A Survey of Machine Unlearning in Generative AI Models: Methods, Applications, Security, and Challenges

An Huang, Zhipeng Cai, *Fellow, IEEE*, Zuobin Xiong, *Member, IEEE*

***Abstract*—Generative AI has flourished over the past decade, with generative models advancing in both the industrial and academic sectors. Given various applications, some scenarios have seen the misuse of generative AI, particularly in the integration with the Internet of Things (IoT). IoT devices often handle personal and sensitive data, raising serious concerns about privacy leakage and security breaches when generating data. As a promising countermeasure, machine unlearning has emerged to solve the problems posed by these generative models by effectively removing specific concepts or sensitive information from trained models. In this survey, anchored in generative models, machine unlearning approaches are reviewed, categorized, and discussed comprehensively and systematically. Existing unlearning approaches are classified into gradient-based techniques, task vectors, knowledge distillation, data sharding, and reliable unlearning methods. Apart from previous works, this survey extends the review of attack methods that aim to exploit the vulnerability in generative models and assess the robustness of these unlearning methods. In addition, popular metrics and datasets in machine unlearning research are summarized and evaluated based on effectiveness, efficiency, and security. Finally, we shed light on the future directions of this emerging research topic by discussing applications, highlighting challenges, and exploring research frontiers for the current machine unlearning community and the new investigators to come.**

***Index Terms*—Machine Unlearning, Generative AI, the Internet of Things, Security & Privacy.**

I. Introduction

The rise of Artificial Intelligence Generated Content (AIGC) has substantially transformed the landscape of digital content creation [1]–[3]. Generative models have become one of the most dynamic and rapidly advancing research areas in AIGC, thanks to significant advances in deep learning models over recent years. From generating photorealistic images to creating a meta-universe, diverse applications are mainly derived from the three prominent generative model families: Generative Adversarial Networks (GANs) [3]–[6], AutoEncoder (AE) [7]–[11], and Diffusion Models (DMs) [12]–[15]. Although AIGC offers new opportunities, these generative models pose significant challenges to generative AI in terms of data privacy, security, and ethical accountability when misused. For example, by taking crafted prompts, these models may generate content that infringes on copyright or contains private information, including sensitive intelligence and secrets [16]–[19].

An Huang and Zuobin Xiong are with the Department of Computer Science, University of Nevada Las Vegas, Las Vegas, NV, 89154 USA (email: huanga7@unlv.nevada.edu; zuobin.xiong@unlv.edu).

Zhipeng Cai is with the Department of Computer Science, Georgia State University, Atlanta, GA, 30303 USA (email: zcai@gsu.edu).

Meanwhile, the Internet of Things (IoT) received considerable benefits from generative AI along with emerging challenges [20]–[23]. The enhanced capability of generative models enables various IoT devices, including mobile phones [24], autonomous vehicles [25], robots [26], and applications in the metaverse [27]. However, IoT devices are prime targets for attacks as they handle large amounts of sensitive information from terminal users. Malicious attacks can easily compromise these IoT devices and generative models, revealing sensitive or private information on which the models are trained.

Although traditional approaches to content erasing involve retraining from scratch, these processes are computationally expensive and often impractical for large-scale generative models, especially on resource-constrained devices. As a result, the concept of *machine unlearning* has recently gained attention. It addresses these problems by selectively eliminating specific information or concepts from an existing model (e.g., generative AI models) without compromising its general performance. Machine unlearning [28]–[31], inspired by the legal and ethical imperatives of the "**right to be forgotten**", has become a critical technique in machine learning research. Many countries/regions have promulgated relevant laws and provisions to restrict the use of data in models, such as the European Union's General Data Protection Regulation (GDPR) [32], the California Consumer Privacy Act (CCPA) [33], and the Act on the Protection of Personal Information (APPI) [34]. Machine unlearning provides an effective solution by selectively removing specific data or concepts. This capability helps align models with privacy laws and reduces potential biases in generated outputs. Due to the complex output and high-dimensional latent spaces, machine unlearning presents unique challenges and opportunities in the generative models. Successful unlearning within these models requires accurately pinpointing the concepts to be removed while carefully preserving the model's general performance and stability. For instance, when removing the concept "car" from a model, it is crucial to eliminate car-specific characteristics from the generated images, while retaining unrelated capabilities, such as accurately generating images of "trains."

This survey comprehensively explores machine unlearning techniques applied to generative models, focusing on image generation domains. Specifically, to the best of our knowledge, this survey is the first to focus on the security of machine unlearning methods, as unlearned models may still be vulnerable to various adversarial attacks as techniques emerge.

A. Summary of Contributions

Table I exhibits the contributions and highlights the differences between our survey and some prior art. Compared

with [35]–[41], our paper conducts a comprehensive survey on machine unlearning, which underlines the combination of generative models and machine unlearning, with a focus on the security implications of these methods. The main contributions of our paper are as follows:

- We provide a detailed taxonomy of existing methods for machine unlearning in generative models, covering a broad spectrum of techniques.
- We explicitly focus on the security aspects of machine unlearning and thoroughly review potential vulnerabilities and attacks against unlearning methods.
- Mainstream metrics and datasets for evaluating unlearning methods are collected and summarized in terms of effectiveness, efficiency, generation quality, and security.
- We emphasize practical applications and highlight current challenges and potential future research directions of machine unlearning in this field.

TABLE I: Comparison with existing surveys with topics: T1 - Generative Models, T2 - Machine Unlearning, T3 - Security.

Survey	Covered Topics			Key Content
	T1	T2	T3	
[35]	✓	-	-	Image generation models
[36]	-	✓	-	Machine unlearning methods
[37]	-	-	✓	Security of deep learning
[38]	✓	✓	-	Machine unlearning in generative AI
[39]	-	✓	✓	Security of machine unlearning
[40]	✓	-	✓	Security of diffusion models
[41]	✓	-	✓	Security of multimodal generative models
Ours	✓	✓	✓	Machine unlearning in generative models and the security of these methods

B. Systematic Literature Review Method

We employ a systematic literature review to survey machine unlearning in generative models. The method [42] introduces an appropriate protocol design approach to help computer science and engineering researchers perform rigorous reviews of current empirical evidence. The primary search process for this review protocol, which informs our survey, is presented as follows:

The review protocol consists of several steps, including selecting keywords, identifying relevant databases, defining inclusion criteria, and extracting review data. (i) Our search concentrates on the keywords "machine unlearning" and "generative models". Thus, we define the advanced search rule as "(machine unlearning) AND (generative models OR diffusion models OR AEs OR GANs)". (ii) We employ this rule to search for relevant papers in IEEE Xplore, the ACM Digital Library, arXiv of Computer Science, and Google Scholar, limiting the publication years from 2020 to 2025. (iii) After filtering and removing duplicates, we reviewed 151 references, 19 of which pertained to various survey topics, 41 to machine unlearning techniques, and 91 to related issues. (iv) Data extraction was performed based on insights from these reviewed references.

C. Organization

Fig. 1 highlights the structure of this survey, where the core parts are Section III, IV, and V, which study the unlearning taxonomy, unlearning security, and evaluation methods. The remainder of this paper is organized as follows. Section II introduces the background of related work in machine unlearning and the preliminaries of techniques in generative models. Section III introduces the method taxonomy, systematically categorizing existing machine unlearning approaches in generative models. Section IV reviews the security aspects of unlearning, focusing on existing attack methods and the corresponding defense strategies. Section V introduces the evaluation methods for the performance of unlearning methods. The critical applications of machine unlearning in real-world scenarios are explored in Section VI. Section VII delves into the challenges and future work, and opens research directions in this emerging field. Finally, Section VIII draws a conclusion of this survey.

II. Background

A. Machine Unlearning

Machine unlearning [43] can be broadly defined as the process of selectively removing the influence of specific training data from a machine learning model without retraining from scratch. Since many models use user data collected from the Internet, when models are trained with sensitive or personal data, users have the right to request that the model owner remove their data to ensure compliance with privacy regulations. In this case, machine unlearning methods are essential to ensure that the model complies with the law requirements. Machine unlearning [28]–[31], [36], [44] has been widely applied to various models, such as classification, segmentation, and generative models. However, it poses unique challenges in generative models due to their complexity and size. Generative models often learn intricate relationships in the data, and removing specific influences can be difficult without compromising the model's overall performance.

The evolution of machine unlearning started in 2014 with the "Right to be Forgotten" and has been investigated in various areas. In 2015, the first formal definition of machine unlearning was proposed, laying the foundation for subsequent research. Regulatory frameworks such as the GDPR and CCPA in 2018 further solidified the legal necessity for machine unlearning. Certified unlearning was explored in 2019, providing assurances for data deletion. In 2021, the SISA framework was introduced to enhance the scalability of unlearning in distributed scenarios. From 2023, machine unlearning has been introduced into generative AI, with a range of new methods covered in this survey. Table II summarizes the significant milestones in the development of machine unlearning, highlighting key signals in the research, legal, and industry domains.

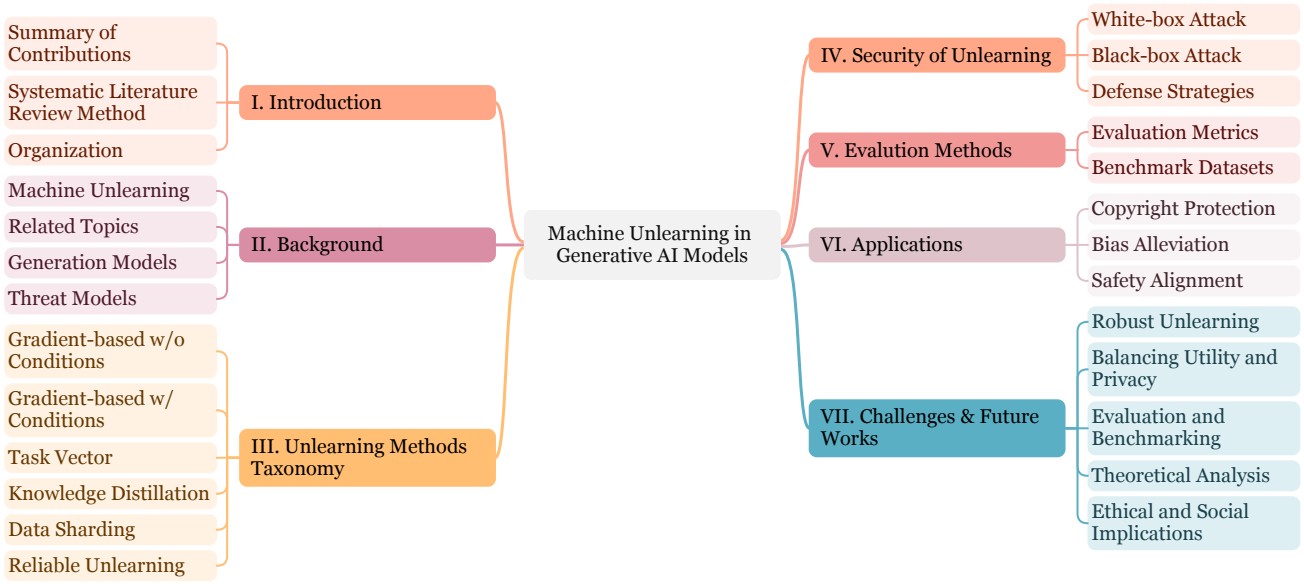

Fig. 1: The overview structure of this survey

TABLE II: The milestone timeline of machine unlearning development

Year	Milestone	Area	Key Innovation
2014	Right to be Forgotten [45]	Legal	Introduced concept of forgetting digital traces
2015	First Machine Unlearning Paper [43]	Research	Defined concept of machine unlearning formally
2018	GDPR [32], CCPA [33]	Legal	Formalized the right to be forgotten in law
2019	Certified Unlearning [46]	Research	Guarantees for data deletion
2021	SISA Framework [28]	Research	Sharded training to speed up unlearning
2023	Machine Unlearning in Generative AI [18], [19]	Research	Fine-tuning generative models to forget knowledge
2023	Action in Tech Companies [47]–[50]	Industry	More unlearning methods are explored in the industry

B. Related Topics

1) Knowledge Editing: Knowledge editing [51]–[53] refers to techniques used to directly modify a model's parameters or internal representations to achieve a desired change in behavior. Techniques [54]–[56], such as mechanistic localization and conflict-free editing, have been explored as potential solutions to unlearn specific data in generative models. Many researchers employ knowledge-editing methods to mitigate the influence of specific training data without requiring complete retraining.

2) Poisoning Attacks: Poisoning attacks [57]–[59] involve injecting malicious data into the training process to compromise the integrity or behavior of a model. Such attacks produce undesirable output features or biased embedding for generative models. The poisoning attacks [60]–[62] and machine unlearning can confuse the model with a specific concept, causing the failure to generate a text-image-aligned pair. However, poisoning attacks introduce dirty data during model training, whereas machine unlearning is typically applied after the model has been trained.

