hat{b}_j)]. \tag{8}$$

Therefore, CustomMark is trained in an end-to-end manner to minimize the objective $L_{attr} = L_{Sty} + L_{BCE} + \beta L_{Reg}$ during training, where $\beta = 10$ for our experiments.

During inference, if the random Gaussian noise and the input prompt are perturbed, the diffusion model embeds a watermark within the generated image. This watermark can be decoded using $\mathcal{D}_C$ to the concept-specific bit-secret, functioning as hidden signatures for attribution.

Concept Attribution in Inference. To attribute the generated images, we extract the bit secret embedded by the LDM using $\mathcal{D}_C$. Using this predicted bit-secret $\hat{\boldsymbol{s}} = \mathcal{D}_C(\hat{\boldsymbol{X}})$ and the bit-secret $\boldsymbol{s}_i$ corresponding to the concept c_i, we define the attribution mapping function f as:

$$f(\hat{\boldsymbol{X}}) = \underset{i \in [1,M]}{\operatorname{argmax}}\, g(\mathcal{D}_C(\hat{\boldsymbol{X}}), \boldsymbol{s}_i), \tag{9}$$

where,

$$g(\mathcal{D}_C(\hat{\boldsymbol{X}}), \boldsymbol{s}_i) = g(\hat{\boldsymbol{s}}, \boldsymbol{s}_i) = \sum_{k=1}^{l}[\hat{b}_k = b_{ik}], \tag{10}$$

and $[\hat{b}_k = b_{jk}]$ is an indicator function that returns 1 if the bits match, and 0 otherwise. Thus, using the predicted bit-sequence, we assign the generated images to the concept whose bit-sequence matches the best, *i.e.*, the i^{th} concept that maximizes $g(\hat{\boldsymbol{s}}, \boldsymbol{s}_i)$.

3.3. Sequential Learning

In real-world scenarios, the number of concepts requiring attribution is not always fixed. The set of concepts can change frequently, making it impractical to retrain the attribution model from scratch each time new concepts are introduced. To address this challenge, we propose the idea of sequential learning with CustomMark.

Figure 3. Comparison with ProMark [6] on ImageNet. ProMark produces low-quality images with bubble-like artifacts from its encryption, whereas CustomMark enables LDMs to generate high-quality images that closely match the original training concepts.

For example, if CustomMark is initially trained on M concepts, denoted as $\hat{\mathcal{C}} = \{c_1, c_2, \ldots, c_M\}$, and a new concept c_{M+1} needs to be attributed, the model can be fine-tuned on the expanded set $\hat{\mathcal{C}} \cup c_{M+1}$, starting from the model pretrained on $\hat{\mathcal{C}}$. This approach allows the model to adapt to new concepts without requiring a predefined set during initial training. Our experiments demonstrate that learning new concepts in this manner requires only about 10% additional iterations, making it significantly more efficient than retraining CustomMark from scratch.

3.4. Multi-Concept Learning

In real-world text-to-image generation, multiple concepts are often combined within a single prompt, such as *"a painting of a dog in the style of Van Gogh."* To enable concept attribution in such cases, CustomMark extends its attribution mechanism to handle multiple concepts simultaneously.

Given two concepts, c_i and c_j, from the attributed set $\hat{\mathcal{C}}$, their respective token embeddings $\boldsymbol{p}_{c_i}$ and $\boldsymbol{p}_{c_j}$ are perturbed using the concept encoder $\mathcal{E}_C$. This results in the perturbed embeddings:

$$\hat{\boldsymbol{p}}_{c_i} = \mathcal{E}_C(\boldsymbol{p}_{c_i}, \boldsymbol{s}_i), \quad \hat{\boldsymbol{p}}_{c_j} = \mathcal{E}_C(\boldsymbol{p}_{c_j}, \boldsymbol{s}_j). \tag{11}$$

The perturbed prompt embeddings $\hat{\boldsymbol{P}} = \{\boldsymbol{p}_1, \ldots, \hat{\boldsymbol{p}}_{c_i}, \ldots, \hat{\boldsymbol{p}}_{c_j}, \ldots, \boldsymbol{p}_n\}$ are then used in the LDM to generate a watermarked image $\hat{\boldsymbol{X}}$. During decoding, the secret decoder $\mathcal{D}_C$ is designed to recover the **concatenated secret** associated with both concepts:

$$\hat{\boldsymbol{s}} = \mathcal{D}_C(\hat{\boldsymbol{X}}) = [\boldsymbol{s}_i; \boldsymbol{s}_j]. \tag{12}$$

The concatenation ensures that both concept-specific secrets are extracted from the generated image, thereby enabling attribution for multiple concepts simultaneously. The attribution function f is then applied independently for each concept:

$$f(\hat{\boldsymbol{X}}) = \underset{i,j \in [1,M]}{\operatorname{argmax}}\, g(\mathcal{D}_C(\hat{\boldsymbol{X}}), [\boldsymbol{s}_i; \boldsymbol{s}_j]). \tag{13}$$

This approach ensures that CustomMark can reliably attribute both concepts in a multi-concept image, allowing for effective auditing of GenAI models even when multiple stylistic or semantic elements are present in the generated content.

