Rethinking Machine Unlearning in Image Generation Models

Renyang Liu*
Institute of Data Science, National University of Singapore
Singapore, Singapore
ryliu@nus.edu.sg

Wenjie Feng*†
School of Artificial Intelligence and Data Science, University of Science and Technology of China
Hefei, Anhui, China
fengwenjie@ustc.edu.cn

Tianwei Zhang
College of Computing and Data Science, Nanyang Technological University
Singapore, Singapore
tianwei.zhang@ntu.edu.sg

Wei Zhou
National Pilot School of Software, Yunnan University
Kunming, Yunnan, China
zwei@ynu.edu.cn

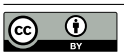
Xueqi Cheng
AI Safety of Chinese Academy of Sciences, Institute of Computing Technology, Chinese Academy of Sciences
Beijing, China
cxq@ict.ac.cn

See-Kiong Ng
Institute of Data Science, National University of Singapore
Singapore, Singapore
seekiong@nus.edu.sg

Abstract

With the surge and widespread application of image generation models, data privacy and content safety have become major concerns and attracted great attention from users, service providers, and policymakers. Machine unlearning (MU) is recognized as a cost-effective and promising means to address these challenges. Despite some advancements, image generation model unlearning (IGMU) still faces remarkable gaps in practice, e.g., unclear task discrimination and unlearning guidelines, lack of an effective evaluation framework, and unreliable evaluation metrics. These can hinder the understanding of unlearning mechanisms and the design of practical unlearning algorithms. We perform exhaustive assessments over existing state-of-the-art unlearning algorithms and evaluation standards, and discover several critical flaws and challenges in IGMU tasks. Driven by these limitations, we make several core contributions, to facilitate the comprehensive understanding, standardized categorization, and reliable evaluation of IGMU. Specifically, (1) We design CatIGMU, a novel hierarchical task categorization framework. It provides detailed implementation guidance for IGMU, assisting in the design of unlearning algorithms and the construction of testbeds. (2) We introduce EvalIGMU, a comprehensive evaluation framework. It includes reliable quantitative metrics across five critical aspects. (3) We construct DataIGM, a high-quality unlearning dataset, which can be used for extensive evaluations of IGMU, training content detectors for judgment, and benchmarking the state-of-the-art unlearning algorithms. With EvalIGMU and DataIGM, we discover that most existing IGMU algorithms cannot handle the unlearning well across different evaluation dimensions, especially for preservation and robustness. Data, source code, and models are available at https://github.com/ryliu68/IGMU.

*These authors made equal contributions to the paper.
†Wenjie Feng is the corresponding author and the project lead.



CCS '25, Taipei, Taiwan

ACM ISBN 979-8-4007-1525-9/2025/10
https://doi.org/10.1145/3719027.3744793

Warning: This paper includes explicit sexual content and other material that may be disturbing or offensive to certain readers.

CCS Concepts

• **Security and privacy → Social network security and privacy**; **Privacy protections**; **Social aspects of security and privacy**; **Software and application security**.

Keywords

Image Generation Model, Machine Unlearning, AI Safety, Unsafe Mitigation, Benchmarking

ACM Reference Format:
Renyang Liu, Wenjie Feng, Tianwei Zhang, Wei Zhou, Xueqi Cheng, and See-Kiong Ng. 2025. Rethinking Machine Unlearning in Image Generation Models. In *Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security (CCS '25), October 13–17, 2025, Taipei, Taiwan.* ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3719027.3744793

1 Introduction

Recent advancements in image generation models (IGMs) have garnered widespread attention for their ability to produce images from textual, visual, or multimodal prompts. Among these, Stable Diffusion (SD) [35] represents a groundbreaking innovation and has emerged as a leading choice for generating diverse, high-fidelity images, ranging from photorealistic scenes to imaginative artworks. Despite this impressive potential, IGMs also pose new ethical, societal, and safety concerns: they can produce unsafe or undesirable content, significantly hindering their practical deployment [24] and legal compliance [42]. For instance, generating copyrighted artistic styles without authorization has fueled debates around intellectual property and copyright infringement [55]; generating harmful or biased content poses risks to societal norms and safety standards [38]. These highlight the urgent need for robust solutions to ensure the ethical and responsible use of IGMs.

Machine Unlearning (MU) [1] emerges as a promising solution to mitigate the generation of unsafe content. MU has been extensively studied in classification models [14]. When applied to IGMs,

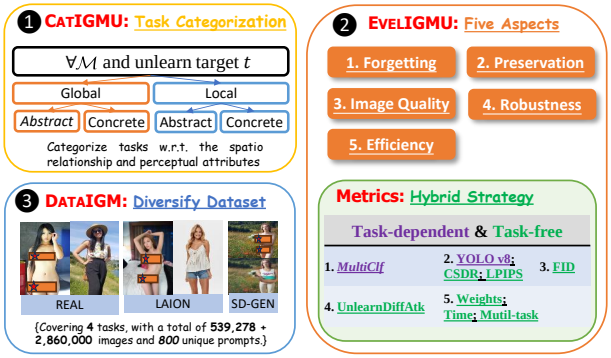


Figure 1: Core components of IGMU. ❶ CatIGMU: a framework for unlearning task categorization and definition. ❷ EvalIGMU: a framework for evaluating IGMU algorithms with various metrics at both task-specific and general-purpose measurement. ❸ DataIGM: a new dataset for exploring existing evaluation measures, training new content detectors, and benchmarking SOTA unlearning algorithms.

MU is typically instantiated as a *concept-removal* task that seeks to erase sensitive or objectionable targets, such as specific styles, objects, or harmful content while preserving the model's ability to produce high-quality, benign outputs [13, 23, 30]. For instance, image generation model unlearning (IGMU) can eliminate artistic styles like *Van Gogh* from a SD model. Then the output of a prompt p = *"A Van Gogh style picture about a man walking through wheat fields"* will align with that of p' = *"A picture about a man walking through wheat fields"* with similar image quality.

A number of image generation model unlearning (IGMU) algorithms have been proposed. To gain a deep understanding of these solutions, we conduct a series of empirical studies, to assess their implementations, measurements, and effectiveness. Unfortunately, we discovered several major limitations. (1) *Non-specific categorization of unlearning tasks.* Existing methods address either broad concepts or specific unlearning tasks, but do not carefully distinguish them [11, 52, 61]. The lack of a systematic definition and categorization of unlearning tasks can result in great inconsistencies and ambiguity in task analysis. (2) *Undetermined goals of unlearned model.* Improper expectations of the unlearned model greatly hinder the distinction of unlearning tasks and the design and evaluation of accurate unlearning algorithms. (3) *Unreliable evaluation metrics.* The metrics adopted in these works, including both task-specific and general-purpose ones, cannot accurately reflect the unlearning impact. These three limitations undermine the understanding and evaluations of existing unlearning solutions, and could mislead the design of new methods.

To address these challenges, we present a systematic and comprehensive investigation towards IGMU. As shown in Figure 1, we make three major contributions, to facilitate a comprehensive understanding, standardized categorization, implementation guidance, and reliable evaluation of IGMU. First, we design **CatIGMU**, a hierarchical framework that provides fine-grained unlearning task categorization and definitions of unlearning goals. It categorizes unlearning targets from two perspectives: spatial-scope relationship and perceptual attributes. This provides detailed guidance for the design and implementation of unlearning algorithms. Second, we design **EvalIGMU**, a holistic and robust evaluation framework. It integrates more accurate metrics across five critical aspects: *Forgetting*, *Preservation*, *Image Quality*, *Robustness*, and *Efficiency*. Third, we curate a dataset, **DataIGM**, from diverse sources. It contains high-quality samples tailored to IGMU, covering different scenarios. This dataset serves as an important foundation for unlearning performance benchmark, and constructing content detectors. Leveraging **EvalIGMU** and **DataIGM**, we benchmark ten state-of-the-art unlearning methods, demonstrating that current IGMU methods struggle to achieve satisfactory performance across these evaluation dimensions, particularly in preservation and robustness.

Our contributions can be summarized as follows:

- We design **CatIGMU**, a systematic framework for categorizing unlearning tasks based on spatial-scope relationships and perceptual attributes. It benefits unlearning algorithm design and evaluation benchmark construction.
- We propose **EvalIGMU**, a holistic evaluation framework equipped with refined metrics, including Multi-head Classifiers, CSDR, and other reliable measures, to evaluate unlearning performance across five critical aspects.
- We curate a high-quality dataset, **DataIGM**, incorporating multi-source data, including real-world and generated images, to train the more reliable multi-classification content detector and to evaluate the effectiveness of widely-used unlearning methods across various tasks.
- Leveraging **EvalIGMU** and **DataIGM**, we conduct extensive re-evaluations of ten state-of-the-art unlearning methods. Our results reveal critical shortcomings of these methods, particularly in achieving accurate unlearning, preserving unabridged benign content, maintaining high image quality, and ensuring robustness.

2 Background and Related Work

2.1 Conditional Image Generation

Multimodal large models have revolutionized artificial intelligence by seamlessly integrating multiple modalities (e.g., text, images, audio), driving unprecedented advancements in understanding and generating diverse content [56]. By leveraging cross-modal interactions, these models produce high-quality outputs tailored to complex scenarios and applications. As an important member, image generation models [35] excel at transforming noise into high-quality images guided by signals from various modalities.

Recently, diffusion-based models, represented by Stable Diffusion [35], have achieved unparalleled efficiency and quality by performing diffusion in lower-dimensional latent spaces. Leveraging pre-trained encoders like CLIP [39], Stable Diffusion outperforms in a variety tasks like image synthesis [35], and super-resolution [49]. Its adaptability, photorealistic quality, and output diversity have solidified its dominance in image-generation research [10].

Proxy-based methods are commonly adopted to assess the performance of $\mathcal{M}$ via specific metrics on a subset of generated images over prompts $\{p_i\}$. `CLIP Score` [20] measures the alignment by the cosine similarity between the CLIP embeddings of text prompts and generated images; `FID` (Fréchet Inception Distance) [21] quantifies the distributional similarity between generated and real images in

the feature space; LPIPS (Learned Perceptual Image Patch Similarity) [58] evaluates the perceptual similarity based on deep feature representations and correlates well with human judgments.

2.2 Image Generation Model Unlearning

Model Unlearning (MU) is a technique to erase the influence of specific subsets of training data from a trained model in an effective and economical manner [1, 5, 40]. For image generation models, it can remove specific styles, objects, or harmful content (e.g., sexual or violent elements) from a well-trained model, making it incapable of generating images containing such content without affecting its ability to produce other target-free images.

Various machine unlearning algorithms have been developed specifically for image generation models. Early efforts, such as adversarial training, aim to reduce the model's sensitivity to specific features. For instance, Wang et al. [50] modified latent representations to diminish the influence of specific text embeddings. More recent methods, such as target concept forgetting [19, 57] and model editing [13, 15], aim to disentangle and suppress undesired content in the latent space by fine-tuning models with counterfactual prompts or images explicitly designed to exclude target elements.

ESD [12] fine-tunes U-Net using negative guidance, aligning the probabilities of the target concept with a null string to steer predictions away from the erased concept. By focusing on the local components in U-Net for higher unlearning efficiency, researchers designed several methods that only edit the cross-attention layers [13, 15, 30, 57]. Specifically, UCE [13] optimizes the projection matrices using closed-form editing techniques, which encourage the model to refrain from embedding residual information of the target phrase into other words, thereby removing traces of the given target in the prompt. RECE [15] achieves concept erasure by iteratively performing model editing and embedding derivation. FMN [57] minimizes the attention weights corresponding to the target concept, gradually making the model disregard the concept during image generation. To improve the robustness of model unlearning against adversarial attacks that induce regenerating forgotten content via crafted prompts, AdvUnlearn [60] integrates adversarial training for the text-encoder layer. Similarly, SafeGen [29] targets internal model representations to mitigate explicit content and ensure the ethical alignment of outputs.