C. Generative Models

By learning complex data distributions, generative models can generate high-quality synthetic images. The architectures of these popular models, such as GAN, AE, and DM, are shown in Fig. 2.

GAN consists of two neural networks: a generator G and a discriminator D, which are trained simultaneously in a minimax game. The generator learns to map a latent variable $\mathbf{z} \sim p_{\mathbf{z}}(\mathbf{z})$ to a data distribution $G(\mathbf{z})$. The discriminator aims to distinguish real data samples from generated ones. The objective function for GANs is formulated as:

$$\begin{aligned}\min_G \max_D V(D,G) = \mathbb{E}_{\mathbf{x}\sim p_{\text{data}}(\mathbf{x})}[\log D(\mathbf{x})] \\ +\mathbb{E}_{\mathbf{z}\sim p_{\mathbf{z}}(\mathbf{z})}[\log(1-D(G(\mathbf{z})))],\end{aligned} \tag{1}$$

where $p_{\text{data}}(\mathbf{x})$ is the real data distribution, and $p_{\mathbf{z}}(\mathbf{z})$ is a prior distribution (e.g., Gaussian) used to sample $\mathbf{z}$. During training, the generator improves by producing more realistic samples to fool the discriminator. In contrast, the discriminator learns to better distinguish real from fake samples. This adversarial training process ultimately leads the generator to approximate the actual data distribution, allowing the generation of high-quality synthetic data.

AE aims to learn a compact representation of the input data by encoding it into a lower-dimensional latent space and subsequently reconstructing it. It consists of two components: an encoder E that maps input data $\mathbf{x}$ to a latent representation $\mathbf{z}$, and a decoder D that reconstructs $\hat{\mathbf{x}}$ from $\mathbf{z}$. This process can be expressed as:

$$\mathbf{z} = E(\mathbf{x}), \quad \hat{\mathbf{x}} = D(\mathbf{z}). \tag{2}$$

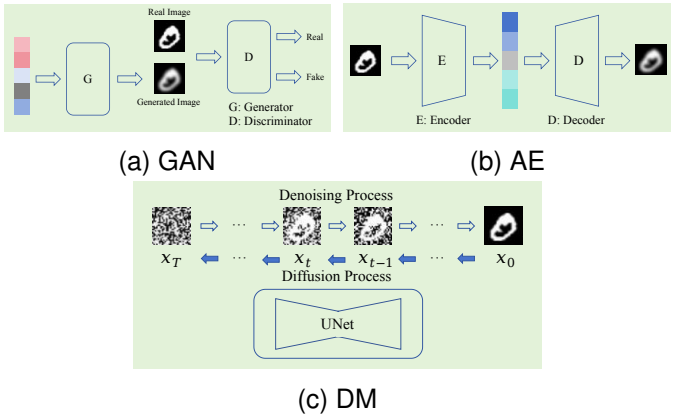

Fig. 2: Overview of different generative models.

The model is trained by minimizing the reconstruction loss, typically measured as the mean squared error between the original and the reconstructed input. Based on this basic architecture, various models have been developed, including the variable autoencoder (VAE) [63] and the mask autoencoder (MAE) [64].

DM learns to synthesize data by gradually denoising a variable sampled from a Gaussian distribution. These models are based on a forward diffusion process, which incrementally adds Gaussian noise to a data sample $\mathbf{x}_0$ over T timesteps, following a Markovian process. The reverse process aims to learn a parameterized denoising function $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$, typically modeled as a neural network that predicts the original sample given a noisy observation. The training is to minimize the variational bound on the negative log-likelihood, which simplifies to a noise prediction loss:

$$\mathcal{L} = \mathbb{E}_{t,\mathbf{x}_0,\boldsymbol{\epsilon}}\left[\|\boldsymbol{\epsilon} - \epsilon_\theta(\mathbf{x}_t, t)\|^2\right], \tag{3}$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is the added noise, and $\epsilon_\theta(\mathbf{x}_t, t)$ is the predicted noise. By iteratively optimizing the learned reverse process, diffusion models can generate high-quality samples, which makes them particularly effective for image, audio, and text synthesis.

D. Threat Models

1) White-Box Attack: The white-box setting [65] means that attackers have full access to the target generative model. They can mostly use parameters, architecture, source codes, and other internal information. This setting aligns with standard practices, as the models are shared in the open source community, such as Github [66] and Hugging Face [67].

2) Black-Box Attack: The black-box setting [68] represents the most challenging and demanding scenario. In this case, attackers have limited access to the target generative models. Mostly, they use the designed input queries and the generated output distribution to guess the model's information.

III. Unlearning Method Taxonomy

This section classifies and discusses various unlearning methods based on their techniques. A concise reference table is provided in Table III.

A. Gradient-based without Conditions

Gradient-based methods utilize model gradients to efficiently eliminate the impact of specific data. Due to their different architectures, each unlearning method has its own limitations and applicable models. Gradient-based methods without conditions mean that they directly use images as input and do not require additional text prompts as conditional input.

Adapt-then-Unlearn [69] is a two-stage unlearning method for pre-trained generative adversarial networks (GANs) by leveraging parameter space semantics. The first stage fine-tunes the pre-trained GAN generator G_{θ_G} to generate samples that only contain the undesired feature, using negative samples for training. This adapted generator is trained using an adversarial loss with an adaptation regularization term. The adaptation loss penalizes deviations in important parameters using Fisher information. In the second stage, unlearning is performed by optimizing the generator on positive samples while ensuring that its parameters diverge from those of the adapted generator. Sharing a similar idea, **Cascaded Unlearning** [70] introduces a substitute mechanism to maintain latent space continuity and a fake label regularization. The substitute mechanism reassigns the latent embedding to a meaningful alternative by replacing the target unlearning image x_0 with an alternative representation $S(x_0)$. This method prevents abrupt changes in the latent space, while ensuring that the model no longer generates the original image. Fake label regularization defines a criterion for the discriminator, trained to assign a low-confidence score to target images. These multistage methods encourage the generator to move away from the adapted parameters while maintaining high-quality image generation, although with weak time efficiency.

Based on the feature discriminator, **Feature Unlearning** [71] unlearns specific features in GANs and VAEs. This method involves identifying the target feature within the latent space, which is achieved by using vector arithmetic. The target vector z_e is calculated by subtracting the mean latent representation of the negative dataset (without the target feature) from the mean latent representation of the positive dataset (with the target feature). When the target feature is absent, a reconstruction loss ensures that the output remains unchanged. The target-erased output is changed for latent vectors to unlearn the target feature. This framework offers an effective method for feature unlearning that relies on a powerful discriminator.

Considering the privacy leakage, **Generative Unlearning for Any Identity (GUIDE)** [72] erases specific identities from pre-trained EG3D and StyleGAN2 models. The framework identifies the target latent embedding by extrapolating between the source and the average latent embedding, which ensures that the target identity is distinct from the source identity while maintaining proximity to the latent distribution. This method designs loss functions to optimize the pre-trained model, which includes local and adjacency unlearning loss, as well as global preservation loss. However, the loss functions make it hard to find the best balance between utility and privacy.

I2I Generator Unlearning [73] optimizes the encoder-decoder model to control the distributions of generated images for both the remaining and forgotten sets. The objective of

unlearning involves minimizing the KL divergence in the remaining set D_R while maximizing it in the forgotten set D_F. The objective of unlearning involves minimizing the KL divergence in the remaining set D_R while maximizing it in the forgotten set D_F with the following loss function,

$$\begin{aligned}\mathcal{L}_{\text{unlearn}} = \mathbb{E}_{x_r\sim D_R, x_f\sim D_F, n\sim N(0,\Sigma)}[\|E_\theta(T(x_r))- \\ E_{\theta_0}(T(x_r))\|_2^2 + \alpha\|E_\theta(T(x_f)) - E_{\theta_0}(T(n))\|_2^2],\end{aligned} \tag{4}$$

where $T(\cdot)$ is an operation such as cropping or masking applied to x. This framework achieves efficient unlearning by targeting only encoder parameters while preserving computational efficiency. Moreover, it is compatible with various generative models, including the GAN and MAE.

B. Gradient-based with Conditions

In this category, when fine-tuning the model to unlearn specific concepts, text prompts are used as conditional input. In some cases, these unlearning methods do not require additional auxiliary datasets and assume that the model has a built-in conditioning mechanism.

DMs face additional challenges due to the flexibility of text prompt conditions. **Safe Latent Diffusion (SLD)** [74] constructs a safety guidance by modifying the classifier-free guidance, which additionally uses an inappropriate concept prompt. Additionally, a warm-up parameter ensures that guidance is applied after the initial image structure emerges. Lastly, a momentum term is introduced to accelerate changes consistently, guiding them in the same direction across timesteps. Thus, SLD modifies latent vectors in the diffusion process, effectively mitigating inappropriate image content without retraining the model.

As a pioneer in the field, **Erased Stable Diffusion (ESD)** [18] introduces a fine-tuning approach to erase specific concepts. This method leverages the model's knowledge, thereby avoiding the need for external datasets. The method modifies the conditional noise prediction $\epsilon_\theta(x_t, c, t)$ for a given concept c to ensure that it is guided away from the undesired concept. The optimization objective of ESD is expressed as:

$$\epsilon_\theta(x_t, c, t) \leftarrow \epsilon_{\theta^*}(x_t, t) - \eta\left[\epsilon_{\theta^*}(x_t, c, t) - \epsilon_{\theta^*}(x_t, t)\right], \tag{5}$$

where ϵ_{θ^*} represents the frozen pre-trained model and η is the strength of the negative guidance. This modified score function restricts the erasure of the target concept. By fine-tuning the model's parameters with this objective, ESD achieves permanent concept erasure while maintaining minimal interference with other concepts. However, when erasing entire object classes, this method may partially fail, removing only specific distinctive attributes rather than the entire class concept.

Selective Amnesia (SA) [75] is a method inspired by continuous learning to achieve controlled forgetting in conditional generative models. The approach utilizes Elastic Weight Consolidation and Generative Replay to strike a balance between forgetting and retaining essential information. The goal is to maximize the logarithmic likelihood of the forgotten set to obtain a maximum a posteriori estimate. The objective for forgetting is derived as:

$$\begin{aligned}\mathcal{L}_{\text{SA}} = \mathbb{E}_{q(x|c)p_f(c)}[\log p(x|\theta, c)] - \lambda\sum_i F_i\frac{1}{2}(\theta_i - \theta_i^*)^2 \\ +\mathbb{E}_{p(x|c)p_r(c)}[\log p(x|\theta, c)],\end{aligned} \tag{6}$$

where x is a sample image, c is the corresponding concept, θ^* represents the parameters of the pre-trained model, F_i is the Fisher Information Matrix. SA provides controlled and interpretable forgetting with minimal impact on remembered concepts.

Forget-Me-Not (FMN) [76] modifies cross-attention scores to diminish the representation of undesired concepts while preserving the model's generative abilities for other content. The cross-attention mechanism refines visual features using textual features, and attention scores determine the influence of textual tokens on visual features. The attention re-steering loss function minimizes the attention scores associated with the target concept across the index of textual tokens via Eq. (7):

$$\mathcal{L}_{\text{Attn}} = \sum_{a\in A[:,i:j]}\|a\|^2, \tag{7}$$

where $A[:, i : j]$ indexes the attention scores of the target concept in textual tokens. Then, a visual denoising loss is introduced to refine the generated samples, which ensures consistent generation quality during forgetting. The final loss function combines both objectives, balancing the contributions of the two losses. FMN can eliminate the need for user-defined replacement concepts, allowing the model to revert to its inherent knowledge. This approach is lightweight, practical, and applicable to real-world scenarios. However, it requires an auxiliary dataset to unlearn specific concepts.

Saliency Unlearning (SalUn) [77] targets the weight of the selected model using weight saliency. This method identifies the salient weights by computing a saliency map based on the gradient of the forgetting loss $\ell_f(\theta; D_f)$ with respect to the model weights θ. The saliency map is defined as: $m_S = \mathbb{I}(|\nabla_\theta\ell_f(\theta; D_f)|_{\theta=\theta_o} \geq \gamma)$, where $\mathbb{I}$ is the indicator function, γ is a threshold, and θ_o represents the pre-trained model weights. This map highlights the weights that are influential for the forgotten dataset D_f. The model weights are decomposed into salient and non-salient components as follows,

$$\theta_u = m_S \odot \theta + (1 - m_S) \odot \theta_o, \tag{8}$$

where θ_u represents the updated model, and $\odot$ denotes element-wise multiplication. This formulation ensures that updates are focused only on salient weights. The weight saliency approach achieves efficient and targeted updates, reducing computational overhead compared to retraining while maintaining the model's generative performance for non-forgotten data.