4. Experiments

Implementation Details For training CustomMark, a predefined list of prompts is used per concept (see supplement). For concepts, we use $1,000$ ImageNet [18] classes, 23 WikiArt [56] artists, and a custom 200 list of artists (see supplement). For text-to-image LDM, Stable Diffusion 1.5 is used. Unless stated, we use a bit-sequence of size 16. We evaluate CustomMark using four metrics: bit accuracy, attribution accuracy, CSD [54] score, and CLIP [64] score as described. For attribution assessment, bit accuracy is the maximum percentage of bits matched between the predicted bit-secret and any of the concept-specific secrets, and attribution accuracy is the percentage of times the predicted bit-secret matches the correct concept-specific secret. For quality assessment, the CSD score is the cosine similarity between CSD descriptors, which assesses the style match between two images, and the CLIP score is the cosine similarity between CLIP image embeddings. For all evaluations, we report average results on 100 generated and/or 100 clean images. For 10 concepts, CustomMark is trained for $20K$ iterations. All experiments are conducted on 8 A100 NVIDIA GPUs with a batch size of 8 per GPU.

4.1. Results

Comparison with Attribution Methods We evaluate various passive and proactive attribution methods on images generated by LDMs trained on the ImageNet and WikiArt datasets, which contain 1000 and 23 classes, respectively. Here, each class is treated as a unique concept. For fair comparison, we generate 100 images per class for both ProMark [6] and CustomMark. Since ProMark and CustomMark embed different watermarks, their accuracy is reported only on their respective 100 watermarked images. Whereas for passive methods, including ALADIN [48], CLIP [42], AbC [64], SSCD [40], and EKILA [8], that rely on embeddings, the evaluation is done on images generated by both proactive models *i.e.*, an average over a total of 200 generated images per concept. As shown in Tab. 1, the passive methods exhibit relatively low attribution accuracy.

In contrast, the proactive methods, ProMark and CustomMark, significantly outperform the passive methods,

Method	Type	Attribution Accuracy (%) ↑	
		ImageNet	Wikiart
ALADIN [48]	Passive	5.55	18.58
CLIP [42]	Passive	42.61	52.60
AbC [64]	Passive	53.51	56.03
SSCD [40]	Passive	25.50	45.34
EKILA [8]	Passive	30.98	43.03
ProMark [6]	Proactive	**87.30**	87.19
CustomMark	Proactive	87.12	**89.25**

Table 1. Comparison with passive and proactive methods on images generated by a conditional model trained on ImageNet and Wikiart dataset. CustomMark outperforms the passive methods on both datasets significantly. Both proactive methods have similar performance on ImageNet, but for Wikiart, CustomMark performs better than ProMark.

Figure 4. Attribution results of three concept artists: VanGogh, Monet, and Picasso, sampled from LDM before and after applying the attribution capability of customization-based method [21] and CustomMark. [21] makes the LDM sample images far apart from the original style of artists, while CustomMark-watermarked images are much closer to the original style.

with much higher accuracy in both datasets. Although ProMark trains on an entirely watermarked dataset with all LDM parameters learnable in training, its performance is still comparable to CustomMark. Further, ProMark adversely impacts image quality, as shown in Fig. 3, where the generated ImageNet samples of ProMark are of lower quality and display visible artifacts. To quantify the quality, we calculate the FID score [25, 50] between the original ImageNet images (from a pretrained model without watermarks) and the watermarked images from each proactive model. The pretrained model achieves an FID score of 13.28. ProMark yields an FID score of 17.63, while CustomMark achieves an FID score of 14.73, indicating substantially better image quality. Thus, CustomMark not only maintains robust attribution performance but also generates higher-quality images than ProMark, making it a more effective solution for practical applications.

Comparison with Customization-Based Watermarking Methods We compare our method with [21], which also leverages textual token perturbations to guard personalized concepts. However, in [21], authors train a new concept encoder-decoder pair for each personalization; an impractical solution in the real world. For a fair comparison, we adapt [21] by training a single encoder-decoder pair for 3 artists' styles as concepts, namely VanGogh, Monet, and Picasso. As shown in Tab. 2, CustomMark surpasses this baseline in all metrics, achieving higher watermark detection accuracy (99.29), attribution accuracy (94.29), and generation quality (CSD score 0.81 and CLIP score 0.77). These results demonstrate the effectiveness of CustomMark for concept watermarking in GenAI.