2.3 Unlearning Evaluation

Existing works adopt diverse standards or principles to evaluate the effectiveness of unlearning algorithms, which are normally specific to the erased target (e.g., nudity, artist style, objects). They utilize specific deep learning classifiers or detectors to measure the unlearning effects and adopt some commonly used metrics for assessing the model's ability.

Content Detectors. The task-related measurements include Style Classifier [61], Nude Detector [37], Q16 [45], GCD [7], et al.

Specifically, Style Classifier is fine-tuned from the ViT model [8] on the WikArt [43] dataset to recognize artist styles; Nude Detector is trained on a custom-collected dataset to identify various nudity types, e.g., "BUTTOCKS_EXPOSED" and "ANUS_EXPOSED"; Q16 leverages the zero-shot capability of the CLIP model [39] to detect harmful content, including sexual and violence; GCD is an open-source detector for identifying celebrities. Additionally, ResNet-50 [18] and YOLO [25] are also employed to recognize common objects. Their detection accuracy of the images from the unlearned model represents the unlearning performance.

Metrics. Some metrics adopted in image generation are used to evaluate the performance of the unlearned model. E.g., Frechet Inception Distance (FID) [22] assesses the quality of generated images; LPIPS [59] evaluates the perceptual consistency between the generated images and given anchor images; CLIP Score [20] measures the semantic alignment between the generated images and text prompts; CLIP Accuracy [53] quantifies the model's ability to distinguish outputs between the target and anchor prompts.

It is also important to assess the robustness of the unlearning algorithm, i.e., whether the forgotten content can re-emerge or is still retained. This can be achieved with adversarial attacks [4, 61] and membership inference attacks [51]. Some methods, like UnlearnDiffAtk [61], P4D [4], PUND [16], Ring-A-Bell [48] and CCE [34], systematically examine the residual traces of the forgotten concepts. Those studies verify whether the forgotten content still reappears when triggered by carefully crafted adversarial prompts.

Benchmark. UnlearnCanvas [62] and CPDM [32] evaluate the performance of the unlearned models by constructing a benchmark dataset to evaluate how unlearned models can forget certain targets they've learned. Recently, Ren et al. [41] introduced a Six-CD benchmark to evaluate existing unlearning methods across six tasks: "harm", "nudity", "identities of celebrities", "copyrighted characters", "objects", and "art styles". However, this benchmark only focuses on the unlearning and retention aspects (including in-prompt and out-prompt) while overlooking others, such as image quality, robustness, etc. Its evaluation relies heavily on existing detectors and classifiers. Therefore, the evaluation reliability and confidence are limited by the accuracy or potential issues of these measurements, which will be validated by our subsequent empirical study.

3 Preliminary & Formalization

3.1 Image Generation

Formally, a deep image generative model can be represented as $\mathcal{M} : \mathcal{P} \to \mathbb{P}(\mathcal{I})$, where the input $\mathcal{P}$ can be text strings, latent codes (e.g., random noise) or conditional signals (e.g., gender, class label, or reference images), and $\mathbb{P}(\mathcal{I})$ is the power set of the set $\mathcal{I}$, for each $p \in \mathcal{P}$. $\mathcal{M}(p)$ generates an image subset of $\mathcal{I}$ satisfying specific requirements, including content alignment and image quality.

3.2 Image Generation Model Unlearning

Consider an image generation model $\mathcal{M}$ with learnable parameters Θ and the target content to be forgotten $\mathcal{T} \subset \mathcal{P}$, MU intends to prevent $\mathcal{M}$ from generating images or content related to $\mathcal{T}$ achieved by applying some unlearning algorithm $\mathcal{A}_u$: $\mathcal{A}_u(\mathcal{M}, \mathcal{T}) \to \mathcal{M}_u$. The unlearned model $\mathcal{M}_u$ should satisfy the following requirements:

R1 (**Forgetting**): $\forall t \in \mathcal{T},\ \mathcal{M}_u(t) \cap \mathcal{M}(t) = \emptyset$;
R2 (**Preservation**): $\forall p \in \mathcal{P} \setminus \mathcal{T},\ \mathcal{M}_u(p) \subseteq \mathcal{M}(p)$,

In many scenarios, **R2** can be relaxed as $sim(\mathcal{M}_u(p), \mathcal{M}(p)) \geq \sigma$, where $sim(\cdot, \cdot)$ is a similarity function and σ is a constant threshold.

The above requirements provide the conceptual formulation grounded in the fundamental principles of these two aspects. **Forgetting** matches with the fundamental expectation of MU and adopts the formulation adopted by recent concept-removal studies [15, 19, 29, 30, 57, 60], that is, there is no overlapping ($\emptyset$) for the generated images of the unlearned model $\mathcal{M}_u$ and the original model $\mathcal{M}$ in terms of the forgotten target concept $t \in \mathcal{T}$. **Preservation** establishes norms for the unlearned model from the perspective of the model generation and generalization abilities by considering the non-target objects. In practice, task-specific detectors (e.g., style or nudity classifiers) provide a feasible quantitative proxies for these criteria; their detection accuracy provides an intuitive measure of forgetting and preservation quality that aligns with real-world visual-perception and quantitative-evaluation needs.

However, the practical implementation of the ideal conditions (**R1** & **R2**) varies greatly across tasks due to properties of unlearning, e.g., it is non-unique, task-wise, and even subjective. Therefore, following such basic requirements, we provide detailed categorization, analysis, and guidance in Sec. 3.2 and sampling-based evaluation implementations in Sec. 5.3.

Particularly for Text-to-Image models in our primary considerations, we provide more fine-grained and complete notions for the long prompt for MU, where the target unlearning content t is only part of the prompt text: it can be denoted as $S \oplus T$ where S is the remaining part of the prompt not related to the target forget, T is a placeholder for text-described target forget, and $\oplus$ denotes the union of strings; it corresponds to the complete target prompt when $T \leftarrow t$, i.e., $S \oplus t$. In the rest of this paper, we use t to denote $S \oplus t$ for simplicity if this does not cause any ambiguity. We use BROWN to highlight the target content in the text to be unlearned/erased. These could include *Artist Style* with the prompt *'A Van Gogh style picture about a man walking through wheat fields'*, *Object* with the prompt *'A red apple on the table'*, or *Harmful Content* with the prompt *'A naked girl playing on a beach'*, etc.

4 Empirical study

4.1 Machine unlearning tasks for IGM

Current methods [12, 13, 30, 44, 47, 57] cover different unlearning tasks for IGM, including harmful content, copyright-related (e.g., artistic styles), and privacy (e.g., individual faces like 'Donald Trump'). However, the same unlearning tasks may be referred to and categorized differently across various works. For example, ESD [12], SPM [31], and AdvUnlearn [60] refer to IGMU as "Concept Erasure" and encompass the unlearning of Style (e.g., Van Gogh), Object (e.g., Church or Parachute), and Concept (e.g., Nudity). Besides, other works adopt varied terminologies and classifications. For instance, AC [27] refers to objects as "specific object instances"; FMN [57], ConceptPrune [3], and MACE [30] label nudity unlearning as "Explicit Content", while other methods, such as KPOP [2] describes it as "unethical content", RECE [15] and SLD [44] classify it as "inappropriate concepts", SDD [26] and CCE [34] group it under "NSFW". Therefore, such diversity of tasks and inconsistency in task naming and categorization result in great confusion in the definition and understanding of unlearning tasks, and fail to reveal and explain their nature and differences accurately.

What is the relationship between the unlearning of "Van Gogh style", "nude girl", and "Blue Sky"? From the perspective of forgetting and preservation, can the same unlearning method be applied to different unlearning tasks like an "Apple", "Donald Trump", a "Spiderman", and a "Rabbit"? As we can see, those case-by-case and task-dependent methods fail to capture the essential relationships (differences and similarities) behind a wide range of different task instances, either making the unlearning methods not universally applicable or generalizable, or causing unnecessary costs of repeated discovery due to a serious underestimation of their capabilities. Furthermore, it hinders the guidance and implementation of the unified design and consistent evaluation of unlearning algorithms.

> **Observation** 1. *Real unlearning requirements for IGM are diverse, and existing machine unlearning methods cover various unlearning tasks. However, there are lots of inconsistencies and even conflicts in naming and categorization for those tasks; existing methods usually solve them in a case-by-case and task-dependent manner, without considering the relationship between tasks and the generalization of the methods.*

4.2 Unlearning Achievement and Expectation

R1-**R2** in Sec. 3.2 provide the basic requirements for the goal of the unlearned models $\mathcal{M}_u$s; their implementation counterparts show the content and form of the actual image, and can correspond to the user's expectations of $\mathcal{M}_u$s behavior, which can be a resource (e.g., training data) for those content detectors and the ground truth as a reference for accurately evaluating $\mathcal{M}_u$'s performance.

Taking the possible expectations we can conceive & design and the possible outputs of the existing works into consideration, Figure 2 exhibits specific examples for some random selected unlearning tasks, including "nude girl", "Van Gogh style", and "parachute", and five state-of-the-art unlearned models $\mathcal{M}_u$ for the case study. Specifically, Fig. 2(a) gives the prompt '$S \oplus t$' highlighted with unlearning targets 't' and the corresponding output of the original model $\mathcal{M}$; Fig. 2(b) exhibits some possible outcomes for $\mathcal{M}_u$ that we can conceive and construct that satisfy the previous requirements, the detailed explanation is deferred to Sec. 5.2; Fig. 2(c) shows that the output of the unlearned model based on the unlearning methods, including ESD [12], MACE [30], Receler [23], UCE [13], and SPM [31]. The possible outcomes in Fig. 2(b) include:

- **Expectation**: They are designed to erase the target unlearning content related to 't' while preserving other elements in the original image as much as possible. Multiple different results meet the requirements for each case.
- **Default**: They are the pre-set default placeholder of a model. Here, we take a black image as an example.
- **NULL**: They are the generated images of $\mathcal{M}$ corresponding to the prompt 'S', i.e., removing the target 't' and its associations parts in the original prompt.
- **Replace**: They are any real images unrelated to 't' or any generated images of $\mathcal{M}$ corresponding to the prompt '$S \oplus T$' with $T \neq t$, i.e., replacing the target t in the original prompt with other unrelated content.
- **Any**: They are any real images unrelated to 't' or generated images of $\mathcal{M}$ for any prompt that does not contain 't'.

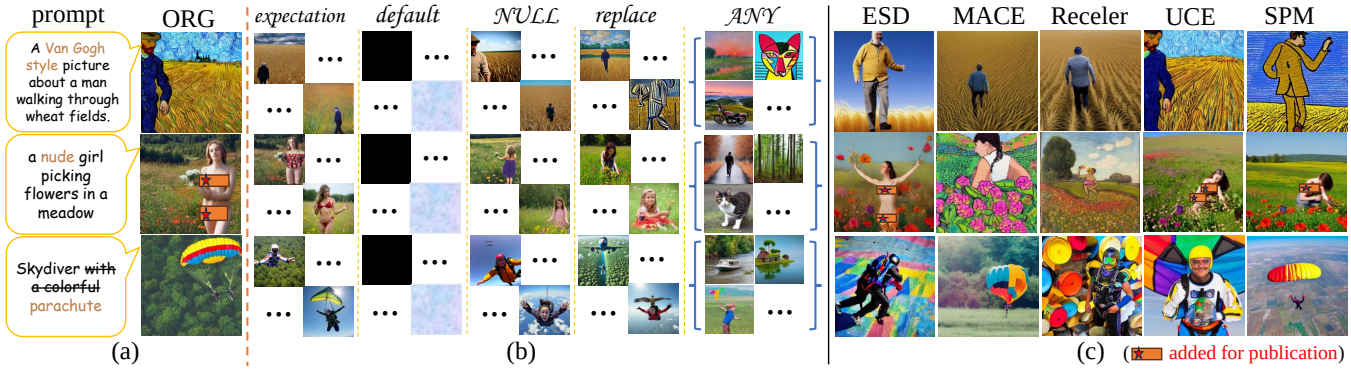


Figure 2: Case study showcasing various post-unlearning candidates and images from existing unlearned models. Columns from left to right: (a) prompts and corresponding images generated by $\mathcal{M}$; (b) the expected outputs alongside other potential outputs for post-unlearning; (c) images generated by different unlearned models. From top to bottom, these cases include Nudity, Van Gogh, and Object (parachute) unlearning. Text highlights the target word(s) t to be unlearned and ~~Text~~ indicates the associations that should be erased together with the target word(s) during the unlearning process, i.e., 'S'.