Based on geometric and textual information, **GEOM-ERASING** [78] achieves a precise and controlled removal of implicit concepts. Implicit concepts, such as watermarks or toxic content, are inherently learned during training without explicit textual representation, making their removal particularly challenging. The method begins by identifying implicit concepts in the generated images using an external classifier. The classifier outputs confidence scores p_i and coordinates $o_i = [a_1^i, b_1^i, a_2^i, b_2^i]$, representing the bounding box of the implicit concept within the image, which is converted into location tokens for the text prompt. The updated text condition incorporates both the concept prompt and its spatial location, represented by discrete location tokens. To reduce emphasis on regions containing implicit concepts, a region-specific weight map is applied during optimization. GEOM-ERASING achieves the precise erasure of implicit concepts by integrating

spatial information with textual prompts and reweighting the loss in clean regions.

EraseDiff [79] formulates the objective function based on KL divergence but focuses on diffusion model unlearning. A constraint optimization problem is considered by deviating from the learnable reverse process of the ground-truth denoising procedure. EraseDiff modifies the backward diffusion process to prevent the generation of meaningful images related to forgotten concepts. To enforce effective unlearning, the learned noise predictor $\epsilon_{\hat{\theta}}$ is trained to produce noise from a different distribution, such as uniform noise. The final optimization problem integrates both objectives under a constrained formula:

$$\min_{\theta} L(\theta, D_r) \quad \text{s.t.} \quad f(\theta, D_f) - \min_{\phi|\theta} f(\phi, D_f) \leq 0. \tag{9}$$

This method balances forgetting and retention, ensuring the model maintains high utility while eliminating information.

Concept Ablation (CA) [80] also minimizes the KL divergence between the model's distribution for a target concept and that of an anchor concept, preventing the generation of undesired styles, instances, or memorized images. CA offers two optimization strategies: model-based ablation and noise-based ablation. Anchor distributions are derived from conditioning the pre-trained model on anchor prompts or by pairing anchor images with modified prompts. Regularization using the standard diffusion loss ensures that surrounding concepts are preserved. This method is efficient in ablating instances, artistic styles, and memorized images.

Separable Multi-Concept Erasure (SepME) [81] addresses the challenges of multi-concept erasure and subsequent restoration through two key components: generation of concept-irrelevant representations (G-CiRs) and weight decoupling (WD). G-CiRs prevent the erasure of substantial but irrelevant information to maintain DMs' generative capabilities for remaining concepts. WD decomposes weight increments into independent components to enable flexible erasure and restoration of concepts. SepME aims to integrate G-CiRs and WD, which provides a scalable and flexible solution for multi-concept erasure and restoration, ensuring minimal interference with remaining concepts. Compared to weight decoupling, **Target-aware Forgetting (TARF)** [82] addresses challenges in machine unlearning by decoupling the class label and the target concept. It introduces a dynamic optimization process that integrates annealed forgetting and target-aware retaining to selectively erase undesired concepts. Concretely, TARF employs a three-phase process, including target identification to isolate forgetting and retaining data, target separation to decouple entangled representations, and retraining approximation to align with the retraining objective.

Challenging Forgets (CF) [83] proposes a bi-level optimization (BLO) framework to evaluate machine unlearning effectiveness under challenging scenarios. The approach focuses on identifying the most difficult subset of data, known as the "worst-case forget set", to unlearn while preserving the utility of the remaining data. BLO has two levels, where the upper-level optimization selects the worst-case forget set, and the lower-level optimization solves the unlearning problem. To efficiently solve optimization problems, gradient unrolling is applied in conjunction with sign-based stochastic gradient descent, thereby reducing computational complexity by eliminating the need for second-order derivative calculations. The combined optimization ensures the selection of the worst-case forget set and the robustness of the unlearning methods.

Few-shot unlearning [84] unlearns specific concepts from text-to-image diffusion models by modifying the text encoder. This approach ensures that the model maintains image fidelity while preventing the generation of undesired concepts. The method is inspired by textual inversion, updating the text encoder's concept representation with a small perturbation: $c_\theta \leftarrow c_\theta + \Delta c$. The method applies a loss function similar to textual inversion, but in the reverse direction, to compute Δc. To minimize disruption to unrelated concepts, the method limits training iterations and updates only specific layers of the text encoder, such as the feed-forward layers and the final self-attention layer. This maintains the overall text-image alignment while erasing the target concept. Few-shot unlearning offers a rapid and efficient approach to machine unlearning in diffusion models. However, without adjusting U-Net parameters, this method retains the potential for generating NSFW content.

C. Task Vector

Task Vector approaches introduce specific directional changes in the latent space to suppress undesired data representations, which are conducted by weight displacements in the model's parameter space resulting from fine-tuning a specific task or concept. **Moderator** [85] introduces a Task Vector-based method for fine-grained content moderation in text-to-image diffusion models. The system employs self-reverse fine-tuning (SRFT) to generate task vectors, which represent weight displacements fine-tuned for specific moderation tasks. The process begins by prompting the model to generate self-supervised datasets corresponding to the policy objectives, which are then computed as the weight difference between the fine-tuned and original models. These vectors are subtracted or adjusted to achieve moderation while preserving the generative ability for unrelated content. The Task Vector-based approach offers a scalable, efficient, and adaptable solution for content moderation in diffusion models, balancing robustness and performance preservation.

Similar to Moderator, **Robust Concept Erasure (RCE)** [86] uses task vectors for robust concept unlearning in diffusion models. This method erases undesirable concepts in an input-independent manner by subtracting the task vector from the original model weights. The authors propose the Diverse Inversion technique to optimize the trade-off between concept erasure and maintaining the model's generative performance. This method generates diverse adversarial prompt embeddings targeting the undesired concept, ensuring that the model is robust to unexpected prompts.

D. Knowledge Distillation

Knowledge Distillation facilitates unlearning by transferring knowledge from the original model to a new model, eliminating the undesired information. **Safe self-Distillation Diffusion (SDD)** [87] is a self-distillation method that fine-tunes the

model by enforcing the noise estimate conditioned on a target concept to align with the unconditional one. Thus, it eliminates the problematic concept without requiring explicit negative examples. To enforce the removal of a specific concept c_s, SDD modifies the loss function by ensuring that the noise estimate conditioned on c_s is indistinguishable from the unconditional noise estimate. In addition, an exponential moving average teacher model is introduced to stabilize the optimization and prevent catastrophic forgetting. The final objective integrates self-distillation loss with standard diffusion loss to balance concept removal and image quality.

Score Forgetting Distillation (SFD) [88] is also a distillation-based method, using cross-class score distillation. This method aligns the conditional scores of the target and replacement classes to facilitate the unlearning process. As it is difficult to solve the loss function directly, the authors introduce an alternative denoising score matching loss function.

$$\mathcal{L}_{\text{dsm}}(\psi;\theta,c)=\mathbb{E}_{z_t,t,x\sim\mathcal{D}_{\theta,c}}\left[\gamma_t\frac{a_t^2}{\sigma_t^4}\left\|x_\psi(z_t,c)-x\right\|_2^2\right]. \tag{10}$$

SFD incorporates a data-free loss function into the distillation objective of a pre-trained diffusion model, mitigating the need for real data.

E. Data Sharding

Data Sharding splits the training data into shards, allowing the removal of specific shards without affecting the integrity of the remaining model. **Diffusion Soup** [89] introduces a data sharding method that merges model weights trained on different data shards for text-to-image diffusion models. This approach enables efficient training-free continual learning and unlearning by simply averaging the model weights corresponding to data shards. The soup weight is easy to update based on the formula $w_{soup}=\frac{w_{soup}-k_i w_i}{1-k_i}$. The method ensures flexibility in handling dynamically changing datasets, enabling the addition or removal of shards without retraining. Diffusion Soup approximates the geometric mean of distributions across data shards by averaging weights, providing robust anti-memorization properties and zero-shot style blending capabilities.

Compared to the weight distribution, **Encoded Ensembles** [90] learns data attribution in diffusion models. This approach trains multiple models on engineered subsets of the training data, allowing for the efficient evaluation of the training data's influence on model outputs. Each subset encodes specific training data characteristics, enabling temporary unlearning through ensemble ablation. This method circumvents the computational costs of traditional unlearning by removing models associated with the target data from the ensemble. It also introduces a Jacobian approximation to expedite counterfactual generation. However, these data sharding methods all have substantial space complexity due to the storage of models or data.

F. Reliable Unlearning

Reliable unlearning uses explainable or stable training methods to eliminate the influence of specific data or concepts by exploring closed-form solutions for unlearning.

Direct Unlearning Optimization (DUO) [91] utilizes paired image data and incorporates output-preserving regularization. The method treats unlearning as a preference optimization problem. Given paired datasets of unsafe images x_0^- and their safe counterparts x_0^+, the preference model learns a reward function $r(x_0)$, which is optimized using a binary cross-entropy loss. To ensure the model does not excessively diverge from its prior distribution, DUO employs KL-constrained optimization:

$$\max_{p_\theta}\mathbb{E}_{x_0\sim p_\theta(x_0)}[r(x_0)]-\beta D_{\text{KL}}[p_\theta(x_0)\|p_\phi(x_0)], \tag{11}$$

where $p_\phi(x_0)$ is the pretrained model distribution, and β regulates divergence. The combined objective of DUO is to balance unlearning and prior preservation. This approach enables robust unlearning of unsafe concepts while maintaining high-quality generation for unrelated content. However, this method overlooks the intersection between the text encoder and U-Net.

Concentrating on cross-attention projection matrices, **Unified Concept Editing (UCE)** [92] introduces a closed-form editing framework to simultaneously moderate multiple concepts in text-to-image diffusion models. The UCE framework formulates unlearning as an optimization problem, modifying the projection matrices W_k and W_v of the attention mechanism. To erase an undesired concept c_i, UCE aligns its projection to a new target embedding c_i^*, ensuring the model no longer associates c_i with its prior representation $v_i^*\leftarrow W^{\text{old}}c_i^*$. The optimization is formulated as follows,

$$\min_W\sum_{c_i\in E}\|Wc_i-v_i^*\|^2+\sum_{c_j\in P}\|Wc_j-W^{\text{old}}c_j\|^2, \tag{12}$$

where E is the set of concepts to edit and P is the set of concepts to preserve. This function has a closed-form solution, which allows UCE to efficiently apply multiple edits in one step while preserving model quality, enabling large-scale modifications without the need for costly retraining.

Building upon UCE, **Reliable and Efficient Concept Erasure (RECE)** [93] is also a closed-form approach, which improves erasure by iteratively deriving new embeddings that regenerate erased concepts and then reapply erasure. Given an edited projection matrix W^{new}, RECE finds a derived embedding c_i' that maximizes similarity to the original erased concept. If c' is found to elicit the undesired concept, then a subsequent step erases it. To ensure model performance is preserved while removing c', RECE introduces a regularization term that prevents excessive alteration of model behavior. The final concept erasure is iteratively performed using Eq. (13)

$$c'=(\lambda I+\sum_i W_i^{\text{new}T}W_i^{\text{new}})^{-1}(\sum_i W_i^{\text{new}T}W_i^{\text{old}})c. \tag{13}$$

RECE significantly improves concept erasure efficiency by leveraging an iterative closed-form approach that refines the removed concept embeddings while preserving model quality. It ensures more thorough unlearning than previous methods while maintaining high fidelity in generated images.

IV. Security of Unlearning

Although the above methods show great potential to eliminate undesirable concepts, many works [108]–[110] reveal the vulnerability of machine unlearning to different attacks. Consequently, defense strategies are demonstrated to achieve stronger unlearning against the attacks. Table III provides a concise overview of these attack and defense methods.

TABLE III: Overview of different methods and applications in machine unlearning for generative models. Applications: A1 - Copyright Protection, A2 - Bias Alleviation, A3 - Safety Alignment.