Method	Bit Acc. (%)↑	Attribution Acc. (%) ↑	CLIP Score ↑	CSD Score ↑
Feng *et al.* [21]	90.87	74.14	0.57	0.51
CustomMark	**99.29**	**94.29**	**0.81**	**0.77**

Table 2. Comparison with customization-based method by Feng *et al.* [21]. [KEYS: Acc.=Accuracy]

Shown in Fig. 4 are some qualitative results for comparison. Unlike [21], which struggles to preserve individual artistic styles like brushstrokes and color palettes, CustomMark accurately captures each artist's unique nuances. For example, for Picasso (second row, last col), [21] generates Van Gogh-style brushstrokes.

Sequential Learning In Fig. 5, we showcase CustomMark's sequential learning capability, where the model begins attribution with three concepts and subsequently integrates additional concepts one at a time. This setup reflects a dynamic, real-world setting where the need for concept attribution evolves over time as new styles are added. Instead of retraining the model from scratch for each new concept, CustomMark employs sequential learning to incrementally learn attributions for new concepts without erasing previously learned styles.

Starting with three initial concepts, CustomMark finetunes the model as new concepts are introduced, updating attribution while preserving distinct stylistic features. This is evident in the similarity between clean and watermarked images in each column, where CustomMark maintains high fidelity to the original style. With sequential learning, it attributes new concepts with only 10% additional iterations per concept, avoiding full retraining. These results demonstrate CustomMark's scalability and efficiency in preserving style-consistent, high-quality outputs for GenAI models.

Unseen Artists Watermarking We demonstrate CustomMark's ability to attribute both seen and unseen concepts using textual inversion. As shown in Fig. 4, known concepts are watermarked by perturbing their token embeddings. However, in real-world scenarios, generative models often encounter novel concepts outside the initial training set, requiring adaptability beyond predefined attributions.

To address this, we leverage textual inversion to derive token embeddings for unseen concepts. Once obtained, we

Figure 5. **Sequential learning of new concepts.** CustomMark starts with three initial concepts and incrementally learns new attributions without retraining from scratch. Each column displays clean and watermarked images, demonstrating CustomMark's efficiency in adapting to new styles with only about 10% extra training iterations per concept while maintaining high stylistic fidelity. We only show the concept used to create the image. A list of all the prompts used is given in the supplement.

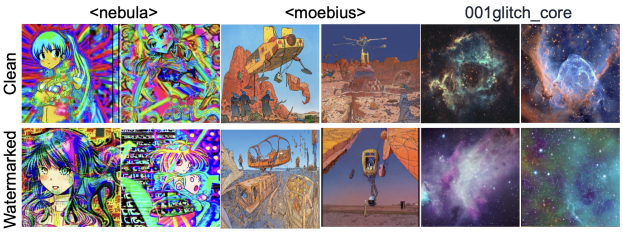

Figure 6. **Attribution of Unseen Concepts with CustomMark.** Shown is the CustomMark's ability to handle attribution for unseen concepts. The consistent style between clean and watermarked images across new styles demonstrates CustomMark's robustness in preserving artistic fidelity while achieving scalable attribution. We only show the concept used to create the image. A list of all the prompts used is given in the supplement.

apply watermark perturbations, enabling attribution without significant model retraining. Fig. 6 illustrates this by showing stylistic consistency between clean and watermarked images, preserving unique attributes of each new style. This demonstrates CustomMark's adaptability, allowing it to generalize to new styles while maintaining fidelity and stylistic integrity.

Multi-Concept Watermarking For this scenario, we take 20 concepts into consideration (10 objects, and 10 artists). Each concept is associated with an 8-bit secret. The decoder extracts a 16-bit secret for the generated image. The qualitative results in Fig. 7 demonstrate that CustomMark successfully embeds attribution signatures for both object (e.g., "dog," "tree") and style (e.g., "Van Gogh," "Picasso") concepts within a single image while preserving visual quality. Quantitatively, the attribution and bit accuracy evaluated on 100 clean and generated images are 89.14% and 95.47%.

4.2. Ablations

Unless stated otherwise, we use a model trained for 10 concepts for all the ablation experiments (see supplement).

Nearby Concepts and Clean Images. CustomMark provides the flexibility to easily switch from the watermarked image generation to a non-watermarked version, which we define as clean image generation. To do this, we use the non-perturbed original text tokens, while keeping the mapper network $\mathcal{M}_C$ and the fine-tuned attention weights of the model. An all-zero bit secret is used as an input to $\mathcal{M}_C$ and the secret decoder is expected to output the same for these clean images. We evaluate CustomMark's ability to generate clean images for 1) attributable concepts: that are fine-tuned with CustomMark and 2) nearby concepts: that are related to attributable concepts but not exactly the same. For example, if CustomMark can attribute paintings of Van Gogh, then paintings from other artists are considered nearby concepts. For this evaluation, we use three attributable artists (first three columns of Fig. 5) and seven random nearby artists (see supplement).