Expectation is carefully constructed, following the requirements seriously, for each generated image of interest; it is task-dependent and has high quality; **Default** is some easy-configured but trivial outcomes; **NULL** can generate diverse images with a high probability, but it is still possible to generate images related to '$S \oplus t$'; **Replace** explicitly excludes content related to 't' and keeps the remaining for 'S', but it may not completely faithfully retain or present the elements related to 'S' in the original generated image in the same way, because it is not as finely controlled as Expectation; **Any** can explicitly exclude content related to 't', but does not make clear constraints on the preservation content. It can be seen that ***Expectation*** *is the only one that perfectly and accurately meets the expectation for the behavior of the corresponding unlearned model* $\mathcal{M}_u$. Besides, the above analysis's conclusions are consistent with our observations for our considered cases in Fig. 2(b), and the **Expectation**, as the best one, also has great differences for different tasks: its effect (forgetting or preservation) on the image may be on local elements (e.g., "nude girl") or all elements of the entire image (e.g., "Van Gogh style").

Those existing works in Fig. 2(c) covers different unlearning strategies, including remapping and alignment of the embedding spaces or directed updates of the model parameters. From the multi-perspective evaluation of their results, it can be seen that the average performance on different tasks is different overall, perhaps reflecting the difficulty of the task. For example, judging from the visual, the performance of most methods for the unlearning of "Van Gogh style" is better than the other two tasks; and the output results of different methods for the same task are quite different, and each method has inconsistent performance on different tasks. In addition, they suffer from the following issues: either the target content of 't' is not forgotten, such as ESD, UCE, and SPM on "nude girl" unlearning, or the content related to S is not accurately preserved, such as ESD and Receler on "parachute" unlearning, or the quality of the generated image is seriously damaged, such as Receler and UCE on "nude girl" and "parachute" unlearning tasks.

> **Observation** 2. *The goal of the unlearned model* $\mathcal{M}_u$ *and expectation for its behavior are task-dependent, and '**Expectation**' becomes the only one that well satisfies all requirements (**R1-R2**). Existing unlearning methods have inconsistent performance over different tasks and suffer from various issues related to forgetting, preservation, image quality, etc.*

4.3 Evaluation of the IGMU Evaluation

For those unlearning tasks commonly used in existing SOTA works [12, 15, 30, 31, 53, 57, 60], that is, unlearning of Nudity (e.g., nude girl), Artist Style (e.g., Van Gogh style), and Object (e.g., parachute and church), as demonstrated in Figure 2, we quantitatively evaluate the reliability of the widely used evaluation methods (i.e., Content Detectors) and metrics. To this end, based on the discussion about the expectation of an unlearned model in Sec. 4.2, we build a well-curated test bed, DataIGM (detailed information is given in Sec. 5.4), by integrating multi-source data.

Evaluation testbed. For each task, the construction of DataIGM consists of the following three parts:

- *REAL*: They are the real dataset used to train the corresponding detectors (including training and test/validation sets), e.g., WikiArt for Style Classifier.
- *LAION*: They are (part of and randomly selected) the real dataset used to train the original model $\mathcal{M}$.
- *SD-GEN*: They are the generated images corresponding to $\mathcal{M}(S \oplus t)$ and $\mathcal{M}(S)$.

In terms of data split, the training (test/validation) set of REAL belongs to the training (test) of DataIGM. In the following subsections, only the corresponding test set for each task of DataIGM is used for evaluation.

4.3.1 Style Unlearning. Style Classifier (SC) is widely adopted to detect specific artist styles [16, 61, 62]. We use Accuracy, Precision, Recall, and F1 Score as metrics to evaluate SC for the unlearning of Van Gogh style on DataIGM, and the result is shown in Fig. 3.

SC achieves the best results on *REAL* across all metrics, but performs poorly on *LAION* and *SG-GEN* except for Precision, although

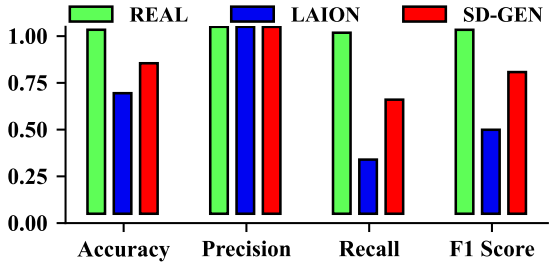


Figure 3: The performance of Style Classifier on DataIGM.

Table 1: The evaluation results of NudeNet and Q16 for Nudity unlearning task on DataIGM.

Evaluator	Data	Accuracy	Precision	Recall	F1 Score
NudeNet	REAL	74.68	98.96	49.89	66.34
	LAION	75.18	98.65	51.07	67.30
	SD-GEN	78.31	99.58	56.86	72.89
Q16	*REAL*	60.76	69.24	38.74	49.68
	LAION	59.98	87.48	23.30	36.80
	SD-GEN	52.51	88.38	5.78	10.85

the results on SG-GEN are slightly better than those of *LAION*. Therefore, although SC maintains its performance when applied to REAL that is I.I.D. with its training data, it does not have strong generalization ability over other real data (i.e., *LAION*) and generated images (i.e., *SD-GEN*), which may result from its overfitting to training data or the distribution shift between its training set and test sets here. Thus, SC has serious drawbacks when applied to a wide range of unlearning model evaluations, which will lead to unreliable results.

4.3.2 Nudity Unlearning. **NudeNet** [37] as detector and **Q16** [45] as classifier are widely adopted for evaluation [4, 12, 15, 34, 36, 44, 48, 54, 61]. **NudeNet**[1] can detect explicit body parts[2]; **Q16** is designed to label images as either safe or harmful (including nudity). Table 1 summarizes the evaluation results of them on DataIGM[3]

NudeNet achieves near-best Precision but has poor performance across other metrics, where its recall is only near 50%; there is no significant difference in its performance between different data. In addition, it is also challenging to detect all sensitive parts in the image accurately and completely. **Q16** has inferior performance across all metrics and all data, its Accuracy is no more than 61% and Recall is less than 40%. Among the three data, it performs best on REAL and worst on SG-GEN — its Accuracy is near to random guessing and Recall less than 6% although with the best Precision, such a result may be attributed to the distribution shift of the generated data compared to the real data. Therefore, **NudeNet** and **Q16** have significant limitations that prevent their applicability in evaluating the performance of the unlearned model for the Nudity unlearning task. Similarly, the evaluation conclusions about the Nudity unlearning tasks are unreliable and inaccurate.

[1]https://github.com/notAI-tech/NudeNet.

[2]Following common settings in existing works, the sensitive parts include "MALE_BREAST_EXPOSED", "MALE_GENITALIA_EXPOSED", "FEMALE_BREAST_EXPOSED", "FEMALE_GENITALIA_EXPOSED", "BUTTOCKS_EXPOSED", and "ANUS_EXPOSED".

[3]Here, REAL of DataIGM only contains the training and test set used for NudeNet training, while the data corresponding to Q16 are not publicly available.

Table 2: The detection results of ResNet-50 on DataIGM.

Task	Data	Accuracy	Precision	Recall	F1 Score
Parachute	*REAL*	99.00	100.00	98.00	98.99
	LAION	91.75	100.00	83.50	91.01
	SD-GEN	85.05	99.86	70.20	82.44
Church	*REAL*	87.00	100.00	74.00	85.06
	LAION	78.10	100.00	56.20	71.96
	SD-GEN	87.90	100.00	75.80	86.23

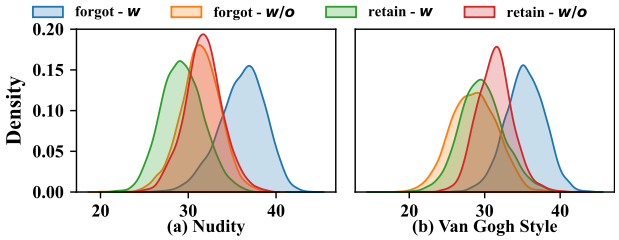


Figure 4: CLIP Score distribution for Nudity unlearning and Van Gogh style unlearning. Here, *w* indicates the presence of the target word(s), while *w/o* denotes its absence.

4.3.3 Object Unlearning. **ResNet-50** [18] pre-trained on ImageNet [6] is used to detect the target object in an image for the success of erasing [11, 13, 51, 52, 61]. The unlearning of church and parachute are considered here, and the results are listed in Table 2.

ResNet-50 achieves near the best Precision for both tasks, but there exist obvious differences in its performance on these two tasks across other metrics — performance for parachute unlearning is better than for church unlearning. In addition, for church, the results for *REAL* and *SD-GEN* are similar, but significantly better than *LAION*, which indicates that the previous two data resources have similar distribution while *LAION* has a great distribution shift leading to poor adaption; for parachute, due to the distribution shift between different real data and between real data and generated data, there are differences in performance on different data: REAL performs best (i.e., well-generalized for I.I.D. data), *LAION* is second, and *SD-GEN* is the worst. Therefore, **ResNet-50** has inconsistent performance across different tasks and cannot be well suited for out-of-distribution data and generated data from Objects unlearning, which will lead to unreliable evaluation results.

4.3.4 Task-Free Metrics. **CLIP Score** and **CLIP Accuracy** is widely adopted by existing works[27–31, 52, 53, 60] and can be used to evaluate the unlearning effectiveness *w.r.t.* **R1**-**R2**.

For CLIP Score, we conducted experiments with four types of text-image pairs for each unlearning task. Specifically, we label the images generated by $\mathcal{M}$ with $S \oplus t$ and S as "forgot" and "retain", respectively. We then calculated the CLIP Scores for two text prompts, $S \oplus t$ and S, with the "forgot" and "retain" images, respectively. Thus, it leads to four categories: forgot-w, forgot-w/o, retain-w, and retain-w/o, where w means with the target word(s) t, w/o means without the target word(s) t. Fig. 4 shows the distributions of these four categories for the unlearning tasks of "Nudity" and "Van Gogh style", and there is substantial overlap among them,

Table 3: The CLIP Accuracy results on DataIGM.

Task	Data	Accuracy	Precision	Recall	F1 Score
Van Gogh style	*REAL*	66.40	68.75	60.10	64.16
	LAION	70.24	71.63	67.00	69.26
	SD-GEN	87.03	94.37	78.80	85.86
Nudity	REAL	66.05	68.54	59.30	63.61
	LAION	82.07	85.03	77.80	81.27
	SD-GEN	77.92	69.54	99.40	81.82
Parachute	*REAL*	100.00	100.00	100.00	100.00
	LAION	99.40	98.81	100.00	99.40
	SD-GEN	98.55	98.12	99.00	98.56
Church	*REAL*	100.00	100.00	100.00	100.00
	LAION	99.90	99.90	99.90	99.90
	SD-GEN	99.45	99.80	99.10	99.45

which indicates that the CLIP Score is insufficient for reliably determining whether generated images still retain target content.