Section Category	Method Category	Methods	Model Category	Key Idea	Applications		
					A1	A2	A3
Unlearning Method Taxonomy	Gradient-based without Conditions	Adapt-then-Unlearn [69]	GAN	Two-stage Unlearning	✓		
		Cascaded Unlearn [70]	GAN	Label Substitute Mechanism	✓		
		Feature Unlearning [71]	GAN, VAE	Latent Space Projection	✓		
		GUIDE [72]	GAN	Latent Space Manipulation	✓		
		I2I [73]	GAN, MAE, DM	Encoder Mutual Information	✓		
	Gradient-based with Conditions	SLD [74]	DM	Safety Guidance	✓		✓
		ESD [18]	DM	Log Probability Gradient	✓		✓
		SA [75]	VAE, DM	Continual Learning to Forget	✓		✓
		FMN [76]	DM	Attention Re-steering	✓		✓
		SalUn [77]	DM	Weight Saliency	✓		✓
		GEOM-ERASING [78]	DM	Geometric Information	✓		✓
		EraseDiff [79]	DM	Constraint Optimization	✓		✓
		CA [80]	DM	KL Divergence Minimization	✓		
		SepME [81]	DM	G-CiRs, Weight Decouple	✓		
		TARF [82]	DM	Target-aware Gradient Ascent	✓		
		CF [83]	DM	Worst-case Forget Set Selection	✓	✓	
		Few-shot Unlearn [84]	DM	Text Encoder Few-shot unlearning	✓		✓
	Task Vector	Moderator [85]	DM	Self-reverse Fine-tuning	✓	✓	✓
		RCE [86]	DM	Task Vector	✓		✓
	Knowledge Distillation	SDD [87]	DM	Self-Distillation	✓		✓
		SFD [88]	DM	Score Distillation	✓		
	Data Sharding	Diffusion Soup [89]	DM	Model Merging	✓		
		Encoded Ensembles [90]	DM	Ensemble Unlearning	✓		
	Reliable Unlearning	DUO [91]	DM	Direct Preference Optimization			✓
		UCE [92]	DM	Closed-form Solution	✓	✓	✓
		RECE [93]	DM	Closed-form Solution	✓	✓	✓
Security of Unlearning	White-box Attack	CCE [94]	DM	Textual Inversion	✓		✓
		P4D [95]	DM	Latent Space Manipulation			✓
		UnlearnDiffAtk [96]	DM	Generation as Classification	✓		✓
		RECORD [97]	DM	Coordinate Descent	✓		
	Black-box Attack	RAB [98]	DM	Concept Extraction	✓		✓
		JPA [99]	DM	Gradient Masking			✓
		DiffZOO [100]	DM	Query-Based Attack			✓
	Defense Strategies	Receler [101]	DM	Lightweight Eraser	✓		✓
		AdvUnlearn [102]	DM	Adversarial Training	✓		✓
		RACE [103]	DM	Adversarial Training		✓	✓
		AdvAnchor [104]	DM	Adversarial Anchor	✓		✓
		DoCo [105]	DM	Concept Domain Correction	✓		✓
		Meta-Unlearning [106]	DM	Meta-learning Objective	✓		✓
		SAFREE [107]	DM	Self-Validating Filtering	✓		✓

A. White-box Attack

In the white-box attack setting, attackers are assumed to have full access to the pre-trained models. **Circumventing Concept Erasure (CCE)** [94] is an attack method designed to circumvent concept erasure techniques in text-to-image diffusion models. This method exploits the hypothesis that concept erasure in generative models often behaves as input filtering rather than actual erasure. After applying concept unlearning, the modified diffusion model alters its prediction to suppress erased concepts. CCE circumvents this by learning a placeholder string c^* that reactivates the erased concept. The learned embedding v^* is optimized to ensure that c^* effectively restores the erased concept. The learned embedding v^* is optimized as follows:

$$v^* = \arg\min_v \mathbb{E}_{z_t,t,\epsilon\sim\mathcal{N}(0,1)} \left[\|\epsilon - \epsilon_\theta(z_t, c^*, t)\|_2^2 \right]. \quad (14)$$

To bypass inference-based erasure techniques, such as SLD, CCE modifies its optimization to counteract the additional guidance term. The learned embedding v^* can then be substituted in the model vocabulary, allowing the user to generate images of the erased concept without modifying the model weights. CCE demonstrates that post hoc concept erasure in diffusion models does not entirely remove concepts but redirects them within the model's embedding space.

Prompting4Debugging (P4D) [95] optimizes prompts in the latent space to bypass concept removal methods, negative prompting techniques, and safety-guided diffusion models. Specifically, this method optimizes the noise prediction in the unlearned model G' conditioned on the adversarial prompt P^* to closely match that of the unconstrained model G

when conditioned on the original forbidden prompt P. The optimization objective is as follows:

$$\mathcal{L}_{\text{P4D}} = \|\epsilon_\theta(z_t, W(P), t) - \epsilon_{\theta'}(z_t, P^*, t)\|_2^2, \tag{15}$$

where $W(P)$ is the fixed text encoder mapping prompt P to an embedding, and θ' denotes the parameters of the safety-equipped model. To improve transferability and interpretability, P4D adopts a discrete optimization approach, where the soft prompt P^* is continuously updated and then projected onto a discrete vocabulary space. P4D uncovers new vulnerabilities in safety-enhanced T2I models using latent-space optimization and red-teaming strategies.

UnlearnDiffAtk [96] exploits the inherent classification ability of diffusion models, allowing for the generation of adversarial prompts that circumvent unlearning mechanisms without requiring auxiliary models. It formulates the attack as a classification problem within the diffusion process:

$$p_{\theta^*}(c'|x) \propto \frac{\exp\left\{-\mathbb{E}_{t,\epsilon}[\|\epsilon - \epsilon_{\theta^*}(z_t|c')\|_2^2]\right\}}{\sum_j \exp\left\{-\mathbb{E}_{t,\epsilon}[\|\epsilon - \epsilon_{\theta^*}(z_t|c_j)\|_2^2]\right\}}, \tag{16}$$

where c_j are competing prompts to compare classification confidence. UnlearnDiffAtk minimizes the noise prediction error for adversarial prompts, ensuring they are indistinguishable from legitimate prompts while successfully bypassing the unlearning mechanism. Using the intrinsic classification capabilities of DMs, the attack enables efficient adversarial prompt discovery without requiring auxiliary classifiers or diffusion models.

RECORD [97] discovers prompts that can generate erased content by minimizing a loss function that measures the deviation of the erased model's noise predictions from those of the original model. This method searches for an adversarial prompt y^* that minimizes the distance between their noise estimates on the unlearned diffusion model $\epsilon_{\theta'}$ and the original pre-unlearning model ϵ_θ. To efficiently search for y^* in the discrete token space, RECORD employs a coordinate descent algorithm that iteratively optimizes individual tokens while keeping the others fixed. The loss function is approximated over a batch of sampled latent embeddings. At each step, RECORD selects the token position s to update by computing token gradients: $g_j = \nabla_{c_j} \hat{L}(y(c_j, s), Z)$, where c_j represents candidate tokens. The algorithm evaluates the top K candidate tokens and selects the one that minimizes $\hat{L}$, updating the prompt sequence accordingly. A greedy update strategy is employed to ensure stable convergence in the discrete token space: $\hat{L}(y(c^*, s), R) < \hat{L}(y, R)$, where R is a reference set used to track optimization progress. RECORD significantly outperforms prior adversarial attacks, demonstrating that existing machine unlearning methods in diffusion models do not fully unlearn knowledge.

B. Black-box Attack

In the black-box attack setting, adversaries can only use limited information to attack machine unlearning models. Instead of requiring access to the model's parameters, **Ring-A-Bell (RAB)** [98] is a model-agnostic red-teaming framework and searches adversarial prompts that can reactivate unlearned concepts through an optimization process using text encoders. To construct adversarial prompts, RAB extracts the semantic representation of the target concept c by comparing text encodings of semantically similar prompt pairs:

$$\hat{c} = \frac{1}{N} \sum_{i=1}^{N} \left(f(P_i^c) - f(P_i^{\neg c})\right), \tag{17}$$

where $f(\cdot)$ is the text encoder, P_i^c is a prompt containing the concept, and $P_i^{\neg c}$ is a concept-neutral counterpart. The adversarial prompt optimization then modifies an initial prompt embedding $f(P)$ by injecting the extracted concept representation. Thus, RAB bypasses unlearning techniques in diffusion models effectively.

Based on the black-box adversarial attack, **Jailbreaking Prompt Attack (JPA)** [99] exploits vulnerabilities in the text embedding space of high dimensions. JPA identifies adversarial prompts using antonym-based contrastive learning. For a restricted concept c, JPA constructs an embedding representation by computing the differences between antonym pairs. The adversarial prompt is then optimized by injecting this concept embedding into a target prompt p_t, $T(p_r) = T(p_t) + \lambda \cdot r$. JPA optimizes a prefix prompt by maximizing the cosine similarity between embeddings $T(p_t)$ and $T(p_r)$ to project this modified representation back into the discrete text space. This optimization is performed using the projected gradient descent method. Since discrete token updates are non-differentiable, JPA introduces a soft assignment strategy. To prevent safety checkers from detecting sensitive words, JPA applies gradient masking, setting high gradient penalties for restricted tokens $v_{ik} \leftarrow v_{ik} - M_k$, where M_k is a large penalty term for sensitive words. The method maintains semantic fidelity while systematically evading text-based, image-based, and text-image-based safety filters.

Unlike a model-based attack, **DiffZOO** [100] is a purely query-based black-box attack that circumvents safety mechanisms in text-to-image diffusion models. The goal of DiffZOO is to find an adversarial prompt p^* that maximizes the probability of generating an erased concept while only querying the model. To optimize within the discrete prompt space, DiffZOO introduces Continuous Position Replacement Vectors, which learn token replacement probabilities. To update these replacement probabilities, this method employs zeroth-order optimization to estimate gradient updates without requiring access to the model. By computing the estimated gradient for each token replacement, DiffZOO iteratively optimizes the C-PRV parameters using

$$z_i \leftarrow z_i - \eta \frac{\partial L}{\partial z_i}, \quad u_{i,j} \leftarrow u_{i,j} - \eta \frac{\partial L}{\partial u_{i,j}}. \tag{18}$$

DiffZOO demonstrates that T2I diffusion models remain vulnerable to black-box query-only adversarial prompting.

C. Defense Strategies

These attack methods demonstrate the limitations of current unlearning techniques in threat models. Therefore, it is important to understand how these limitations can be improved and to be defensive. To defend against these attacks, **Receler** [101], **AdvUnlearn** [102], **Robust Adversarial Concept Erasure (RACE)** [103], **AdvAnchor** [104], and **Domain Correction (DoCo)** [105] use adversarial training to explore defense machine unlearning strategies from different aspects.

Receler introduces a concept erasure method employing additional learnable layers. This approach adds a lightweight eraser module, which constitutes only 0.37% of the model parameters, to remove specific concepts from the model outputs. The method integrates the eraser after each cross-attention layer in the diffusion U-Net architecture, ensuring precise manipulation of textual and visual features associated with the target concept. In particular, the objective is defined as:

$$\mathcal{L}_{Erase} = \mathbb{E}_{x_t,t}\left[\|\epsilon_{\theta'}(x_t, e_c, t) - \epsilon_E\|^2\right] \tag{19}$$

where $\epsilon_E = \epsilon_\theta(x_t, t) - \eta\left[\epsilon_\theta(x_t, e_c, t) - \epsilon_\theta(x_t, t)\right]$. Receler ensures two key properties: locality and robustness. Locality is achieved through concept-localized regularization, and robustness is reinforced by adversarial learning.

AdvUnlearn executes defense by integrating adversarial training into the unlearning process, formulating it as a bi-level optimization problem as shown in Eq. (20):

$$\begin{aligned} \min_{\theta} \quad & \mathcal{L}_u(\theta, c^*) \\ s.t. \quad & c^* = \underset{\|c'-c_e\|_0 \le \epsilon}{\arg\min}\ \mathcal{L}_{\text{atk}}(\theta, c'), \end{aligned} \tag{20}$$

where c^* represents the optimal adversarial prompt found by maximizing the likelihood of regenerating the erased concept, and $\mathcal{L}_u$ is the upper-level unlearning objective. Instead of fine-tuning the entire diffusion model, AdvUnlearn optimizes the text encoder using adversarial prompts generated by a fast attack generation method, which is more effective than modifying the U-Net.

RACE bridges adversarial perturbations to the concept erasure process, preventing the reconstruction of unlearned concepts. It formulates an adversarial perturbation δ that maximizes the attack success rate.

$$\delta^* = \arg\max_{\|\delta\| \le \epsilon} \mathcal{L}_{\text{DM}}(\theta, c_e + \delta), \tag{21}$$

where ϵ constrains the magnitude of the perturbation. This approach ensures that the model does not just forget a specific concept token and its adversarially derived variations. The method reduces the attack success rate of red-teaming techniques and strengthens the integrity of the erased concepts, increasing the failure rate through prompt engineering.

AdvAnchor introduces an adversarial anchor-based framework to improve the reliability of machine unlearning in text-to-image diffusion models. AdvAnchor generates adversarial anchors $e_{\text{adv-anchor}}$ by introducing universal perturbations to the target concept embedding. To ensure that the adversarial anchor effectively erases the concept, AdvAnchor optimizes e_{adv} using the adversarial objective. This objective allows the model to dissociate $e_{\text{adv-anchor}}$ from e_{c_e} by maximizing their divergence in the noise prediction space.

Based on the anchor concept, DoCo proposes two complementary components: concept domain correction and concept-preserving gradient. Through adversarial training, concept domain correction aligns the distribution of the target concept c^* with that of an anchor concept c. A discriminator D distinguishes between outputs conditioned on c^* and c, while the generator (diffusion model) is optimized to fool D. The concept-preserving gradient is employed to resolve the conflicts between the objectives of unlearning and preservation.