For attributable concepts, the model achieves high bit accuracy (96.13%) and attribution accuracy (85.45%) with an all-zeros bit secret, indicating effective attribution of clean concepts. For nearby concepts, it maintains strong bit accuracy (92.36%) and attribution accuracy (81.90%), showcasing the adaptability of CustomMark for practical applications, allowing selective watermarking for certain styles while not watermarking concepts that don't specifically request it. The generation quality with CustomMark is comparable to the pretrained LDM, with an FID score of 14.51

Figure 7. Attribution for multiple Concepts present in a single prompt with CustomMark.

Method	Bit Acc. (%)↑	Attribution Acc. (%) ↑	CLIP Score ↑	CSD Score ↑
CSD	98.6	88.15	0.65	0.73
CSD + L2 (latent)	99.12	90.94	0.70	0.67
CSD + L2 (image)	99.17	92.35	0.73	0.74
CSD + L2 + LDM atte.	**99.29**	**94.29**	**0.81**	**0.77**

Table 3. Ablation study for various style losses. [KEYS: Acc. Accuracy, Att. Attribution, Atte.: Attention]

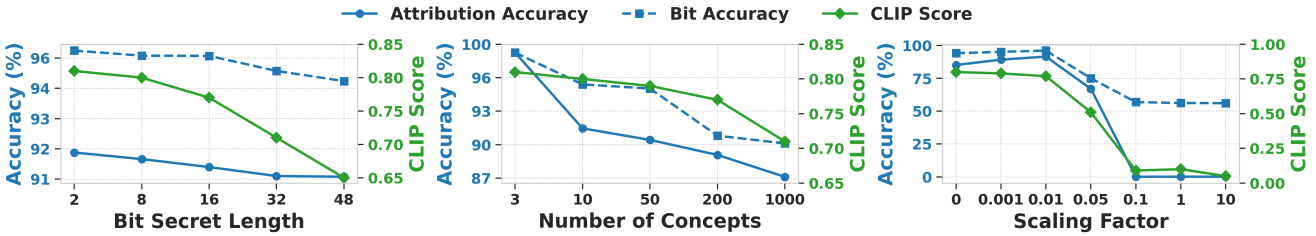

Figure 8. Ablation study for varying different parameters of CustomMark. We show the performance variation by varying the bit secret length, the number of concepts, and the scaling factor.

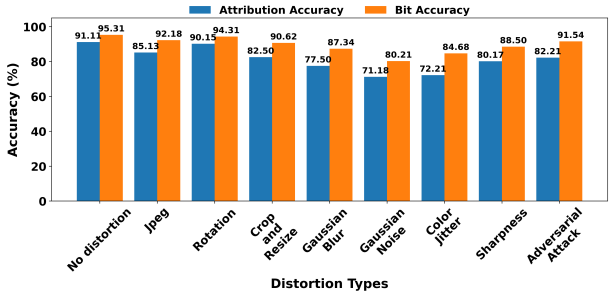

Figure 9. Robustness evaluation of decoder by applying distortion to generated images.

between original and clean images.

Style Loss. Tab. 3 presents an ablation study on different style loss combinations and their impact on bit accuracy, attribution accuracy, and qualitative metrics. The baseline using only CSD performs well, but adding L2 loss in LDM's latent space improves accuracy, with a slight drop in the CSD score. Further applying L2 loss in image space enhances overall performance, boosting attribution accuracy, CLIP, and CSD scores. The best results are achieved by CustomMark, which combines CSD, L2 loss, and attention layer training, yielding the highest gains across all metrics and validating our design choice.

Robustness. Fig. 9 demonstrates CustomMark's robustness against various post-processing distortions, including JPEG compression, rotation, cropping, resizing, Gaussian blur, noise, color jitter, and sharpness (see supplement). CustomMark maintains high attribution and bit accuracy, with minimal impact from common distortions like JPEG compression and rotation, while stronger distortions (e.g., Gaussian blur, noise) cause slight accuracy drops. Against adversarial attacks [78], it retains 82.21% attribution accuracy, only slightly lower than the original 91.11%. These results highlight CustomMark's resilience in real-world scenarios.