For CLIP Accuracy, we conduct experiments on the test set of DataIGM across four tasks and specify the used (target, anchor) prompt pair of each task as: ("nude", "properly dressed") for Nudity unlearning, ("with Van Gogh style", "without Van Gogh style") for Style unlearning, ("church", "bird") and ("parachute", "bird") for Object unlearning. As the results in Table 3 show, CLIP Accuracy performs well for the Object unlearning tasks and clearly distinguishes between the generated images and the target and anchor prompts. In contrast, it becomes less effective for the rest of the unlearning tasks, particularly on the *REAL* data. For Van Gogh style, among three data, *SD-GEN* performs best, *LAION* is second, and *REAL* is the worst; while for nude, *LAION* performs best, *SD-GEN* is second, and *REAL* is the worst. Such results reflect the inconsistent influence of different types of data on CLIP Accuracy and its instability for different tasks. Therefore, it is also not a widely applicable and stable evaluation metric.

Observation 3. *Those task-dependent unlearning evaluation measurements (content detectors) have major flaws to varying degrees, and their performance is inconsistent across tasks or even for different instances of the same task; the resultant evaluation results based on them are inaccurate and unreliable due to the distribution shift of the test data. Those task-free metrics (CLIP Score and CLIP Accuracy) also cannot provide reliable evaluation for unlearned models; their results are neither sufficient nor consistent between different unlearning tasks, which makes them not widely applicable.*

5 Our Contributions

As summarized in the key observations in Sec. 4, those issues of the existing works involve the unlearning task itself, the behavioral expectations of the unlearned models, and the measurement and evaluation metrics. They will cause great obstacles and difficulties for researchers to accurately understand the unlearning mechanisms and evaluate the proposed algorithms, seriously undermining the reliability of conclusions drawn from past studies.

It becomes increasingly urgent to establish a comprehensive and standardized study for the analysis and evaluation of IGMU. We introduce two complementary frameworks to address this need and build a carefully curated dataset as a reference for the common unlearning tasks. Specifically, (1) CatIGMU is a hierarchical framework that categorizes and unifies different types of unlearning tasks and their interrelationships (Sec. 5.1) and provides detailed implementation guidance (Sec. 5.2). (2) EvalIGMU is a holistic evaluation framework for accurately assessing diverse unlearning algorithms (Sec. 5.3) and provides a detailed implementation by incorporating effective metrics. (3) DataIGM is a multi-sourced dataset across commonly-used unlearning tasks as a test-bed implementation based on CatIGMU. It could be used for multiple purposes: analyzing content detectors and quantitative metrics in existing IGMU works (Sec. 4.3); training new and reliable detectors for unlearning tasks; and evaluating existing IGMU algorithms (Sec. 5.4). Below, we present detailed descriptions for the design of each contribution.

5.1 CatIGMU Framework

We propose CatIGMU, a comprehensive hierarchical framework to categorize and unify those complicated unlearning tasks for IGM, Figure 5 illustrates the overview of the framework and specific task instances, additional cases are presented in Table 5.

Design Principle: To accurately determine and differentiate those complicated tasks of unlearning for IGM and to explore the essential 'order world', we decompose and categorize them following a hierarchical structure based on the following criteria,

(1) **Spatial Relationship**: investigate whether the forget target fills [4] the entire limited image canvas and thereby divides it into *Global* and *Local* at the first tier.
(2) **Perceptual Attribute**: according to the perceivable nature of the forget target, it can be divided into *Abstract* and *Concrete* at the second tier.
(3) **Tasks & Semantic Association**: aligning with existing unambiguous and explicit tasks in IGM, it can be further refined into *Style, Entity, Status, Property, Collective Concept* [5] at the third tier. Meanwhile, it considers the semantic associations between the forget target as an adjunct and its associated subjects in the grammar of the prompt text.

Therefore, we not only consider the context in which the target is located but also include the properties of the target itself as matter and the semantic associations for nature language in text prompts, which derives distinctive unlearning tasks for different targets in different environments within a limited space. As a result, we achieve a hierarchical categorical framework through the combination of mutually exclusive dimensions, which can also be flexibly expanded to include other new tasks that may arise in the future. Consequently, such a comprehensive categorization framework will benefit the unification & differentiation of various unlearning tasks, the design of new unlearning algorithms, and the comparison of different unlearned models for accurate performance evaluation.

Framework Interpretation: Here, we elaborate on the details of the CatIGMU framework and its components. Given any target content with prompt $t \in \mathcal{T} \subset \mathcal{P}$ to unlearn, we determine the category of

[4] In practice, we can use a threshold-based approach to determine the proportion of the target content in the entire image canvas for its belonging here.
[5] Although we have tried our best to avoid overlap for this category division, there may still be some task that belongs to multiple categories, which stems from the complexity and subjective judgment of natural language semantics. Besides, those unlearning tasks can also be flexibly extended and added.

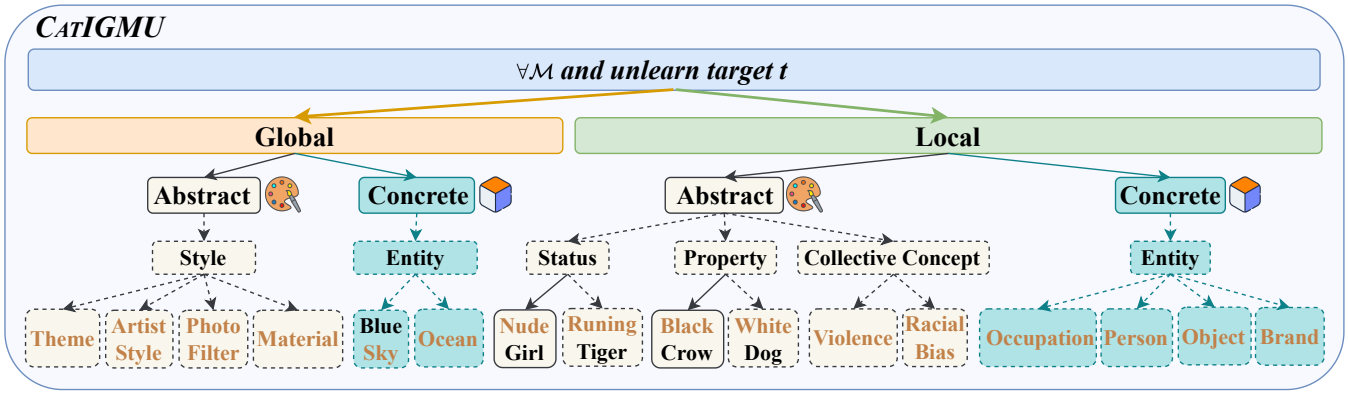


Figure 5: Overview of CatIGMU framework. Based on the spatial relationship between the unlearning target t and image canvas (global vs. local), the perceptual attributes of the unlearning target t (abstract vs. concrete), and different unlearning tasks (style, object, identity, etc.), CatIGMU constructs a hierarchical structure for systematic analysis and clear differentiation. Each task is instantiated at the leaf node, and the unlearning target t is highlighted with the BROWN text; solid (dashed) frames indicate strong (weak/near-independent) semantic association between the forget target and its associated subjects in prompt text, especially for *'Local'-'Abstract'*.

the corresponding unlearning task by traversing CatIGMU from top to bottom to facilitate the subsequent process, e.g., evaluation.

Investigating the spatial relationship, it will divide into two branches: *Global* and *Local*. The *Global* section considers the situation where the target content (almost, if not all) covers the entire canvas, and its forgetting would have an overall impact on the entire image. At the next layer down this branch, the *Abstract* can be used to represent a kind of *Style*, a general class, including `Theme` and `Style`, `Material`, `Photo Filter`, etc.; the *Concrete* covers tangible *Entity* that span almost the whole scene reflected in the image, which can be `Blue-Sky`, `Ocean`, `Desert`, etc.

The *Local* section considers more common cases where the target content is only part of the image and usually accompanies other elements in the image; thus, its removal has a limited impact on the main content of the original image. At the next layer down here, the *Abstract* corresponds to some *descriptive Status*, *Properties*, *Virtual or Collective Concepts*, for example, `Nude Girl`, `White Dog`, and `Violence`. In addition, further analysis in terms of grammar and semantics shows that the target unlearning content here as an adjunct may be strongly bound to the associated object, which could be default missing; for example, *Nudity* can usually only refer to people; while other target content and the associated object are weakly related or nearly independent, such as the relationship between white and dog for `White Dog` unlearning. Therefore, we use solid and dashed frames to differentiate the above two cases. For the *Concrete*, it includes one type of perceptible things or people, such as *Occupation*, *Person*, specific *Objects*, *Brand*, and so on. For instance, it can be Doctor, Donald Trump, Church, Apple logo, etc.

Note that in addition to the above text description and definition, the differences and impacts of such categorization on subsequent processing will be elaborated in the next subsections.

5.2 Detailed Implementation Guidelines

We have the fundamental questions: *How should the unlearned model $\mathcal{M}_u$ perform? Does it really align with our expectations?*

Although the basic concept and requirements for machine unlearning for the image generation models are introduced in Sec. 3, their specific implementation and correspondence to specific tasks are still unclear, especially the determination and measurement of **forgetting** (**R1**) and **preservation** (**R2**) — this is *the essential difference among the above categories about unlearning tasks.*

In terms of forgetting, the format of the generated image of $\mathcal{M}_u(t)$ for any $t \in \mathcal{T}$ could be diverse across different tasks and even become non-unique; as for preservation, it could be defined at both set-level (images for other target-unrelated prompts) and sample-level (maintain remaining elements in the image except for the target unlearning content *w.r.t.* t). Therefore, these two aspects jointly determine the expected outcomes of $\mathcal{M}_u$ for various tasks.

Our hierarchical framework CatIGMU enables the task-dependent expectations for model $\mathcal{M}_u$'s behavior to become categorically integrated and consistent; it would also be conducive to the standardized solution of existing issues of model outcomes (i.e., Obs. 1-Obs. 2), including its complexity and non-uniqueness, as well as the inappropriateness and ambiguity in existing work.

For all categories, as one of the *trivial outputs*, $\mathcal{M}_u(t)$ can directly use *"cannot generate the content related to* [t]*"* as output content of the generated images, which is an over-simplified approach and will cause the loss of detailed information. Another *trivial one* is to indiscriminately *replace* the target content in the generated image *with Any* other element that is not related to t, which may cause great confusion in use and damage the user-friendliness of the model. Therefore, we have the following refined version.

❶ For the *'Global-Abstract'* category, w.l.o.g., we assume the target unlearning content plays a role of *transforming the real scene* [6]. Thus, as the preservation content, $\mathcal{M}_u(t)$ will be expected to the real scene before t's transforming for $t \in \mathcal{T}$ or after applying the other transforming except for t; for each sample, $\mathcal{M}_u$ should behave consistently with $\mathcal{M}$ for other prompts $p \in \mathcal{P} \setminus \mathcal{T}$.

[6] Admittedly, it is extremely difficult to determine whether such a transformation, especially the artist's style, is realism or abstraction — perhaps the artist himself cannot define herself, let alone those who have passed away. Moreover, the formation of a work of art is a complex process, which may be the result of the joint action of factors such as physical inspiration and artistic imagination. Therefore, the assumption here for the *Artist Style* is just a simplification.

❷ For the *'Global-Concrete'* category, compared to 'Abstract', the target unlearning content corresponds to *the overall physical object in the original real scene* in an image. Thus, when $\mathcal{M}_u$ really forgets, at the sample level, each output image of $\mathcal{M}_u(S \oplus t)$ should not contain the related content/objects corresponding to $t \in \mathcal{T}$, so it can be replaced by any other default placeholder image (e.g., any solid color image, random image, mosaic, blank, etc.). For the preservation, $\mathcal{M}_u(S \oplus t)$ should maintain the remaining part originally associating with/accompanying target forget content if it exists, e.g., *Ship* in the Ocean and *Cloud* in Sky can be generated in other backgrounds, except Ocean and Sky, respectively; $\mathcal{M}_u(p)$ follows those of ❶ for other prompts $p \in \mathcal{P} \setminus \mathcal{T}$.