Besides adversarial training, **Meta-Unlearning** [106] introduces an additional meta-objective to enforce long-term concept forgetting. This method prevents diffusion models from relearning unlearned concepts after malicious fine-tuning. Typically, unlearning is formulated as an optimization problem in which the model forgets concepts from the forgotten set D_{F} while preserving the performance in the retained set D_{R}. Meta-unlearning introduces a secondary objective to prevent such relearning by simulating the fine-tuning process during unlearning, which is defined as:

$$L_{\text{meta}}(\theta^{\text{F}}) = -L_{\text{DM}}(\theta^{\text{F}}; D_{\text{F}}) - \zeta\left(L_{\text{DM}}(\theta^{\text{F}}; D_{\text{R}}) - L_{\text{DM}}(\theta; D_{\text{R}})\right). \tag{22}$$

where θ^{F} represents the model parameters after fine-tuning on the forget subset D_{F}. The first term discourages the model from reducing loss on the forget set after fine-tuning. The second term ensures that fine-tuning on the forget set degrades performance on the retain set, causing benign concepts related to the forget set to self-destruct. Meta-unlearning ensures that erased concepts remain challenging to recover even after adversarial attempts by backpropagating through the simulated fine-tuning process. Therefore, it can prevent reexposure to the forgotten set from erased knowledge, providing a more resilient approach to safe and compliant generative models.

Unlike conventional unlearning-based methods that fine-tune model weights to remove unsafe concepts, **SAFREE** [107] is a training-free approach to ensure safe generation models. SAFREE uses adaptive filtering mechanisms at both the textual embedding and visual latent space levels to detect toxic concepts in the input prompt embedding space. First, it identifies a toxic concept subspace C, represented as a matrix of unsafe keyword embeddings. SAFREE computes a residual vector to assess the conceptual proximity of an input token embedding p_i to this subspace. A larger residual distance indicates a stronger association with the toxic concept. To suppress unsafe content, SAFREE integrates a self-validating filtering mechanism that dynamically adjusts the number of denoising steps based on the similarity between the filtered and original prompt embeddings. Additionally, SAFREE extends filtering into the visual latent space using an adaptive latent re-attention mechanism. It attenuates unsafe features in the frequency domain via a spectral filtering operation. Filtering at both the textual and visual levels provides a strong, adaptive, and training-free safeguard for responsible generative models.

However, these robust unlearning methods are not meant to be completely secure, as they are still at risk of emerging attacks in the arms race.

V. Evaluation Methods

This section provides a detailed overview of the evaluation methods used in unlearning, including metrics and datasets to evaluate unlearning approaches. Table IV and V provide an overview of these evaluation metrics and benchmark datasets.

A. Evaluation Metrics

1) Generation Quality: Generation quality metrics evaluate the realism of the generated data and its fidelity and diversity after unlearning. Frechet Inception Distance (FID) [111], Kernel Inception Distance (KID) [112], Inception Score (IS) [113], and Learned Perceptual Image Patch Similarity

TABLE IV: Overview of different evaluation metrics in machine unlearning for generative models.

Category	Metrics	Brief Introduction
Generation Quality	FID [111], KID [112], IS [113], LPIPS [114]	Measures distribution similarity of two image sets using feature statistics.
	CLIPS, CLIPA [115], TIFA [116]	Measures alignment between generated images and textual descriptions.
	IR [117], VR [118]	Measures image quality and coherence using reward models.
Unlearning Effectiveness	GCDS [119]	Detecting celebrity likeness in generated images.
	ES [86]	Quantifies how effectively a concept is erased from the model's generation.
	ICR [78]	Measures the ratio of implicit concepts that persist in generations.
	CE [75]	Likelihood of the erased concept still appearing in generated images.
	TFR [71]	Target Feature Ratio measures the proportion of removed features in outputs.
	ID [120]	Evaluates whether erased identities remain in generated images.
	MR [80], MS [76]	Measures how much the model retains memorized unlearned examples.
	PUL [69]	Percentage of Un-Learning, indicating the fraction of successful forgetting.
	LHR [85]	Measures how effectively harmful concepts are removed based on LLM.
	WD [121]	Weight distance between the retrain models and other unlearned models.
	UA, RA, TA [122]	Measures overall performance of classification tasks on different classes.
	AUC [123]	Measures classification performance over thresholds.
Efficiency and Security	UT, RTE [77]	Computational efficiency of unlearning methods.
	MIA [124]	Membership Inference Attack success rate, assessing data privacy risk.
	ASR [125]	Measures percentage of adversarial attacks recovering unlearned content.
Subjective Evaluation	VC [15]	Side-by-side visual comparisons of pre- and post-unlearning outputs.
	QA [84]	Qualitative assessment of image differences before and after unlearning.
	ME [85]	Human assessment of image generation quality and concept removal.

(LPIPS) [114] measure the distribution similarity of two image sets using feature statistics to quantify image realism. CLIP Score (CLIPS) and CLIP Accuracy (CLIPA) [115], along with Text-to-Image Faithfulness evaluation with Question Answering (TIFA) [116], evaluate the alignment between the generated images and the textual descriptions, ensuring semantic consistency. Image Reward (IR) [117] and Vision Reward (VR) [118] use reward models to measure image quality and coherence.

2) Unlearning Effectiveness: These metrics assess the effectiveness of unlearning techniques in removing targeted concepts from generated images. Giphy Celebrity Detection Score (GCDS) [119] detects the presence of celebrity likeness in generated images. Erasure Score (ES) [86], Implicit Concept Ratio (ICR) [78], and Classifier Entropy (CE) [75] evaluate how effectively a concept is erased from the model's output and determine the likelihood of erased concepts reappearing in generated images. Target Feature Ratio (TFR) [71] and similarity of identities (ID) [120] measure the extent to which specific features or identities remain in unlearned images. The Memorization Rate (MR) [80] and the Memorization Score (MS) [76] assess how much the model retains unlearned examples. The Percentage of Un-Learning (PUL) [69] quantifies the fraction of successful forgetting. LLM-based Harm Rate (LHR) [85] evaluates the effectiveness of removing harmful concepts using LLM analysis. Weight Distance (WD) [121] computes the deviation between retrained models and unlearned models. Unlearning Accuracy (UA), Remaining Accuracy (RA), and Testing Accuracy (TA) [122] measure the classification accuracy across different classes. Area Under Curve (AUC) [123] assesses classification performance across various thresholds.

3) Efficiency and Security: These metrics focus on the computational efficiency and security risks associated with unlearning methods. Unlearning Time (UT) and Retraining Efficiency (RTE) [77] measure the computational efficiency of unlearning techniques. The Membership Inference Attack (MIA) [124] evaluates the success rate of Membership Inference Attacks, assessing the risk to data privacy. Attack Success Rate (ASR) [125] measures the attack success rate, indicating the percentage of adversarial attacks that recover unlearned content.

4) Subjective Evaluation: Subjective evaluation metrics involve human assessments to ensure the visual and conceptual effectiveness of unlearning. As mentioned in [84], quantitative analysis can not always reflect the unlearning performance. Thus, qualitative analysis is essential to provide another assessment of generation differences before and after unlearning. The most common metric for this part is the visual comparison (VC) [15], which refers to side-by-side comparisons of generated images from pre- and post-unlearning outputs. In [85], the Manual Evaluation (ME) is used for human evaluation of image generation quality and concept removal effectiveness.

B. Benchmark Datasets

Machine unlearning in generative models is often evaluated using specific tasks in the following datasets.

1) Object Generation Datasets: In the model training, object generation datasets provide diverse categories of images for generation models to learn the distribution of different classes. Typically, these datasets are utilized in unlearning tasks to assess the model's performance in unlearning objects or classes. MNIST [126] is a handwritten digit dataset widely used for benchmarking classification models. SVHN [127] is a dataset of street view house numbers used primarily for digit recognition tasks. CIFAR [128], which includes CIFAR-10 and CIFAR-100, consists of small object images with different categories. STL [129] is a high-resolution variant of CIFAR-10, offering a higher-quality alternative for evaluating the

TABLE V: Overview of different benchmark datasets in machine unlearning for generative models.

Category	Dataset	Brief Introduction
Object Generation Datasets	MNIST [126]	Handwritten digit dataset used for classification tasks.
	SVHN [127]	Street View House Numbers dataset for digit recognition.
	CIFAR [128]	Small object classification dataset (CIFAR-10, CIFAR-100).
	STL [129]	High-resolution variant of CIFAR-10 for classification.
	Places-365 [130]	Scene recognition dataset for classification tasks.
	ImageNet [131], [132]	Large-scale image classification dataset with diverse categories.
Identity Recognition Datasets	CelebA [133]	Large-scale dataset for celebrity face recognition.
	FFHQ [122]	High-quality face dataset used in generative model training.
	AFHQ [134]	Animal face dataset for generative modeling and unlearning.
Generative & Artist Style Datasets	MSCOCO [135]	Large dataset with text-image pairs used for captioning and generation.
	LAION [136], [137]	Large-scale open-source text-image dataset for training models.
	WIKIArt [138]	Artistic style dataset used for evaluating style removal and unlearning.
Unlearning-Specific Datasets	I2P [74]	Dataset containing NSFW or biased image prompts for unlearning safety.
	ICD [78]	Dataset designed for testing implicit concept unlearning.
	UnlearnCanvas [139]	Dataset specifically designed for evaluating machine unlearning techniques.

classification model. Places-365 [130] is a scene recognition dataset designed for classification tasks. ImageNet [131] is a large-scale dataset comprising diverse image categories and serves as a standard benchmark for evaluating deep learning models. Imagenette [132] is a smaller subset of ImageNet that enables quick testing of unlearning methods.

2) Identity Recognition Datasets: Identity recognition datasets are commonly used to evaluate unlearning methods that focus on removing or modifying identity-related features in generated images. CelebA [133] is a large-scale dataset containing celebrity faces, frequently used for face recognition and attribute classification. FFHQ [122] provides high-quality human face images and serves as a key dataset for training and evaluating generative models. AFHQ [134] contains images of animal faces, making it worthwhile to evaluate identity removal and domain adaptation techniques in generative models.

3) Generative & Artist Style Datasets: Datasets in this category are used to evaluate the effectiveness of unlearning generative content and artistic styles. Here, generative content specifically refers to the use of text prompts to generate images. MSCOCO [135] contains text-image pairs and is widely used for captioning and generation tasks. LAION [136], [137] is an open-source large-scale dataset that contains a vast collection of text-image pairs, providing a valuable resource for training and evaluating generative tasks. WIKIArt [138] is a large dataset of artistic paintings commonly used for tasks such as style transfer, style removal, and training generative models.

4) Unlearning-Specific Datasets: These datasets are designed to evaluate the effectiveness of machine unlearning techniques in removing specific concepts from the generated output. The Inappropriate Image Prompts dataset [74] consists of NSFW or biased image prompts, making it suitable for studying safety-focused unlearning. The Implicit Concept Dataset [78] is created to test the removal of implicit concepts embedded in models. UnlearnCanvas [139] is a dataset specifically designed to benchmark unlearning performance, offering customized challenges to evaluate different techniques.

VI. Applications

Machine unlearning can address critical concerns related to copyright protection, bias alleviation, and safety alignment, which underscore the importance of deploying ethical, legal, and responsible generative AI models. The applications are summarized in Table III with a feature demonstration.

A. Copyright Protection

Copyright protection is essential in generative models due to the risks associated with unauthorized reproduction, potential legal liability, and ethical issues related to intellectual property (IP). These models, trained on vast datasets that often contain copyrighted material, can generate images that closely resemble protected content, including artistic styles, logos, and celebrity likenesses, potentially leading to infringement. This phenomenon raises legal and ethical concerns, as AI-generated content can devalue original works and result in lawsuits against AI developers and users. Additionally, businesses and creators face reputational risks if their AI-generated outputs inadvertently violate copyrights. To address these concerns, researchers have introduced some novel machine unlearning methods [18], [73]–[82], [91]–[93]. These methods can selectively erase copyrighted concepts from generative models without degrading overall performance by redirecting the model from generating protected content, allowing it to revert to alternative representations learned during training. For example, the techniques efficiently erase specific styles (e.g., Van Gogh paintings) or characters (e.g., Mickey Mouse). Therefore, it prevents the unauthorized generation of protected content while maintaining general image synthesis capabilities.

B. Bias Alleviation

Bias alleviation in generative models is essential to prevent the amplification of societal stereotypes, the underrepresentation of marginalized groups, and ethical concerns related to AI-driven discrimination. These models [74], [140]–[143] often learn biases from large-scale datasets, resulting in skewed outputs such as male-dominated depictions of leadership roles

or racially biased portrayals in law enforcement contexts. These biases not only harm fairness and inclusion, but also have sustained consequences in media representation, hiring tools, and digital marketing [144]. Biased output can violate anti-discrimination laws or platform guidelines, particularly in high-stakes contexts such as healthcare and education. Users and stakeholders increasingly demand algorithmic fairness in generative models to maintain trust and prevent reputational damage.

To address this, bias mitigation techniques [91]–[93] aim to modify the internal associations of the model, thus ensuring more diverse and equitable content generation. Many of these approaches aim to strike a balance between fairness and preserving image quality or creative freedom. Looking ahead, user-centered bias correction, where users can flag or adjust biased outputs, may become a key direction for the responsible deployment of generative AI.

