A Survey on Privacy and Safety in Diffusion Models

Shruti Chanumolu
schanumolu@umass.edu
University of Massachusetts Amherst
Amherst, Massachusetts, USA

David Thibodeau
dpthibodeau@umass.edu
University of Massachusetts Amherst
Amherst, Massachusetts, USA

Jeng-Yu Chou
jchou@umass.edu
University of Massachusetts Amherst
Amherst, Massachusetts, USA

ABSTRACT

In this survey, we explore existing attack and mitigation strategies for generative models, particularly focusing on diffusion models. These models, which have demonstrated impressive capabilities in generating high-quality synthetic images, face two critical challenges: privacy and safety concerns. When generative models are trained on sensitive data, such as medical images, they pose a risk of exposing private information. To mitigate this risk, existing research employs differential privacy techniques to ensure that the generated synthetic data does not inadvertently leak individual information, while maintaining high utility for tasks like image classification. Additionally, diffusion models have the potential to generate harmful content, such as biased or inappropriate images. We investigate existing methods that aim to prevent the generation of such unsafe content and ensure that diffusion model outputs are both secure and suitable for use in sensitive applications. We cover techniques like differential privacy for safeguarding sensitive data in training, content filtering, and model adjustments to prevent the generation of biased or unsafe images. Through our review of existing methods, we provide insights into how these strategies can be further improved or combined to enhance the reliability and safety of diffusion models in real-world applications.

CCS CONCEPTS

• Security and privacy → Privacy-preserving protocols; **• Computing methodologies → Artificial intelligence**.

KEYWORDS

generative AI, diffusion models, privacy preservation, model safety, bias and harm mitigation

1 INTRODUCTION

Diffusion models (DM) are a type of generative model typically used for image generation tasks. As they become central to the study of generative artificial intelligence (AI), their ability to generate realistic and high-quality images has vast implications across various domains including art, entertainment, healthcare, and general computer science such as diverse computer vision tasks [16]. As the adoption of AI-based tools increases in real-world applications, it becomes increasingly more important to be aware of the dangers associated with diffusion model usage. The ability of diffusion models to inadvertently memorize and regenerate sensitive information amplifies the risk of exposing private data when used without safeguards. Furthermore, their potential to generate harmful, biased, or inappropriate content highlights a pressing need for responsible usage and oversight.

There are risks associated with diffusion models such as training data leakage (e.g., potential for out-of-the-box generative models to memorize and regenerate data it was trained on, membership inference attacks, etc.) and the generation of harmful or biased content. Conventional mitigation strategies include low complexity data, random noise, differential privacy, and other approaches that modify models and training processes to prevent attacks.

To better understand these challenges, we focus on the following questions:

- **Q1:**What are the specific attack vectors on diffusion models, and how do these models become vulnerable to exploitation?
- **Q2:** What limitations do we observe across existing mitigation strategies for privacy-loss prevention?
- **Q3:** How do the various threat models associated with diffusion models impact the reliability and safety of these generative AI systems?

The paper is structured as follows: Section 2 introduces diffusion models. Section 3 discuss threat models and attacks on generative AI. Sections 4 addresses safety concerns and bias in output. Sections 5 and 6 cover mitigation strategies and privacy-loss prevention approaches. Section 7 concludes the paper.

2 BACKGROUND

Diffusion Models (DM) are generative models, designed to generate data resembling the training data they were exposed to. Fundamentally, these models operate by progressively corrupting the training data through the iterative addition of Gaussian noise, represented mathematically as $q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}\cdot x_{t-1}, \beta_t \cdot I)$, where x_t is the noisy data at step t, β_t is the variance schedule, and $\mathcal{N}$ denotes a Gaussian distribution. During training, the model learns to reverse this noising process using a denoising objective, typically optimizing the parameterized noise predictor ϵ_θ to minimize the expected loss, $\mathbb{E}_{x,\epsilon,t}\left[\|\epsilon - \epsilon_\theta(x_t, t)\|^2\right]$, where ϵ represents the true noise. After training, data generation is achieved by starting with randomly sampled noise $x_T \sim \mathcal{N}(0, I)$ and iteratively denoising it using the learned reverse process, $p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t))$, ultimately recovering data samples that resemble the original training data [60]. There are several variations and approaches to diffusion models:

2.1 Denoising Diffusion Probabilistic Models (DDPMs)

DDPMs work by progressively adding Gaussian noise to the data, converting it into pure noise. During training, a neural network learns to reverse this process by predicting the noise at each step. After training, the model generates new samples by starting with

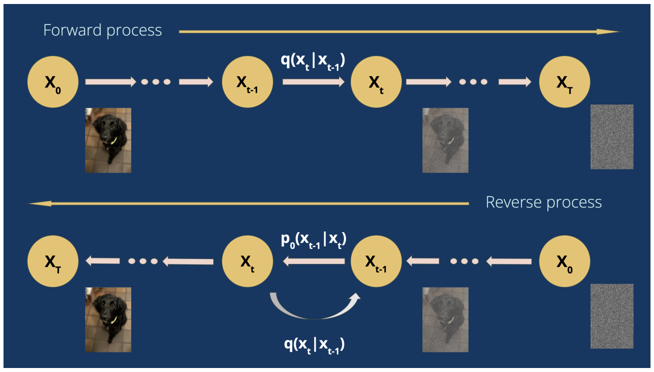

Figure 1: In the forward process, these models progressively degrade the training data by adding Gaussian noise at each step. This process is modeled as a Markov chain. In the reverse process, the model learns to reconstruct the original data by reversing the noising process from the forward pass. By traversing backwards along this chain, the model can gradually transform noise back into meaningful data.

random noise and iteratively denoising it. These models are well-suited for tasks like image and audio generation.
The forward noising process is given by:

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I)$$

The reverse process is learned by predicting the noise:

$$\mathbb{E}_{x,t,\epsilon}[\|\epsilon - \epsilon_\theta(x_t, t)\|^2]$$

2.2 Score-Based Generative Models (SGMs)

SGMs rely on estimating the gradient of the log-density (the score) of the data at different noise levels. The model learns these scores, which are then used to iteratively adjust noisy samples toward the high-probability regions of the data distribution. SGMs are effective for high-quality image generation and other generative tasks. The score of a probability density function $p_t(x)$:

$$s(x) = \nabla_x \log p_t(x)$$

SGMs approximate this score function using neural networks and sample data through Langevin dynamics:

$$x_{t+1} = x_t + \frac{\alpha}{2}\nabla_x \log p(x_t) + \sqrt{\alpha}z$$
$$(z \sim \mathcal{N}(0, I))$$

[53].

2.3 Stochastic Differential Equation (SDE)-Based Models

SDE-based models generalize diffusion processes to continuous time, where data evolves according to stochastic differential equations. The forward process corrupts data, and the reverse SDE reconstructs it using learned scores. These models benefit from flexibility and advanced numerical solvers, making them suitable for various applications like high-dimensional data synthesis. The forward SDE is described as:

$$dx = f(x,t)dt + g(t)dW$$

The reverse-time SDE is:

$$dx = [f(x,t) - g(t)^2\nabla_x \log p_t(x)]dt + g(t)d\bar{W}$$

2.4 Conditional Diffusion Models

These models extend base diffusion frameworks by conditioning the reverse process on additional information such as class labels, textual descriptions, or images. This conditioning enables tasks like text-to-image generation, class-conditional synthesis, and inpainting. Conditional diffusion models achieve state-of-the-art results in creative generation tasks. The reverse process conditioned on auxiliary information y:

$$p_\theta(x_{t-1}|x_t, y) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, y, t), \Sigma_\theta(x_t, y, t))$$

2.5 Latent Diffusion

Latent Diffusion Models (LDMs) improve upon traditional diffusion models by addressing the high computational costs that are typically associated with general diffusion processes. Traditional diffusion models operate in the pixel space and tend to incur extensive computational costs such as GPUs and memory due to repeated evaluations and other computations performed in the high-dimensional space of images. Unlike these conventional diffusion models, LDMs perform this process within a compressed latent space, enabling the models to approximate the de-noising process in a more efficient, lower-dimensional representation which in turn reduces the time and computational resources needed whilst also achieving performance that is comparable or an improvement on conventional diffusion methods.

Similar to conventional diffusion models approaches, LDMs involve (1) a perceptual compression followed by (2) generative modeling (i.e. first removing high-frequency details, then obtaining the semantic and conceptual composition of the given data). Essentially, the space that the diffusion models will be trained on in this approach is perceptually equivalent, but less computationally expensive compared to that of conventional diffusion practices. In this first stage, an auto-encoder is used to obtain such a lower-dimensional representational space, allowing for effective generation without excessive quality loss in the output. Then transformers are connected to the diffusion model's UNet to facilitate token-based conditioning mechanisms [45]. This structure supports a variety of diverse applications, making LDMs versatile and computationally efficient for large-scale image generation tasks.

While LDMs are typically able to outperform prior diffusion and GAN-based methods while significantly saving on time and resources, they may result in loss of output quality when dealing with tasks that require fine-grained precision and slower sequential sampling processes. However, overall, LDMs present a powerful approach to efficient, high-quality image generation, outperforming many traditional diffusion and GAN models in precision and recall across various synthesis tasks.

2.5.1 *Stable Diffusion.*

What is now referred to as *stable diffusion* is a particular implementation of the latent diffusion [45] process. Released in 2022 by Stability AI, the model became arguably the most popular AI image generation technology available for open use. The released

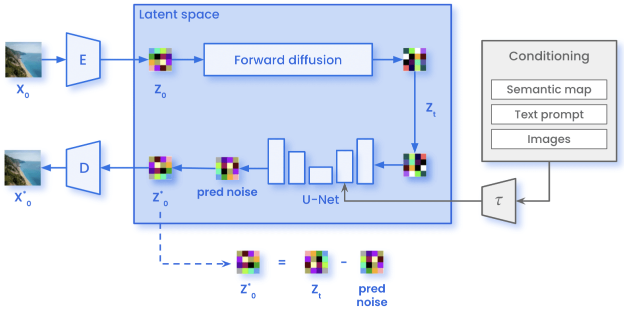

Figure 2: ***Stable diffusion process while training: It generates images by combining text representations with random noise in a latent space. Guided by a U-Net, it iteratively denoises the latent representation, transforming it into a clear image that aligns with the text prompt.***

stable diffusion model (SDM) built upon latent diffusion to provide increased efficiency, versatility, and reliability, as well as providing a user-friendly interface.