Bit Secret Length Fig. 8(a) analyzes the effect of secret length on bit accuracy, attribution accuracy, and CLIP score. As the secret length increases, both accuracy metrics decline, suggesting that longer secrets are harder for the decoder to recover, impacting attribution performance. Additionally, the CLIP score drops, indicating stylistic deviations. This trade-off suggests that a moderate bit length, such as 16, balances attribution accuracy and stylistic fidelity.

Number of Concepts. Fig. 8(b) examines how the number of unique artist concepts affects attribution and stylistic fidelity. As concepts increase, bit and attribution accuracy decline, likely due to the growing challenge of distinguishing among them. Similarly, the CLIP score drops, suggesting that maintaining stylistic consistency becomes harder with a broader range of styles in watermarked images.

Scaling Factor. Fig. 8(c) shows the impact of the scaling factor in Eq. (6) on attribution and stylistic similarity. Increasing the scaling factor sharply reduces both bit and attribution accuracy, likely due to overpowering the sampled Gaussian noise. Conversely, decreasing it causes the LDM to diverge, generating noise images, as reflected in the declining CLIP score. This underscores the need for a low scaling factor to balance attribution accuracy and stylistic preservation, leading us to select 0.01 for our experiments.

5. Conclusion

We propose CustomMark, an efficient and flexible technique for enabling concept attribution in pre-trained text-to-image LDMs. Addressing the growing demand for ethical content generation in GenAI models, CustomMark provides a customization-based approach to embed concept-specific watermarks, allowing artists to request attribution for their work. Unlike previous methods, CustomMark allows selective attribution without requiring all concepts to be predefined before training, and entire watermarking of the training data. It supports sequential learning to add new concepts in an online way. We demonstrate that CustomMark can handle hundreds of artist styles and diverse ImageNet classes while maintaining image quality and ensuring robust attribution. By fine-tuning the model for new concepts with minimal computational overhead, CustomMark streamlines the attribution process. This helps bridge the gap between GenAI developers and the creative community and promotes the responsible use of GenAI in content creation.

References

[1] Yossi Adi, Carsten Baum, Moustapha Cisse, Benny Pinkas, and Joseph Keshet. Turning your weakness into a strength: Watermarking deep neural networks by backdooring. In *USENIX-S*, 2018. 2

[2] Vishal Asnani, Xi Yin, Tal Hassner, Sijia Liu, and Xiaoming Liu. Proactive image manipulation detection. In *CVPR*, 2022. 2

[3] Vishal Asnani, Abhinav Kumar, Suya You, and Xiaoming Liu. PrObeD: Proactive object detection wrapper. In *NeurIPS*, 2023. 2

[4] Vishal Asnani, Xi Yin, Tal Hassner, and Xiaoming Liu. MaLP: Manipulation localization using a proactive scheme. In *CVPR*, 2023. 2

[5] Vishal Asnani, Xi Yin, Tal Hassner, and Xiaoming Liu. Reverse engineering of generative models: Inferring model hyperparameters from generated images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 2

[6] Vishal Asnani, John Collomosse, Tu Bui, Xiaoming Liu, and Shruti Agarwal. Promark: Proactive diffusion watermarking for causal attribution. In *CVPR*, 2024. 1, 2, 3, 5, 6

[7] Vishal Asnani, Xi Yin, and Xiaoming Liu. Proactive schemes: A survey of adversarial attacks for social good. *arXiv preprint*, 2024. 2

[8] Kar Balan, Shruti Agarwal, Simon Jenni, Andy Parsons, Andrew Gilbert, and John Collomosse. EKILA: Synthetic media provenance and attribution for generative art. In *CVPR*, 2023. 1, 2, 5, 6

[9] Oliver Bown. AI doesn't like to credit its sources. for artists, that's a problem. *Tatlor*, 2024. 1

[10] Jonathan Brokman, Omer Hofman, Roman Vainshtein, Amit Giloni, Toshiya Shimizu, Inderjeet Singh, Oren Rachmil, Alon Zolfi, Asaf Shabtai, Yuki Unno, and Hisashi Kojima. Montrage: Monitoring training for attribution of generative diffusion models. In *ECCV*, 2024. 2

[11] Tu Bui, Ning Yu, and John Collomosse. RepMix: Representation mixing for robust attribution of synthesized images. In *ECCV*, 2022. 2

[12] Xirong Cao, Xiang Li, Divyesh Jadav, Yanzhao Wu, Zhehui Chen, Chen Zeng, and Wenqi Wei. Invisible watermarking for audio generation diffusion models. In *TPS-ISA*, 2023. 2

[13] Chang Chen, Zhiwei Xiong, Xiaoming Liu, and Feng Wu. Camera trace erasing. In *CVPR*, 2020. 2

[14] Huili Chen, Bita Darvish Rouhani, Cheng Fu, Jishen Zhao, and Farinaz Koushanfar. Deepmarks: A secure fingerprinting framework for digital rights management of deep learning models. In *ICMR*, 2019. 2