C. Safety Alignment

Aligning generative models with safety standards, particularly in filtering NSFW content, is vital to ensure ethical deployment, avoid legal liabilities, and protect user welfare across platforms and devices. Google published a report on the adversarial misuse of generative AI [145], calling for pressing attention to the security threats posed by such models. Without adequate safeguards, generative models [94]–[97] risk producing explicit or harmful content, which can violate community guidelines, age restrictions, or national content regulations such as GDPR or CCPA.

Recent methods [18], [146]–[150] propose to remove harmful concepts at the model level by fine-tuning, reinforcement learning, or encoder pruning, so that prompts cannot trigger them. These solutions help align models with regulatory frameworks and platform accountability, especially in public-facing or IoT-driven applications. However, future research must also consider the ethical transparency and explainability of these safety filters to ensure that suppression decisions are not arbitrary, biased, or overly opaque to end users.

VII. Challenges and Future Works

Despite significant advances in machine unlearning techniques, applying these methods to generative models, such as diffusion models, presents several challenges. The following discussion identifies critical issues and provides information on potential research directions and opportunities to advance the field further.

A. Robust Unlearning

A major challenge in generative model unlearning is achieving robust unlearning to ensure that erased concepts cannot be recovered under adversarial prompts, distribution shifts, or fine-tuning. Current methods [94]–[97], such as modifying attention layers or latent representations, may leave residual traces of the removed concept, making them vulnerable to prompt engineering attacks or inversion techniques that reconstruct forgotten knowledge. Recent studies [96], [98], [102] demonstrate that adversarially crafted text inputs can bypass unlearning mechanisms, forcing models to regenerate content that was supposed to be erased. Techniques such as adversarial training have been proposed to enhance robustness, but these methods often face a tradeoff between the effectiveness of unlearning and maintaining generation quality. Future work should focus on adversarially robust unlearning mechanisms, such as incorporating certifiable removal techniques, self-supervised feedback loops, and meta-learning approaches that adaptively refine the model after unlearning. Additionally, integrating diffusion model interpretability techniques can help analyze how unlearning affects the generative process at different denoising stages, leading to more effective and verifiable concept erasure.

B. Balancing Utility and Privacy

With increasing integration of generative models into IoT environments, privacy and security concerns [20], [21], [151] become even more pronounced due to the extensive collection and use of sensitive and environmental data. If attackers steal and misuse this information, they can deploy generative models to reconstruct realistic data from individuals without their consent, posing threats to personal privacy, identity theft, and the dissemination of misinformation. However, to defend against this evasion, unlearning methods should also strike a balance between removing unlearned concepts and preserving generation quality for utility. To address this, researchers introduce methods like GUIDE to enable identity removal from pre-trained generative models. These functions effectively erase a specific identity while preserving the model's general generative capabilities, ensuring stronger privacy protection in AI-generated content while maintaining model performance. However, straightforward unlearning approaches [102], such as fine-tuning or parameter editing, often result in degraded generation quality for benign concepts, raising concerns about utility loss. Achieving this balance requires further investigation of regularization techniques and modular optimization strategies, such as focusing on text encoders rather than the entire model architecture.

C. Evaluation and Benchmarking

Current evaluation metrics for text-to-image model unlearning are insufficient and inconsistent, often relying on qualitative comparisons or indirect similarity measures such as CLIP-based retrieval. Existing metrics primarily focus on generation quality or safety in non-adversarial scenarios, which fail to capture the robustness of models under adversarial conditions. Tools [96], [98] such as *Ring-A-Bell* and *UnlearnDiffAtk* offer promising directions, but require further standardization and widespread adoption. However, systematic and reliable evaluation frameworks for assessing the efficacy of unlearning methods are still lacking. Future work should establish standardized benchmarks for unlearning performance, measuring both removal effectiveness (e.g., how well the erased concept is suppressed across diverse prompts) and model integrity (e.g., whether unrelated content generation remains unaffected). Novel evaluation methods could include counterfactual testing

frameworks, adversarial prompting techniques to detect residual knowledge, and user perception studies to assess whether unlearning aligns with real-world expectations. In addition, designing scalable automated metrics, such as concept purity scores or entropy-based divergence from the original model, would enable more objective and reproducible evaluations of unlearning techniques in content generation scenarios.

D. Theoretical Analysis

Solid work in machine unlearning for text-to-image models requires a steady theoretical foundation to quantify and formalize unlearning efficiency, convergence, and guarantees. Current approaches often rely on heuristic-based optimization without rigorous mathematical proofs on whether the unlearned concept is permanently removed or suppressed. Developing provable bounds on unlearning effectiveness, such as information-theoretic measures of concept removal, privacy guarantees under differential privacy frameworks, or adversarial testing to detect residual knowledge, would provide a stronger theoretical foundation and more convincing results. Additionally, exploring connections to catastrophic forgetting in continual learning could lead to a deeper understanding of how selective unlearning affects model generalization, ensuring that erasing one concept does not unintentionally degrade unrelated knowledge.

E. Ethical and Social Implications

Beyond technical considerations, machine unlearning in AI-generated content raises profound ethical and social concerns. Central to these concerns is the tension between individual data rights and public accountability. Users may request the removal of their data or associated concepts, but this must be balanced against the public interest and the need for historical preservation. For example, attempts to unlearn politically sensitive content could be used to selectively erase critical narratives, posing risks of censorship and historical revisionism. A notable case is the controversy surrounding the Stability AI opt-out program, which allowed artists to remove their work from the training data [47]. Although this was seen as a step toward respecting the rights of creators, some critics noted a lack of transparency and accountability in the verification and implementation of these requests. Similarly, OpenAI has been under scrutiny for how it moderates and forgets controversial or copyrighted information in ChatGPT [152], especially under pressure from authors and publishers who have filed lawsuits demanding the removal of proprietary content.

Moreover, unlearning itself can unintentionally introduce or amplify model biases if not applied uniformly in different data domains [102], [153]. If certain groups disproportionately request removals (e.g., due to higher privacy concerns), this could lead to representational imbalances in model behavior, ultimately affecting fairness in generated output. These challenges underscore the need for governance frameworks that integrate both technical and ethical oversight. Moreover, transparent deletion protocols, audit mechanisms, and cross-disciplinary collaboration are critical to ensure that machine unlearning is conducted responsibly. Future research should actively engage with this tradeoff between society and technology to enable ethical deployment of unlearning systems.

VIII. Conclusion

Machine unlearning in generative models is an imperative and rapidly evolving field to address several critical concerns. The comprehensive survey explores various unlearning methods and categorizes the literature based on their unlearning techniques, adversarial attacks, and the corresponding defenses. In addition, a variety of evaluation metrics, benchmarking datasets, and critical applications are presented, along with a thoughtful discussion for future researchers. Finally, we discuss the future challenges and research directions in this field and beyond, to explore the research frontier of machine unlearning and contribute to the research community in this field. In conclusion, this survey will serve as a helpful recipe for any interested party embarking on artificial intelligence, machine unlearning, generative models, and security & privacy related topics.

IX. Acknowledgments

This work was supported by the National Science Foundation under grants No. 2429960, 2434899, 2416872, 2315596, 2244219, and 2146497.

References

[1] X. Chen, L. Luo, F. Tang, M. Zhao, and N. Kato, "Aigc-based evolvable digital twin networks: a road to the intelligent metaverse," *IEEE Network*, 2024.

[2] S. Long, J. Tan, B. Mao, F. Tang, Y. Li, M. Zhao, and N. Kato, "A survey on intelligent network operations and performance optimization based on large language models," *IEEE Communications Surveys & Tutorials*, 2025.

[3] Z. Cai, Z. Xiong, H. Xu, P. Wang, W. Li, and Y. Pan, "Generative adversarial networks: A survey toward private and secure applications," *ACM Computing Surveys (CSUR)*, vol. 54, no. 6, pp. 1–38, 2021.

[4] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[5] Z. Xiong, W. Li, and Z. Cai, "Federated generative model on multi-source heterogeneous data in iot," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 9, 2023, pp. 10 537–10 545.

[6] M. Megahed and A. Mohammed, "A comprehensive review of generative adversarial networks: Fundamentals, applications, and challenges," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 16, no. 1, p. e1629, 2024.

[7] A. Xiong, C. Qiao, W. Li, D. Wang, D. Li, B. Gao, and W. Wang, "Block-chain abnormal transaction detection method based on generative adversarial network and autoencoder," *High-Confidence Computing*, p. 100313, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2667295225000170

[8] H. Huang, R. He, Z. Sun, T. Tan *et al.*, "Introvae: Introspective variational autoencoders for photographic image synthesis," *Advances in neural information processing systems*, vol. 31, 2018.

[9] M. El-Kaddoury, A. Mahmoudi, and M. M. Himmi, "Deep generative models for image generation: A practical comparison between variational autoencoders and generative adversarial networks," in *Mobile, Secure, and Programmable Networking: 5th International Conference, MSPN 2019, Mohammedia, Morocco, April 23–24, 2019, Revised Selected Papers 5*. Springer, 2019, pp. 1–8.

[10] H. Xu, Z. Cai, D. Takabi, and W. Li, "Audio-visual autoencoding for privacy-preserving video streaming," *IEEE Internet of Things Journal*, vol. 9, no. 3, pp. 1749–1761, 2021.

[11] K. Berahmand, F. Daneshfar, E. S. Salehi, Y. Li, and Y. Xu, "Autoencoders and their applications in machine learning: a survey," *Artificial Intelligence Review*, vol. 57, no. 2, p. 28, 2024.

[12] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Advances in Neural Information Processing Systems*, 2020.

[13] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," *arXiv preprint arXiv:2010.02502*, 2020.

[14] S. Li, L. Yang, X. Jiang, H. Lu, D. An, Z. Di, W. Lu, J. Chen, K. Liu, Y. Yu *et al.*, "Swiftdiffusion: Efficient diffusion model serving with add-on modules," *arXiv preprint arXiv:2407.02031*, 2024.

[15] H. Cao, C. Tan, Z. Gao, Y. Xu, G. Chen, P.-A. Heng, and S. Z. Li, "A survey on generative diffusion models," *IEEE Transactions on Knowledge and Data Engineering*, 2024.

[16] Z. Xiong, W. Li, Y. Li, and Z. Cai, "Exact-fun: an exact and efficient federated unlearning approach," in *2023 IEEE International Conference on Data Mining (ICDM)*. IEEE, 2023, pp. 1439–1444.

[17] Z. Xiong, W. Li, and Z. Cai, "Appro-fun: Approximate machine unlearning in federated setting," in *2024 33rd International Conference on Computer Communications and Networks (ICCCN)*. IEEE, 2024, pp. 1–9.

[18] R. Gandikota, J. Materzyńska, J. Fiotto-Kaufman, and D. Bau, "Erasing Concepts from Diffusion Models," in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. Paris, France: IEEE, Oct. 2023, pp. 2426–2436.

[19] M. Pawelczyk, S. Neel, and H. Lakkaraju, "In-context unlearning: Language models as few shot unlearners," *arXiv preprint arXiv:2310.07579*, 2023.

[20] J. Wen, J. Nie, J. Kang, D. Niyato, H. Du, Y. Zhang, and M. Guizani, "From generative ai to generative internet of things: Fundamentals, framework, and outlooks," *IEEE Internet of Things Magazine*, vol. 7, no. 3, pp. 30–37, 2024.

[21] W. Jiang, Y. Zhang, H. Han, and J. Mu, "Generative ai for consumer internet of things: Challenges and opportunities," *IEEE Consumer Electronics Magazine*, 2025.

[22] F. Tang, L. Luo, Z. Guo, Y. Li, M. Zhao, and N. Kato, "Semi-distributed network fault diagnosis based on digital twin network in highly dynamic heterogeneous networks," *IEEE Transactions on Mobile Computing*, 2024.

[23] H. Tyagi, R. Kumar, and S. K. Pandey, "A detailed study on trust management techniques for security and privacy in iot: Challenges, trends, and research directions," *High-Confidence Computing*, vol. 3, no. 2, p. 100127, 2023.

[24] X. Zhou, W. Liang, K. Yan, W. Li, K. I.-K. Wang, J. Ma, and Q. Jin, "Edge-enabled two-stage scheduling based on deep reinforcement learning for internet of everything," *IEEE Internet of Things Journal*, vol. 10, no. 4, pp. 3295–3304, 2022.

[25] X. Zhou, X. Xu, W. Liang, Z. Zeng, and Z. Yan, "Deep-learning-enhanced multitarget detection for end–edge–cloud surveillance in smart iot," *IEEE Internet of Things Journal*, vol. 8, no. 16, pp. 12 588–12 596, 2021.