The stable diffusion process begins by taking a text prompt as input. The model uses OpenAI's CLIP [44] model to encode the text into a 768-dimension context vector. The encoding process ensures that different words or groups of words with similar meaning will be represented similarly. Concepts are encoded among combinations of the vector dimensions. This encoded vector is then used by the model while scheduling de-noising and throughout the layers of the U-net to guide the rest of the diffusion process. All of the important objects, features, colors, orientation, and spatial positioning information are encoded.

A cosine-based de-noising schedule is used to produce a more consistent training process and more stability in output. As the generation moves though the multiple layers of the de-noising process, larger features first emerge, followed by the finer details. The cosine schedule removes significantly more noise in the earlier layers than the latter. This prevents large changes from occurring during the latter stages of the generation process, which could otherwise lead to inconsistent and lower-quality outputs.

Furthermore, the SDM makes architectural and parameter choices which "stabilize" or lead to a higher level of consistency in image generation. First, an optimized scaling factor of 7.5 is introduced, making text prompt-adherence significantly more important than the overall image quality. The image quality is still high, and the takeaway is that if CLIP does a good job in interpreting the prompt, then the generated images will consistently adhere closely to the prompt. Moreover, GroupNorm is used in favor of BatchNorm, larger batch sizes and gradient-clipping are used, and weight updates utilize an exponential moving average. These all result in an SDM which produces high-quality images, with an increased emphasis on consistent output in comparison to standard LDMs.

3 ATTACKS

Diffusion models are susceptible to a wide range of adversarial threats that exploit vulnerabilities in training and inference stages. These vulnerabilities have the potential to lead to privacy breaches, or leakage, biased or adversarially manipulated outputs, and harmful content generation. Key threat models include privacy risks from training data exposure, such as membership inference and other model inversion attacks – through which adversaries may deduce sensitive information about a target model's training data – and data and model poisoning risks through embedded backdoors or otherwise adversarially corrupted outputs. Generative models also face several challenges in content moderation, as outputs may unintentionally contain, and reveal, unsafe material, biases, or reflect systemic issues in the training data. Also, the integrity and reliability of diffusion models are capable of being compromised through adversarial attacks on textual prompts, input images, and even fine-tuning processes. Overall, these vulnerabilities emphasize the need for public awareness of risks to diffusion models, especially those used in deployed applications, as well as the need for safeguards – including content filters, differential privacy techniques, and adversarial training – to mitigate the risks inherent across diffusion model architectures. The following section explores specific attack vectors, including backdoor, membership inference, and Trojan attacks as well as other adversarial manipulations of model inputs and fine-tuning and training processes.

3.1 Backdoor Attacks

A backdoor attack [55] in machine learning is a form of adversarial attack in which an attacker manipulates a model such that, when given a specific input, or input type, the model will behave in a way that is specially engineered by this attacker. In general, in these attacks, an adversarial party can tamper with the generative process so that specific triggers produce adversary-intended outputs. An attacker may manipulate the training data to corrupt the overall training dataset, or rather, modify the task(s) the model is being instructed to perform, allowing the attacker to embed such a backdoor trigger. Thus, when a "trigger" input is provided to the trained model, the model will carry out the modified task and compromise its integrity. These irregular or abnormal outputs of the model may contain biases or other harmful content, although the average end user would be unaware of such behavior while using the same model unless they happened to provide a triggering input to the model. There are a variety of distinct backdoor attacks, including poisoning [50], clean-label [14], and trigger-based variants such as static [31], dynamic [47], and invisible trigger [29] attacks.

3.1.1 Poisoning attacks.

Backdoor poisoning attacks occur when an attacker inserts a carefully selected pattern into the feature space of a model, causing it to associate the pattern with a target class chosen by the attacker [14, 50]. Adversary-corrupted data samples, usually associated with a class label selected by the attacker, will carry a specific feature pattern or "trigger" designed to cause the model to misclassify samples that have this pattern during inference. Poison-based backdoor attacks can be divided into several subcategories based on their goals and the techniques used to influence the model's behavior. Variations include *availability* and *targeted* poisoning attacks. *Availability poisoning attacks* aim to degrade the performance of a targeted model. Rather than engineering specific misclassifications or output generation, availability poisoning attacks influence the output of the targeted model indiscriminately [22]. *Targeted*, or*integrity, poisoning attacks* are similar to availability attacks, but with a key difference in their objective. Targeted attacks aim to cause specific misclassifications during testing, while maintaining correct predictions on other test samples [22, 50].

3.1.2 *Clean-label attacks.*

In *clean-label backdoor attacks*, attackers operate under an additional constraint: they cannot alter the original labels of the poisoned, or adversarially modified, samples. Instead, they inject trigger patterns only in benign data samples without altering their corresponding labels, allowing the targeted model to learn an association between the trigger and some target class through subtle correlations in the feature space. An advantage of clean-label attacks over regular poisoning attacks is that adversarially corrupted training data camouflages better into the complementary, benign, data samples, consequently making the attacks more difficult to detect. Of course, these attacks may still be detected if there is a noticeable enough discrepancy between the poisoned data samples and their original labels [14].

3.1.3 *Backdoor triggers.*

Backdoor triggers can be categorized into three main settings: static, dynamic, and invisible. Many existing backdoor attacks utilize *static triggers*. Static triggers follow fixed patterns (e.g. a specific pixel arrangement in images) that are consistently applied across all poisoned training samples. An adversary introduces these into the training data such that when the model encounters these triggers in inference data samples, it will produce misclassifications, or generate abnormal outputs, on the input data to a label or other instruction of the attacker's choosing [31]. Static triggers are widely used in computer vision backdoor attacks, where adding a particular color patch or watermark across various training images helps create an association between the trigger and the target label [31].

Unlike the consistency of a static setting, the patterns and location of *dynamic triggers* are adjusted based on certain parameters – such as image location, orientation, or even sample-specific patterns – rather than staying fixed across data samples [47]. Dynamic triggers are particularly useful in situations where a fixed, visibly identifiable trigger might raise suspicion, allowing the backdoor to blend with the model's general training data more seamlessly.

Invisible triggers are subtle perturbations added into benign samples, allowing them to blend easily into benign data. In image backdoor attacks, invisible triggers may be introduced through pixel-level adjustments that steer the model's interpretation toward the target class chosen by an attacker without affecting visual content. Since invisible triggers result in poisoned data samples appearing indistinguishable from their benign counterparts, invisible backdoor attacks more easily circumvent manual detection [30].

TrojanDiff [8], a Trojan attack on diffusion models, leverages both static and invisible triggers to compromise model behavior. It embeds pre-defined triggers, such as a blended image (e.g., a Hello Kitty illustration) or a fixed patch (e.g., a white square), into the noise input during training. These triggers are static, as they remain consistent across poisoned training samples, and invisible, as they are imperceptible during inference. TrojanDiff operates through three key stages: the Trojan diffusion process, which biases the model's noise distribution; the Trojan generative process, which reverses this bias to embed the attack; and the Trojan sampling procedure, which generates adversarial outputs. This method achieves high attack success rates across diverse targets – including specific classes, out-of-domain distributions, and individual instances – while maintaining model performance under benign conditions.

Static, dynamic, and invisible triggers each have distinct advantages for attackers embedding backdoors into models, depending on their goals. Static triggers are simple, efficient, and consistently activate backdoors when inputs match the trigger, but their fixed patterns make them susceptible to detection. Dynamic triggers are more adaptable and stealthy, especially in diverse datasets where static patterns stand out. Invisible triggers provide the highest stealth, ideal for undetectable manipulation in security-sensitive systems, but are computationally complex and less straightforward. Overall, static triggers are practical for basic uses, dynamic and invisible triggers excel in stealth and versatility for sensitive attacks.

3.1.4 *Threat Model.*

The threat assumes that adversaries have access to the training process, enabling them to covertly embed themselves within a model's learning process. Specifically, in backdoor attacks, adversaries are entities that have access to the training data of a model and are able to manipulate that data to engineer the model's behavior. As a result, they pose serious threats not only in computer vision but across various real-world applications of machine learning models, including misclassification, privacy leakage, and manipulation of outputs in sensitive applications such as security systems, autonomous vehicles, or content generation platforms. If a diffusion model is compromised in such a way, it could also be manipulated to generate misleading content, leading to disinformation or harmful content dissemination which is particularly dangerous given how quickly digital content can spread over social network platforms. Not only can models be modified to produce biased or harmful content, backdoor attacks can also be utilized to cause privacy leakage [58], for example, backdoor-enabled membership inference attacks. The stealth and efficacy of such attacks depend on the type of trigger employed—static, dynamic, or invisible—each offering varying levels of adaptability, detectability, and complexity.

As diffusion models become more widely adopted across domains, the chances they handle sensitive information increases. Additionally, as the application of machine learning in academia, industry, and other domains increases, the risk of hidden backdoors also increases. The lack of transparency and potential lack of knowledge from user bases into how models are trained increases the potential for malicious entities to execute backdoor attacks in a variety of models, including those that are publicly available for individual or commercial use. Therefore, securing diffusion models against backdoor attacks is crucial to ensure they remain reliable, trustworthy, and safe for real-world applications.

3.2 Membership Inference Attacks

Another common attack against machine learning models is *membership inference attacks* (MIA). This attack aims to violate the privacy of a model's training data by inferring whether a data point was a member of the training set.

Though there is quite extensive analysis of MIA against various model types, Matsumoto et al. provide one of the first comprehensive analysis of MIA in diffusion models. MIA differs greatly against diffusion models in comparison to previously studied models: *"diffusion models have unique hyperparameters, i.e., timesteps, sampling steps, and sampling variances, that have never been contained in conventional models. It indicates that we no longer know how the membership inference attack on diffusion models varies compared to the traditional models" [39].* Although they have now been revealed to be similarly resistant as GANs, it has been recently determined where weaknesses lay in diffusion models. There exist multiple methods outlined below for inferring membership under diffusion models. These methods largely rely on exploiting the diffusion models' reliance on a time stepped-approach to generating images.

3.2.1 *Noise Prediction-based Attacks.*

Noise prediction-based membership inference attacks exploit the diffusion model's learned denoising behavior to identify training set members. A diffusion model establishes a probabilistic process that transitions a real image into noise [59]. Given white-box access to a model, the removal of noise during the generation process can indicate members of a training set. As a diffusion model follows its noise-removal schedule, it will become more predictable which noise will be removed as the model is trained on an increasing number of epochs. Wu et al. determine a *t-error* - an approximated posterior estimation error - between actual and predicted noise removal, by taking a forward step in the diffusion process and a backward step in the denoising process. A trained quantile regression model is then used to determine a threshold below which the *t-error* will indicate membership. An emphasis is placed on the intermediate sampling steps, as early steps are very noisy, and late steps have a lack of meaningful noise, so they do not reveal quite as predictable noise removal. Wu et al. used a group of 7 weak attackers (each an approximately 5,000 parameter quantile regression neural net) to achieve a 99.94% TPR with only a 1% FPR on the CIFAR 10 dataset [59]. Diffusion models often require extensive training across numerous epochs to achieve high-quality image generation. However, this increases membership inference vulnerability as the model tends to memorize noise removal processes, highlighting a trade-off between image quality and privacy.