[15] John Collomosse and Andy Parsons. To Authenticity, and Beyond! Building safe and fair generative AI upon the three pillars of provenance. *IEEE Computer Graphics and Applications*, 2024. 1

[16] Yingqian Cui, Jie Ren, Han Xu, Pengfei He, Hui Liu, Lichao Sun, and Jiliang Tang. DiffusionShield: A watermark for copyright protection against generative diffusion models. *arXiv preprint*, 2023. 2

[17] Bita Darvish Rouhani, Huili Chen, and Farinaz Koushanfar. DeepSigns: An end-to-end watermarking framework for ownership protection of deep neural networks. In *ASPLOS*, 2019. 2

[18] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, 2009. 5

[19] Ning Ding, Yujia Qin, Guang Yang, Fuchao Wei, Zonghan Yang, Yusheng Su, Shengding Hu, Yulin Chen, Chi-Min Chan, Weize Chen, et al. Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nature Machine Intelligence*, 2023. 2

[20] Jiahua Dong, Wenqi Liang, Hongliu Li, Duzhen Zhang, Meng Cao, Henghui Ding, Salman Khan, and Fahad Shahbaz Khan. How to continually adapt text-to-image diffusion models for flexible customization? *arXiv preprint*, 2024. 2

[21] Weitao Feng, Jiyan He, Jie Zhang, Tianwei Zhang, Wenbo Zhou, Weiming Zhang, and Nenghai Yu. Catch you everything everywhere: Guarding textual inversion via concept watermarking. *arXiv preprint*, 2023. 1, 2, 4, 6

[22] Weitao Feng, Wenbo Zhou, Jiyan He, Jie Zhang, Tianyi Wei, Guanlin Li, Tianwei Zhang, Weiming Zhang, and Nenghai Yu. AquaLoRA: Toward white-box protection for customized stable diffusion models via watermark lora. *arXiv preprint*, 2024. 2

[23] Pierre Fernandez, Guillaume Couairon, Hervé Jégou, Matthijs Douze, and Teddy Furon. The stable signature: Rooting watermarks in latent diffusion models. In *ICCV*, 2023. 2

[24] Rohit Gandikota, Joanna Materzynska, Jaden Fiotto-Kaufman, and David Bau. Erasing concepts from diffusion models. In *ICCV*, 2023. 1

[25] 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. *NeurIPS*, 2017. 6

[26] Gangyang Hou, Bo Ou, Min Long, and Fei Peng. Separable reversible data hiding for encrypted 3d mesh models based on octree subdivision and multi-msb prediction. *IEEE Transactions on Multimedia*, 2023. 2

[27] Youngdong Jang, Dong In Lee, MinHyuk Jang, Jong Wook Kim, Feng Yang, and Sangpil Kim. WateRF: Robust watermarks in radiance fields for protection of copyrights. In *CVPR*, 2024. 2

[28] Hyeonho Jeong, Geon Yeong Park, and Jong Chul Ye. VMC: Video motion customization using temporal attention adaption for text-to-video diffusion models. In *CVPR*, 2024. 2

[29] Ruiqi Jiang, Hang Zhou, Weiming Zhang, and Nenghai Yu. Reversible data hiding in encrypted three-dimensional mesh models. *IEEE Transactions on Multimedia*, 2017. 2

[30] Youngeun Kim, Yuhang Li, Abhishek Moitra, Ruokai Yin, and Priyadarshini Panda. Do we really need a large number of visual prompts? *Neural Networks*, 2024. 2

[31] John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. A watermark for large language models. In *ICML*, 2023. 2

[32] Nupur Kumari, Binazeiang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *CVPR*, 2023. 2

[33] Nilakshan Kunananthaseelan, Jing Zhang, and Mehrtash Harandi. LaViP: Language-grounded visual prompting. In *AAAI*, 2024. 2

[34] Yugeng Liu, Zheng Li, Michael Backes, Yun Shen, and Yang Zhang. Watermarking diffusion model. *arXiv preprint*, 2023. 2

[35] Kang Ma, Ying Fu, Chunshui Cao, Saihui Hou, Yongzhen Huang, and Dezhi Zheng. Learning visual prompt for gait recognition. In *CVPR*, 2024. 2

[36] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *CVPR*, 2023. 2

[37] Travis Munyer and Xin Zhong. DeepTextMark: Deep learning based text watermarking for detection of large language model generated text. *arXiv preprint*, 2023. 2

[38] Sungho Park and Hyeran Byun. Fair-VPT: Fair visual prompt tuning for image classification. In *CVPR*, 2024. 2