For the *'Local'* section, one of the *trivial forgetting ways* is to *directly remove/erase the correspondence to $t \in \mathcal{T}$ completely* in the generated image, including the adjunct object itself and its affiliated attributes if it exists, especially for the long prompt, the case where the target unlearning t is only part of the *long prompt text*, for instance, *'A naked girl playing on a beach'* and *'A red apple on the table.'*, ... However, direct removal, in a simple and coarse-grained way, will result in significant changes in image content and might destroy the complete expression of the original prompt semantics. So we have a specific discussion under the different categories for *'Local'* as follows.

❸ For the *'Local-Abstract'* category, except for direct complete removal, we would like to provide a more detailed and fine-grained analysis at the sample level: for each image generated by $\mathcal{M}(S \oplus t)$, the forgetting can be implemented by removing or modifying (editing) the parts that are directly related to t while maintaining the remaining parts about S. For example, `Status` unlearning for the prompt *'A naked girl playing on a beach.'*, as the strong association case, can be achieved by dressing sensitive exposed parts [7] of the human body, thus satisfying both forgetting (i.e., naked) and preservation (i.e., a girl playing on a beach.). `White Dog` unlearning, as the independent association case, can be achieved by simply editing and changing the hair color of the object (Dog) in the image while maintaining the properties of all other aspects of the dog (i.e., actions, expressions, etc.) and other elements (i.e., background, scenery) in the image; `Violence` unlearning can be achieved by replacing or modifying the corresponding attributes or elements in the original generated image of $\mathcal{M}(S \oplus t)$ with some benign ones without affecting other elements and parts.

❹ For the *'Local-Concrete'* category, the target content related to $t \in \mathcal{T}$ is just (a small part(s) of) something concrete of a generated image in $\mathcal{M}(S \oplus t)$. Therefore, for each generated image, besides the direct removal, $\mathcal{M}_u(S \oplus t)$ can use another unrelated counterpart to replace the target corresponding to t or modify the content so that it is no longer (can't be identified as) the original entity while preserving other elements about S in the original image. For example, *'a church next to a river'*. Nonetheless, it is important to note that it is extremely difficult to determine which features are unique to the target content and which characteristics define and decide the target itself, which may also depend on society and human common but dynamic cognition, as well as the model's status; thus, this makes it difficult to determine the minimum scope of modification in the image in some cases. For instance, `Object` unlearning can be achieved by replacing the target Entity in the image in $\mathcal{M}(S \oplus t)$ with *various different alternatives*; `Person` unlearning may be influenced by the bias status of $\mathcal{M}$ for $t \in \mathcal{T}$ and *user's specification and the training data* of $\mathcal{M}$.

However, it must be admitted that in the above discussion, we are considering unlearning tasks that are widely covered in existing work, and there might be some that are not fully considered or unknown that are beyond the coverage of our CatIGMU. In addition, determining whether to completely forget or preserve is subjective to a certain extent and might be a matter of opinion; How to precisely implement the above expectations at the sample level, what measurement metrics should be used, and how to measure them accurately are also challenging problems. Although CatIGMU is carefully designed and strives to be perfected, it is still rudimentary and immature. We also expect that it can be continuously improved and enriched to include enough different unlearning tasks and scenarios for image generative models flexibly.

Furthermore, such data pair (the prompt $t \in \mathcal{T}$ and the expected content in the generated image) with clear correspondence for target unlearning content, as discussed above, will be critical for those fine-tuning-based unlearning approaches as feedback; it is also fundamental as a test bed or benchmark for the evaluation of different unlearning methods across various unlearning tasks.

Lemma 5.1. *Our CatIGMU framework and the above implementation guidelines can provide insights and specifications into the nature of the IGMU problem, and thus help in the design of unlearning algorithms $\mathcal{A}_u$ and the construction of related test beds or benchmarks.*

Prompts Variations. Except for the above-expectation behavior of $\mathcal{M}_u$ for the prompt "$S \oplus t$", under the causal inference paradigm [9], we probe other prompt variations by applying intervention operator to T of the text prompts $S \oplus T$. Consequently, we have,

- $\hat{P}$-1 $do(T =$ 'none'): just replace the target unlearning t and its associations with 'none,' i.e., only S remains in the prompt.
- $\hat{P}$-2 $do(T \neq t)$: just like the counterfactual setting by replacing t with others $\hat{t} \neq t$, leading to $S \oplus \hat{t}$, in the prompt.

$\hat{P}$-1 corresponds to an *Automatic Intervention*; $\hat{P}$-2 can be interpreted as a *Soft Intervention*. They are similar to the setting of **NULL** and **Replace** in Sec. 4.2, respectively. Given the defects and possible ambiguity of the outputs of $\mathcal{M}$ for the **NULL** and **Replace** strategies discussed previously, an ideal unlearned model $\mathcal{M}_u$ should forget the content about t completely and will no longer generate related images, no matter what prompt is used. Therefore, $\mathcal{M}_u(\hat{P}\text{-}1)$ and $\mathcal{M}_u(\hat{P}\text{-}2)$ should keep the same as the above-discussed $\mathcal{M}_u(S \oplus t)$.

Case Study. Taking specific downstream unlearning tasks corresponding to CatIGMU and Figure 5 as examples, Tab. 4 lists some instantiation of the above four categories (*Van Gogh style, Blue Ocean, Nudity (naked girl)*, and *Apple*) and summarizes the text-described expectations of the generated images of $\mathcal{M}_u$.

We enumerate examples (the target forgetton t) as the instantialization of each unlearning task outlined in EvalIGMU framework in Tab. 5.

[7] Note that the definition of "sensitive parts" here is complex and non-uniform, which depends on the constraints of different cultural, religious, social customs and other factors, so it also relies on the specific situation.

Table 4: Guidelines for what the expected outputs should be when implementing unlearning for different cases.

Target	Original Prompt: $S \oplus t$	Forgot	Output of $\mathcal{M}_u(S \oplus t)$	
			Expectation	Trivial solutions
Global Abstract Style	*A Van Gogh style picture about a man walking through wheat fields.*	Van Gogh style	A realistic or other artistic style picture about a man walking through wheat fields.	*"I cannot generate images with Van Gogh style."*
Global Concrete Object	*Blue oceans.*	The entire image content	A default image, i.e., a solid color or random noise.	*"I cannot generate the oceans."*
Local Abstract Status	*A naked girl playing on a beach.*	Nudity elements	A **clothed** girl playing on a beach.	*"I cannot generate the naked girl.".*
Local Concrete Entity	*A red apple on the table.*	Apple	1) Table[remove]. 2) Table with any other objects[replace].	"I cannot generate the apple".

Table 5: Examples of unlearning tasks and categories under the CatIGMU framework.

Tasks	Examples
Art Style	Van Gogh, Monet, Picasso, Cubism, Realism, ...
Theme	Nature, Fantasy, Sci-fi, Mythological Scenes, ...
Material	Wood, Stone, Steel, Glass, Plastic, Marble, ...
Photo Filter	Vintage, Black & White, Sepia, Warm Tone, ...
Color Tone	Warm, Cool, Monochrome, Pastel, Vibrant, ...
Descriptive Status	Blue Sky, Starry Sky, Sea, Desert, Clouds, ...
Descriptive Status	Nude Girl, Naked Person, Running Tiger, ...
Properties	Black Crow, Striped Zebra, White Dog, ...
Collective Concept	Racial Bias, Violence, Drugs, Harassment, ...
Occupation	Doctor, Nurse, Police Officer, Firefighter, ...
Person	Donald Trump, Elon Musk, Taylor Swift, ...
Object	Church, Parachute, Bird, Cup, Airplane, Car, ...
Brand Icon	Nike, Apple, Samsung, Google, Coca-Cola, ...

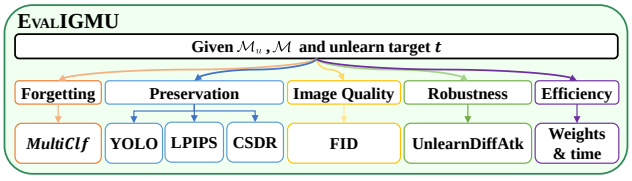


Figure 6: The overview and implementation of the proposed evaluation framework, EvalIGMU, for machine unlearning for IGM. It considers five aspects: *Forget, Preserve, Image Quality, Robustness,* and *Efficiency*. Some quantitative metrics are listed for each of them as one implementation of EvalIGMU.

5.3 Evaluation Framework & Implementation

Considering the major flaws in current measurements as shown in Sec. 4.3 and the urgent need for a comprehensive and accurate evaluation of machine unlearning methods, we propose EvalIGMU, *a holistic and systematic evaluation framework* for the unlearning algorithms for IGM and the unlearned model, i.e., $\mathcal{A}_u$ and $\mathcal{M}_u$.

Based on the requirements **R1**-**R2** of IGMU tasks in Sec. 3 and other requirements of IGMU, EvalIGMU considers the following five aspects. Figure 6 provides an overview and possible metric instances about EvalIGMU.

- **Forgetting:** Whether the forgetting requirement **R1** is met, i.e., whether the target forget content is completely removed in the generated images of $\mathcal{M}_u(t)$.
- **Preservation:** Whether the preservation requirements are met, i.e., whether the generated images of $\mathcal{M}_u(t)$ and $\mathcal{M}_u(p)$ retain the corresponding parts that should be retained accurately and completely.
- **Image Quality:** Evaluate the quality of the generated images from both $\mathcal{M}_u(t)$ and $\mathcal{M}_u(p)$, including the fidelity, aesthetics, and text-image alignment, etc. [17].
- **Robustness:** Measure $\mathcal{M}_u$'s resistance against regenerating any unlearned content related to $t \in \mathcal{T}$ for some malicious inputs design, e.g., text-perturbation or jailbreak.
- **Efficiency:** Evaluate the computational efficiency of the unlearning algorithm $\mathcal{A}_u$ in deriving $\mathcal{M}_u$ from $\mathcal{M}$, as well as the scalability of $\mathcal{M}_u$. For instance, assess whether it supports simultaneous multi-target unlearning.

5.3.1 Implementation. Relying on the empirical verification and key observations in Sec. 4, to provide a feasible and reliable implementation for the EvalIGMU framework, we carefully examine the existing relevant quantitative metrics for their effectiveness and select some reliable ones or design some more effective ones for the above five aspects. The detailed metrics are summarized as follows.

Forgetting: Under the model-based paradigm, we adopt the Classification Accuracy of *a new multi-head classifier, MultClf*, to reflect the unlearning performance. The *MultClf* is fine-tuned from the CLIP-ViT model [39] over DataIGM. With the shared feature extractor, different heads of *MultClf* can be used for various IGM unlearning tasks. More details of *MultClf* will be introduced later.

Preservation: The following metrics are used together:

- **CLIP Score Difference Rate (CSDR)**[8]: It calculates the difference rate between CLIP Score pairs. i.e., the CLIP Score pairs are computed using p with the corresponding image generated by $\mathcal{M}_u(t)$ and $\mathcal{M}(p)$, respectively. We use the average CSDR (%) to quantify discrepancies between images generated by $\mathcal{M}_u(t)$ and $\mathcal{M}(p)$. For an arbitrary task t and the corresponding p, we generate N images using $\mathcal{M}_u(t)$ and $\mathcal{M}(p)$, respectively. This process can be formulated as: average CSDR = $\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{|CS(\mathcal{M}(p)_i,p)-CS(\mathcal{M}_u(t)_j,p)|}{CS(\mathcal{M}(p)_i,p)}\times 100$.
- **LPIPS** [9]: we use average LPIPS to quantify the perceptual similarity between images generated by $\mathcal{M}_u(t)$ and $\mathcal{M}(p)$. For an arbitrary unlearning task t and the corresponding p, we generate N images for t using $\mathcal{M}_u(t)$ and K images for p using $\mathcal{M}(p)$, respectively. This process can be formulated as: average LPIPS = $\frac{1}{N\cdot K}\sum_{i=1}^{N}\sum_{j=i+1}^{K} LPIPS(\mathcal{M}_u(t)_i, \mathcal{M}(p)_j)$, where N can equal to K.
- **YOLO**: In particular, we use the Detection Rate (%) about humans by employing the YOLO v8 [25] model to detect the presence of individuals to evaluate the preservation of humans in $\mathcal{M}_u$ for those unlearning tasks related to humans, e.g., Nudity and Person.