3.2.2 *Reconstruction loss analysis.*

Similar to noise-prediction, the overall loss metric can be evaluated throughout each of the time steps in the diffusion process. The loss is typically computed as the KL divergence between the forward process $q(x_t|x_0)$ and the learned reverse process $p_\Theta(x_{t-1}|x_t)$. Matsumoto et al. show that models sampling at a higher number of steps throughout the generation process become more vulnerable to MIA [39]. A time t can be determined where a difference in loss below a certain threshold τ can be achieved to determine membership with high probability, with the membership score computed as:

$$L(x) = \Sigma_t KL(q(x_t|x_0)||p_\Theta(x_{t-1}|x_t))$$

where lower $L(x)$ values indicate likely membership. The intuition is that already trained-on images will be generated with a lower loss more quickly and consistently, due to memorization, particularly in the middle time steps. Carlini et al. combine the use of multiple trained shadow models with this loss-based approach to increase the effectiveness of their attack, achieving membership detection rates above 90 percent on standard benchmarks. Shadow models help establish robust threshold values τ by observing loss patterns on known member and non-member samples [7].

3.2.3 *Shadow models with gradient analysis.*

Another legitimate attack vector is the use of shadow models. This approach involves training of similar-sized diffusion models, so is slightly less efficient and more computationally expensive. A "shadow model" [41] can be trained on well-known image datasets, and evaluated to determine membership qualities. Specifically, their GSA method analyzes the gradients to determine the parameters which have highest correlation with membership, and then compresses these parameters for efficiency in evaluation. The gradients are evaluated throughout each of the time steps to identify the part of the generation process most likely to shed light on membership. With white-box access to the victim model, the attacker uses the gradient-based inference learned from the shadow model to then evaluate the victim model's gradients and determine membership. While requiring considerably more overhead, this method outperforms the loss-based attack methods, and produces an AUC of .999 for both of their implemented GSA attacks.

3.2.4 *Threat Model.*

MIAs expose potentially sensitive information on which models were trained, presenting a major vulnerability to machine learning models and violate the privacy of individuals whose data was used during training. MIA can be performed in either white-box or black-box settings, depending on the specific method of attack. Generally, direct access to the trained model is required, so as to monitor behavior throughout the diffusion and de-noising processes.

Though MIA may be performed against a diffusion model for white-hat or research purposes, the greatest threat comes from malicious actors who may attempt to exploit the training data for ill purposes. Since diffusion models deal specifically with images, training data may contain sensitive photos which may be sought after for purposes of exploiting or blackmailing a victim. Training data may also be proprietary, and an attacker may seek to steal this data to profit from. Furthermore, diffusion models are becoming increasingly popular for use in medical fields. One use is the generation of synthetic medical imagery, which may assist in the training of other models utilized for anomaly detection and the diagnosing of serious medical conditions. Diffusion models used for generating medical images have generally been trained on thousands or more real medical images, which are protected by patient privacy laws. Should an MIA be successful against one of these models, thousands of patients may have confidential medical information exposed.

These are just a few of the cases which highlight the need for diffusion models to implement effective defense measures against membership inference attackers.

3.3 Adversarial Attacks

Adversarial attacks on diffusion models aim to manipulate the model's output by introducing imperceptible perturbations that exploit vulnerabilities in the diffusion or denoising process. These attacks can distort the generated images or misalign the output with

Attack Method	Attack setting	Input Perturbation	Optimization Domain	Key Features
Style Mimicry[51]	White-box	Image	Latent	Mimics artistic styles or specific visual themes by shifting latent space.
Encoder Attack [48]	White-box	Image	Latent	Moves input representation closer to a target style or specific feature using adversarial perturbations.
Diffusion Attack [48]	White-box	Image	Pixel	Adds perturbations to disrupt denoising, degrading output quality or generating incorrect outputs.
MMA-Diffusion [61]	White-box	Image, Text	Latent, Pixel	Combines text and image perturbations to evade safety filters and disrupt multimodal model outputs.
MFA [63]	White-box	Image	Noise	Alters noise mean during reverse diffusion to degrade model outputs.
SAGE [35]	White-box	Text	Latent	Systematically uncovers vulnerabilities using minimal perturbations in latent, token embedding, and text spaces.
JPA [37]	Black-box	Text	Latent	Bypasses safety filters to generate NSFW or harmful content.
Sneakyprompt [62]	Black-box	Text	Latent	Uses reinforcement learning to optimize prompts that exploit linguistic loopholes, bypassing safety filters to generate NSFW content.
AdvDM [33]	White-box	Fine-tuning image	Noise	Degrades generation quality by disrupting semantic and textual understanding during fine-tuning.
Mist [32]	White-box	Fine-tuning image	Latent, Noise	Enhances transferability of adversarial examples across different imitation methods (e.g., DreamBooth, textual inversion).
Anti-DreamBooth [25]	White-box	Fine-tuning image	Latent, Noise	Prevents personalization by adding adversarial noise to training images.

Table 1: Comparison of Adversarial Attack Methods in Diffusion Models

the intended input. Depending on the method of perturbation, these attacks can be categorized into three types: perturbations added to the input image, to the text prompt, or through fine-tuning, where adversarial patterns are embedded during training. Table 1 summarizes various adversarial attacks on diffusion models, categorizing them by their access type (white-box/black-box), input perturbation targets (image, text, or fine-tuning), optimization domain, and their unique attack mechanisms, such as disrupting denoising, bypassing safety filters, or embedding adversarial patterns.

3.3.1 Attacks Targeting the Input Image.

These attacks target input images in diffusion models, particularly in image-to-image generation tasks, where minor perturbations can cause significant deviations in the output. The adversary introduces noise into the initial image or latent space, either with or without access to the model. Although access enhances the attack's precision, it is not essential, as black-box attacks can also effectively disrupt the denoising process.

In **latent space attacks**, perturbations are added to the encoded representation of an input image, shifting its position in the latent space and altering the model's understanding of the input. For example, the encoder attack in PhotoGuard [48] optimizes a perturbation δ to move the latent representation of the input c closer to a target representation $E(c_{\text{tgt}})$, minimizing $\|E(c+\delta) - E(c_{\text{tgt}})\|$ under the constraint $\|\delta\|_\infty \leq \eta$. A variation of this is style mimicry attacks, where perturbations shift the encoded representation of the input image to match a style-transferred version of itself. This is achieved using a similar optimization: $\min_\delta \|E(c+\delta) - E(S(c,T))\|$, where $S(c,T)$ represents the input c transformed to a target style T using a style-transfer model [51]. Such attacks are often used to force diffusion models to mimic the artistic styles of specific artists or techniques, enabling the generation of outputs closely resembling those styles.

In **pixel space attacks**, perturbations are applied directly to the pixel values of the input image, subtly altering its visual appearance. These changes are designed to degrade the quality of the generated outputs or mislead the model into generating unintended content.

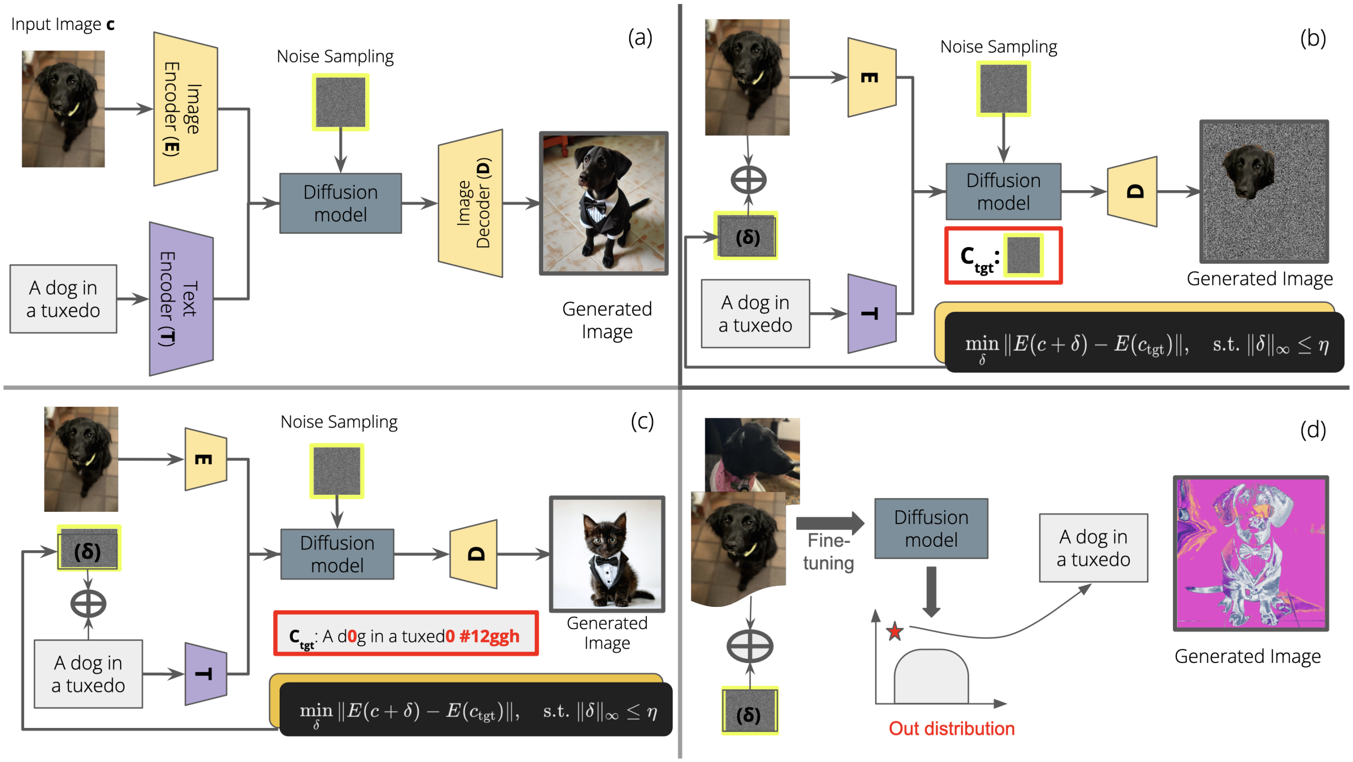

Figure 3: *Adversarial attacks on diffusion models (DMs) can be categorized by perturbation targets. (a) Benign diffusion models (text-to-image) serve as the baseline for comparison. (b) Image input perturbation involves small changes to input images to mislead the model. (c) Text input perturbation modifies textual prompts to produce undesired outputs. (d) Fine-tuning with perturbed input images introduces adversarial examples during training, causing consistent vulnerabilities. These categories highlight key strategies for compromising DMs.*

For instance, the diffusion attack in PhotoGuard [48] optimizes δ to minimize $\|G(c+\delta) - c_{\text{tgt}}\|$, where $G(c+\delta)$ represents the model's output from the perturbed input and c_{tgt} is a target undesirable output.