[39] Sen Peng, Yufei Chen, Cong Wang, and Xiaohua Jia. Protecting the intellectual property of diffusion models by the watermark diffusion process. *arXiv preprint*, 2023. 2

[40] Ed Pizzi, Sreya Dutta Roy, Sugosh Nagavara Ravindra, Priya Goyal, and Matthijs Douze. A self-supervised descriptor for image copy detection. In *CVPR*, 2022. 5, 6

[41] Ryan Po, Guandao Yang, Kfir Aberman, and Gordon Wetzstein. Orthogonal adaptation for modular customization of diffusion models. In *CVPR*, 2024. 2

[42] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 5, 6

[43] Arezoo Rajabi, Rakesh B Bobba, Mike Rosulek, Charles Wright, and Wu-chi Feng. On the (im)practicality of adversarial perturbation for image privacy. *PETS*, 2021. 2

[44] Anna Rogers. The attribution problem with generative ai. *Hacking Semantics*, 2022. 1

[45] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 3

[46] Nataniel Ruiz, Sarah Adel Bargal, and Stan Sclaroff. Disrupting deepfakes: Adversarial attacks against conditional image translation networks and facial manipulation systems. In *ECCVW*, 2020. 2

[47] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *CVPR*, 2023. 2

[48] Dan Ruta, Saeid Motiian, Baldo Faieta, Zhe Lin, Hailin Jin, Alex Filipkowski, Andrew Gilbert, and John Collomosse. ALADIN: All layer adaptive instance normalization for fine-grained style similarity. In *ICCV*, 2021. 1, 2, 5, 6

[49] Alexandre Sablayrolles, Matthijs Douze, Cordelia Schmid, and Hervé Jégou. Radioactive data: tracing through training. In *ICML*, 2020. 2

[50] Maximilian Seitzer. pytorch-fid: FID Score for PyTorch. https://github.com/mseitzer/pytorch-fid, 2020. Version 0.3.0. 6

[51] Shawn Shan, Jenna Cryan, Emily Wenger, Haitao Zheng, Rana Hanocka, and Ben Y Zhao. Glaze: Protecting artists from style mimicry by {Text-to-Image} models. In *USENIX-S*, 2023. 1

[52] Jing Shi, Wei Xiong, Zhe Lin, and Hyun Joon Jung. Instantbooth: Personalized text-to-image generation without test-time finetuning. In *CVPR*, 2024. 2

[53] Kihyuk Sohn, Huiwen Chang, José Lezama, Luisa Polania, Han Zhang, Yuan Hao, Irfan Essa, and Lu Jiang. Visual prompt tuning for generative transfer learning. In *CVPR*, 2023. 2

[54] Gowthami Somepalli, Anubhav Gupta, Kamal Gupta, Shramay Palta, Micah Goldblum, Jonas Geiping, Abhinav Shrivastava, and Tom Goldstein. Measuring style similarity in diffusion models. *arXiv preprint*, 2024. 2, 4, 5

[55] Sebastian Szyller, Buse Gul Atli, Samuel Marchal, and N Asokan. Dawn: Dynamic adversarial watermarking of neural networks. In *ACM-MM*, 2021. 2

[56] Wei Ren Tan, Chee Seng Chan, Hernan Aguirre, and Kiyoshi Tanaka. Improved ArtGAN for conditional synthesis of natural image and artwork. *IEEE Transactions on Image Processing*, 2019. 5

[57] Long Tang, Dengpan Ye, Yunna Lv, Chuanxi Chen, and Yunming Zhang. Once and for all: Universal transferable adversarial perturbation against deep hashing-based facial image retrieval. In *AAAI*, 2024. 2

[58] Li Tang, Qingqing Ye, Haibo Hu, Qiao Xue, Yaxin Xiao, and Jin Li. Deepmark: A scalable and robust framework for deepfake video detection. *ACM Transactions on Privacy and Security*, 2024. 2

[59] Yuan-Yu Tsai and Hong-Lin Liu. Integrating coordinate transformation and random sampling into high-capacity reversible data hiding in encrypted polygonal models. *IEEE Transactions on Dependable and Secure Computing*, 2022. 2

[60] Thanh Van Le, Hao Phung, Thuan Hoang Nguyen, Quan Dao, Ngoc N Tran, and Anh Tran. Anti-DreamBooth: Protecting users from personalized text-to-image synthesis. In *ICCV*, 2023. 1

[61] Feifei Wang, Zhentao Tan, Tianyi Wei, Yue Wu, and Qidong Huang. SimAC: A simple anti-customization method for protecting face privacy against text-to-image synthesis of diffusion models. In *CVPR*, 2024. 2

[62] Jiangfeng Wang, Hanzhou Wu, Xinpeng Zhang, and Yuwei Yao. Watermarking in deep neural networks via error back-propagation. *Electronic Imaging*, 2020. 2