Image Quality: We measure the change in image quality before and after applying the machine unlearning algorithm $\mathcal{A}_u$ (i.e., relative quality), which equals evaluating whether it degrades the models' generative ability. Here, we adopt $FID(\mathcal{M}_u(p^*), \mathcal{M}(\hat{p}))$ to measure the *distribution discrepancy* between two generated image sets, where *FID* refers to FID [22] and

- For *Abstract* category: $p^* = t$, $\hat{p} = do(T = \text{'none'})$ as in $\hat{P}$-1.
- For *Concrete* category: $p^* = \hat{p} = do(T \neq t)$ as in $\hat{P}$-2.

Lower *FID* scores indicate that the two image sets have more similar distributions, suggesting better preservation of image quality.

Robustness: We use the Attack Success Rate (ASR) of *UnlearnDiffAtk* [61], a mechanism that generates adversarial prompts to attack $\mathcal{M}_u$ and induce it to regenerate forgotten content, to assess the robustness of the unlearned models.

Efficiency: To ensure a fair and comprehensive evaluation, we consider several factors: the size of the editing module, the editing technique employed, time consumption, and the ability to handle multiple tasks simultaneously. Among these, time consumption and the ability to handle multiple tasks are particularly critical. An effective IGMU algorithm should address these aspects with a well-balanced approach.

Discussion: Admittedly, our EvalIGMU framework, especially for the *Forgetting* and *Robustness* aspects, is currently designed to deal with the unlearning request with an explicit keyword-based term, following the commonly adopted setting in existing work [12, 31, 57, 60]; such a simplified threat model does not yet cover cases with descriptive-based forgotten targets (as semantically equivalent or paraphrased expressions to the keyword term), which also bring certain limitations to the generalizability of EvalIGMU. Solving this problem is beyond our scope here, but with the help of natural language processing technology, it is also possible to build a mapping from descriptive case to keyword terms or to deal with descriptive cases directly, to alleviate and solve this limitation; in the future, corresponding modules can be included in our framework as extensions or plugins.

[8]Original CLIP Score [20] computes cosine similarity between image embedding and text embedding, denoted as $CS(i, t)$ for the image i and text t.

[9]LPIPS (Learned Perceptual Image Patch Similarity) [59], computes the Euclidean distance between two images, denoted as $LPIPS(x_1, x_2)$, for image x_1 and x_2.

Table 6: The performance of *MultClf* on DataIGM dataset.

Task	Data	Accuracy	Precision	Recall	F1 Score
Van Gogh style	*REAL*	93.12	92.67	93.65	93.16
	LAION	96.65	96.98	96.30	96.64
	SD-GEN	98.77	98.55	99.00	98.77
Nudity	*REAL*	92.89	88.07	99.24	93.32
	LAION	93.00	93.52	92.40	92.96
	SD-GEN	99.16	99.34	98.98	99.16
Church	*REAL*	100.00	100.00	100.00	100.00
	LAION	99.85	99.70	100.00	99.85
	SD-GEN	99.10	98.81	99.40	99.10
Parachute	*REAL*	100.00	100.00	100.00	100.00
	LAION	98.95	99.59	98.30	98.94
	SD-GEN	97.75	98.77	96.70	97.73

5.3.2 MultClf Classifier for 'Forget' evaluation. To address the issues of the existing evaluator that adopts task-specific classifiers or detectors, that is, they are typically trained on specific datasets (i.e., only *REAL* part of DataIGM) and struggle to generalize to generated data (i.e., *SD-GEN* part of DataIGM), such as the Style Classifier [61] and Nude Detector[10], we propose a multi-head classifier, *MultClf*, by fine-tuning the CLIP Vision Transformer (CLIP-ViT) on the DataIGM based on its default setting in Table 7. Considering its generality, especially for generated images, DataIGM, as detailed described in Table 7 (Sec. 5.4), incorporates diverse data sources designed for various unlearning tasks, including Van Gogh, Nudity, and Objects (Church and Parachute), with balanced representation across *REAL*, *LAION*, and *SD-GEN* datasets. We freeze the CLIP-ViT backbone and fine-tune four classifier heads with the listed four tasks, respectively.

To validate the effectiveness of *MultClf* model, we evaluated with various metrics, including Accuracy, Precision, Recall, and F1 score, and the test results are summarized in Table 6. The empirical results consistently demonstrate its high performance across all metrics, *MultClf* notably excels in handling generated datasets (*SD-GEN*), a domain largely overlooked by prior classifiers. Therefore, given its reliability and superior consistent performance, *MultClf* is more capable of evaluating the **'Forget'** aspect about unlearned models in IGMU tasks for the above included common ones.

Discussion: For tasks beyond those considered, evaluation can be effectively conducted by collecting task-specific data from diverse sources, fine-tuning the classifier head, and applying it to assess new tasks efficiently.

5.4 DataIGM

We construct a comprehensive dataset, **DataIGM**, focusing on 4 representative unlearning tasks: *Nudity*, *Style*, *Church*, and *Parachute*.

[10]https://github.com/notAI-tech/NudeNet

Table 7: The details of the proposed DataIGM dataset, where *Pair* means counterpart.

		REAL		*LAION*		*SD-GEN**	
		target	pair	target	pair	target	pair
Train	Van Gogh	1,322	1,322	4,000	4,000	16,000	16,000
	Nudity	190,000	190,000	4,000	4,000	16,000	16,000
	Church	1,300	1,300	4,000	4,000	4,000	4,000
	Parachute	1,300	1,300	4,000	4,000	4,000	4,000
Test	Van Gogh	567	567	1,000	1,000	4,000	4,000
	Nudity	3,800	3,800	1,000	1,000	4,000	4,000
	Church	100	100	1,000	1,000	1,000	1,000
	Parachute	50	50	1,000	1,000	1,000	1,000

* Here, we include only the *SD-GEN* data generated by the original $\mathcal{M}$ to align with the "train" and "test" objectives. The additional 2,860,000 *SD-GEN* images produced by $\mathcal{M}$ and $\mathcal{M}_u$ are an extension used exclusively for benchmarking purposes.

It comprises three distinct sources, i.e., *REAL*, *LAION*, and *SD-GEN*, selected to capture diverse scenarios and enable rigorous analysis.

REAL consists of original data used to train specific classifiers. For style unlearning, we use samples (with or without Van Gogh style) from WikiArt [43]; for nudity, we adopt the NudeNet dataset[11]; and for object removal, we use images of 'church' and 'parachute' from ImageNet-1k. *LAION* includes samples from Stable Diffusion's training sets (e.g., LAION-5B [46]), accessed via Hugging Face. *SD-GEN* contains images generated by the base model $\mathcal{M}$ and ten unlearned variants $\mathcal{M}_u$, using targeted prompts constructed via ChatGPT-4 [33], covering all four tasks.

As discussed in Sec. 4.2 and Sec. 5.2, achieving perfect unlearning is inherently difficult, e.g., reliably "dressing" nudity or removing artistic styles and specific objects while preserving all other visual elements. To mitigate this, we generate reference and comparison images using both $\mathcal{M}$ and $\mathcal{M}_u$ across target prompts t and their modified variants defined in Secs. $\hat{P}$-1 and $\hat{P}$-2. Table 7 summarizes DataIGM, where *target* refers to the content to be forgotten, and *pair* denotes its retained counterpart.

The multi-sourced DataIGM dataset plays three key roles: ❶ It enables the identification of discrepancies in detector performance across different data sources and, more importantly, facilitates evaluation on generated data (*SD-GEN*) aligned with the goals of IGMU. ❷ It serves as a high-quality resource for training reliable content detectors required by various evaluation tasks in IGMU. ❸ It provides a standardized test bed for benchmarking state-of-the-art unlearning algorithms across diverse evaluation dimensions.

5.5 Discussion

Based on the above categorization, decomposition, and analysis in this section, we can see that the implementation that meets the basic requirements **R1**-**R2** would be complicated, non-unique, and even somewhat subjective, and there is no unified once-and-for-all solution for varying unlearning tasks about the IGMs. Meanwhile, the diversity and non-exhaustiveness of the unlearning tasks, the uncertainty and semantic ambiguity of the natural language description (as the prompt for generation), and the limitation of sampling-based verification and existing quantitative metrics pose inherent challenges to accurate evaluation for the unlearned models.

The proposed IGMU framework consists of CatIGMU, EvalIGMU, and DataIGM. It aims to provide a structured insight, a systematic solution, and a reliable and practical methodology for image generative model unlearning. Specifically, **CatIGMU** introduces a hierarchical taxonomy for categorizing unlearning tasks, along with fine-grained, case-wise implementation guidance applicable to both idealized and practical settings; thus it provides principles about unlearning expectation behaviors across task types and reduces ambiguity in designing or interpreting experimental setups. Grounded in the fundamental requirements **R1**-**R2** and the flaws disclosed by empirical study of existing methods and evaluations, **EvalIGMU** forms a holistic and systematic evaluation framework — it covers five critical aspects as comprehensively as we know it is possible about IGMU and contains different quantitative metrics, which are meticulously selected (or trained over DataIGM) based on the extensive (ex-ante and ex-post) experimental validation. Also, the qualitative and quantitative results illustrated in Sec. 6 verify the effectiveness and reliability of those metrics. **DataIGM** consists of a multi-source, manually-selected, high-quality dataset tailored to IGMU, covering different common scenarios; it facilitates the construction of more accurate and reliable detectors/evaluators (e.g., the training of *MultClf*) and subsequent benchmarking of the state-of-the-art unlearning approaches.

Overall, our flexible framework will inspire more reliable and practical unlearning algorithms for IGMs, help to build comprehensive and precise task-wise ground-truth datasets, and extend to design more effective evaluation metrics or indicators, which can also be used to enrich and expand our framework in turn.

6 Re-evaluation of IGMU methods

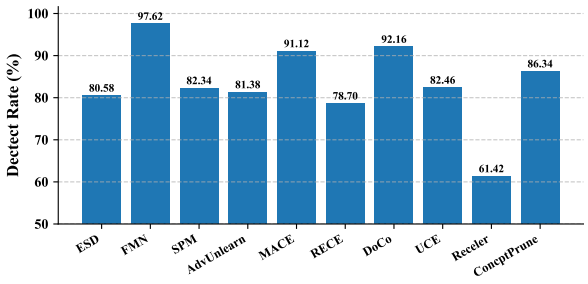


Figure 7: The human detection results of YOLO v8 on the images generated by unlearned models for the Nudity unlearning task.

We employ the proposed EvalIGMU framework and its implementation explained in Sec. 5.3 to evaluate the performance of existing IGMU algorithms systematically.