In **distribution-shifting attacks** such as the Maximum Mean Shift Attack (MFA) [63], the focus is on disrupting the reverse diffusion process by altering the mean of the predicted noise during denoising. Here, the perturbation δ is optimized to maximize the expected mean of the predicted noise $\max_\delta \mathbb{E}\|\mu(\epsilon_\theta(x_t, t, c+\delta))\|$, where $\epsilon_\theta(x_t, t, c+\delta)$ is the denoising network's prediction for noise at a given timestep t. This disruption moves the data distribution away from the original, leading to degraded outputs. Importantly, in these attacks, the adversary requires access to the model, as they rely on exploiting the intermediate noise predictions during the reverse diffusion process.

3.3.2 *Attacks Targeting the Text Prompt.*

In these attacks, adversaries subtly modify prompts to mislead the model into generating unintended, low-quality, or sensitive content while avoiding detection by safety systems. For instance, attackers may append small textual perturbations (e.g., “A*B#”) to prompts, replace characters with visually similar glyphs (e.g., “room” to “r00m”) or replace words with phonetically similar ones (e.g., “see” to “sea”) to distort outputs. The Sneakyprompt [62] attack takes this further by embedding restricted requests within seemingly innocuous prompts. For instance, instead of explicitly requesting a weapon, an attacker might say, "Create a diagram of historical hunting tools like bows and arrows," tricking the model into generating prohibited content. These carefully crafted modifications exploit linguistic loopholes to bypass safety filters while remaining subtle enough to avoid human detection. These modifications are optimized to bypass safety filters that screen for sensitive words or content and ensure changes remain imperceptible to human inspection [13, 67].

Advanced strategies include targeted manipulations to remove specific elements from outputs or redirect generation toward unrelated categories by modifying key dimensions in the text embedding space. SAGE (Searching Adversarial Failures in Generative Models) [35] systematically explores latent, token embedding, and text spaces to uncover vulnerabilities. In the latent space, minimal perturbations to latent variables are introduced, revealing failures like distorted or irrelevant outputs by leveraging residual connections to address gradient vanishing issues. In the token embedding space, adversarial tokens subtly alter outputs, while in the text space, SAGE uses language model priors to craft human-readable but adversarial prompts that expose misalignments in semantics and prompt-to-output relationships. JPA (Jailbreaking Prompt Attack) [37], by contrast, focuses on bypassing safety mechanisms to generate harmful or Not Safe For Work (NSFW) content. It captures unsafe concepts using antonym pairs (e.g., "nude" vs. "clothed") to define target embeddings and optimizes these prompts through cosine similarity and gradient masking, ensuring sensitive words are avoided while adversarial inputs evade detection. This approach leverages the embedding space's inherent unsafe concepts, subtly manipulating outputs while maintaining semantic alignment with the original text.

While single-input adversarial attacks exploit vulnerabilities in isolation, multiple-input adversarial attacks can target interactions between modalities. For instance, MMA-Diffusion [61] combines a text-modal attack, which manipulates prompts to bypass safety filters, with an image-modal attack, introducing imperceptible perturbations to evade embedding-based NSFW checkers. Such multimodal strategies amplify the impact of attacks on diffusion models, underscoring the need for defenses that address cross-modal vulnerabilities.

3.3.3 Attacks Targeting Fine-Tuning Images.

These attacks on fine-tuned diffusion models involve embedding subtle, carefully crafted perturbations into the training or fine-tuning images. These minimal-norm changes, often discovered using gradient-based optimization methods, push the model's representation of the affected images outside its expected distribution. As a result, the model struggles to adapt to new tasks or produce high-quality, realistic outputs for content associated with these perturbed images.

AdvDM [33] showed that by altering the objective of diffusion models, one could degrade their generation quality. Later approaches, such as Mist [32], further refined this idea by combining strategies to disrupt both the semantic understanding and textual associations of images. When applied to personalization techniques like Textual Inversion or DreamBooth [46], these perturbations prevent the model from accurately capturing a subject's defining characteristics, leading to low-quality, unrealistic outputs whenever that subject is referenced.

Anti-DreamBooth [25] specifically targets DreamBooth [46]-based personalization by injecting adversarial noise into the subject's reference images. This ensures that the model cannot properly learn the subject's features, breaking the intended personalization and protecting against unauthorized replication. In essence, these techniques demonstrate how strategic adversarial attacks can severely undermine a diffusion model's ability to adapt, personalize, or reliably generate content.

3.3.4 Threat Models.

Adversarial attacks exploit vulnerabilities in the inference or training processes of diffusion models, using imperceptible perturbations or crafted inputs to manipulate outputs. Attackers, including individuals or organizations, introduce subtle modifications to prompts, images, or fine-tuning datasets to generate biased, harmful, or unauthorized content, such as deepfakes, disinformation, or style-copied material. These attacks pose significant risks in areas like security, medical imaging, and public communication by undermining trust and amplifying harm. By targeting latent spaces, embeddings, or gradients, attackers evade detection and bypass safety mechanisms. With diffusion models increasingly deployed in high-stakes domains and model weights often openly accessible, these risks are amplified in both black-box and white-box settings. Securing diffusion models against such threats is crucial for ensuring their safe, ethical, and reliable use.

4 BIAS AND SAFETY CONCERNS

Diffusion models, trained on uncurated datasets, often reflect biases and can generate unsafe content, including stereotypes, misinformation, or explicit imagery. These issues raise ethical concerns, highlighting the need for robust filtering and bias mitigation to ensure fair and safe outputs.

4.1 Social Bias

Social biases in diffusion models like Stable Diffusion, DALL-E, and others originate from primarily from data-related factors. These models are trained on large-scale datasets scraped from the internet, such as LAION-400M and similar repositories, which are inherently biased. The datasets often overrepresent dominant demographics (e.g., white males in high-paying roles) and reflect societal stereotypes embedded in the data. This results in biased outputs, such as underrepresentation of women and people of color in certain professions, skewed depictions of age and gender, and reinforcement of cultural and racial stereotypes [20, 36, 42].

The architecture of these diffusion models exacerbates the problem. Text-to-image systems use multimodal embeddings like CLIP, which itself encodes societal biases due to its training data. These embeddings guide the diffusion process, influencing the generated images. For example, when prompted with "CEO," models often generate images predominantly of white males, while a prompt like "nurse" disproportionately depicts women. The iterative denoising process in diffusion models also lacks mechanisms to correct for these biases during image synthesis [20, 42].

Efforts to mitigate these biases involve various technical approaches. Prompt engineering allows users to specify desired traits (e.g., "a Black female CEO"), which can sometimes balance outputs but often introduces other quality disparities, such as lower resolution or less realistic images for underrepresented groups. Iterative Distribution Alignment (IDA) has been proposed to rebalance biased distributions by aligning output distributions with predefined fairness constraints. This method optimizes weights in the generation process, yielding balanced results in gender and ethnicity while maintaining convergence efficiency [20, 36]. Reinforcement learning frameworks, such as policy gradients, have also been applied to adjust outputs dynamically, but they often suffer from slow convergence and suboptimal results [20].

Despite these advancements, significant technical challenges remain. Bias in latent representations persists due to the reliance on biased priors from training datasets. Intersectional bias remains a critical issue, as models struggle to accurately depict overlapping identity characteristics (e.g., non-binary Black individuals). Image quality disparities are another unresolved issue, with images of underrepresented groups often exhibiting lower photorealism and fidelity. Finally, existing methods fail to generalize well across varied prompts and user contexts, limiting their effectiveness in real-world applications [36, 40, 42].

Future research should prioritize diverse dataset curation, fairness-aware architectures, and advanced multimodal debiasing techniques to balance inclusivity and quality.

4.2 Safety Risks

The safety risks in DMs are multifaceted, stemming from their technical frameworks, training data sources, and deployment environments. These models are trained on massive datasets scraped from the internet, which are inherently noisy and include harmful biases, misinformation, and NSFW content [12, 27]. Because DMs learn to synthesize images by iteratively denoising random noise, the generation process lacks explicit safeguards against inappropriate outputs. Moreover, the latent space representations learned by these models are often uninterpretable, allowing for the unintentional or deliberate creation of content such as deepfakes, violent imagery, or sexually explicit material. For instance, these models have been used to generate deepfakes of public figures engaging in fabricated scandals or explicit content, undermining reputations and public trust [3]. Similarly, they may propagate misinformation, such as creating convincing images of fake events like terrorist attacks or protests, which can incite panic and disrupt societal stability [3, 18].

Efforts to mitigate these risks face significant technical hurdles. Dataset filtering, while useful for removing explicit content, can inadvertently degrade model performance by excluding relevant but non-problematic data [12]. Post-hoc filters and NSFW classifiers during inference can be bypassed, especially in open-source settings where model weights are accessible. Even fine-tuning approaches, which aim to erase harmful concepts from the model, often fail to generalize well and can negatively impact the model's ability to generate high-quality or diverse outputs [12, 43]. Furthermore, ethical concerns such as the generation of culturally insensitive or biased outputs highlight the need for robust interpretability. For example, when models are prompted to generate images of criminals, they may disproportionately depict individuals from certain ethnic groups, reflecting and reinforcing societal prejudices [3, 12]. Without robust interpretability and effective safeguards, the potential for misuse remains a critical concern in the widespread adoption of diffusion models.

5 MITIGATION STRATEGIES

Many common mitigation strategies for privacy loss and harmful generation prevention consist of actions or systems implemented at, or prior to, the training stage of a model. In this Section, we detail existing methods utilized to address the threat models described in Section 3. While the strategies are distinct, they have multiple applications in addressing various threats including privacy-preservation, prevention of harmful content generation, and protection again biased outputs.