[63] Run Wang, Felix Juefei-Xu, Meng Luo, Yang Liu, and Lina Wang. FakeTagger: Robust safeguards against deepfake dissemination via provenance tracking. In *ACM MM*, 2021. 2

[64] Sheng-Yu Wang, Alexei A Efros, Jun-Yan Zhu, and Richard Zhang. Evaluating data attribution for text-to-image models. In *ICCV*, 2023. 1, 2, 5, 6

[65] Sheng-Yu Wang, Aaron Hertzmann, Alexei A Efros, Jun-Yan Zhu, and Richard Zhang. Data attribution for text-to-image models by unlearning synthesized images. *arXiv preprint*, 2024. 1

[66] Xiaoshuai Wu, Xin Liao, and Bo Ou. SepMark: Deep separable watermarking for unified source tracing and deepfake detection. *arXiv preprint*, 2023. 2

[67] Xiaoshuai Wu, Xin Liao, Bo Ou, Yuling Liu, and Zheng Qin. Are watermarks bugs for deepfake detectors? rethinking proactive forensics. *arXiv preprint*, 2024. 2

[68] Zihao Xiao, Xianfeng Gao, Chilin Fu, Yinpeng Dong, Wei Gao, Xiaolu Zhang, Jun Zhou, and Jun Zhu. Improving transferability of adversarial patches on face recognition with generative models. In *CVPR*, 2021. 2

[69] Yang Yang, Wen Wang, Liang Peng, Chaotian Song, Yao Chen, Hengjia Li, Xiaolong Yang, Qinglin Lu, Deng Cai, Boxi Wu, et al. LoRA-Composer: Leveraging low-rank adaptation for multi-concept customization in training-free diffusion models. *arXiv preprint*, 2024. 2

[70] Chin-Yuan Yeh, Hsi-Wen Chen, Shang-Lun Tsai, and Sheng-De Wang. Disrupting image-translation-based deepfake algorithms with adversarial attacks. In *WACVW*, 2020. 2

[71] Innfarn Yoo, Huiwen Chang, Xiyang Luo, Ondrej Stava, Ce Liu, Peyman Milanfar, and Feng Yang. Deep 3D-to-2D watermarking: Embedding messages in 3D meshes and extracting them from 2D renderings. In *CVPR*, 2022. 2

[72] Ziyang Yuan, Mingdeng Cao, Xintao Wang, Zhongang Qi, Chun Yuan, and Ying Shan. CustomNet: Zero-shot object customization with variable-viewpoints in text-to-image diffusion models. *arXiv preprint*, 2023. 2

[73] Jie Zhang, Dongdong Chen, Jing Liao, Weiming Zhang, Huamin Feng, Gang Hua, and Nenghai Yu. Deep model intellectual property protection via deep watermarking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 2

[74] Yimeng Zhang, Xin Chen, Jinghan Jia, Sijia Liu, and Ke Ding. Text-visual prompting for efficient 2d temporal video grounding. In *CVPR*, 2023. 2

[75] Yuxin Zhang, Nisha Huang, Fan Tang, Haibin Huang, Chongyang Ma, Weiming Dong, and Changsheng Xu. Inversion-based style transfer with diffusion models. In *CVPR*, 2023. 2

[76] Yushu Zhang, Jiahao Zhu, Mingfu Xue, Xinpeng Zhang, and Xiaochun Cao. Adaptive 3d mesh steganography based on feature-preserving distortion. *IEEE Transactions on Visualization and Computer Graphics*, 2023. 2

[77] Xuandong Zhao, Yu-Xiang Wang, and Lei Li. Protecting language generation models via invisible watermarking. *arXiv preprint*, 2023. 2

[78] Xuandong Zhao, Kexun Zhang, Zihao Su, Saastha Vasan, Ilya Grishchenko, Christopher Kruegel, Giovanni Vigna, Yu-Xiang Wang, and Lei Li. Invisible image watermarks are provably removable using generative AI. In *NeurIPS*, 2024. 8

[79] Yunqing Zhao, Tianyu Pang, Chao Du, Xiao Yang, Ngai-Man Cheung, and Min Lin. A recipe for watermarking diffusion models. *arXiv preprint*, 2023. 2

[80] Zhengyue Zhao, Jinhao Duan, Kaidi Xu, Chenan Wang, Rui Zhang, Zidong Du, Qi Guo, and Xing Hu. Can protective perturbation safeguard personal data from being exploited by stable diffusion? In *CVPR*, 2024. 1

[81] Yaoyao Zhong and Weihong Deng. Towards transferable adversarial attack against deep face recognition. *IEEE Transactions on Information Forensics and Security*, 2020. 2