Unlearning Methods: We include ten state-of-the-art methods: ESD [12], FMN [57], SPM [31], AdvUnlearn [60], MACE [30], RECE [15], DoCo [53], Receler [23], ConceptPrune [3], and UCE [13]. Model weights are sourced from three main channels: (1) the AdvUnlearn GitHub repository[12], as referenced in [60]; (2) official

[11] https://github.com/notAI-tech/NudeNet/tree/v2

[12] https://github.com/OPTML-Group/AdvUnlearn

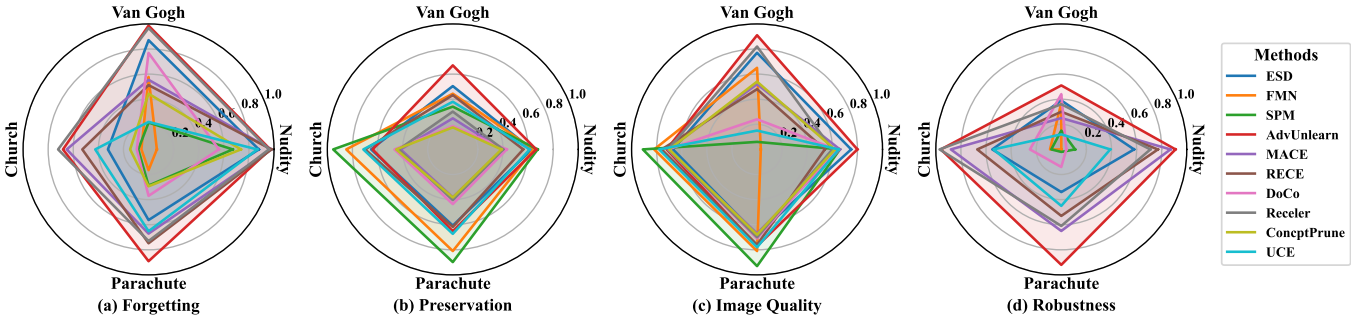


Figure 8: Performance evaluation for ten unlearning methods on four unlearning tasks (*Nudity*, *Van Gogh style*, *Church*, and *Parachute*) across four evaluation aspects: (a) Forgetting, (b) Preservation, (c) Image Quality, and (d) Robustness.

releases by authors (e.g., MACE [30], RECE [15], DoCo [53]); and (3) in-house training using their official implementations.

6.1 Unlearning Tasks

We evaluate unlearning performance on four representative tasks: *Van Gogh style*, *Nudity*, *Church*, and *Parachute*. These categories are commonly used and supported in prior IGMU works and span diverse unlearning types—including Global-abstract-style, Local-abstract-status, and Local-concrete-entities—while avoiding subjective or overly curated categories. All evaluations (and empirical studies in Sec.4) are conducted using publicly available checkpoints or official code with default generation settings to ensure reproducibility and minimize bias from human intervention.

Dataset: We adopt the proxy-based methods for evaluation. For the four unlearning tasks as Sec. 4.3, we sampled 286,000 paired images generated by $\mathcal{M}$ and $\mathcal{M}_u$s from the above unlearning methods by using both target prompts t and its corresponding variants of $\hat{P}$-1 and $\hat{P}$-2; the number of images for each model for each prompt is the same.

Implementation: We conduct systematic evaluations of existing IGMU algorithms using the proposed EvalIGMU framework and its implementation explained in Sec. 5.3. All experiments are implemented in PyTorch and run on two NVIDIA A6000 GPUs.

6.2 Quantitative Results

Figure 8 summarizes the performance of existing unlearning algorithms across four unlearning tasks: *Nudity*, *Van Gogh style*, *Church*, and *Parachute* on four aspects: (a)*Forgetting*, (b)*Preservation*, (c) *Image Quality*, and (d) *Robustness*. All values in Figure 8 are normalized to $[0, 1]$, with higher values indicate better performance (for smaller-is-better metrics, $1 -$ value is used). Based on those reliable metrics, we have the following findings,

(1) **Forgetting:** *MultClf*'s results in Figure 8(a) indicate that while most methods perform effectively in nudity and parachute unlearning, they struggle significantly with church unlearning. These results indicate their inconsistencies in different tasks, even in the same type of task, i.e., church and parachute. AdvUnlearn exhibits the best and most balanced performance across all tasks. In contrast, FMN and SPM fail to achieve meaningful unlearning, particularly for the church unlearning task.

(2) **Preservation:** Figure 8(b) shows the averaged "CSDR + LPIPS" for their general evaluation. It shows that existing methods perform poorly in preservation regarding semantic alignment (CSDR) and perceptual similarity (LPIPS). Only SPM and FMN perform modestly in church and parachute unlearning tasks. Additional results from YOLO v8 in Figure 7 indicate that existing unlearning methods struggle to preserve 'human' in strongly associated unlearning tasks (i.e., nude and human). FMN achieves the best performance, while some methods (e.g., Receler) fail to generate images containing human in up to 38.58% of cases. Therefore, existing unlearning algorithms fail to remove target content accurately while preserving unrelated elements effectively.

(3) **Image Quality:** The FID results in Figure 8(c) indicate that existing methods perform better in church and parachute unlearning but struggle with abstract unlearning tasks, particularly in the Van Gogh style. Among these methods, AdvUnlearn demonstrates balanced performance across all four tasks. In contrast, methods like SPM perform well on Church but poorly on Van Gogh style, while FMN performs well on church but struggles with Nudity unlearning.

(4) **Robustness:** The results of UnlearnDiffAtk in Figure 8(d) reveal that existing unlearned models are highly susceptible to adversarial attack on prompts, often being guided to regenerate content that should have been forgotten, particularly in Van Gogh style unlearning tasks. Similarly, AdvUnlearn demonstrates the best robustness across most tasks, except for the Van Gogh style. In contrast, SPM shows minimal robustness, performing poorly across all four tasks.

6.3 Efficiency Discussion

Efficiency is evaluated based on runtime and the ability to handle multiple tasks, as outlined in Sec. 5.3. Table 8 compares **TEN** SOTA IGMU methods in terms of modified modules, employed techniques, runtime, and multi-task supportiveness. Among these methods, FMN, MACE, ConceptPrune, and UCE demonstrate a commendable balance between efficiency and versatility. They complete unlearning tasks within approximately $42 \sim 50$ seconds by modifying only a small portion of model parameters ($0.12\% \sim 0.37\%$) while effectively supporting multi-task unlearning. In contrast, methods such as Receler and AdvUnlearn, which depend on adversarial training to achieve unlearning, exhibit significantly longer runtimes, taking

approximately 2 hours and 7 hours, respectively. This extended runtime severely limits their scalability. These findings underscore the importance of developing methods that achieve efficient unlearning and support versatility across multiple tasks better to meet the growing demands of IGMU applications.

Table 8: Comparison of unlearning methods with respect to modified modules, applied techniques, runtime, and multi-concept unlearning capability.

Method	Module	Technique	Runtime*	Multi-task
ESD	Cross-attention	Finetuning	1.5 h	True
FMN	Cross-attention	Finetuning	42 s	True
SPM	SPM	Latent anchoring	1.2 h	True
AdvUnlearn	Text encoder	Adv training	7 h	False
MACE	Cross-attention & Multi-LoRA	Closed-form & Finetuning	50 s	True
RECE	Cross-attention	Multi-epoch-Closed-form	12 min	True
DoCo	Cross-attention	Adv training	45 min	True
Receler	Cross-attention	Adv training	2 h	True
ConceptPrune	FFN	Pruning	40 s	True
UCE	Cross-attention	Closed-form	45 s	True

Note: This information is sourced from the original papers and officially provided code. Runtime* values are estimated based on the Van Gogh style unlearning task using default configurations on a single A6000 GPU.

6.4 Summary

The findings reveal the following insights: ❶ Significant limitations exist in current unlearning algorithms; they fail to achieve balanced performance aligned with EvalIGMU's expectations, particularly in preservation and robustness against adversarial prompts. ❷ Performance varies across tasks; existing methods struggle on global abstract unlearning (e.g., Van Gogh style) but perform better on other unlearning tasks (this is aligned with previous work, Six-CD [41]). ❸ The same method demonstrates inconsistent performance across different tasks and evaluation aspects, e.g., SPM has varying preservation performance in Figure 8(b). ❹ Even within the same evaluation aspect and unlearning task type, performance varies; for instance, unlearned models are more robust for church unlearning than parachute unlearning. ❺ In terms of efficiency, significant variation exists in the time required for the unlearning process, even among methods that edit the same module using the same technique. With a comprehensive evaluation framework aligned with refined measurements, our results offer a detailed and accurate comparison of these methods, providing valuable insights and guidance for future research.

Discussion: As ❷ highlights and prior studies [12, 13, 30, 53, 60], unlearning performance exhibits substantial variation across tasks, suggesting that a single algorithm may behave inconsistently depending on the semantic and structural properties of the target content. This observation reinforces the fundamental limitations of existing approaches, as discussed in Sec. 4.

7 Conclusion

In this work, we systematically addressed key challenges in image generation model unlearning, including the lack of clear task definitions and implementation guidance, the absence of a comprehensive evaluation framework, and unreliable evaluation metrics. To tackle these issues, we proposed CatIGMU, a hierarchical task categorization framework with detailed implementation guidelines; EvalIGMU, a multi-dimensional evaluation framework supported by refined metrics; DataIGM, a large-scale, high-quality dataset tailored for IGMU. Leveraging EvalIGMU and DataIGM, we conducted a benchmark study of ten state-of-the-art unlearning algorithms. Empirical results reveal that current methods struggle to achieve balanced performance across the evaluation dimensions defined in EvalIGMU, particularly in preserving benign content, maintaining image quality, and resisting adversarial prompts. These contributions are intended to advance and promote both theoretical research and practical applications in IGMU, paving the way for more effective and trustworthy unlearning solutions for IGMs.

Future Work Despite their comprehensiveness, the proposed frameworks present several limitations that merit further investigation. First, more fine-grained case studies of unlearning tasks are needed to better align CatIGMU with emerging real-world requirements. Additionally, while EvalIGMU evaluates five critical aspects now, future research could expand its scope to include dimensions such as explainability or fairness. Moreover, extending these frameworks to encompass more intricate tasks and developing lightweight metrics for scalable benchmarking are promising directions. As demonstrated in this work, existing IGMU methods still face limitations in robustness, fidelity, and generalization, so designing new unlearning algorithms that perform consistently well across all evaluation aspects remains an open challenge.

Acknowledgments

This research is supported by A*STAR, CISCO Systems (USA) Pte. Ltd and National University of Singapore under its Cisco-NUS Accelerated Digital Economy Corporate Laboratory (Award I21001E0002), and the start-up grants from the University of Science and Technology of China (USTC).

References

[1] Lucas Bourtoule, Varun Chandrasekaran, Christopher A. Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. 2021. Machine Unlearning. In *S&P*. IEEE, 141–159.

[2] Anh Bui, Khanh Doan, Trung Le, Paul Montague, Tamas Abraham, and Dinh Phung. 2024. Removing Undesirable Concepts in Text-to-Image Generative Models with Learnable Prompts. *arXiv preprint arXiv:2403.12326* (2024).

[3] Ruchika Chavhan, Da Li, and Timothy M. Hospedales. 2024. ConceptPrune: Concept Editing in Diffusion Models via Skilled Neuron Pruning. *CoRR* abs/2405.19237 (2024).

[4] Zhi-Yi Chin, Chieh-Ming Jiang, Ching-Chun Huang, Pin-Yu Chen, and Wei-Chen Chiu. 2024. Prompting4Debugging: Red-Teaming Text-to-Image Diffusion Models by Finding Problematic Prompts. In *ICML*. OpenReview.net.

[5] A. Feder Cooper, Christopher A. Choquette-Choo, Miranda Bogen, Matthew Jagielski, Katja Filippova, Ken Ziyu Liu, and et al. 2024. Machine Unlearning Doesn't Do What You Think: Lessons for Generative AI Policy, Research, and Practice. *CoRR* abs/2412.06966 (2024).

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *CVPR*. IEEE Computer Society, 248–255.

[7] Ihor Kroosh Dmitry Voitekh Nick Hasty and Dmytro Korduban. 2019. Giphy's open source celebrity detection deep learning model and code. https://github.com/Giphy/celeb-detection-oss.

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *ICLR*. OpenReview.net.

[9] Frederick Eberhardt and Richard Scheines. 2007. Interventions and causal inference. *Philosophy of science* 74, 5 (2007), 981–995.