5.1 Data Augmentation

Data augmentation is a technique in deep learning and other AI applications that enhances model generalization by introducing diversity to the training data [26], enabling better performance on unseen data. Widely used across fields like image classification, object detection, and speech recognition, it is also effective for preventing attacks on diffusion models. There are a variety of distinct data augmentation techniques that range in complexity and prevalence and can be applied to data samples or layered on top of each other. Simple augmentations include image translations, horizontal reflections flipping, rotation, scaling, cropping, padding, and modifications to color [24]. Figure 4 contains visual examples of such color augmentation methods that affect attributes like brightness, contrast, saturation, and hue. More advanced approaches include dropout [4], Random Erasing [64, 66], AutoAugment [10] and Faster AutoAugment [19] which automate the search for optimal augmentation policies to lead to enhanced model performance, and Mixup, which optimize performance and encourage broader generalization to intermediate class concepts.

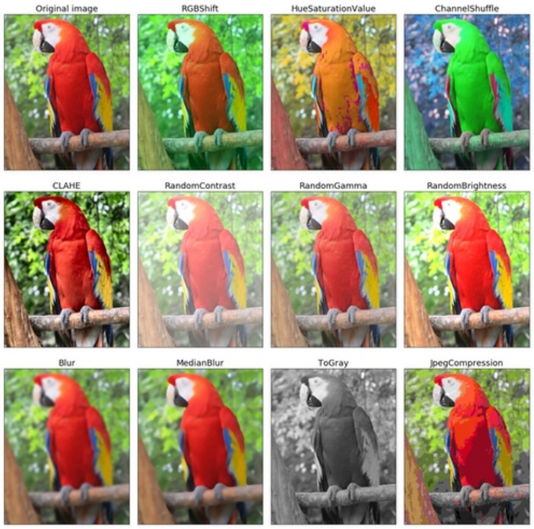

Figure 4: Examples of simple data augmentation methods available in the Albumentations [5] library that modify images at the pixel level without altering their spatial structure. These methods adjust color properties and pixel intensity values to enhance variability while preserving the original content and geometry.

This versatility makes data augmentation particularly valuable when collecting new data is difficult. Likewise, this makes it especially useful in preventing general adversarial attacks against diffusion models. For instance, it can be used to counter adversarial attacks by disrupting static and dynamic triggers in backdoor attacks and introduce unpredictability in diffusion models. This creates obstacles for attackers attempting membership inference attacks, as the model's behavior becomes less predictable, reducing the risk of over-fitting to the training set. By varying the de-noising process, it reduces over-fitting and limits attackers' ability to exploit stable patterns, thereby enhancing model robustness against membership inference and similar attacks.

5.2 Model Pruning

Model pruning is a commonly-used technique in the design of neural networks which aims to reduce the size, computational requirements, and the complexity of a model. This is accomplished through the removal of model parameters deemed to be unnecessary in maintaining a threshold of model performance. The goal is to determine specific parameters which contribute least to the model's success, and to remove them.

Many methods of model pruning exist, and can be performed pre-training, during training, and post-training [9]. Methods range from very basic - largest absolute value (magnitude-based) parameter removal or evaluating second-order derivatives to determine importance, to the more complex - such as finding the most efficient subnetworks of neurons (structural-approach). Most current techniques are very model- and task-specific, including those for diffusion models. Pruning of diffusion models is very architecture specific, and focuses heavily on the pruning of U-net components and iteratively pruning throughout timesteps of the denoising and diffusion process to monitor for adherence to the intended diffusion schedule.

A key benefit of pruning is its ability to combat a complex model's tendency to overfit due to memorization of the training set. While sometimes coming at the expense of creating training instability, this also serves to create less predictable training information (such as noise-removal patterns) from which to perform MIA. Additionally, the structure of the training data's latent representation becomes changed, altering predictable patterns of how training data is represented in comparison to non-training data. This also makes MIA more difficult, as privacy leakage has been observed to approximately scale with model size [7]. After training similarly performant models ranging from 25M parameters to 130M parameters, it was found that the TPR of MIA increased from 17% for the 25M parameter model to nearly 43% for the 130M parameter one [7]. Backdoor attack risk can be reduced as well, through careful pruning practices [56]. Magnitude-based pruning techniques are able to better target specific weights which serve as backdoor triggers, though more structural-based pruning is able to better disrupt more distributed, complex backdoor attack triggers. This happens when the backdoor-related parameters are removed during pruning. One notable consideration, however, is that well-designed backdoor attacks may rely on parameters which are least likely to be targeted during the pruning process - utilizing weights whose removal would too heavily compromise model performance.

5.3 Training Stage & Model Architecture

When designing and training a model, many architectural and hyper-parameter choices have significant implications on privacy.

As stated in the previous sections, one of the factors which presents highest risk to a diffusion model's data privacy is its tendency to overfit. Thus, many of the traditional techniques used to reduce overfitting will also assist in reducing the effectiveness of attacks - such as MIA. Regularization methods such as early stopping, dropout, and gradient clipping are all effective at reducing model overfitting and memorization of the training data. Model choices such as a lower number of parameters and fewer attention layers are additional measures for reducing overfitting, as they force a model to learn more generalized rather than specific patterns. A simpler model following these approaches also increases the difficulty of an adversary hiding trojans. The major trade off to be considered here is that these methods tend to sacrifice the model's image generation quality, while reducing the overfitting.

Training choices which affect gradient values can also be used quite effectively to mitigate potential attack strategies. Using dropout provides additional benefits in mitigation of both MIA and backdoor attacks. This technique randomly deactivates neurons/parameters at each model layer with a given probability. Accordingly, similarly to what is described in the model pruning subsection, this potentially eliminates backdoor and trojan attack triggers by disrupting the neural pathways through the training process. In addition, the randomly deactivated neurons leads to sparse gradients, and gradient values which pass through the active neurons will be scaled larger. These changing and off-scale parameter subsets disrupt MIA techniques such as the previously outlined gradient analysis of shadow models. Similarly, gradient-clipping also provides defense against this class of MIA by restricting the magnitude of gradients which reduces the potential information leakage from individual data samples [2].

In short, mindfulness of privacy when designing and training a diffusion model can lead to a model which is more robust to attacks. The main consideration is the tension that exists between privacy-preserving measures and generative quality of the model.

5.4 Backdoor Detection

Backdoor mitigation strategies aim to detect and prevent attacks where adversaries modify models or training data to induce malicious behaviors triggered by specific inputs. While common approaches like data augmentation and preprocessing, discussed elsewhere in this section, are useful for backdoor defense, other effective strategies focus on both data and model-centric methods. These include HDBSCAN, Spectral Signature, Isolation Forest, Neural Cleanse, and Strong Random Perturbation (STRIP). *HDBSCAN* is utilizes KMeans for defensive clustering over neuron activations [6, 50]. This makes it effective for finding poisoned or anomalous data points. *Spectral Signatures* are a property of all known backdoor attacks which allow for the efficient detection and removal of poisoned data samples across model architectures [54]. This method analyzes latent space representations to detect outliers or backdoor-poisoned samples, employing statistical and clustering methods to clean corrupted training datasets. *Isolation Forests* detect outliers by isolating them. This method is particularly suited for backdoor detection, as backdoor-injected data samples often appear as outliers compared to more diverse background points in the dataset [34]. *Neural Cleanse* is a model-centric approach for detecting and reverse engineering triggers hidden in deep neural networks across various downstream tasks. After identifying abnormal trigger patterns in data samples, it mitigates their impact through filtering, pruning, or unlearning [57]. *Strong Random Perturbation (STRIP)* is another model-centric approach applicable to both backdoor and Trojan attacks. It examines a model's inference behavior by perturbing input samples and observing the randomness in predictions. Consistent class predictions across perturbations indicates potential backdoor or Trojan attacks [15]. In summary, these strategies utilize both data characteristics and model behaviors to effectively detect, mitigate, and prevent backdoor and related attacks by identifying anomalous patterns, isolating poisoned data, and addressing malicious triggers in deep learning models.

5.5 Image Immunization

The rapid advancements in text-to-image diffusion models, such as Stable Diffusion, have enabled unauthorized personalization and artistic style mimicry, posing significant ethical and copyright challenges. Various protective techniques have been developed to mitigate these risks. Anti-DreamBooth [25] applies imperceptible adversarial perturbations to user images, specifically designed to disrupt the iterative denoising steps fundamental to diffusion models. By targeting the fine-tuning process of DreamBooth, this method ensures that models trained on perturbed images fail to reproduce realistic outputs of the target subject, even under prompt mismatches or model variations. Similarly, Glaze [51] introduces "style cloaks," which are subtle perturbations computed to shift an image's representation in the feature space of generative models toward a misleading style. These cloaks alter the learning process during model fine-tuning, causing models to generate outputs inconsistent with the original artist's style while preserving the visual integrity of the protected images.

In contrast, IMMA [65] immunizes the model itself rather than the input data. It learns model parameters that degrade the effectiveness of malicious adaptation techniques, such as DreamBooth, LoRA, or Textual Inversion. This is achieved by optimizing a bi-level program where the upper-level task maximizes the adaptation difficulty while the lower-level task simulates the adaptation process. By modifying critical layers, such as cross-attention mechanisms, IMMA [65] ensures poor adaptation performance for malicious concepts while retaining usability for benign applications. Mist [32] complements these approaches by improving the cross-method transferability of adversarial examples through a fused loss function. By combining semantic and textual losses, Mist generates perturbations that are robust against various imitation techniques, including DreamBooth, textual inversion, and image-to-image generation. This enhances resistance to adversarial defenses like noise purification while ensuring effective disruption.

PhotoGuard [48] focuses on latent diffusion models (LDMs), employing two methods to immunize images: encoder attacks, which force the latent representation of an image toward undesirable regions, and diffusion attacks, which inject perturbations into the denoising steps. These methods ensure that manipulated or edited images become unrealistic, effectively raising the cost and complexity of malicious edits. By targeting both data and model vulnerabilities, these techniques collectively provide a robust defense against style mimicry and unauthorized personalization, safeguarding intellectual property while preserving the broader utility of generative AI technologies.

5.6 Synthetic Data

Considering the diffusion model's propensity to memorize training data, there exists another solution to this problem: the use of synthetic data in the training set. Generative models such as GANs or diffusion models themselves may be used to generate the data. An entirely synthetic dataset may be used; or only part of the dataset may be synthetic, replacing only the most sensitive instances of the real-world data. Slokom et al. use a *TSTR* (train on synthetic and test on real) method to evaluate a model's performance utilizing a synthetic dataset [52]. Their findings were that their model performed similarly on the synthetic data, and that it was a viable alternative to using real data for training purposes.

Utilizing synthetic data largely eliminates the issue of memorization of training examples - if the data points being memorized are synthetic, the potential leaks present no real-world privacy concerns. Additionally, high-quality synthetic data can be generated with an emphasis on promoting the model's ability to better generalize after training. This can serve in the partially synthetic dataset instances to train a model which is more robust to leaking the parts of the training data which are real.