[26] X. Zhou, W. Liang, I. Kevin, K. Wang, Z. Yan, L. T. Yang, W. Wei, J. Ma, and Q. Jin, "Decentralized p2p federated learning for privacy-preserving and resilient mobile robotic systems," *IEEE Wireless Communications*, vol. 30, no. 2, pp. 82–89, 2023.

[27] X. Zhou, W. Liang, W. Li, K. Yan, S. Shimizu, and K. I.-K. Wang, "Hierarchical adversarial attacks against graph-neural-network-based iot network intrusion detection system," *IEEE Internet of Things Journal*, vol. 9, no. 12, pp. 9310–9319, 2021.

[28] L. P. Bourtoule, V. Chandrasekaran, C. Marchant *et al.*, "Machine unlearning," *IEEE Symposium on Security and Privacy*, 2021.

[29] A. Ginart, M. Guan, G. Valiant, and J. Zou, "Making ai forget you: Data deletion in machine learning," in *Advances in Neural Information Processing Systems*, 2019, pp. 3513–3526.

[30] O. M. Safa, M. M. Abdelaziz, M. Eltawy, M. Mamdouh, M. Gharib, S. Eltenihy, N. M. Ghanem, and M. M. Ismail, "A comparative study of machine unlearning techniques for image and text classification models," *arXiv preprint arXiv:2412.19583*, 2024.

[31] W. Wang, Z. Tian, C. Zhang, and S. Yu, "Machine unlearning: A comprehensive survey," *arXiv preprint arXiv:2405.07406*, 2024.

[32] European Parliament and Council of the European Union. Regulation (EU) 2016/679 of the European Parliament and of the Council. [Online]. Available: https://data.europa.eu/eli/reg/2016/679/oj

[33] (2018) California consumer privacy act. California State Legislature. [Online]. Available: https://leginfo.legislature.ca.gov/faces/codes_displayText.xhtml?division=3.&part=4.&lawCode=CIV&title=1.81.5

[34] (2015) Act on the protection of personal information. Japan. [Online]. Available: https://www.ppc.go.jp/en/legal/

[35] N. Zhang and H. Tang, "Text-to-image synthesis: A decade survey," *arXiv preprint arXiv:2411.16164*, 2024.

[36] T. T. Nguyen, T. T. Huynh, Z. Ren, P. L. Nguyen, A. W.-C. Liew, H. Yin, and Q. V. H. Nguyen, "A survey of machine unlearning," *arXiv preprint arXiv:2209.02299*, 2022.

[37] J. C. Costa, T. Roxo, H. Proença, and P. R. Inácio, "How deep learning sees the world: A survey on adversarial attacks & defenses," *IEEE Access*, 2024.

[38] Z. Liu, G. Dou, Z. Tan, Y. Tian, and M. Jiang, "Machine unlearning in generative ai: A survey," *arXiv preprint arXiv:2407.20516*, 2024.

[39] Z. Liu, H. Ye, C. Chen, Y. Zheng, and K.-Y. Lam, "Threats, attacks, and defenses in machine unlearning: A survey," *arXiv preprint arXiv:2403.13682*, 2024.

[40] V. T. Truong, L. B. Dang, and L. B. Le, "Attacks and defenses for generative diffusion models: A comprehensive survey," *arXiv preprint arXiv:2408.03400*, 2024.

[41] X. Liu, X. Cui, P. Li, Z. Li, H. Huang, S. Xia, M. Zhang, Y. Zou, and R. He, "Jailbreak attacks and defenses against multimodal generative models: A survey," *arXiv preprint arXiv:2411.09259*, 2024.

[42] S. Keele *et al.*, "Guidelines for performing systematic literature reviews in software engineering," Technical report, ver. 2.3 ebse technical report. ebse, Tech. Rep., 2007.

[43] Y. Cao and J. Yang, "Towards making systems forget with machine unlearning," in *2015 IEEE symposium on security and privacy*. IEEE, 2015, pp. 463–480.

[44] C. Li, H. Jiang, J. Chen, Y. Zhao, S. Fu, F. Jing, and Y. Guo, "An overview of machine unlearning," *High-Confidence Computing*, p. 100254, 2024.

[45] "Art. 17 GDPR - Right to erasure ('right to be forgotten') - GDPR.eu — gdpr.eu," https://gdpr.eu/article-17-right-to-be-forgotten/?cn-reloaded=1, 2014.

[46] C. Guo, T. Goldstein, A. Hannun, and L. Van Der Maaten, "Certified data removal from machine learning models," *arXiv preprint arXiv:1911.03030*, 2019.

[47] J. Kemper, "Artists remove 80 million images from Stable Diffusion 3 training data," https://the-decoder.com/artists-remove-80-million-images-from-stable-diffusion-3-training-data/, 2023.

[48] R. Eldan and M. Russinovich, "Who's harry potter? approximate unlearning in llms," 2023. [Online]. Available: https://arxiv.org/abs/2310.02238

[49] Google, "Announcing the first Machine Unlearning Challenge — research.google," https://research.google/blog/announcing-the-first-machine-unlearning-challenge/, 2023.

[50] S. R. Kadhe, F. Ahmed, D. Wei, N. Baracaldo, and I. Padhi, "Split, unlearn, merge: Leveraging data attributes for more effective unlearning in llms," 2024. [Online]. Available: https://arxiv.org/abs/2406.11780

[51] N. Si, H. Zhang, H. Chang, W. Zhang, D. Qu, and W. Zhang, "Knowledge unlearning for llms: Tasks, methods, and challenges," *arXiv preprint arXiv:2311.15766*, 2023.

[52] S. Wang, Y. Zhu, H. Liu, Z. Zheng, C. Chen, and J. Li, "Knowledge editing for large language models: A survey," *ACM Computing Surveys*, vol. 57, no. 3, pp. 1–37, 2024.

[53] V. Mazzia, A. Pedrani, A. Caciolai, K. Rottmann, and D. Bernardi, "A survey on knowledge editing of neural networks," *IEEE Transactions on Neural Networks and Learning Systems*, 2024.

[54] P. Guo, A. Syed, A. Sheshadri, A. Ewart, and G. K. Dziugaite, "Mechanistic unlearning: Robust knowledge unlearning and editing via mechanistic localization," *arXiv preprint arXiv:2410.12949*, 2024.

[55] Y. Wang, M. Chen, N. Peng, and K.-W. Chang, "Deepedit: Knowledge editing as decoding with constraints," *arXiv preprint arXiv:2401.10471*, 2024.

[56] D. Jung, J. Seo, J. Lee, C. Park, and H. Lim, "Come: An unlearning-based approach to conflict-free model editing," *arXiv preprint arXiv:2502.15826*, 2025.

[57] B. Biggio, B. Nelson, and P. Laskov, "Poisoning attacks against support vector machines," in *Proceedings of the 29th International Coference on International Conference on Machine Learning*, 2012, pp. 1467–1474.

[58] Z. Tian, L. Cui, J. Liang, and S. Yu, "A comprehensive survey on poisoning attacks and countermeasures in machine learning," *ACM Computing Surveys*, vol. 55, no. 8, pp. 1–35, 2022.

[59] Z. Liu, H. Ye, C. Chen, Y. Zheng, and K.-Y. Lam, "Threats, attacks, and defenses in machine unlearning: A survey," *IEEE Open Journal of the Computer Society*, 2025.

[60] S. Zhai, Y. Dong, Q. Shen, S. Pu, Y. Fang, and H. Su, "Text-to-image diffusion models can be easily backdoored through multimodal data poisoning," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 1577–1587.

[61] S. Shan, W. Ding, J. Passananti, S. Wu, H. Zheng, and B. Y. Zhao, "Nightshade: Prompt-specific poisoning attacks on text-to-image generative models," in *2024 IEEE Symposium on Security and Privacy (SP)*. IEEE Computer Society, 2024, pp. 212–212.

[62] W. Ding, C. Y. Li, S. Shan, B. Y. Zhao, and H. Zheng, "Understanding implosion in text-to-image generative models," in *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, 2024, pp. 1211–1225.

[63] D. P. Kingma, M. Welling *et al.*, "Auto-encoding variational bayes," 2013.

[64] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 16 000–16 009.

[65] V. Porkodi, M. Sivaram, A. S. Mohammed, and V. Manikandan, "Survey on white-box attacks and solutions," *Asian Journal of Computer Science and Technology*, vol. 7, no. 3, pp. 28–32, 2018.

[66] GitHub, Inc., "Github - where the world builds software," 2025, accessed: March 13, 2025. [Online]. Available: https://github.com

[67] Hugging Face, Inc., "Hugging face - the ai community building the future," 2025, accessed: March 13, 2025. [Online]. Available: https://huggingface.co

[68] S. Bhambri, S. Muku, A. Tulasi, and A. B. Buduru, "A survey of black-box adversarial attacks on computer vision models," *arXiv preprint arXiv:1912.01667*, 2019.

[69] P. Tiwary, A. Guha, S. Panda *et al.*, "Adapt then unlearn: Exploiting parameter space semantics for unlearning in generative adversarial networks," *arXiv preprint arXiv:2309.14054*, 2023.

[70] H. Sun, T. Zhu, W. Chang, and W. Zhou, "Generative adversarial networks unlearning," *arXiv preprint arXiv:2308.09881*, 2023.

[71] S. Moon, S. Cho, and D. Kim, "Feature Unlearning for Pre-trained GANs and VAEs," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 19, pp. 21 420–21 428, Mar. 2024.

[72] J. Seo, S.-H. Lee, T.-Y. Lee, S. Moon, and G.-M. Park, "Generative unlearning for any identity," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 9151–9161.

[73] G. Li, H. Hsu, C.-F. Chen, and R. Marculescu, "Machine unlearning for image-to-image generative models," *arXiv preprint arXiv:2402.00351*, 2024.

[74] P. Schramowski, M. Brack, B. Deiseroth, and K. Kersting, "Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 22 522–22 531.

[75] A. Heng and H. Soh, "Selective amnesia: A continual learning approach to forgetting in deep generative models," *Advances in Neural Information Processing Systems*, vol. 36, pp. 17 170–17 194, 2023.

[76] G. Zhang, K. Wang, X. Xu, Z. Wang, and H. Shi, "Forget-me-not: Learning to forget in text-to-image diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 1755–1764.

[77] C. Fan, J. Liu, Y. Zhang, E. Wong, D. Wei, and S. Liu, "SalUn: Empowering Machine Unlearning via Gradient-based Weight Saliency in Both Image Classification and Generation," in *The Twelfth International Conference on Learning Representations*, Oct. 2023.

[78] Z. Liu, K. Chen, Y. Zhang, J. Han, L. Hong, H. Xu, Z. Li, D.-Y. Yeung, and J. T. Kwok, "Implicit concept removal of diffusion models," in *European Conference on Computer Vision*. Springer, 2024, pp. 457–473.

[79] J. Wu, T. Le, M. Hayat, and M. Harandi, "Erasediff: Erasing data influence in diffusion models," *arXiv preprint arXiv:2401.05779*, 2024.

[80] N. Kumari, B. Zhang, S.-Y. Wang, E. Shechtman, R. Zhang, and J.-Y. Zhu, "Ablating Concepts in Text-to-Image Diffusion Models," in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. Paris, France: IEEE, Oct. 2023, pp. 22 634–22 645.

[81] M. Zhao, L. Zhang, T. Zheng, Y. Kong, and B. Yin, "Separable multi-concept erasure from diffusion models," *arXiv preprint arXiv:2402.05947*, 2024.

[82] J. Zhu, B. Han, J. Yao, J. Xu, G. Niu, and M. Sugiyama, "Decoupling the class label and the target concept in machine unlearning," *arXiv preprint arXiv:2406.08288*, 2024.

[83] C. Fan, J. Liu, A. Hero, and S. Liu, "Challenging forgets: Unveiling the worst-case forget sets in machine unlearning," in *European Conference on Computer Vision*. Springer, 2024, pp. 278–297.

[84] M. Fuchi and T. Takagi, "Erasing concepts from text-to-image diffusion models with few-shot unlearning," *arXiv preprint arXiv:2405.07288*, vol. 2, 2024.

[85] P. Wang, Q. Li, L. Yu, Z. Wang, A. Li, and H. Jin, "Moderator: Moderating text-to-image diffusion models through fine-grained context-based policies," in *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, 2024, pp. 1181–1195.

[86] M. Pham, K. O. Marshall, C. Hegde, and N. Cohen, "Robust concept erasure using task vectors," *arXiv preprint arXiv:2404.03631*, 2024.

[87] S. Kim, S. Jung, B. Kim, M. Choi, J. Shin, and J. Lee, "Towards safe self-distillation of internet-scale text-to-image diffusion models," *arXiv preprint arXiv:2307.05977*, 2023.

[88] T. Chen, S. Zhang, and M. Zhou, "Score forgetting distillation: A swift, data-free method for machine unlearning in diffusion models," *arXiv preprint arXiv:2409.11219*, 2024.