[10] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell,

Tim Dockhorn, Zion English, and Robin Rombach. 2024. Scaling Rectified Flow Transformers for High-Resolution Image Synthesis. In *ICML*. OpenReview.net.
[11] Chongyu Fan, Jiancheng Liu, Yihua Zhang, Eric Wong, Dennis Wei, and Sijia Liu. 2024. SalUn: Empowering Machine Unlearning via Gradient-based Weight Saliency in Both Image Classification and Generation. In *ICLR*. OpenReview.net.
[12] Rohit Gandikota, Joanna Materzynska, Jaden Fiotto-Kaufman, and David Bau. 2023. Erasing Concepts from Diffusion Models. In *ICCV*. IEEE, 2426–2436.
[13] Rohit Gandikota, Hadas Orgad, Yonatan Belinkov, Joanna Materzynska, and David Bau. 2024. Unified Concept Editing in Diffusion Models. In *WACV*. IEEE, 5099–5108.
[14] Tony Ginart, Melody Guan, Gregory Valiant, and James Zou. 2019. Making AI Forget You: Data Deletion in Machine Learning. In *NeurIPS*. 3518–3529.
[15] Chao Gong, Kai Chen, Zhipeng Wei, Jingjing Chen, and Yu-Gang Jiang. 2024. Reliable and Efficient Concept Erasure of Text-to-Image Diffusion Models. In *ECCV*, Vol. 15111. Springer, 73–88.
[16] Xiaoxuan Han, Songlin Yang, Wei Wang, Yang Li, and Jing Dong. 2024. Probing Unlearned Diffusion Models: A Transferable Adversarial Attack Perspective. *CoRR* (2024).
[17] Sebastian Hartwig, Dominik Engel, Leon Sick, Hannah Kniesel, Tristan Payer, Timo Ropinski, et al. 2024. Evaluating Text to Image Synthesis: Survey and Taxonomy of Image Quality Metrics. *arXiv preprint arXiv:2403.11821* (2024).
[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *CVPR*. IEEE Computer Society, 770–778.
[19] Alvin Heng and Harold Soh. 2023. Selective Amnesia: A Continual Learning Approach to Forgetting in Deep Generative Models. In *NeurIPS*.
[20] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. CLIPScore: A Reference-free Evaluation Metric for Image Captioning. *arXiv preprint arXiv:2104.08718* (2021).
[21] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In *NeurIPS*. 6626–6637.
[22] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In *NeurIPS*. 6626–6637.
[23] Chi-Pin Huang, Kai-Po Chang, Chung-Ting Tsai, Yung-Hsuan Lai, Fu-En Yang, and Yu-Chiang Frank Wang. 2024. Receler: Reliable Concept Erasing of Text-to-Image Diffusion Models via Lightweight Erasers. In *ECCV*. Springer, 360–376.
[24] Montreal AI Ethics Institute. 2023. Unstable Diffusion: Ethical Challenges and Some Ways Forward. https://montrealethics.ai/unstable-diffusion-ethical-challenges-and-some-ways-forward/. Accessed: 2024-12-31.
[25] Glenn Jocher, Abhiram Chaurasia, Jiuan Qiu, and Robby Stoken. 2023. YOLOv8: The Next Generation of YOLO. https://github.com/ultralytics/ultralytics.
[26] Sanghyun Kim, Seohyeon Jung, Balhae Kim, Moonseok Choi, Jinwoo Shin, and Juho Lee. 2023. Towards Safe Self-Distillation of Internet-Scale Text-to-Image Diffusion Models. *CoRR* (2023).
[27] Nupur Kumari, Bingliang Zhang, Sheng-Yu Wang, Eli Shechtman, Richard Zhang, and Jun-Yan Zhu. 2023. Ablating Concepts in Text-to-Image Diffusion Models. In *ICCV*. IEEE, 22634–22645.
[28] Senmao Li, Joost van de Weijer, Taihang Hu, Fahad Shahbaz Khan, Qibin Hou, Yaxing Wang, and Jian Yang. 2024. Get What You Want, Not What You Don't: Image Content Suppression for Text-to-Image Diffusion Models. In *ICLR*. OpenReview.net.
[29] Xinfeng Li, Yuchen Yang, Jiangyi Deng, Chen Yan, Yanjiao Chen, Xiaoyu Ji, and Wenyuan Xu. 2024. SafeGen: Mitigating Unsafe Content Generation in Text-to-Image Models. In *CCS*.
[30] Shilin Lu, Zilan Wang, Leyang Li, Yanzhu Liu, and Adams Wai-Kin Kong. 2024. MACE: Mass Concept Erasure in Diffusion Models. In *CVPR*. Computer Vision Foundation / IEEE Computer Society, 6430–6440.
[31] Mengyao Lyu, Yuhong Yang, Haiwen Hong, Hui Chen, Xuan Jin, Yuan He, Hui Xue, Jungong Han, and Guiguang Ding. 2024. One-dimensional Adapter to Rule Them All: Concepts, Diffusion Models and Erasing Applications. In *CVPR*. Computer Vision Foundation / IEEE Computer Society, 7559–7568.
[32] Rui Ma, Qiang Zhou, Bangjun Xiao, Yizhu Jin, Daquan Zhou, Xiuyu Li, Aishani Singh, Yi Qu, Kurt Keutzer, Xiaodong Xie, et al. 2024. A Dataset and Benchmark for Copyright Protection from Text-to-Image Diffusion Models. *arXiv preprint arXiv:2403.12052* (2024).
[33] OpenAI. 2023. ChatGPT-4: A Large-Scale Multimodal Language Model. https://openai.com. Accessed: 2024-08-31.
[34] Minh Pham, Kelly O. Marshall, Niv Cohen, Govind Mittal, and Chinmay Hegde. 2024. Circumventing Concept Erasure Methods For Text-To-Image Generative Models. In *ICLR*. OpenReview.net.
[35] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2024. SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. In *ICLR*. OpenReview.net.
[36] Samuele Poppi, Tobia Poppi, Federico Cocchi, Marcella Cornia, Lorenzo Baraldi, Rita Cucchiara, et al. 2024. Safe-CLIP: Removing NSFW Concepts from Vision-and-Language Models. In *ECCV*.
[37] Bedapudi Praneeth. 2023. NudeNet: Deep Learning Model for Nudity Detection. https://github.com/notAI-tech/NudeNet.
[38] Yiting Qu, Xinyue Shen, Xinlei He, Michael Backes, Savvas Zannettou, and Yang Zhang. 2023. Unsafe Diffusion: On the Generation of Unsafe Images and Hateful Memes From Text-To-Image Models. In *CCS*. ACM, 3403–3417.
[39] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning Transferable Visual Models from Natural Language Supervision. In *ICML*. 8748–8763.
[40] Protection Regulation. 2016. Regulation (EU) 2016/679 of the European Parliament and of the Council. *Regulation (eu)* 679 (2016), 2016.
[41] Jie Ren, Kangrui Chen, Yingqian Cui, Shenglai Zeng, Hui Liu, Yue Xing, Jiliang Tang, and Lingjuan Lyu. 2024. Six-CD: Benchmarking Concept Removals for Benign Text-to-image Diffusion Models. *CoRR* (2024).
[42] Kevin Roose. 2022. A.I. Generated Art Won an Art Prize. Artists Aren't Happy. *The New York Times* (2022). https://www.nytimes.com/2022/09/02/technology/ai-artificial-intelligence-artists.html Accessed: 2024-12-31.
[43] Babak Saleh and Ahmed Elgammal. 2015. WikiArt: Visual Art Dataset for Recognition and Aesthetics Analysis. In *ECCV*. Springer, 3–10.
[44] Patrick Schramowski, Manuel Brack, Björn Deiseroth, and Kristian Kersting. 2023. Safe Latent Diffusion: Mitigating Inappropriate Degeneration in Diffusion Models. In *CVPR*. Computer Vision Foundation / IEEE Computer Society, 22522–22531.
[45] Patrick Schramowski, Christopher Tauchmann, and Kristian Kersting. 2022. Can Machines Help Us Answering Question 16 in Datasheets, and In Turn Reflecting on Inappropriate Content?. In *FAccT*. ACM, 1350–1361.
[46] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. 2022. LAION-5B: An open large-scale dataset for training next generation image-text models. In *NeurIPS*.
[47] Juwon Seo, Sung-Hoon Lee, Tae-Young Lee, Seungjun Moon, and Gyeong-Moon Park. 2024. Generative Unlearning for Any Identity. In *CVPR*. IEEE, 9151–9161.
[48] Yu-Lin Tsai, Chia-Yi Hsu, Chulin Xie, Chih-Hsun Lin, Jia-You Chen, Bo Li, Pin-Yu Chen, Chia-Mu Yu, and Chun-Ying Huang. 2024. Ring-A-Bell! How Reliable are Concept Removal Methods For Diffusion Models?. In *ICLR*. OpenReview.net.
[49] Hongyu Wang, Qing Li, Xiangyu Liu, Tong Lu, and Hao Zhou. 2023. Zero-Shot Image Restoration Using Denoising Diffusion Models. In *ICCV*. 2029–2038.
[50] Peng Wang, Lingzhi Zhang, Yanghua Li, Yuming Jiang, Huachun Yang, and Li Liu. 2021. Text-Driven Image Manipulation by Predicting Image Representations in Textual Semantic Space. *IEEE Transactions on Image Processing* 30 (2021), 7213–7228.
[51] Jing Wu and Mehrtash Harandi. 2024. Scissorhands: Scrub Data Influence via Connection Sensitivity in Networks. *CoRR* (2024).
[52] Jing Wu, Trung Le, Munawar Hayat, and Mehrtash Harandi. 2024. EraseDiff: Erasing Data Influence in Diffusion Models. *CoRR* (2024).
[53] Yongliang Wu, Shiji Zhou, Mingzhuo Yang, Lianzhe Wang, Heng Chang, Wenbo Zhu, Xinting Hu, Xiao Zhou, and Xu Yang. 2025. Unlearning Concepts in Diffusion Model via Concept Domain Correction and Concept Preserving Gradient. In *AAAI*. AAAI Press, 8496–8504.
[54] Yijun Yang, Ruiyuan Gao, Xiaosen Wang, Tsung-Yi Ho, Nan Xu, and Qiang Xu. 2024. MMA-Diffusion: MultiModal Attack on Diffusion Models. In *CVPR*. Computer Vision Foundation / IEEE Computer Society, 7737–7746.
[55] Chenshuang Zhang, Chaoning Zhang, Mengchun Zhang, and In So Kweon. 2023. Text-to-image Diffusion Models in Generative AI: A Survey. *CoRR* (2023).
[56] Duzhen Zhang, Yahan Yu, Jiahua Dong, Chenxing Li, Dan Su, Chenhui Chu, and Dong Yu. 2024. MM-LLMs: Recent Advances in MultiModal Large Language Models. In *ACL*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, 12401–12430.
[57] Gong Zhang, Kai Wang, Xingqian Xu, Zhangyang Wang, and Humphrey Shi. 2024. Forget-Me-Not: Learning to Forget in Text-to-Image Diffusion Models. In *CVPR*. IEEE, 1755–1764.
[58] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *CVPR*. Computer Vision Foundation / IEEE Computer Society, 586–595.
[59] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *CVPR*. 586–595.
[60] Yimeng Zhang, Xin Chen, Jinghan Jia, Yihua Zhang, Chongyu Fan, Jiancheng Liu, Mingyi Hong, Ke Ding, and Sijia Liu. 2024. Defensive Unlearning with Adversarial Training for Robust Concept Erasure in Diffusion Models. *CoRR* (2024).
[61] Yimeng Zhang, Jinghan Jia, Xin Chen, Aochuan Chen, Yihua Zhang, Jiancheng Liu, Ke Ding, and Sijia Liu. 2024. To Generate or Not? Safety-Driven Unlearned Diffusion Models Are Still Easy to Generate Unsafe Images ... For Now. In *CVPR*. Computer Vision Foundation / IEEE Computer Society, 385–403.
[62] Yihua Zhang, Yimeng Zhang, Yuguang Yao, Jinghan Jia, Jiancheng Liu, Xiaoming Liu, and Sijia Liu. 2024. Unlearncanvas: A stylized image dataset to benchmark machine unlearning for diffusion models. *arXiv preprint arXiv:2402.11846* (2024).