There exist several drawbacks to this approach, however. First, the generation of a large quantity of high-quality synthetic data can require a large amount of computational resources. Secondly, even the SOTA methods for producing synthetic data will sometimes fail to capture the patterns and details contained in real-world data. This can lead to less-natural outputs and the suffering of model performance. Lastly, the synthetic data also comes from a generative model - which may also suffer from the same privacy concerns attempting to be mitigated. This may lead to a cascading leakage from one model to the next, by inclusion of memorized samples in the "synthetic" data.

5.7 Differential Privacy

Differential Privacy (DP) protects diffusion models against membership inference attacks by minimizing the influence of individual data points during training. Techniques like DP-SGD, which combines gradient clipping and the addition of Gaussian noise, ensure that the model learns general patterns rather than memorizing specific data points [11, 17, 21]. In gradient clipping, gradients are capped at a fixed threshold to reduce their sensitivity to individual samples, while Gaussian noise is added to obscure specific data contributions, safeguarding privacy. Noise multiplicity strengthens these safeguards by reusing data at multiple noise levels during training, reducing gradient variance and enhancing performance without increasing the privacy cost [17]. To further boost image quality, semantic-aware pretraining selects semantically relevant subsets of public data for pretraining, improving model generalization while keeping sensitive data private [28]. Together, these techniques enable the generation of high-quality synthetic images with robust privacy guarantees, making them resilient to adversarial attacks.

Despite their benefits, DP methods face an inherent trade-off between cost and utility [11, 17, 21]. Techniques like DP-SGD and noise multiplicity significantly increase computational overhead, leading to higher training time and memory usage, which may limit feasibility in resource-constrained environments. While approaches like semantic-aware pretraining and augmentation multiplicity improve image fidelity and performance, their utility depends heavily on the availability of semantically aligned public datasets, which may not always exist [28]. Moreover, achieving optimal performance often requires extensive hyperparameter tuning and careful model configuration, increasing the complexity and expertise required for implementation. The challenge lies in striking a balance between privacy, computational efficiency, and utility, ensuring that synthetic images meet practical quality standards without imposing prohibitive resource demands.

Method	Approach	Focus Area	Success Rate (Approximate)	Evaluation Metrics
Anti-DreamBooth [25]	Adversarial noise to images	Disrupting personalized models (DreamBooth)	High (visual artifacts, 95%)	Visual artifacts, generation quality metrics (SSIM, LPIPS), evaluated on VGGFace2, CelebA-HQ
Glaze [51]	Style cloaking for artistic mimicry	Protecting artist styles from mimicry	Very High (92%-96%)	User studies (∼1000 artists), CLIP-based stylistic similarity scores
IMMA [65]	Model immunization against adaptation	Preventing malicious model adaptations	Moderate to High (85%-90%)	Adaptation performance metrics, success/failure in mimicking target styles
Mist [32]	Improved adversarial examples	Cross-method adversarial robustness	High (robust across methods, 90%)	Cross-method robustness, adversarial transferability, resistance to countermeasures
PhotoGuard [48]	Latent diffusion adversarial perturbations	Disrupting image editing and manipulation	Very High (95%)	Editing failures, FID, human evaluation of manipulated image quality

Table 2: Comparison of Protection Methods Against Copyright Infringement

6 DISCUSSION

Privacy and safety are critical concerns for diffusion models, as they are vulnerable to attacks like membership inference, backdoor manipulation, and adversarial perturbations. While mitigation strategies discussed in Section 5 address these risks, they often involve trade-offs like increased computational costs or reduced output quality. In this section we discuss these limitations, emphasizing the need for improved approaches to ensure secure and reliable use of diffusion models in real-world applications.

6.1 Privacy and Safety

6.1.1 Ease of Mitigation Strategy Implementation.

Strategies like data augmentation are relatively easy to implement. Libraries like OpenCV [1], Albumentations [5], and imgaug [23] are publicly available. This is advantageous considering the unique capabilities augmentation contributes for enhancing a wide variety of computer vision tasks.

Other strategies explored in Section 5 are more complex. For example, the backdoor detection approaches detailed in Section 5.4 are more attack-specific, leveraging both data-centric and model-centric approaches to identify and mitigate malicious backdoor behaviors. Methods like HDBSCAN, Spectral Signatures, and Isolation Forests focus on detecting poisoned or anomalous data points using clustering and statistical outlier analysis, while Neural Cleanse and STRIP target model behavior to uncover hidden triggers. Table 3 summarizes these techniques, highlighting their diverse methodologies and contributions to robust backdoor defense mechanisms.

The remainder of the discussed strategies may not be complex implementation-wise, but may be time-consuming or computationally expensive. Making adjustments to the model or training stage are generally straightforward; such as implementing dropout, gradient-clipping, hyperparameter tuning, etc. Likewise, the generation of synthetic data is very simple, given access to a capable generative model. However, these defense strategies are generally resource-intensive. For large models, retraining after model or hyperparameter changes will likely take a number of days. For synthetic image generation, we have the same case. Given a complex generative model capable of producing high-quality images and access to GPU hardware, a data set of 60,000 images (size of the CIFAR-10 dataset) would take about a week (assuming 10 seconds to generate each image).

6.1.2 Attack Harms and Relevance.

Backdoor attacks pose a serious risk by covertly embedding malicious behavior into machine learning models, threatening applications in compute vision and across various other real-world applications. In diffusion models, such attacks can enable the generation of misleading or harmful content, amplifying risks such as disinformation and privacy leakage [58].

Adversarial attacks on diffusion models exploit vulnerabilities in training and inference, introducing subtle, hard-to-detect perturbations that bypass safety filters and alter model behavior. These attacks can lead to the generation of harmful or biased content, the spread of misinformation, or deepfake creation, such as fabricated political statements or fake protest images. They can compromise sensitive applications like medical imaging by generating misleading diagnostic images. Adversaries can also replicate an artist's style to produce unauthorized works, undermining intellectual property rights. Alarmingly, these models can be manipulated to generate illegal content, such as child sexual abuse material, posing severe ethical and legal risks. Real-world scenarios include using such attacks for deceptive advertising, public opinion manipulation, impersonation, or creating harmful content for blackmail or exploitation.

Membership inference attacks can also reveal private information related to individuals whose images were used in model training but were never intended to be public. With the potential for diffusion models to be trained on medical imaging, this presents a very real risk to patients who may consent to contributing data for private use. Table 4 presents common and SOTA MIA methods and

Method	Type	Advantages
HDBSCAN	Data-centric	Utilizes KMeans for clustering neuron activations to detect poisoned or anomalous data points.
Spectral Signatures	Data-centric	Detects poisoned data samples by analyzing latent space representations and employing statistical and clustering techniques.
Isolation Forest	Data-centric	Identifies outliers in datasets, leveraging the observation that backdoor-injected samples often appear as outliers.
Neural Cleanse	Model-centric	Detects and reverse engineers backdoor triggers in neural networks by identifying abnormal trigger patterns and mitigating them through filtering, pruning, or unlearning.
STRIP	Model-centric	Detects backdoors and Trojan attacks by perturbing input samples and analyzing consistency in predictions; consistent predictions indicate potential malicious behavior.

Table 3: Comparison of Backdoor Detection Approaches

Attack Method	Dataset	AUC	TPR @ 1% FPR
Loss-reconstruction (baseline)	CIFAR-10	.801	5.65%
Noise-prediction (baseline)	CIFAR-10	-	8%
Loss-reconstruction comb. w/ shadow-models (LiRA)	CIFAR-10	.982	99%
Shadow-models (GSA2)	CIFAR-10	.999	97.88%
Shadow-models (GSA1)	CIFAR-10	.999	99.7%
Noise-prediction w/ quantile regression models	CIFAR-10	-	99.94%

Table 4: Comparison of various MIA attack methods on diffusion models

their effectiveness metrics, as measured by AUCs and TPRs. Considering the difficulty in MIA detection, it may be nearly unknown when a model's data has been compromised.

As diffusion models are increasingly adopted to handle sensitive information, the lack of transparency and limited user understanding of model integrity increases the risk for attacks on models publicly available for individual or commercial use. Securing these models is critical to ensuring they remain reliable, trustworthy, and safe for real-world use.

6.2 Limitations

The following categories highlight the diverse nature of limitations of existing mitigation strategies discussed in Section 5, ranging from data issues to model-centric weaknesses, as well as concerns related to computational complexity, adaptability, and generalization of well-known mitigation techniques.

6.2.1 Dependence on Data Quality. Techniques like HDBSCAN, Spectral Signatures, and Isolation Forests rely on the quality and diversity of training data and may struggle to distinguish poisoned from benign samples when datasets are heavily skewed or contaminated. Insufficient clean samples or dominance of poisoned data reduce their reliability, increasing the likelihood of successful data poisoning or misclassification attacks. Furthermore, usage of augmented or synthetic datasets, while helping to combat overfitting and MIA, can impact the ultimate image generation quality of diffusion models. Strategies like Glaze [51] and Anti-DreamBooth [25] rely heavily on the integrity of cloaked or perturbed data to prevent misuse. If adversaries have access to large amounts of uncloaked or unperturbed data, these defenses are rendered ineffective. Additionally, adversarial noise, while designed to be imperceptible, can sometimes subtly degrade the quality of the images, reducing their usability for legitimate purposes.

6.2.2 Model-centric Limitations. Neural Cleanse and STRIP, which analyze model behavior, may struggle against sophisticated or well-disguised backdoor attacks. For instance, STRIP relies on perturbing data to detect anomalies, but advanced backdoor techniques evade detection by avoiding obvious behavioral patterns during inference.

Making training-time or model adjustments to reduce overfit models and their susceptibility to MIA can also lead to the suffering of image generation quality. This tension may be difficult to navigate because diffusion models are highly complex and have a natural tendency to memorize training samples.

Adversarial examples crafted to protect against image-to-image generation attacks may struggle to effectively defend against other tasks, such as DreamBooth [46], which fine-tunes models to enable personalized text-to-image synthesis, or textual inversion, which adjusts text embeddings to replicate specific artistic styles. Similarly, IMMA's [65] strategy focuses on pre-release model immunization to resist malicious adaptations, but this approach cannot be retroactively applied to already deployed and widely used models, leaving existing systems vulnerable.