[89] B. Biggs, A. Seshadri, Y. Zou, A. Jain, A. Golatkar, Y. Xie, A. Achille, A. Swaminathan, and S. Soatto, "Diffusion soup: Model merging for text-to-image diffusion models," in *European Conference on Computer Vision*. Springer, 2024, pp. 257–274.

[90] Z. Dai and D. K. Gifford, "Training data attribution for diffusion models," *arXiv preprint arXiv:2306.02174*, 2023.

[91] Y.-H. Park, S. Yun, J.-H. Kim, J. Kim, G. Jang, Y. Jeong, J. Jo, and G. Lee, "Direct unlearning optimization for robust and safe text-to-image models," *arXiv preprint arXiv:2407.21035*, 2024.

[92] R. Gandikota, H. Orgad, Y. Belinkov, J. Materzyńska, and D. Bau, "Unified concept editing in diffusion models," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 5111–5120.

[93] C. Gong, K. Chen, Z. Wei, J. Chen, and Y.-G. Jiang, "Reliable and efficient concept erasure of text-to-image diffusion models," in *European Conference on Computer Vision*. Springer, 2024, pp. 73–88.

[94] M. Pham, K. O. Marshall, N. Cohen, G. Mittal, and C. Hegde, "Circumventing concept erasure methods for text-to-image generative models," *arXiv preprint arXiv:2308.01508*, 2023.

[95] Z.-Y. Chin, C.-M. Jiang, C.-C. Huang, P.-Y. Chen, and W.-C. Chiu, "Prompting4debugging: Red-teaming text-to-image diffusion models by finding problematic prompts," *arXiv preprint arXiv:2309.06135*, 2023.

[96] Y. Zhang, J. Jia, X. Chen, A. Chen, Y. Zhang, J. Liu, K. Ding, and S. Liu, "To generate or not? safety-driven unlearned diffusion models are still easy to generate unsafe images... for now," in *European Conference on Computer Vision*. Springer, 2024, pp. 385–403.

[97] L. Beerens, A. D. Richardson, K. Zhang, and D. Chen, "On the vulnerability of concept erasure in diffusion models," *arXiv preprint arXiv:2502.17537*, 2025.

[98] Y.-L. Tsai, C.-Y. Hsu, C. Xie, C.-H. Lin, J.-Y. Chen, B. Li, P.-Y. Chen, C.-M. Yu, and C.-Y. Huang, "Ring-a-bell! how reliable are concept removal methods for diffusion models?" *arXiv preprint arXiv:2310.10012*, 2023.

[99] J. Ma, A. Cao, Z. Xiao, Y. Li, J. Zhang, C. Ye, and J. Zhao, "Jailbreaking prompt attack: A controllable adversarial attack against diffusion models," *arXiv preprint arXiv:2404.02928*, 2024.

[100] P. Dang, X. Hu, D. Li, R. Zhang, Q. Guo, and K. Xu, "Diff-zoo: A purely query-based black-box attack for red-teaming text-to-image generative model via zeroth order optimization," *arXiv preprint arXiv:2408.11071*, 2024.

[101] C.-P. Huang, K.-P. Chang, C.-T. Tsai, Y.-H. Lai, and Y.-C. F. Wang, "Receler: Reliable concept erasing of text-to-image diffusion models via lightweight erasers," *arXiv preprint arXiv:2311.17717*, 2023.

[102] Y. Zhang, X. Chen, J. Jia, Y. Zhang, C. Fan, J. Liu, M. Hong, K. Ding, and S. Liu, "Defensive unlearning with adversarial training for robust concept erasure in diffusion models," *Advances in Neural Information Processing Systems*, vol. 37, pp. 36 748–36 776, 2024.

[103] C. Kim, K. Min, and Y. Yang, "Race: Robust adversarial concept erasure for secure text-to-image diffusion model," in *European Conference on Computer Vision*. Springer, 2024, pp. 461–478.

[104] M. Zhao, L. Zhang, X. Yang, T. Zheng, and B. Yin, "Advanchor: Enhancing diffusion model unlearning with adversarial anchors," *arXiv preprint arXiv:2501.00054*, 2024.

[105] Y. Wu, S. Zhou, M. Yang, L. Wang, H. Chang, W. Zhu, X. Hu, X. Zhou, and X. Yang, "Unlearning concepts in diffusion model via concept domain correction and concept preserving gradient," *arXiv preprint arXiv:2405.15304*, 2024.

[106] H. Gao, T. Pang, C. Du, T. Hu, Z. Deng, and M. Lin, "Meta-unlearning on diffusion models: Preventing relearning unlearned concepts," *arXiv preprint arXiv:2410.12777*, 2024.

[107] J. Yoon, S. Yu, V. Patil, H. Yao, and M. Bansal, "Safree: Training-free and adaptive guard for safe text-to-image and video generation," *arXiv preprint arXiv:2410.12761*, 2024.

[108] N. G. Marchant, B. I. Rubinstein, and S. Alfeld, "Hard to forget: Poisoning attacks on certified machine unlearning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 7, 2022, pp. 7691–7700.

[109] J. Łucki, B. Wei, Y. Huang, P. Henderson, F. Tramèr, and J. Rando, "An adversarial perspective on machine unlearning for ai safety," *arXiv preprint arXiv:2409.18025*, 2024.

[110] M. Bertran, S. Tang, M. Kearns, J. H. Morgenstern, A. Roth, and S. Z. Wu, "Reconstruction attacks on machine unlearning: Simple models are vulnerable," *Advances in Neural Information Processing Systems*, vol. 37, pp. 104 995–105 016, 2024.

[111] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," in *Advances in Neural Information Processing Systems*, 2017, pp. 6626–6637.

[112] M. Bińkowski, D. J. Sutherland, M. Arbel, and A. Gretton, "Demystifying mmd gans," *arXiv preprint arXiv:1801.01401*, 2018.

[113] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," *Advances in neural information processing systems*, vol. 29, 2016.

[114] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018.

[115] J. Hessel, A. Holtzman, M. Forbes, R. L. Bras, and Y. Choi, "Clipscore: A reference-free evaluation metric for image captioning," *arXiv preprint arXiv:2104.08718*, 2021.

[116] Y. Hu, B. Liu, J. Kasai, Y. Wang, M. Ostendorf, R. Krishna, and N. A. Smith, "Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 20 406–20 417.

[117] J. Xu, X. Liu, Y. Wu, Y. Tong, Q. Li, M. Ding, J. Tang, and Y. Dong, "Imagereward: Learning and evaluating human preferences for text-to-image generation," *Advances in Neural Information Processing Systems*, vol. 36, pp. 15 903–15 935, 2023.

[118] J. Xu, Y. Huang, J. Cheng, Y. Yang, J. Xu, Y. Wang, W. Duan, S. Yang, Q. Jin, S. Li *et al.*, "Visionreward: Fine-grained multi-dimensional human preference learning for image and video generation," *arXiv preprint arXiv:2412.21059*, 2024.

[119] "GitHub - Giphy/celeb-detection-oss: GIPHY's Open-Source Celebrity Detection Deep Learning Model — github.com," https://github.com/Giphy/celeb-detection-oss, [Accessed 04-03-2025].

[120] Y. Huang, Y. Wang, Y. Tai, X. Liu, P. Shen, S. Li, J. Li, and F. Huang, "Curricularface: adaptive curriculum learning loss for deep face recognition," in *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 5901–5910.

[121] A. K. Tarun, V. S. Chundawat, M. Mandal, and M. Kankanhalli, "Fast yet effective machine unlearning," *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[122] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4401–4410.

[123] J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (roc) curve." *Radiology*, vol. 143, no. 1, pp. 29–36, 1982.

[124] J. Hayes, L. Melis, G. Danezis, and E. De Cristofaro, "Logan: Membership inference attacks against generative models," *arXiv preprint arXiv:1705.07663*, 2017.

[125] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *2017 ieee symposium on security and privacy (sp)*. Ieee, 2017, pp. 39–57.

[126] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[127] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, A. Y. Ng *et al.*, "Reading digits in natural images with unsupervised feature learning," in *NIPS workshop on deep learning and unsupervised feature learning*, vol. 2011, no. 2. Granada, 2011, p. 4.

[128] A. Krizhevsky, G. Hinton *et al.*, "Learning multiple layers of features from tiny images," 2009.

[129] A. Coates, A. Ng, and H. Lee, "An analysis of single-layer networks in unsupervised feature learning," in *Proceedings of the fourteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings, 2011, pp. 215–223.

[130] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 6, pp. 1452–1464, 2017.

[131] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[132] "GitHub - fastai/imagenette: A smaller subset of 10 easily classified classes from Imagenet, and a little more French — github.com," https://github.com/fastai/imagenette, [Accessed 05-03-2025].

[133] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 3730–3738.

[134] Y. Choi, Y. Uh, J. Yoo, and J.-W. Ha, "Stargan v2: Diverse image synthesis for multiple domains," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 8188–8197.

[135] T.-Y. Lin, M. Maire, S. Belongie, J. Hays *et al.*, "Microsoft coco: Common objects in context," *European Conference on Computer Vision*, pp. 740–755, 2014.

[136] C. Schuhmann, R. Vencu, R. Beaumont, R. Kaczmarczyk, C. Mullis, A. Katta, T. Coombes, J. Jitsev, and A. Komatsuzaki, "Laion-400m: Open dataset of clip-filtered 400 million image-text pairs," *arXiv preprint arXiv:2111.02114*, 2021.

[137] C. Schuhmann, R. Beaumont, R. Vencu, R. W. Gordon *et al.*, "Laion-5b: An open large-scale dataset for training next generation image-text models," *arXiv preprint arXiv:2210.08402*, 2022.

[138] B. Saleh and A. Elgammal, "Large-scale classification of fine-art paintings: Learning the right metric on the right feature," in *International Conference on Machine Learning and Applications*, 2015.

[139] Y. Zhang, Y. Zhang, Y. Yao, J. Jia, J. Liu, X. Liu, and S. Liu, "Unlearncanvas: A stylized image dataset to benchmark machine unlearning for diffusion models," *arXiv preprint arXiv:2402.11846*, 2024.

[140] T. Karras, M. Aittala, S. Laine, E. Härkönen, J. Hellsten, J. Lehtinen, and T. Aila, "Alias-free generative adversarial networks," *Advances in neural information processing systems*, vol. 34, pp. 852–863, 2021.

[141] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695.

[142] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans *et al.*, "Photorealistic text-to-image diffusion models with deep language understanding," *Advances in neural information processing systems*, vol. 35, pp. 36 479–36 494, 2022.

[143] P. Sun, Y. Jiang, S. Chen, S. Zhang, B. Peng, P. Luo, and Z. Yuan, "Autoregressive model beats diffusion: Llama for scalable image generation," *arXiv preprint arXiv:2406.06525*, 2024.

[144] "When AI Gets It Wrong: Addressing AI Hallucinations and Bias," https://mitsloanedtech.mit.edu/ai/basics/addressing-ai-hallucinations-and-bias/, 2023.

[145] "Adversarial Misuse of Generative AI," https://cloud.google.com/blog/topics/threat-intelligence/adversarial-misuse-generative-ai, 2025.

[146] X. Li, Y. Yang, J. Deng, C. Yan, Y. Chen, X. Ji, and W. Xu, "Safegen: Mitigating sexually explicit content generation in text-to-image models," in *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, 2024, pp. 4807–4821.

[147] D. Han, S. Mohamed, and Y. Li, "Shielddiff: Suppressing sexual content generation from diffusion models through reinforcement learning," *arXiv preprint arXiv:2410.05309*, 2024.

[148] L. Yuan, X. Li, C. Xu, G. Tao, X. Jia, Y. Huang, W. Dong, Y. Liu, X. Wang, and B. Li, "Promptguard: Soft prompt-guided unsafe content moderation for text-to-image models," *arXiv preprint arXiv:2501.03544*, 2025.

[149] Y. Wang, J. Chen, Q. Li, X. Yang, and S. Ji, "Aeiou: A unified defense framework against nsfw prompts in text-to-image models," *arXiv preprint arXiv:2412.18123*, 2024.

[150] S. Poppi, T. Poppi, F. Cocchi, M. Cornia, L. Baraldi, and R. Cucchiara, "Safe-clip: Removing nsfw concepts from vision-and-language models," in *European Conference on Computer Vision*. Springer, 2024, pp. 340–356.

[151] X. Wang, Z. Wan, A. Hekmati, M. Zong, S. Alam, M. Zhang, and B. Krishnamachari, "Iot in the era of generative ai: Vision and challenges," *IEEE Internet Computing*, 2024.

[152] "The Authors Guild, John Grisham, Jodi Picoult, David Baldacci, George R.R. Martin, and 13 Other Authors File Class-Action Suit Against OpenAI," https://authorsguild.org/news/ag-and-authors-file-class-action-suit-against-openai/, 2023.
[153] Y. Yang, R. Gao, X. Yang, J. Zhong, and Q. Xu, "Guardt2i: Defending text-to-image models from adversarial prompts," *arXiv preprint arXiv:2403.01446*, 2024.