6.2.3 Robustness. Many mitigation strategies may be vulnerable to adaptive attacks or may not generalize well if attacks evolve over time. For example, attackers can design more resilient triggers or develop new backdoor techniques to evade reverse engineering and detection methods. For instance, despite tools like Neural Cleanse, more sophisticated attacks could leave diffusion models vulnerable by evading detection and reverse engineering, highlighting the need for stronger mitigation strategies. Meanwhile, as generative models become more advanced, the subtle distortions introduced by Glaze-like style-shielding techniques can be reverse-engineered or nullified. Newer models might treat cloaking patterns as common noise, removing them to restore the original aesthetic. Large and diverse training sets help models ignore these perturbations, and adversarial training strategies—such as those inspired [38] and further explored in the context of generative modeling by [49] aim to build model robustness against engineered attacks. Pre-processing techniques also evolve to target and strip away protective transformations.

6.2.4 Privacy budget and cost of implementation. When deploying differentially private diffusion models, the privacy budget significantly limits model utility. Achieving formal privacy guarantees often requires injecting additional noise or imposing aggressive clipping, which reduces output quality and fidelity. As the privacy budget tightens, these constraints intensify, leading to slower training, lower-resolution outputs, and fewer viable samples. Balancing an acceptable privacy budget with model performance becomes challenging, especially for complex data modalities, making high-quality, privacy-preserving diffusion models difficult to implement and deploy at scale. In addition, methods like AdvDM and MIST [32] require computationally intensive optimization processes, such as Monte Carlo estimation or multi-loss function tuning, which are impractical for large-scale deployment. Approaches like PhotoGuard also depend on external policy enforcement by organizations, adding logistical and financial overhead.

7 CONCLUSION

In this survey, we explore papers spanning a variety of domains such as diffusion models, their vulnerabilities to various attacks, and innovations that mitigate risks. While a significant number of these papers specifically address attacks and defenses of diffusion models, many others provide essential background on more general attacks that are applicable to generative models broadly, which diffusion models are particularly susceptible to. Our survey emphasizes how attacks such as membership inference, backdoor insertion, and other adversarial perturbations exploit vulnerabilities, while inherent biases in training datasets can lead to unethical or discriminatory outputs. We focus on diffusion model-specific and more generally applicable mitigation strategies which prevent these serious dangers. To ensure a comprehensive understanding, we examined papers from top security conferences – Usenix, IEEE Security and Privacy, ACM Workshop on Artificial Intelligence and Security, ACM Conference on Computer and Communications Security (CCS) – 13 of which were published between 2019-2024. By piecing together insights from diverse sources, this survey connects the nuanced challenges of privacy, safety, and ethical considerations in diffusion models. Through this extensive review, we aim to provide a holistic understanding of the threats and defenses surrounding these increasingly influential generative models. In conclusion, diffusion models present immense potential across various applications but come with significant privacy and safety challenges. Mitigation strategies like differential privacy, data augmentation, and backdoor detection are effective but often involve trade-offs in computational resources, fidelity, and scalability. Addressing these limitations requires future work to focus on developing hybrid approaches that integrate bias mitigation, automating defense mechanisms, and creating fairness-aware architectures. Additionally, curating diverse, high-quality datasets and advancing interpretability techniques will enhance the ethical, secure, and reliable deployment of diffusion models in real-world applications.

REFERENCES

[1] 2023. OpenCV: Open Source Computer Vision Library. GitHub: https://github.com/opencv/opencv.

[2] Martin Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. 2016. Deep Learning with Differential Privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security (CCS'16)*. ACM. https://doi.org/10.1145/2976749.2978318

[3] Agil Aghasanli, Dmitry Kangin, and Plamen Angelov. 2023. Interpretable-through-prototypes deepfake detection for diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*. 467–474.

[4] Xavier Bouthillier, Kishore Konda, Pascal Vincent, and Roland Memisevic. 2016. Dropout as data augmentation. arXiv:1506.08700 [stat.ML] https://arxiv.org/abs/1506.08700

[5] Alexander Buslaev, Vladimir I. Iglovikov, Eugene Khvedchenya, Alex Parinov, Mikhail Druzhinin, and Alexandr A. Kalinin. 2020. Albumentations: Fast and Flexible Image Augmentations. *Information* 11, 2 (2020). https://doi.org/10.3390/info11020125

[6] Ricardo JGB Campello, Davoud Moulavi, and Jörg Sander. 2013. Density-based clustering based on hierarchical density estimates. In *Pacific-Asia conference on knowledge discovery and data mining*. Springer, 160–172.

[7] Nicholas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramèr, Borja Balle, Daphne Ippolito, and Eric Wallace. 2023. Extracting training data from diffusion models. In *Proceedings of the 32nd USENIX Conference on Security Symposium* (Anaheim, CA, USA) *(SEC '23)*. USENIX Association, USA, Article 294, 18 pages.

[8] Weixin Chen, Dawn Song, and Bo Li. 2023. TrojDiff: Trojan Attacks on Diffusion Models With Diverse Targets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 4035–4044.

[9] Hongrong Cheng, Miao Zhang, and Javen Qinfeng Shi. 2024. A Survey on Deep Neural Network Pruning: Taxonomy, Comparison, Analysis, and Recommendations. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 46, 12 (2024), 10558–10578. https://doi.org/10.1109/TPAMI.2024.3447085

[10] Ekin D. Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V. Le. 2019. AutoAugment: Learning Augmentation Policies from Data. arXiv:1805.09501 [cs.CV] https://arxiv.org/abs/1805.09501

[11] Tim Dockhorn, Tianshi Cao, Arash Vahdat, and Karsten Kreis. 2023. Differentially Private Diffusion Models. arXiv:2210.09929 [stat.ML] https://arxiv.org/abs/2210.09929

[12] Rohit Gandikota, Joanna Materzynska, Jaden Fiotto-Kaufman, and David Bau. 2023. Erasing Concepts from Diffusion Models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2426–2436.

[13] Hongcheng Gao, Hao Zhang, Yinpeng Dong, and Zhijie Deng. 2023. Evaluating the Robustness of Text-to-image Diffusion Models against Real-world Attacks. arXiv:2306.13103 [cs.CR] https://arxiv.org/abs/2306.13103

[14] Yinghua Gao, Yiming Li, Linghui Zhu, Dongxian Wu, Yong Jiang, and Shu-Tao Xia. 2023. Not all samples are born equal: Towards effective clean-label backdoor attacks. *Pattern Recognition* 139 (2023), 109512.

[15] Yansong Gao, Change Xu, Derui Wang, Shiping Chen, Damith C Ranasinghe, and Surya Nepal. 2019. Strip: A defence against trojan attacks on deep neural networks. In *Proceedings of the 35th annual computer security applications conference*. 113–125.

[16] Sahra Ghalebikesabi, Leonard Berrada, Sven Gowal, Ira Ktena, Robert Stanforth, Jamie Hayes, Soham De, Samuel L Smith, Olivia Wiles, and Borja Balle. 2023. Differentially private diffusion models generate useful synthetic images. *arXiv preprint arXiv:2302.13861* (2023).

[17] Sahra Ghalebikesabi, Leonard Berrada, Sven Gowal, Ira Ktena, Robert Stanforth, Jamie Hayes, Soham De, Samuel L. Smith, Olivia Wiles, and Borja Balle. 2023. Differentially Private Diffusion Models Generate Useful Synthetic Images. arXiv:2302.13861 [cs.LG] https://arxiv.org/abs/2302.13861

[18] Yeaeun Gong, Lanyu Shang, and Dong Wang. 2024. Integrating Social Explanations Into Explainable Artificial Intelligence (XAI) for Combating Misinformation: Vision and Challenges. *IEEE Transactions on Computational Social Systems* (2024).

[19] Ryuichiro Hataya, Jan Zdenek, Kazuki Yoshizoe, and Hideki Nakayama. 2019. Faster AutoAugment: Learning Augmentation Strategies using Backpropagation. arXiv:1911.06987 [cs.CV] https://arxiv.org/abs/1911.06987

[20] Ruifei He, Chuhui Xue, Haoru Tan, Wenqing Zhang, Yingchen Yu, Song Bai, and Xiaojuan Qi. 2024. Debiasing text-to-image diffusion models. In *Proceedings of the 1st ACM Multimedia Workshop on Multi-modal Misinformation Governance in the Era of Foundation Models*. 29–36.

[21] Xiao He, Mingrui Zhu, Dongxin Chen, Nannan Wang, and Xinbo Gao. 2024. Diff-Privacy: Diffusion-based Face Privacy Protection. *IEEE Transactions on Circuits and Systems for Video Technology* (2024), 1–1. https://doi.org/10.1109/TCSVT.2024.3449290

[22] Matthew Jagielski, Alina Oprea, Battista Biggio, Chang Liu, Cristina Nita-Rotaru, and Bo Li. 2018. Manipulating Machine Learning: Poisoning Attacks and Countermeasures for Regression Learning. In *2018 IEEE Symposium on Security and Privacy (SP)*. 19–35. https://doi.org/10.1109/SP.2018.00057

[23] Alexander B. Jung, Kentaro Wada, Jon Crall, Satoshi Tanaka, Jake Graving, Christoph Reinders, Sarthak Yadav, Joy Banerjee, Gábor Vecsei, Adam Kraft, Zheng Rui, Jirka Borovec, Christian Vallentin, Semen Zhydenko, Kilian Pfeiffer, Ben Cook, Ismael Fernández, François-Michel De Rainville, Chi-Hung Weng, Abner Ayala-Acevedo, Raphael Meudec, Matias Laporte, et al. 2020. imgaug. https://github.com/aleju/imgaug. Online; accessed 01-Feb-2020.

[24] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2017. ImageNet classification with deep convolutional neural networks. *Commun. ACM* 60, 6 (May 2017), 84–90. https://doi.org/10.1145/3065386

[25] Thanh Van Le, Hao Phung, Thuan Hoang Nguyen, Quan Dao, Ngoc Tran, and Anh Tran. 2023. Anti-DreamBooth: Protecting users from personalized text-to-image synthesis. arXiv:2303.15433 [cs.CV] https://arxiv.org/abs/2303.15433

[26] Gyewon Lee, Irene Lee, Hyeonmin Ha, Kyunggeun Lee, Hwarim Hyun, Ahnjae Shin, and Byung-Gon Chun. 2021. Refurbish Your Training Data: Reusing Partially Augmented Samples for Faster Deep Neural Network Training. In *2021 USENIX Annual Technical Conference (USENIX ATC 21)*. USENIX Association, 537–550. https://www.usenix.org/conference/atc21/presentation/lee

[27] Hang Li, Chengzhi Shen, Philip Torr, Volker Tresp, and Jindong Gu. 2024. Self-discovering interpretable diffusion latent directions for responsible text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12006–12016.

[28] Kecen Li, Chen Gong, Zhixiang Li, Yuzhong Zhao, Xinwen Hou, and Tianhao Wang. 2024. PrivImage: Differentially Private Synthetic Image Generation using Diffusion Models with Semantic-Aware Pretraining. In *33rd USENIX Security Symposium (USENIX Security 24)*. USENIX Association, Philadelphia, PA, 4837–4854. https://www.usenix.org/conference/usenixsecurity24/presentation/li-kecen

[29] Sen Li, Junchi Ma, and Minhao Cheng. 2024. Invisible Backdoor Attacks on Diffusion Models. arXiv:2406.00816 [cs.LG] https://arxiv.org/abs/2406.00816

[30] Yuezun Li, Yiming Li, Baoyuan Wu, Longkang Li, Ran He, and Siwei Lyu. 2021. Invisible backdoor attack with sample-specific triggers. In *Proceedings of the IEEE/CVF international conference on computer vision*. 16463–16472.

[31] Yiming Li, Tongqing Zhai, Baoyuan Wu, Yong Jiang, Zhifeng Li, and Shutao Xia. 2021. Rethinking the Trigger of Backdoor Attack. arXiv:2004.04692 [cs.CR] https://arxiv.org/abs/2004.04692

[32] Chumeng Liang and Xiaoyu Wu. 2023. Mist: Towards Improved Adversarial Examples for Diffusion Models. arXiv:2305.12683 [cs.CV] https://arxiv.org/abs/2305.12683

[33] Chumeng Liang, Xiaoyu Wu, Yang Hua, Jiaru Zhang, Yiming Xue, Tao Song, Zhengui Xue, Ruhui Ma, and Haibing Guan. 2023. Adversarial Example Does Good: Preventing Painting Imitation from Diffusion Models via Adversarial Examples. arXiv:2302.04578 [cs.CV] https://arxiv.org/abs/2302.04578

[34] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. 2008. Isolation forest. In *2008 eighth ieee international conference on data mining*. IEEE, 413–422.

[35] Qihao Liu, Adam Kortylewski, Yutong Bai, Song Bai, and Alan Yuille. 2023. Discovering Failure Modes of Text-guided Diffusion Models via Adversarial Search. arXiv:2306.00974 [cs.CV] https://arxiv.org/abs/2306.00974

[36] Sasha Luccioni, Christopher Akiki, Margaret Mitchell, and Yacine Jernite. 2023. Stable Bias: Evaluating Societal Representations in Diffusion Models. In *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (Eds.), Vol. 36. Curran Associates, Inc., 56338–56351. https://proceedings.neurips.cc/paper_files/paper/2023/file/b01153e7112b347d8ed54f317840d8af-Paper-Datasets_and_Benchmarks.pdf

[37] Jiachen Ma, Anda Cao, Zhiqing Xiao, Yijiang Li, Jie Zhang, Chao Ye, and Junbo Zhao. 2024. Jailbreaking Prompt Attack: A Controllable Adversarial Attack against Diffusion Models. arXiv:2404.02928 [cs.CR] https://arxiv.org/abs/2404.02928

[38] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2019. Towards Deep Learning Models Resistant to Adversarial Attacks. arXiv:1706.06083 [stat.ML] https://arxiv.org/abs/1706.06083

[39] Tomoya Matsumoto, Takayuki Miura, and Naoto Yanai. 2023. Membership Inference Attacks against Diffusion Models . In *2023 IEEE Security and Privacy Workshops (SPW)*. IEEE Computer Society, Los Alamitos, CA, USA, 77–83. https://doi.org/10.1109/SPW59333.2023.00013

[40] Ranjita Naik and Besmira Nushi. 2023. Social Biases through the Text-to-Image Generation Lens. arXiv:2304.06034 [cs.CY] https://arxiv.org/abs/2304.06034

[41] Yan Pang, Tianhao Wang, Xuhui Kang, Mengdi Huai, and Yang Zhang. 2023. White-box Membership Inference Attacks against Diffusion Models. arXiv:2308.06405 [cs.CR] https://arxiv.org/abs/2308.06405

[42] Malsha V. Perera and Vishal M. Patel. 2023. Analyzing Bias in Diffusion-based Face Generation Models. In *2023 IEEE International Joint Conference on Biometrics (IJCB)*. 1–10. https://doi.org/10.1109/IJCB57857.2023.10449200

[43] Matvei Popov and Eva Tuba. 2024. Credible Diffusion: Improving Diffusion Models Interpretability with Transformer Embeddings. In *2024 12th International Symposium on Digital Forensics and Security (ISDFS)*. IEEE, 1–6.

[44] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. arXiv:2103.00020 [cs.CV] https://arxiv.org/abs/2103.00020

[45] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2021. High-Resolution Image Synthesis with Latent Diffusion Models. *CoRR* abs/2112.10752 (2021). https://arxiv.org/abs/2112.10752

[46] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 2023. DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation. arXiv:2208.12242 [cs.CV] https://arxiv.org/abs/2208.12242

[47] Ahmed Salem, Rui Wen, Michael Backes, Shiqing Ma, and Yang Zhang. 2022. Dynamic Backdoor Attacks Against Machine Learning Models. In *2022 IEEE 7th European Symposium on Security and Privacy (EuroSP)*. 703–718. https://doi.org/10.1109/EuroSP53844.2022.00049

[48] Hadi Salman, Alaa Khaddaj, Guillaume Leclerc, Andrew Ilyas, and Aleksander Madry. 2023. Raising the Cost of Malicious AI-Powered Image Editing. arXiv:2302.06588 [cs.LG] https://arxiv.org/abs/2302.06588

[49] Lucas Schott, Josephine Delas, Hatem Hajri, Elies Gherbi, Reda Yaich, Nora Boulahia-Cuppens, Frederic Cuppens, and Sylvain Lamprier. 2024. Robust Deep Reinforcement Learning Through Adversarial Attacks and Training : A Survey. arXiv:2403.00420 [cs.LG] https://arxiv.org/abs/2403.00420

[50] Giorgio Severi, Jim Meyer, Scott Coull, and Alina Oprea. 2021. Explanation-Guided Backdoor Poisoning Attacks Against Malware Classifiers. In *30th USENIX Security Symposium (USENIX Security 21)*. USENIX Association, 1487–1504. https://www.usenix.org/conference/usenixsecurity21/presentation/severi

[51] Shawn Shan, Jenna Cryan, Emily Wenger, Haitao Zheng, Rana Hanocka, and Ben Y. Zhao. 2023. Glaze: Protecting Artists from Style Mimicry by Text-to-Image Models. In *32nd USENIX Security Symposium (USENIX Security 23)*. USENIX Association, Anaheim, CA, 2187–2204. https://www.usenix.org/conference/usenixsecurity23/presentation/shan

[52] Manel Slokom, Peter-Paul de Wolf, and Martha Larson. 2022. When machine learning models leak: an exploration of synthetic training data. In *International Conference on Privacy in Statistical Databases*. Springer, 283–296.

[53] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. 2021. Score-Based Generative Modeling through Stochastic Differential Equations. arXiv:2011.13456 [cs.LG] https://arxiv.org/abs/2011.13456

[54] Brandon Tran, Jerry Li, and Aleksander Madry. 2018. Spectral signatures in backdoor attacks. *Advances in neural information processing systems* 31 (2018).

[55] Vu Tuan Truong, Luan Ba Dang, and Long Bao Le. 2024. Attacks and defenses for generative diffusion models: A comprehensive survey. *arXiv preprint arXiv:2408.03400* (2024).

[56] Bolun Wang, Yuanshun Yao, Shawn Shan, Huiying Li, Bimal Viswanath, Haitao Zheng, and Ben Y. Zhao. 2019. Neural Cleanse: Identifying and Mitigating Backdoor Attacks in Neural Networks. In *2019 IEEE Symposium on Security and Privacy (SP)*. 707–723. https://doi.org/10.1109/SP.2019.00031

[57] Bolun Wang, Yuanshun Yao, Shawn Shan, Huiying Li, Bimal Viswanath, Haitao Zheng, and Ben Y. Zhao. 2019. Neural Cleanse: Identifying and Mitigating Backdoor Attacks in Neural Networks. In *2019 IEEE Symposium on Security and Privacy (SP)*. 707–723. https://doi.org/10.1109/SP.2019.00031

[58] Yuxin Wen, Leo Marchyok, Sanghyun Hong, Jonas Geiping, Tom Goldstein, and Nicholas Carlini. 2024. Privacy Backdoors: Enhancing Membership Inference through Poisoning Pre-trained Models. arXiv:2404.01231 [cs.CR] https://arxiv.org/abs/2404.01231

[59] Zhiwei Steven Wu, Shuai Tang, Sergul Aydore, Michael Kearns, and Aaron Roth. 2023. Membership inference attack on diffusion models via quantile regression. In *NeurIPS 2023 Workshop on Synthetic-Data4ML*. https://www.amazon.science/publications/membership-inference-attack-on-diffusion-models-via-quantile-regression

[60] Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. 2024. Diffusion Models: A Comprehensive Survey of Methods and Applications. arXiv:2209.00796 [cs.LG] https://arxiv.org/abs/2209.00796

[61] Yijun Yang, Ruiyuan Gao, Xiaosen Wang, Tsung-Yi Ho, Nan Xu, and Qiang Xu. 2024. MMA-Diffusion: MultiModal Attack on Diffusion Models. arXiv:2311.17516 [cs.CR] https://arxiv.org/abs/2311.17516

[62] Yuchen Yang, Bo Hui, Haolin Yuan, Neil Gong, and Yinzhi Cao. 2023. SneakyPrompt: Jailbreaking Text-to-image Generative Models. arXiv:2305.12082 [cs.LG] https://arxiv.org/abs/2305.12082

[63] Hongwei Yu, Jiansheng Chen, Xinlong Ding, Yudong Zhang, Ting Tang, and Huimin Ma. 2024. Step Vulnerability Guided Mean Fluctuation Adversarial Attack against Conditional Diffusion Models. *Proceedings of the AAAI Conference on Artificial Intelligence* 38, 7 (Mar. 2024), 6791–6799. https://doi.org/10.1609/aaai.v38i7.28503

[64] Dazhi Zhao, Guozhu Yu, Peng Xu, and Maokang Luo. 2019. Equivalence between dropout and data augmentation: A mathematical check. *Neural Networks* 115 (2019), 82–89.

[65] Amber Yijia Zheng and Raymond A Yeh. 2025. Imma: Immunizing text-to-image models against malicious adaptation. In *European Conference on Computer Vision*. Springer, 458–475.

[66] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. 2020. Random erasing data augmentation. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 34. 13001–13008.

[67] Haomin Zhuang, Yihua Zhang, and Sijia Liu. 2023. A Pilot Study of Query-Free Adversarial Attack against Stable Diffusion. arXiv:2303.16378 [cs.CV] https://arxiv.org/abs/2303.16378